

A bioinformatics approach to structure-based T cell epitope prediction

by

Javed Mohammed Khan

Master of Biotechnology

Macquarie University, Sydney, Australia

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Chemistry and Biomolecular Sciences
Macquarie University
Sydney, Australia

March 2011

IN MEMORY OF MY MOTHER

LATE MRS. SHAHAWAR KHAN

AND

DEDICATED TO MY FAMILY

MR. USMAN KHAN (FATHER), MR. NOOR MOHAMMED KHAN (BROTHER), MRS. BENAZIR
AHMED (SISTER) AND MR. SHOAIB MOHAMMED KHAN (BROTHER).

DECLARATION

This thesis contains original work performed by me. Few aspects of this work have been carried out with help from collaborating researchers; these people have been acknowledged and their contributions recognised in the section in which their assistance was received. This thesis contains no material that has been accepted for the award of any higher degree or diploma at any University or Institution and to the best of my knowledge, contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Javed Mohammed Khan

March 2011

ACKNOWLEDGEMENTS

A few months into my Ph.D. project, I already realised that a researcher cannot complete a Ph.D. alone. Although the list of individuals I wish to thank extends beyond the limits of this format, it is my pleasure to thank the following people who made this thesis possible through their continuous encouragement, prayers and support:

Professional

- My supervisor, Prof. Shoba Ranganathan, for her constant moral support and invaluable suggestions during this work. I thank her for giving me an opportunity to be a part of her group, believing in me and most importantly bestowing her trust in me. I am very grateful for her patience, motivation, enthusiasm, intellectual support and especially for encouraging me to visit Dr. Joo Chuan Tong at the Institute for Infocomm Research, Singapore to acquire additional training in pMHC docking and InCoB 2010 in Tokyo, Japan to participate and interact with leading scientists in the area of computational immunology and structural immunoinformatics. I would also like to thank her for educating me on financial management in academic research. Finally, I thank her for putting up with my routine late arrival to the lab!
- Assoc. Prof. Bridget Mabbutt, my co-supervisor, for support through the first two years of my PhD tenure.
- Dr. Sham Nair, my immunology lecturer (during M. Biotech.), for encouraging me to take up research in Immunology, for thought provoking discussions on TR/pMHC interactions and for his valuable suggestions throughout my PhD.
- Dr. Joo Chuan (Victor) Tong, for invaluable training, guidance and help throughout my Ph.D. and for vital conversations regarding pMHC docking and T cell epitope prediction. I also thank him for providing me with utmost support and the framework for MPID-T2 database.
- Dr. Asif M Khan, for priceless suggestions, significant exchange of ideas during conferences and constructive criticism of my work. I would also like to thank him for suggestions on post PhD plans and his co-operation during my training in Singapore.

- Assoc. Prof. Tin Wee Tan, for useful discussions and allowing me to work in his lab during my initial visit to the National University of Singapore and Institute of Infocomm Research, Singapore.
- Prof. Mark Baker and Assoc. Prof. Joanne Jamie, for helpful suggestions and instructions to that helped me improve the quality of my research.
- Ms. Catherine Wong, Ms. Maria Hyland, Ms. Jane Yang, Dr. Chris McRae, Mr. Michael Baxter and Mr. Doan Lee, for administrative and IT support.
- Macquarie University, for the award of MQRES research scholarship for pursuing a Ph.D. and PGRF funding for attending the InCoB 2010 conference in Tokyo, Japan.
- The Department of Chemistry and Biomolecular Sciences, Faculty of Science and the Higher Degree Research Office at Macquarie University, Sydney, for having provided me with all the facilities for the successful completion of this research project.
- My colleagues, Mrs. Elsa Chacko, Mrs. Ranjeeta Menon, Mr. Gaurav Kumar, Mr. Harish Cheruku, Mr. Gagan Garg, Mr. Varun Khanna, Mr. Jitendra Gaikwad and Mr. Mohammad Islam. Particular thanks to Mr. Harish Cheruku for assistance with various aspects in the final stages of my Ph.D., Mrs. Elsa Chacko for helping me with the MPID-T2 database and Mr. Gaurav Kumar for helping me with statistical computing for the final publication included in this thesis.

Personal

- I am forever indebted to my family (father, brothers, sister and their children), my sister-in-law, Mrs. Sadiya Khan and brother-in-law, Mr. Tahir Saleem Ahmed, whose endless patience, motivation, financial support and encouragement made it possible for me to reach this important milestone in my research career.
- I would like to thank my dear friend Latifa Jabbar for sharing wonderful times and for emotional support that kept me going through the completion of my Ph.D.
- I am deeply thankful to my friends for making life in Australia a memorable experience with their good company in parties, picnics, long drives, cricket matches and for their

moral support: Ramnaresh Gorlamandala, Enoch Nagabyrava, Elsa Chacko, Ranjeeta Menon, Varun Khanna, Harish Cheruku, Akhil Sidharth and Waheed Iqbal.

- My hearty thanks to my best friends, Varun Devata (Melbourne) and Samiyojit Singh (India) for standing by me in good and bad during the course of 15 and 9 long years, respectively. I am forever grateful for the light-hearted, funny and relaxing chats over the phone that helped me maintain my sanity during difficult times in PhD.
- Thanks also to my two other great friends, Sumantra Banerjee (India) and Ashok Bhupathi (India) who always enjoyed my success and all my publications although they never understood a thing about them.
- Last, but by no means the least, my mother, whose death in July 1996 instigated the burning desire to pursue research in me. Throughout my Ph.D., her memory and love has been the guiding light and an intense source of motivation right from when she passed away, especially from 2006, when I arrived in Australia.

TABLE OF CONTENTS

<i>Declaration</i>	<i>i</i>
<i>Acknowledgements</i>	<i>ii</i>
<i>Table of Contents</i>	<i>v</i>
<i>List of Abbreviations</i>	<i>viii</i>
<i>List of Figures</i>	<i>xi</i>
<i>List of Tables</i>	<i>xviii</i>
<i>List of Publications included in this thesis</i>	<i>xix</i>
<i>Abstract</i>	<i>xx</i>
CHAPTER 1 Introduction and literature survey	1
1.1 Overview	1
1.2 Brief history of MHC and TR proteins	3
1.3 Genetic makeup of MHC and TR proteins	4
1.4 Structure and function of MHC	6
1.4.1 MHC-I	7
1.4.2 MHC-II	9
1.5 pMHC binding	10
1.5.1 pMHC-I	11
1.5.2 pMHC-II	12
1.6 Structure and function of TR	13
1.7 First crystal structures of TR/pMHC complexes	16
1.8 TR/pMHC interaction	17
1.9 Issues with T cell epitope prediction	19
1.9.1 Quantity of peptide data	20
1.9.2 Quality of peptide data	21
1.9.3 Bias in peptide data	21
1.10 Databases and resources available	22
1.10.1 Generalized databases and resources	22
1.10.2 Specialized databases and resources	24
1.11 Methods available for T cell epitope prediction	32
1.11.1 Sequence-based approaches	36
1.11.2 Structure-based approaches	47
1.11.3 Sequence-structure-based approaches	54

	<i>Publication 1: Structural Immunoinformatics:</i>	57
	<i>Understanding MHC-peptide-TR binding</i>	
	<i>Publication 2: TR recognition of MHC-peptide complexes</i>	75
1.12	Objectives	87
CHAPTER 2	Methods and Applications	89
CHAPTER 3	pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes	90
3.1	Summary	90
	<i>Publication 3</i>	91
3.2	Conclusions	113
CHAPTER 4	MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures	114
4.1	Summary	114
4.2	Data	114
	<i>Publication 4</i>	163
4.3	Data analysis	195
4.4	Conclusions	210
CHAPTER 5	Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one	212
5.1	Summary	212
	<i>Publication 5</i>	213
5.2	Conclusions	233
CHAPTER 6	<i>In silico</i> prediction of immunogenic T cell epitopes for HLA-DQ8	234
6.1	Summary	234
	<i>Publication 6</i>	235
6.2	Conclusions	257

CHAPTER 7	Conclusions and future directions	258
7.1	Summary	258
7.2	Conclusions	259
7.3	Innovations	260
7.4	Significance and contributions	261
7.5	Future directions	262
References		264

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
ANN	Artificial neural network
APC	Antigen presenting cell
ARB	Average relative binding
A_{ROC}	Area under the receiver operating characteristic curve
ASA	Accessible surface area
ATP	Adenosine triphosphate
β2m	β ₂ -microglobulin
BE	Binding energy or binding free energy
BIMAS	Bioinformatics and molecular analysis section
BLAST	Basic local alignment search tool
Cα	Carbon-alpha
CA	Contact area
CD	Cluster of differentiation
cDNA	Complementary DNA
CDR	Complementarity determining region
CID	Cancer immune database
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
C-terminus	Carboxyl(COOH)-terminus
CTL	Cytotoxic T lymphocyte
DNA	Deoxyribonucleic acid
EMBL	European molecular biology laboratory
ER	Endoplasmic reticulum
FN	False negative
FP	False positive
FPIA	Fusion proteins for immune applications
H	Histocompatibility
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HMM	Hidden markov model
IDDM	Insulin-dependent diabetes mellitus
IEDB	Immune epitope database
IG	Immunoglobulin
IgSF	Immunoglobulin superfamily
IHWG	International histocompatibility working group

Ii	Invariant chain
IMGT	IMGT®, the international ImMunoGeneTics information system®
INN	International nonproprietary names
kDa	kiloDalton
KIR	Killer cell immunoglobulin like receptor
mAb	Monoclonal antibodies
MAPPP	MHC-I antigenic peptide processing prediction
MD	Molecular dynamics
MHC	Major histocompatibility complex
MHC-I	Major histocompatibility complex class I
MHC-II	Major histocompatibility complex class II
MhcSF	MHC superfamily
MIIC	MHC-II compartment
MSEP	Molecular surface electrostatic potential
NCBI	National center for biotechnology information
NIAID	National institute of allergy and infectious diseases
NIH	National institutes of health
NLM	National library of medicine
N-terminus	Amino(NH ₂)-terminus
OMIM	Online mendelian inheritance in man
PDB	Protein data bank
PFR	Peptide flanking residues
pHMM	Profile hidden markov model
PLS	Partial least squares
pMHC	Peptide-MHC complex
pMHC-I	Peptide-MHC-I complex
pMHC-II	Peptide-MHC-II complex
PSCPL	Positional scanning combinatorial peptide libraries
PSSM	Position-specific scoring matrix
PTM	Post translational modifications
QSAR	Quantitative structure-activity relationship
RMSD	Root mean square deviation
ROC	Receiver operating characteristic
RPI	Related proteins of the immune system
SA	Simulated annealing
SBT	Sequencing based typing
SDA	Stepwise discriminating analysis
SE	Sensitivity

SMM	Stabilized matrix method
SP	Specificity
SVM	Support vector machine
TAP	Transporter associated with antigen processing
TN	True negative
TP	True positive
TR	T cell receptor
TrEMBL	Translated EMBL
TR/pMHC	T cell receptor-peptide-MHC complex
TR/pMHC-I	T cell receptor-peptide-MHC-I complex
TR/pMHC-II	T cell receptor-peptide-MHC-II complex
WHO	World health organization

LIST OF FIGURES

- Figure 1.1 A schematic representation of the human MHC genes. 5**
- The diagram shows the location of the genes that encode MHC-I and II proteins. MHC-I gene loci A (orange), B (rose) and C (pink) along with MHC-II gene loci DP (light green), DQ (turquoise) and DR (lavender) are shown. The centromere and the MHC-III loci are coloured yellow and light blue, respectively. The polypeptide chains coded for by the loci are shown within the boxes depicting different loci in MHC-I and II gene complexes.
- Figure 1.2 A simple example of the rearrangement that occurs during TR α (TRA) and β (TRB) chain formation. 5**
- The V, D and J gene segments coding for the variable domain of the TR are shown in green, turquoise and lavender, respectively. The constant gene segment that codes for the constant region of the TR is shown in black.
- Figure 1.3 A cartoon depiction of typical MHC proteins. a. MHC-I. b. MHC-II. 7**
- The α and β -2 microglobulin chains of MHC-I are coloured dark and light green, respectively in **a.** and the α and β chains of MHC-II are coloured dark and light blue, respectively in **b.** The peptide binding cleft, β 2m domain, C-LIKE domains, various regions, plasma membrane and the cytosol are labelled.
- Figure 1.4 A ribbon representation of a MHC-I X-ray crystal structure (Protein Data Bank - PDB [62, 63] code: 1oga [53]). a. The four domains. b. An aerial view of the peptide binding cleft. 8**
- In **a.** the α 1 (G-ALPHA1), α 2 (G-ALPHA2), α 3 (C-LIKE) and the β 2m domains are coloured red, yellow, green and blue, respectively, to clearly show the structure of the MHC-I protein. Highlighted in **b.** is the anatomy of the peptide binding cleft formed by the two α -helices on either side of the β -sheet which forms the floor.

Figure 1.5 **A schematic ribbon representation of a MHC-II X-ray crystal structure (PDB code: 1ymm [78]). a. The four domains. b. An aerial view of the peptide binding cleft.** **10**

In **a.** the $\alpha 1$ (G-ALPHA), $\alpha 2$ (C-LIKE), $\beta 1$ (G-BETA) and $\beta 2$ (C-LIKE) domains are coloured red, blue, yellow and green, respectively, clearly illuminating the structure of the MHC-II protein. Illustrated in **b.** is the anatomy of the peptide binding cleft formed by the two α -helices sitting on either side of the β -sheet which forms the floor.

Figure 1.6 **Different conformations adopted by MHC-I binding peptides. a. A nonameric Tax peptide bound in a flattened conformation to HLA-A*0201 from the PDB structure 1duz [94]. b. A 13-residue peptide bound in a bulged fashion to HLA-B*3508 from the PDB structure 2ak4 [93].** **12**

The peptide and the MHC peptide binding clefts are coloured green and red, respectively. The N and C-terminal peptide ‘anchor’ residues and the ‘pocket’ residues from the MHC peptide binding cleft are shown in ball and stick representation and are portrayed in yellow and blue, respectively, in **a.** to highlight the strong interactions around the peptide termini.

Figure 1.7 **Diversity in the lengths of peptides binding to MHC-II proteins. a. A 6-residue peptidomimetic peptide bound to HLA-DRB1*0401 from the PDB structure 1d5x [95]. b. A 20-residue peptide from myelin basic protein bound to HLA-DR2a heterodimer (composed of an α chain - II-ALPHA from HLA-DRA*0101 and a β chain - II-BETA from HLA-DRB5*0101) from the PDB structure 1fv1 [97].** **13**

The peptide and the MHC peptide binding clefts are coloured green and red, respectively. The peptide residues interacting with the MHC and the ‘pocket’ residues from the MHC peptide binding cleft are shown in ball and stick representation and are portrayed in yellow and blue, respectively, in **b.** to highlight the strong interactions along the length of the peptide nonamer within the

peptide binding cleft. The flattened conformation of the nonamer is clearly evident. The flanking residues extending out of the peptide binding cleft are labelled.

Figure 1.8 Domains in a TR. a. A cartoon depicting a typical $\alpha\beta$ TR, its various regions and domains. b. The variable and constant domains in a 21.30 TR from the TR/pMHC-II X-ray crystal structure 3mbe (PDB code; [114]). c. The $V\alpha$ and $V\beta$ domains in a M67 TR from the TR/pMHC-I X-ray crystal structure 3e2h (PDB code; [115]) rotated 180° along their interacting axis to show the CDR1, 2 and 3 loops. 15

In **a.** the variable and constant domains along with the transmembrane and cytoplasmic regions within the TR α chain (orange) and TR β chain (red), the plasma membrane and the cytosol are labelled. In **b.** the $V\alpha$, $V\beta$, $C\alpha$ and $C\beta$ domains are labelled and coloured red, yellow, blue and green, respectively. In **c.** pMHC interacting CDR1, 2 and 3 loops from $V\alpha$ domain are labelled and coloured pink, turquoise and yellow, respectively. Similarly, the pMHC interacting CDR1, 2 and 3 loops from $V\beta$ domain are labelled and coloured orange, red and green, respectively. The $V\alpha$ and $V\beta$ domains are also labelled in **c.**

Figure 1.9 The TR docking angles for TR/pMHC structures. a. The interacting region of the pMHC from the TR/pMHC-I structure 2e7l (PDB code; [132]) showing the “diagonal” TR docking angle (44° in this case) seen in most TR/pMHC-I complexes [130, 131]. b. The interacting region of the pMHC from the TR/pMHC-II structure 1d9k (PDB code; [130]) showing the “orthogonal” TR docking angle (71° in this case) seen in most TR/pMHC-II complexes [130, 131]. 17

The MHC-I G-ALPHA1 and G-ALPHA2 helices and MHC-II G-ALPHA and G-BETA helices are shown in red ribbon representation in **a.** and **b.** The cognate peptides are depicted in blue ribbon representation. Similarly, the green ellipses portray the orientation of the CDR1, 2 and 3 loops of the TR $V\alpha$ and $V\beta$ domains on the pMHC. The diagonal line (also in green) cutting

across the ellipse (and hence through the centre of the mass of TR V α and V β domains) shows the TR docking angle, with respect to the linear axis of the bound peptide, formed on the pMHC interface.

Figure 1.10 **A pictorial representation of the central CDR3-peptide region surrounded by the CDR1 and 2 loops that interact with the MHC helices in the TR/pMHC-I structure 3h9s (PDB code; [149]).** **19**

The V α and V β domains are labelled. The V α CDR1, 2 and 3 loops are labelled and coloured pink, turquoise and yellow, respectively. Similarly, the V β CDR1, 2 and 3 loops are labelled and coloured orange, red and green, respectively. The dotted blue ellipse represents the central CDR3-peptide region.

Figure 1.11 **A pictorial representation of a subset of the decision tree network utilized by Segal *et al.* [391].** **41**

Represented as each node is the grouping of preferential or non-preferential amino acid residues at various positions for the peptides binding to the murine MHC-I allele H2-Kb. The ellipses denote internal nodes and the rectangles depict terminal nodes. The numbers 0 or 1 signify the predictions non-binding (bright red) or binding (bright green), respectively, at each node.

Figure 1.12 **An example of the three-layer ANN derived by Brusica *et al.* [175] for predicting MHC-I restricted T cell epitopes.** **42**

The first layer (small red circles) represents input nodes with the number of nodes corresponding to the length of the input peptide (in this case 9-mer; AA stands for amino acid). The number of nodes in the second (hidden; blue circles) layer equals the ideal length of the binding peptides (usually set to 9 residues) and a single output node (green circle) predicts binders and non-binders.

Figure 1.13 An illustration of the first HMM topologies implemented for T cell epitope prediction [400]. a. A pHMM and b. A fully connected HMM. 45

The partial order of states and the lack of any given starting or terminating state in **a.** and **b.**, respectively, are evident.

Figure 4.1 A graphical depiction of the correlation between different computed structural interaction parameters for all pMHC-I complexes in MPID-T2. a. pMHC-I interface area vs. pMHC-I gap volume. b. pMHC-I interface area vs. pMHC-I gap index. c. pMHC-I interface area vs. pMHC-I BE. d. pMHC-I interface area vs. pMHC-I H-bonds. e. pMHC-I gap index vs. pMHC-I gap volume. f. pMHC-I gap index vs. pMHC-I H-bonds. g. pMHC-I gap index vs. pMHC-I BE. h. pMHC-I gap volume vs. pMHC-I H-bonds. i. pMHC-I gap volume vs. pMHC-I BE. j. pMHC-I H-bonds vs. pMHC-I BE. k. pMHC-I interface area vs. pMHC-I contact area. l. pMHC-I gap index vs. pMHC-I contact area. m. pMHC-I gap volume vs. pMHC-I contact area. n. pMHC-I H-bonds vs. pMHC-I contact area. o. pMHC-I BE vs. pMHC-I contact area. 201

The respective units are mentioned in the parentheses next to the names of the interaction parameters on the x and y-axes. The corresponding regression coefficients (r^2) are shown within each of the graphs.

Figure 4.2 A graphical illustration of the correlation between different computed structural interaction parameters for all pMHC-II complexes in MPID-T2. a. pMHC-II interface area vs. pMHC-II gap volume. b. pMHC-II interface area vs. pMHC-II gap index. c. pMHC-II interface area vs. pMHC-II BE. d. pMHC-II interface area vs. pMHC-II H-bonds. e. pMHC-II gap index vs. pMHC-II gap volume. f. pMHC-II gap index vs. pMHC-II H-bonds. g. pMHC-II gap index vs. pMHC-II BE. h. pMHC-II gap volume vs. pMHC-II H-bonds. i. pMHC-II gap volume vs. pMHC-II BE. j. pMHC-II H-bonds vs. pMHC-II BE. k. pMHC-II interface area vs. pMHC-II contact area. l. pMHC-II 203

gap index vs. pMHC-II contact area. m. pMHC-II gap volume vs. pMHC-II contact area. n. pMHC-II H-bonds vs. pMHC-II contact area. o. pMHC-II BE vs. pMHC-II contact area.

The respective units are mentioned in the parentheses next to the names of the interaction parameters on the x and y-axes. The corresponding regression coefficients (r^2) are shown within each of the graphs.

Figure 4.3 **A graphical portrayal of the correlation between different** **205**

computed structural interaction parameters for all TR/pMHC-I complexes in MPID-T2. a. TR/pMHC-I interface area vs. TR/pMHC-I gap volume. b. TR/pMHC-I interface area vs. TR/pMHC-I gap index. c. TR/pMHC-I interface area vs. TR/pMHC-I BE. d. TR/pMHC-I interface area vs. TR/pMHC-I H-bonds. e. TR/pMHC-I gap index vs. TR/pMHC-I gap volume. f. TR/pMHC-I gap index vs. TR/pMHC-I H-bonds. g. TR/pMHC-I gap index vs. TR/pMHC-I BE. h. TR/pMHC-I gap volume vs. TR/pMHC-I H-bonds. i. TR/pMHC-I gap volume vs. TR/pMHC-I BE. j. TR/pMHC-I H-bonds vs. TR/pMHC-I BE. k. TR/pMHC-I interface area vs. TR docking angle. l. TR/pMHC-I gap index vs. TR docking angle. m. TR/pMHC-I gap volume vs. TR docking angle. n. TR/pMHC-I H-bonds vs. TR docking angle. o. TR/pMHC-I interface area vs. TR/pMHC-I contact area. p. TR/pMHC-I gap index vs. TR/pMHC-I contact area. q. TR/pMHC-I gap volume vs. TR/pMHC-I contact area. r. TR/pMHC-I H-bonds vs. TR/pMHC-I contact area. s. TR/pMHC-I BE vs. TR/pMHC-I contact area. t. TR/pMHC-I contact area vs. TR docking angle.

The corresponding regression coefficients (r^2) are shown within each of the graphs. The respective units are mentioned in the parentheses next to the names of the interaction parameters along the x and y-axes.

Figure 4.4 **A graphical display of the correlation between different** **207**
computed structural interaction parameters for all TR/pMHC-
II complexes in MPID-T2. a. TR/pMHC-II interface area vs.
TR/pMHC-II gap volume. b. TR/pMHC-II interface area vs.
TR/pMHC-II gap index. c. TR/pMHC-II interface area vs.
TR/pMHC-II BE. d. TR/pMHC-II interface area vs.
TR/pMHC-II H-bonds. e. TR/pMHC-II gap index vs.
TR/pMHC-II gap volume. f. TR/pMHC-II gap index vs.
TR/pMHC-II H-bonds. g. TR/pMHC-II gap index vs.
TR/pMHC-II BE. h. TR/pMHC-II gap volume vs. TR/pMHC-
II H-bonds. i. TR/pMHC-II gap volume vs. TR/pMHC-II BE. j.
TR/pMHC-II H-bonds vs. TR/pMHC-II BE. k. TR/pMHC-II
interface area vs. TR docking angle. l. TR/pMHC-II gap index
vs. TR docking angle. m. TR/pMHC-II gap volume vs. TR
docking angle. n. TR/pMHC-II H-bonds vs. TR docking angle.
o. TR/pMHC-II interface area vs. TR/pMHC-II contact area.
p. TR/pMHC-II gap index vs. TR/pMHC-II contact area. q.
TR/pMHC-II gap volume vs. TR/pMHC-II contact area. r.
TR/pMHC-II H-bonds vs. TR/pMHC-II contact area. s.
TR/pMHC-II BE vs. TR/pMHC-II contact area. t. TR/pMHC-
II contact area vs. TR docking angle.

The corresponding regression coefficients (r^2) are shown within each of the graphs. The respective units are mentioned in the parentheses next to the names of the interaction parameters along the x and y-axes.

LIST OF TABLES

Table 1.1	List of generalized databases and resources used for the study of pMHC and TR/pMHC interactions	23
Table 1.2	List of specialized databases, resources and tools used in the study of pMHC and TR/pMHC interactions	25
Table 1.3	List of available tools and web-servers for T cell epitope prediction	34
Table 2.1	Methods, applications and publications	89
Table 4.1	List of pMHC structures in MPID-T2	115
Table 4.2	List of TR-pMHC structures in MPID-T2	156
Table 4.3	Computed pMHC interaction parameters for pMHC-I structures in MPID-T2	167
Table 4.4	Computed pMHC interaction parameters for pMHC-II structures in MPID-T2	187
Table 4.5	Computed TR/pMHC interaction parameters for TR-pMHC-I structures in MPID-T2	191
Table 4.6	Computed TR/pMHC interaction parameters for TR-pMHC-II structures in MPID-T2	194

LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

The following publications are presented in their published form in this thesis and are referred to from this point onwards as listed in respective sections of the thesis.

1. **Khan JM**, Tong JC, Ranganathan S: Structural Immunoinformatics: Understanding MHC-peptide-TR binding. In *Bioinformatics for Immunomics. Volume 3*. Edited by Davies MN, Ranganathan S, Flower DR. Springer, New York, Immunomics Reviews Series; 2010:77-94. **ISBN**: 978-1-4419-0539-0.
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 50%, JCT 20%, SR 30%; (iii) data analysis: JMK 70%, SR 30%; and (iv) writing: JMK 50%, SR 50%.
2. **Khan JM**, Ranganathan S. TR recognition of MHC-peptide complexes. In *Encyclopedia of Systems Biology, Systems Immunology*. Edited by Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H. Springer, New York, 2011. *In press*.
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 100%; (iii) data analysis: JMK 75%, SR 25%; and (iv) writing: JMK 75%, SR 25%.
3. **Khan JM**, Ranganathan S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res* 2010, **6**(Suppl 1):S2 (pp: 1-16).
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 100%; (iii) data analysis: JMK 75%, SR 25%; and (iv) writing: JMK 50%, SR 50%.
4. **Khan JM**, Cheruku HR, Tong JC, Ranganathan S. MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures. *Bioinformatics* 2011, **27**: 1192-1193.
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 50%, HRC 25%, JCT 25%; (iii) data analysis: JMK 50%, HRC 25%, SR 25%; and (iv) writing: JMK 50%, SR 50%.
5. **Khan JM**, Ranganathan S. Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one. *PLoS One* 2011, **6**(2):e17194 (pp: 1-12).
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 100%; (iii) data analysis: JMK 75%, SR 25%; and (iv) writing: JMK 50%, SR 50%.
6. **Khan JM**, Kumar G, Ranganathan S. *In silico* prediction of immunogenic T cell epitopes for HLA-DQ8. *Manuscript under review*.
Contributions to: (i) concept: JMK 50%, SR 50%; (ii) data gathering: JMK 100%; (iii) data analysis: JMK 50%, GK 25%, SR 25%; and (iv) writing: JMK 50%, SR 50%.

ABSTRACT

The adaptive immune system in higher jawed vertebrates carries out antigen presentation and recognition in two steps. Major histocompatibility complexes (MHC) first bind immunogenic peptide epitopes (p) derived from antigens and present them as peptide-MHC (pMHC) complexes, for subsequent recognition by T cell receptors (TR) leading to T cell activation. A decade after the first TR/pMHC structure was reported, the molecular basis of TR/pMHC interaction is still unknown. Peptide epitopes that bind strongly to MHC proteins are known to elicit T cell response, albeit with ~50% efficiency, forming the basis of T cell-based peptide vaccines. Experimental identification of these epitopes is a tedious, time consuming and expensive process. Computational methods are comparatively inexpensive and efficient in screening numerous peptides against their cognate MHC alleles. Sequence-based prediction methods are well established but are limited by the requirement of large datasets of known MHC-binding peptides. Structure-based prediction approaches, especially docking techniques, are universally applicable and specially suited for alleles with limited data.

For efficient vaccine design and to minimize experimental T cell binding assays, precise computational strategies for rapid prediction of high-binding epitopes for all alleles with a high propensity to activate T cells, are required. Our group has previously developed an accurate structure-based docking protocol, from which prediction models for identifying high-binders have been developed. However, this method is not fast enough to scan an entire proteome for large-scale pathogen screening studies. We also need to understand the physicochemical basis of TR binding to pMHC, to screen high-binders for greater TR binding potential and eliminate those that do not lead to T cell activation. These two specific aims are addressed in this thesis, and applied to predict true T cell epitopes amongst high-binders for a disease-implicated MHC allele.

pDOCK is a new fast, accurate and robust method for high-throughput screening of pathogenic sequences, based on flexible docking of peptides to MHC-I and MHC-II proteins. Compared to our earlier docking methodology, pDOCK shows upto 2.5 fold improvement in accuracy (7-fold compared to earlier published studies) and is ~60% faster. To dissect TR/pMHC interactions, I have collated and analysed 61 TR/pMHC crystallographic structures, available in the new database, MHC Peptide Interaction Database – version T2 (MPID-T2; <http://www.biolinfo.org/mpid-t2>). MPID-T2 is an

updated and extended version of the earlier MPID-T database, augmented with advanced features and new parameters for analysis of pMHC and TR/pMHC structures. Based on this analysis, I have defined criteria for selecting peptides with high probability to activate T cells. These criteria have been validated with published peptide mutation studies, where TR binding has been changed or abolished.

I have applied pDOCK and the TR binding criteria to predict “true” immunogenic epitopes from high MHC-binding peptides for celiac disease and insulin-dependent diabetes mellitus (IDDM) associated HLA-DQ8 allele. Our approach identified potential T cell epitopes, based on MHC and TR specificities, lacking conserved binding motifs, for experimental testing and validation. High prediction accuracy of HLA-DQ8-binding peptides was validated by existing experimental, biochemical and functional data. The bioinformatic approaches developed in this thesis are novel, generic and applicable for the development of effective immunotherapeutic and highly specific peptide vaccines with wide population coverage, capable of eliciting T cell response, thereby cutting down the lead time involved in experimental vaccine development protocols.

Chapter 1: Introduction and literature survey

1.1 Overview

The adaptive immune system plays a vital role in defending higher jawed vertebrates against infectious, allergic and graft vs. host diseases, while malfunctioning of this system leads to autoimmune diseases. The title “adaptive” suggests its ability to adapt and respond to an ever changing variety of new pathogens thereby conferring long-lasting or protective immunity to the host. For maximal immunological protection against this multitude of pathogens, the adaptive immune system carries out antigen presentation and recognition in two steps, where cell surface glycoproteins called major histocompatibility complexes (MHC) or human leukocyte antigens (HLA) in human, first bind antigenic peptide epitopes (p) and present them as peptide-MHC (pMHC) complexes on the surface of antigen-presenting cells (APC), for subsequent recognition by T cell receptors (TR), leading to TR/pMHC complex formation and eventually causing T cell activation [1-4].

The TR/pMHC interaction is relatively feeble compared to other important interactions between the molecules of the immune system [5, 6], yet strong enough to trigger TR mediated activation of T cells, thereby eliciting an immediate immune response to either destroy infected cells directly (*via* CD8⁺ cytotoxic T cells) or activate other immune system cells like B cells and macrophages (*via* CD4⁺ helper T cells) to carry out the immune response. Almost a decade and a half after the first TR/pMHC structure was reported [7], the molecular basis of TR/pMHC interaction is still unknown [8], due in part to the complexities of the proteins involved in this association. Therefore, uncovering the reasons for the specificity of TR/pMHC interactions and their mechanism remain an unsolved problem in understanding the physicochemical basis of TR binding to pMHC.

T cell epitopes are essential subunit peptide sequences that are required to stimulate cellular immune responses, especially the adaptive immune responses [9]. Peptide epitopes can be of endogenous (processed within the cell) or exogenous (processed outside the cell) origins, and these peptide epitopes are presented for surveillance and recognition by the TR in an MHC allele and supertype-dependent manner. Antigenic peptides that bind strongly to MHC alleles are known to elicit T cell responses [9-15]. Hence, their identification is a vital first step in the process of immune epitope prediction. Experimental identification of T cell epitopes is a tedious, time consuming and expensive process owing to the large

number and diversity of both MHC alleles and the antigenic peptides, especially in the light of the extremely low chance of immunogenicity (1 in 2000 peptides) even amongst the peptides that bind strongly to the MHC (50%) [16].

Recently developed computational methods have proven to be vastly efficient in time and cost, in screening the vast numbers of peptides and MHC repertoires [17, 18], as a first step towards T cell epitope prediction. Sequence-based prediction methods are well established but are limited by the requirement of large datasets of known MHC-binding peptides [10, 17, 18]. Structure-based prediction approaches, especially docking techniques, are universally applicable and specially suited for alleles with limited data [10, 11, 17, 18]. Our group has previously developed an efficient structure-based docking protocol [10, 11], from which prediction models for identifying high-binders were developed [11-14]. However, this method is not fast enough to scan an entire proteome for large-scale pathogen screening studies.

Also, a 50% chance of immunogenicity [16] means that only half of any given predicted set of high-binding peptides will eventually function as T cell activators. Hence, identifying the subset of peptides capable of T cell activation *via* TR recognition of pMHC complexes becomes the second step in T cell epitope prediction. Similar to the first step, this step also comes with its own impediments such as the complex structure of TR proteins and the incomplete characterisation of the molecular and physicochemical basis of TR/pMHC interaction. Therefore, for efficient vaccine design and to minimize experimental T cell binding assays, precise computational strategies for the rapid detection of high-binding epitopes, with a high propensity to activate T cells, are required.

In order to address these two steps efficiently using computational methods, a brief history on the discovery of MHC and TR, their genetic makeup, structure and function, pMHC binding, TR/pMHC interaction along with various issues, tools and resources currently available for T cell epitope prediction is first presented. Following this, the significance of studying TR/pMHC interactions in clinical medicine, and research objectives (issues addressed) are presented. The specific aims of this thesis and how they have been addressed forms the rest of the thesis, followed by conclusions and future directions.

1.2 Brief history of MHC and TR proteins

The MHC protein was first discovered in 1936 by the British immunologist, Peter Gorer [19, 20]. He later identified a blood group locus in mice and showed that blood type segregated with susceptibility and resistance to a transplantable tumour [21-23]. This was the first case of individual identification of a histocompatibility locus. He then went on to identify antibody response to tumour inoculation and detect cytotoxic activity of isoantibodies in mice [24, 25]. Later, the American geneticist, George Snell coined the term histocompatibility (H) antigen to describe cell-surface antigens provoking graft rejection [26]. He also demonstrated that differences at the H-2 gene locus provoked the strongest graft rejection of all the potential H antigens seen among various mouse strains [27, 28].

Snell's work on mice led to the discovery of HLA proteins by the French immunologist, Jean Dausset, in early 1950s, when he observed that patients who had multiple blood transfusions had antibodies (alloantibodies) to lymphocytes from other individuals, but not to their own lymphocytes [29]. Dausset went on to define the first HLA determinant in humans, which was the analogue of the murine H-2 complex. In 1969, pioneering research by the Venezuelan immunologist, Baruj Benacerraf, proved that these genes control the body's ability to respond to particular antigens by controlling the cellular responses among immune system cells, thereby, proving the significance of these genes in immune responses [30]. The term MHC was introduced in the early 1970s. Snell, Dausset and Benacerraf shared the 1980 Nobel Prize for physiology or medicine for the discovery of MHC.

Until the mid-1970s, T cell immunology was confounded with hypotheses ranging from the resemblance of a TR to a B cell antigen receptor, to theories about how a TR can recognize the pMHC complex. It was only during the twentieth century that MHC restriction was recognized, proving that the type of antigens recognized by T cells are different, compared to those recognized by B cells and that the scenario in which the former function is fundamentally different from the latter [31, 32]. Hence, discovering the molecular structure of a TR had become an extensively pursued field of research in the early 1980s. Aided by vast improvements in monoclonal T cell production technology, dedicated research groups led by Jim Allison, Ellis Reinherz, John Kappler and Philippa Marrack identified the two-chained $\alpha\beta$ TR protein as early as 1982-83, using murine antibodies [33-36].

Later, Steve Hedrick and Mark Davis together identified the murine TR β -chain employing molecular biology techniques [37-39]. Subsequently, in 1984, the human TR β -chain was identified by the Canadian immunologist, Tak Mak and colleagues [40]. Finally, in the very same year, Davis and co-workers identified the TR α -chain [41], while working on which, they accidentally stumbled upon another type of TR chain, which they labelled the TR γ -chain. The identification of TR γ -chain eventually led to the discovery of a second type of TR, the $\gamma\delta$ TR, which was previously unknown [32]. Identification of all the TR chains consequently resulted in rapid determination of the TR gene loci. The work presented in this thesis focuses on $\alpha\beta$ TR proteins. Therefore, the use of the abbreviation TR is restricted only to $\alpha\beta$ TR proteins.

1.3 Genetic makeup of MHC and TR proteins

The human MHC genes or HLA genes are located on chromosome 6. Due to the vital role played by the MHC proteins in defending against a vast majority of diverse pathogens, the MHC genes themselves must exhibit great variety. This is perhaps the reason as to why the MHC region is one of the densest regions in the mammalian genome. Currently, HLA genes are organized into three major classes or gene complexes, designated class I (MHC-I), II (MHC-II) and III (MHC-III; Fig. 1.1). MHC-III genes, located in between MHC-I and MHC-II genes (Fig. 1.1), primarily encode components of the serum complement system and proteins in other body fluids (e.g. C4, C2, factor B, TNF). MHC-I and MHC-II gene complexes, on the other hand, encode a number of highly polymorphic cell-surface proteins, responsible for antigen presentation.

The MHC-I gene complex is subdivided into three major loci, HLA-A, -B, and -C [1, 42] (Fig. 1.1) and other minor loci. Each major locus codes for a polypeptide; the α -chain of which contains antigenic determinants and is polymorphic. This α -chain, associates with a β -2-microglobulin chain, encoded by a gene outside the MHC complex and is expressed on the cell surface. The MHC-II HLA gene complex, referred to as HLA-D, is sub-divided into at least six loci, namely HLA-DR, -DQ, -DP, -DM, -DO, and -DZ [1, 42, 43], with HLA-DR, -DQ and -DP (Fig. 1.1) being the most expressed and common [44, 45] ones. MHC-I and MHC-II genes are the most polymorphic among all the genes in the human genome. Some of these genes have over 200 allelic variants identified to date. A single human individual expresses a finite number of MHC alleles and is heterozygous for each MHC gene, despite considerable MHC polymorphism.

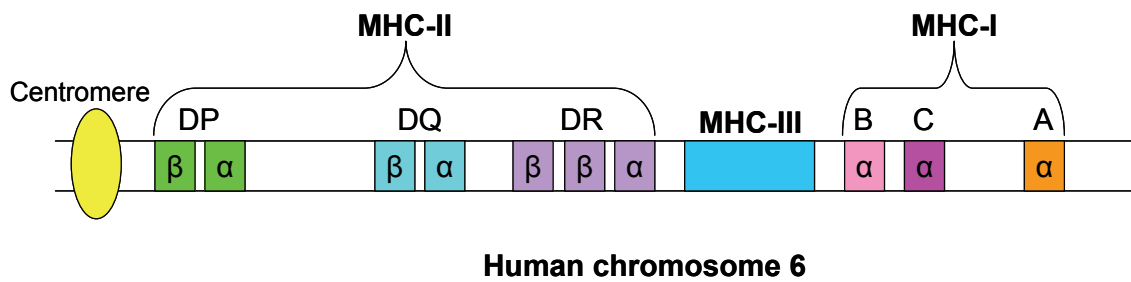


Figure 1.1: A schematic representation of the human MHC genes. The diagram shows the location of the genes that encode MHC-I and II proteins. MHC-I gene loci A (orange), B (rose) and C (pink) along with MHC-II gene loci DP (light green), DQ (turquoise) and DR (lavender) are shown. The centromere and the MHC-III loci are coloured yellow and light blue, respectively. The polypeptide chains coded for by the loci are shown within the boxes depicting different loci in MHC-I and II gene complexes.

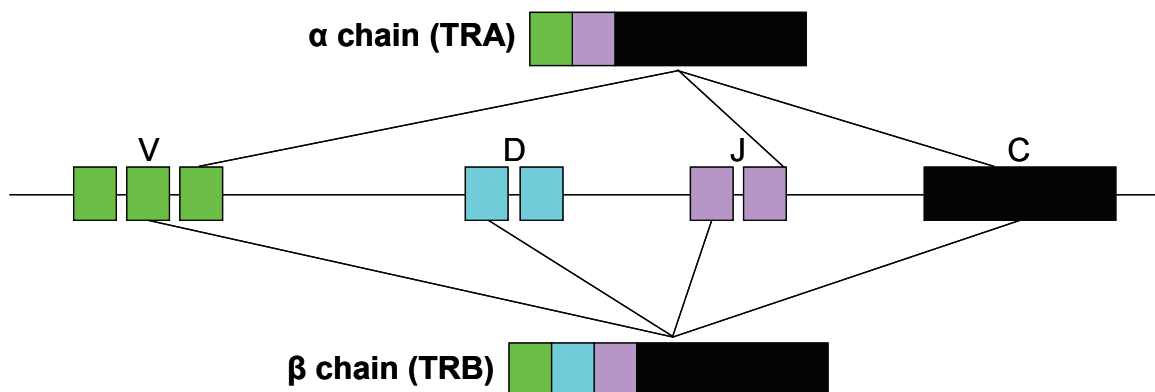


Figure 1.2: A simple example of the rearrangement that occurs during TR α (TRA) and β (TRB) chain formation. The V, D and J gene segments coding for the variable domain of the TR are shown in green, turquoise and lavender, respectively. The constant gene segment that codes for the constant region of the TR is shown in black.

The TR genes are formed by somatic rearrangement of germline gene segments [1] and resemble immunoglobulin (IG) genes in their structure and mechanisms of diversity generation. The array of gene segments that encode the α and β chains in a typical $\alpha\beta$ TR are located on different chromosomes [1, 3]. The TRA (encoding the α chain) and TRB (encoding the β chain) loci in human are located on chromosomes 14 and 7, respectively [3, 46]. Both these chains have constant and variable domains. The constant domains are encoded by the constant (C) gene segment. Similar to the IG heavy-chain (IGH) locus, the TR variable region locus contains separate variable (V), diversity (D) and joining (J) gene segments. These gene segments are brought together by site-specific recombination during T cell development in the thymus [1, 3, 46, 47]. V and J gene segments are present among both TRA and TRB loci. However, the D gene segments are present only in the TRB locus

[3, 48]. Thus, in a seemingly ordered process, one V gene segment, one D gene segment (only for β chain) and one J gene segment are randomly rearranged together, giving rise to a V-(D)-J gene (Fig. 1.2) which represents one of a multiple number of possible sequential recombinations, thereby, generating combinatorial diversity amongst TR proteins.

1.4 Structure and function of MHC

MHC proteins have evolved to protect higher jawed vertebrates from invading pathogens and virtually all substances bearing non-self antigens [1, 4, 15, 49]. As said earlier, peptide fragments of potential antigens are presented to circulating T cells (through TR/pMHC interaction, discussed later) by MHC-I and MHC-II proteins [3, 15, 49]. Hence their role in immune surveillance is extremely crucial. In general, recognition of pMHC complexes by T cells, via TR proteins, is aided by certain structural characteristics, critical for the role of MHC proteins in antigen presentation, shared amongst all MHC proteins [50, 51]. TR/pMHC complex formation, antigen recognition and T cell activation are said to be MHC restricted [8, 52-54], as TR proteins will only bind to antigenic peptides that are associated with MHC proteins. However, understanding how TR proteins recognized the pMHC complex required the first X-ray crystal structure of an MHC protein, which was achieved in 1987 [55, 56].

It is now clear that each MHC protein consists of an extracellular peptide binding cleft (Fig. 1.3) formed by paired binding groove α -helices resting on an eight-stranded anti-parallel β -sheet that forms the floor of the cleft. This peptide binding cleft or groove is above a pair of immunoglobulin (IG)-like regions or C-LIKE domains and is anchored to the cell membrane by transmembrane and cytoplasmic regions (Fig. 1.3) [1, 57, 58]. The binding groove of the MHC protein binds antigenic peptides for presentation on the APC cell surface where TR proteins interact with the displayed antigen and the helices of the MHC proteins [59]. The responsibility for different peptide binding specificities among different MHC alleles rests solely with the highly polymorphic amino acid residues located in and around this cleft [50]. T cell co-receptors, clusters of differentiation molecules, CD4 and CD8, bind to the non-polymorphic IG-like regions or C-LIKE domains of the MHC [1, 54, 60, 61]. These CD4 and CD8 co-receptors are expressed on the membranes of distinct subpopulations of mature T cells. They are known to play a considerably significant role in antigen recognition along with TR proteins. CD8 co-receptors bind specifically to MHC-I proteins and CD4 co-receptors bind to MHC-II proteins. Therefore, CD8⁺ T cells recognize

only pMHC-I complexes and CD4⁺ T cells recognize only pMHC-II complexes. CD8⁺ T cells function as cytotoxic T cells and CD4⁺ T cells are helper T cells.

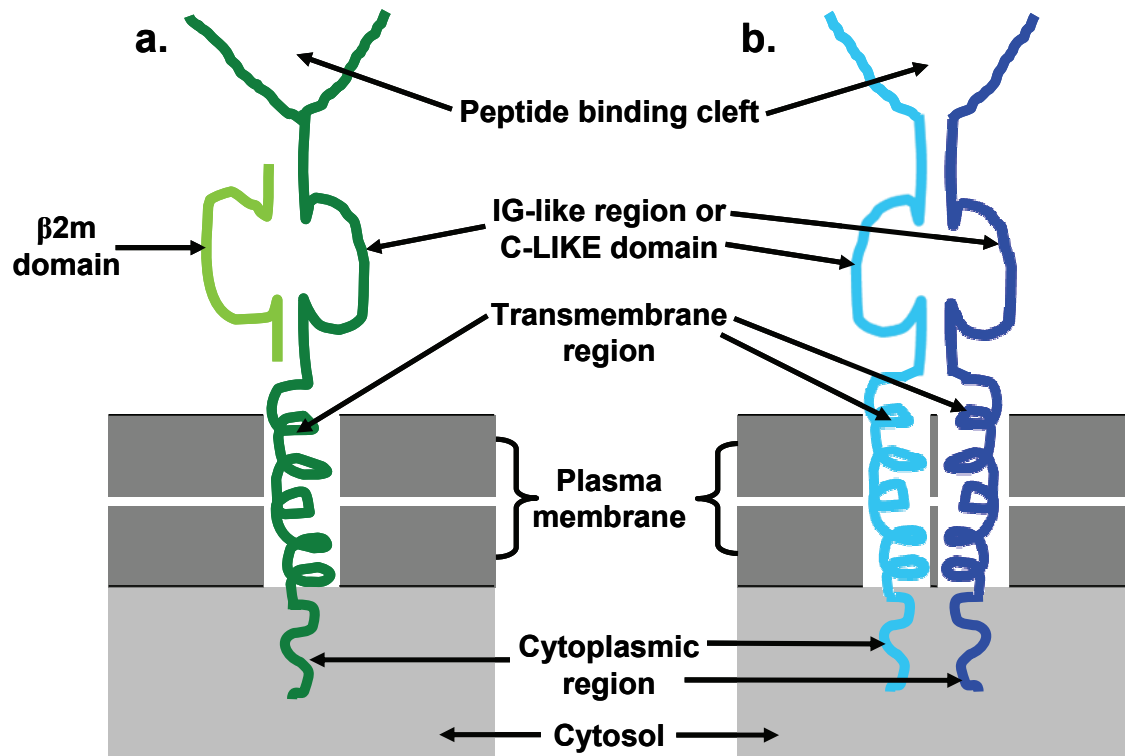


Figure 1.3: A cartoon depiction of typical MHC proteins. a. MHC-I. b. MHC-II. The α and β -2 microglobulin chains of MHC-I are coloured dark and light green, respectively in **a.** and the α and β chains of MHC-II are coloured dark and light blue, respectively in **b.** The peptide binding cleft, β 2m domain, C-LIKE domains, various regions, plasma membrane and the cytosol are labelled.

1.4.1 MHC-I

Typically, MHC-I proteins are ternary heterodimers. They consist of a heavy glycosylated transmembrane α chain (I-ALPHA in IMGT standardized abbreviations [50]) of roughly 45 kDa which is non-covalently linked to a smaller polypeptide light chain, β ₂-microglobulin (β ₂m), of about 12 kDa [10, 11]. The complete protein has four globular extracellular domains (Fig. 4a) and the connecting, transmembrane segment and a short cytoplasmic tail (Fig. 1.3a) that anchor the MHC onto the cell membrane and are usually excluded from the 3D X-ray crystal structures. The heavy α chain consists of α 1 (G-ALPHA1), α 2 (G-ALPHA2) and α 3 (C-LIKE) domains. The G-ALPHA1 and G-ALPHA2 domains form the peptide binding groove or cleft [59], as shown in Figure 1.4a. Both G-ALPHA1 and G-ALPHA2 domains have a similar structure. Beginning from the N-terminus, each domain forms four anti-parallel β -strands followed by a single α -helix across the β -strands. The association of the two domains is such that their β -sheets are

hydrogen-bonded to each other. This hydrogen-bonding results in the formation of a platform of a contiguous eight-stranded anti-parallel β -sheet which acts as the floor of the peptide binding cleft (Fig. 1.4b). There occurs a small propeller twist within this otherwise relatively flat β -sheet. The two α -helices from the G-ALPHA1 and G-ALPHA2 domains appear to form a boundary of the peptide binding groove on either side of the anti-parallel β -sheet (Fig. 1.4b). The C-LIKE α 3 domain is made up of an IG-like region. The β 2m forms the fourth domain and is located close to the C-LIKE domain (shown in Fig. 1.3a and Fig. 1.4a).

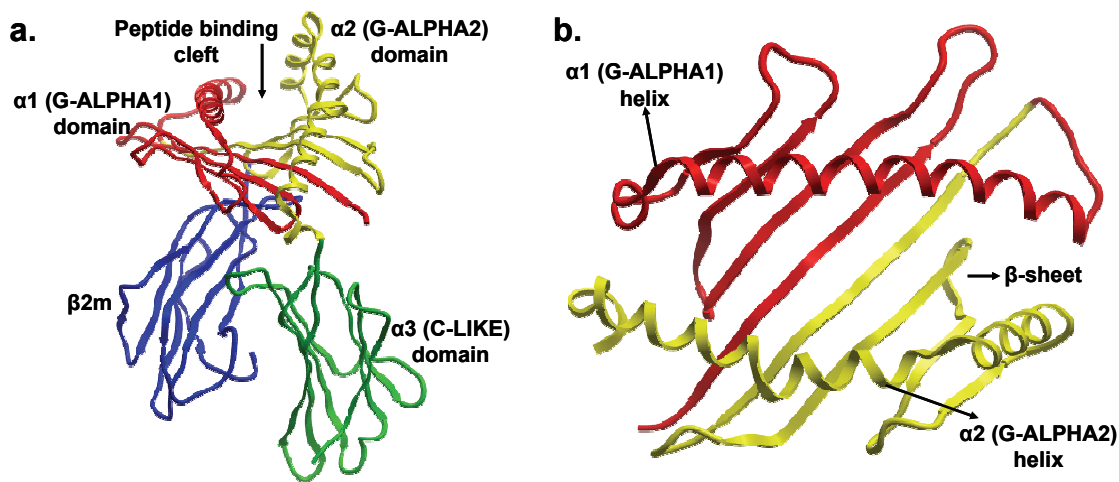


Figure 1.4: A ribbon representation of a MHC-I X-ray crystal structure (Protein Data Bank - PDB [62, 63] code: 1oga [53]). a. The four domains. b. An aerial view of the peptide binding cleft. In a. the α 1 (G-ALPHA1), α 2 (G-ALPHA2), α 3 (C-LIKE) and the β 2m domains are coloured red, yellow, green and blue, respectively, to clearly show the structure of the MHC-I protein. Highlighted in b. is the anatomy of the peptide binding cleft formed by the two α -helices on either side of the β -sheet which forms the floor.

Synthesized in the endoplasmic reticulum (ER) within the cells, MHC-I proteins are present on the surface of almost all nucleated cells, except neurons, in human [43]. Therefore, they are ubiquitously expressed by most cells [59, 64-66]. Aimed at detecting viral infections in cells, the MHC-I-restricted antigen processing and presentation pathway is a sophisticated surveillance mechanism. MHC-I proteins mainly function by binding peptides derived from endogenous antigens and then transporting them to the cell surface where they are presented for surveillance and recognition by the TR proteins of cytotoxic $CD8^+$ T cells. Most peptide ligands that bind MHC-I proteins are sourced from proteins that are degraded by proteases [67]. Exactly how the products of such endopeptidase activity are of such striking precision in terms of the length or size of the peptide ligand

that binds MHC-I proteins, remains an enigma. Perhaps, the proteases directly produce peptides of strikingly similar and appropriate size, or it could be that the proteases may generate longer peptides which are further processed to proportionate size by another biochemical mechanism. A lingering possibility of two short non-continuous peptide fragments being fused together to create the final MHC-I ligand, by means of post-translational protein splicing, also exists [68]. In any case, the transporter associated with antigen processing (TAP) proteins must transport these peptides from the cytosol into the ER and load them onto the MHC-I peptide binding groove in an ATP-dependent manner [67, 69]. What happens in the ER lumen to these transported peptides between their release from the TAP proteins to being loaded onto the MHC-I proteins, is also debatable. However, it is widely believed that the peptides are directly loaded onto MHC-I proteins immediately after release from the TAP proteins [70-73]. This would mean that the loaded peptides are either already of the correct size or they bind as longer peptides and are subsequently trimmed while being bound to the MHC-I proteins.

1.4.2 MHC-II

MHC-II proteins are also transmembrane heterodimeric glycoproteins consisting of two polypeptide chains, namely, an α chain (II-ALPHA; 34 kDa) and a β chain (II-BETA; 29 kDa) held together by non-covalent interactions and with very similar overall quaternary structure to that of MHC-I proteins [12-14, 74-76] (Figure 1.5). Similar to MHC-I proteins, the MHC-II proteins also have four globular extracellular domains, two on each chain, namely $\alpha 1$ (G-ALPHA), $\alpha 2$ (C-LIKE), $\beta 1$ (G-BETA) and $\beta 2$ (C-LIKE) domains (Fig. 1.5a) [59], and the connecting, transmembrane and cytoplasmic regions (Fig. 1.3b) that anchor the MHC onto the APC membrane and are also not present in the 3D X-ray crystal structures. However, their peptide binding groove is formed by the G-ALPHA and G-BETA domains of the α (II-ALPHA) and β (II-BETA) chains, respectively [59]. The G-ALPHA and G-BETA domains mimic the MHC-I G-ALPHA1 and G-ALPHA2 domains by forming the peptide binding cleft with two α -helices, one from each domain, forming the boundary on either side of a β -sheet floor (Fig. 1.5b).

The ER also synthesizes the MHC II proteins with two polypeptide chains α (II-ALPHA) and β (II-BETA) which are assembled and bound by the invariant chain (Ii) [77]. Unlike MHC-I proteins, which are expressed on most cells, MHC-II proteins are expressed on specific APC such as dendritic cells, endothelial cells, monocytes and B cells. MHC-II proteins specialize in binding exogenous antigenic peptides and presenting them at the

APC cell surface for surveillance and recognition by the TR of the CD4⁺ helper T cells. The MHC-II foreign peptide presentation pathway occurs in various steps. At first the antigen is ingested into the APC cytosol and degraded enzymatically into peptide fragments by endosomes and lysosomes. Unlike the MHC-I proteins, in the MHC-II peptide presentation pathway, the binding fragments of Ii prevent the loading of the peptide by binding onto the peptide binding cleft of the MHC-II proteins in the ER.

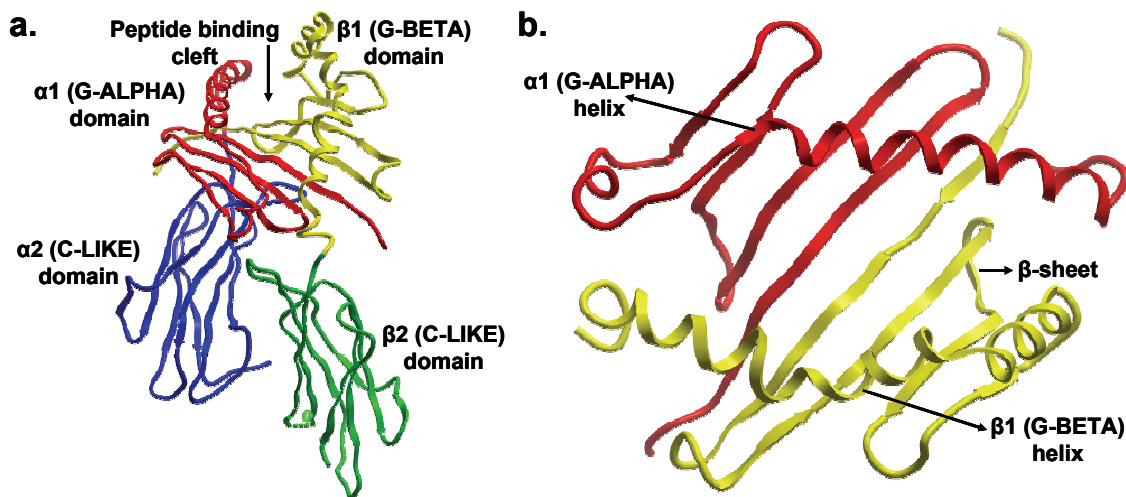


Figure 1.5: A schematic ribbon representation of a MHC-II X-ray crystal structure (PDB code: 1ymm [78]). a. The four domains. b. An aerial view of the peptide binding cleft. In a. the $\alpha 1$ (G-ALPHA), $\alpha 2$ (C-LIKE), $\beta 1$ (G-BETA) and $\beta 2$ (C-LIKE) domains are coloured red, blue, yellow and green, respectively, clearly illuminating the structure of the MHC-II protein. Illustrated in b. is the anatomy of the peptide binding cleft formed by the two α -helices sitting on either side of the β -sheet which forms the floor.

Meanwhile, Ii targets the MHC-II protein to a lysosomal-like compartment termed MHC-II compartment (MIIC) [79, 80]. As a result of the combined action of proteolytic enzymes and HLA-DM protein, Ii is removed from MHC-II proteins within the MIIC. Finally, the degraded peptide antigens bind to the now available peptide binding cleft of the MHC-II proteins. Consequently, the freshly loaded pMHC-II complexes are transported to the APC cell surface, where, recognition by the helper T cells results in the production of cytokines, which stimulate other immune system cells such as B cells and macrophages to carry out the immune response.

1.5 pMHC binding

Cellular immune responses, especially the adaptive immune responses are stimulated by essential subunit peptides called as immunogenic antigens or T cell epitopes. These

immunogenic epitopes are presented for surveillance and recognition by the TR in an MHC allele- (polymorphic MHC proteins) and supertype- (groups of MHC proteins with similar peptide binding properties) dependent manner [81, 82] and can be of endogenous or exogenous origins, as alluded to earlier. Various structural features and interaction parameters have now been characterised for pMHC-I and pMHC-II binding [15, 50, 83-85]. The overall physicochemical properties of these interactions remain the same across pMHC-I and pMHC-II complexes. For example, the nine residues (nonameric core) of the MHC-II peptides that bind within the peptide binding cleft mimic the normal length of the MHC-I peptides [11, 15, 49]. However, the pMHC binding criteria vary slightly for pMHC-I and pMHC-II complexes, particularly in the presence and contribution of the flanking residues (extending outside the peptide binding cleft) in pMHC-II binding [10, 11, 15, 49]. Today, these structural descriptors are widely used in the study of pMHC interactions to decipher the peptide binding preferences of different MHC alleles.

1.5.1 pMHC-I

Residues from both the peptides and the binding groove of the MHC proteins mediate the non-covalent interactions that facilitate peptide binding to MHC-I proteins [50, 51]. Usually, peptides of about eight to eleven amino acids in length are presented by MHC-I proteins [10, 15, 49-51]. In very rare cases, this range extends on either side such that peptides from seven to fourteen residues bind to MHC-I proteins. In any case, most of the peptide residues are bound in an extended conformation within the peptide binding groove (Fig. 1.6a) [49-51]. The polymorphic ‘pocket’ residues within the peptide binding cleft of the MHC-I proteins have side-chains that can accommodate and subsequently bind to the complementary amino acid residues of the peptides (Fig. 1.6a). Hence, the peptide binding cleft can be subdivided into various pockets (A to F) [86].

There are highly polymorphic residues around the N and C-termini of the peptides (Fig. 1.6a). These residues are called anchor residues due to their vital role in ‘anchoring’ the peptide firmly within the peptide binding cleft and thus, contribute greatly not only towards pMHC complex formation but also to their presentation and recognition by TR proteins since strong-MHC-binding peptides are known to elicit T cell responses [9-15, 49, 87, 88]. Therefore, the polymorphic MHC residues that line these pockets within the peptide binding cleft along with the polymorphic peptide residues, determine the individual specificity of a given pMHC interaction. Typically, there are two anchor residues at each terminus. These termini of the peptide are bound by a set of conserved hydrogen bonds

[89] causing them to bury deep into the cleft cavity. However, this burial arrangement of the terminal residues, fascinatingly, does not affect the length of the peptide binding across the cleft. Longer peptides though, may choose from a zigzag conformation [90] to a bulged orientation (Fig. 1.6b) [91-93] within the cleft, to allow peptides of greater length maintain the relative position of their terminal or anchor residues.

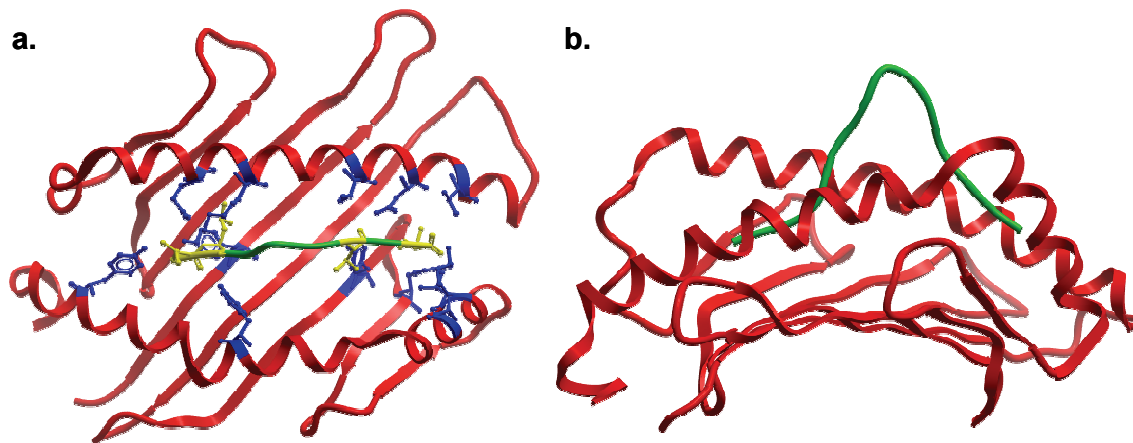


Figure 1.6: Different conformations adopted by MHC-I binding peptides. a. A nonameric Tax peptide bound in a flattened conformation to HLA-A*0201 from the PDB structure 1duz [94]. b. A 13-residue peptide bound in a bulged fashion to HLA-B*3508 from the PDB structure 2ak4 [93]. The peptide and the MHC peptide binding clefts are coloured green and red, respectively. The N and C-terminal peptide ‘anchor’ residues and the ‘pocket’ residues from the MHC peptide binding cleft are shown in ball and stick representation and are portrayed in yellow and blue, respectively, in **a.** to highlight the strong interactions around the peptide termini.

1.5.2 pMHC-II

The peptides presented by MHC-II proteins are generally twelve to twenty amino acids in length [10-15, 49-51]. Similar to MHC-I proteins, MHC-II proteins can also have exceptions in the lengths of the peptides that bind to their peptide binding cleft. Again, in extreme cases the above mentioned range of twelve to 25 amino acids can be extended on either side to accommodate a spectrum of peptides with lengths ranging from six (Fig 1.7a) [95] to 30 amino acids [43]. Again, akin to MHC-I proteins, the peptide binding groove of MHC-II proteins can also be subdivided into a series of pockets (1 to 9) [76, 96, 97]. Specifically, MHC-II proteins form hydrogen bonds with peptide side chain atoms along the length of the peptide nonamer (Fig. 1.7b) that is bound to the peptide binding cleft [49-51, 85, 97], quite unlike pMHC-I binding, where the allele-independent hydrogen bonding between the MHC and the peptide is focused around the N- and C-termini or anchor residues of the peptide. MHC-II proteins also make contacts with the atoms forming the

peptide main chain for the nonamer within the peptide binding cleft [43, 85]. The bound conformation of the nonamer within the groove is usually flattened (Fig 1.7b).

This liberal nature of the MHC-II binding cleft enforces no absolute constraints on the complete length of the peptide that can bind to their grooves. The additional residues of the peptide beyond the nonameric core, called flanking residues, generally extend out of the peptide binding cleft (Fig. 1.7b), on either side in most cases, and do not strictly adopt any particular conformation. Yet, they contribute considerably to pMHC-II binding [10, 11, 15, 49] and differentiate it greatly from pMHC-I binding.

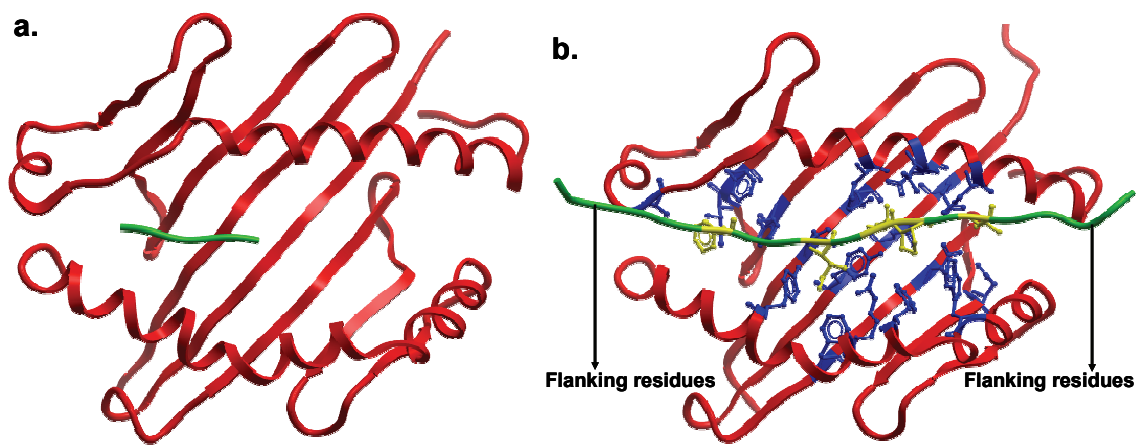


Figure 1.7: Diversity in the lengths of peptides binding to MHC-II proteins. a. A 6-residue peptidomimetic peptide bound to HLA-DRB1*0401 from the PDB structure 1d5x [95]. b. A 20-residue peptide from myelin basic protein bound to HLA-DR2a heterodimer (composed of an α chain - II-ALPHA from HLA-DRA*0101 and a β chain - II-BETA from HLA-DRB5*0101) from the PDB structure 1fv1 [97]. The peptide and the MHC peptide binding clefts are coloured green and red, respectively. The peptide residues interacting with the MHC and the 'pocket' residues from the MHC peptide binding cleft are shown in ball and stick representation and are portrayed in yellow and blue, respectively, in **b**, to highlight the strong interactions along the length of the peptide nonamer within the peptide binding cleft. The flattened conformation of the nonamer is clearly evident. The flanking residues extending out of the peptide binding cleft are labelled.

1.6 Structure and function of TR

TR proteins are arguably as important a part of T cell-dependent immune responses as the MHC proteins. Since the groundbreaking isolation of the genes encoding these vital components of adaptive immune system, more than a quarter of a century ago [40, 41], well over 30,000 articles have been published highlighting their structure, function, interaction with pMHC complexes and various other aspects of their biology [98]. This

vast attention directed towards the TR could be attributed to the fact that specificities within TR proteins render faithful abilities to T cells for distinguishing self-antigens from non-self antigens [98-100]. Thereby, exercising self-tolerance (preventing normal cells from being destroyed) and ensuring a successful immune cascade. Hence, TR proteins are the focal point of research into autoimmune diseases like multiple sclerosis [99] and various other immune system related ailments such as melanoma [101] and multiple myeloma [102-107].

Given these important facts about the vital functionality of a TR protein in T cell mediated immune responses, an essential component of research in immunology is to study the structure of a TR for more insights into its functions and to acquire the knowledge of how exactly it performs its function. As mentioned earlier, a typical $\alpha\beta$ TR has two chains, α and β , each divided into two extracellular domains called constant (encoded by the conserved constant (C) gene segment of the TR coding genes) and variable domains (encoded by rearranged variable (V), diversity (D) and joining (J) gene segments, V-J for α chain and V-D-J gene segments for β chain, respectively) [3, 46-48, 108], which are followed by a transmembrane and a short cytoplasmic region that anchor the respective chains and subsequently the TR onto the T cell surface (Fig. 1.8a).

The constant and variable domains perform specific functions and are generally present in the crystal structures of the TR proteins, which lack the transmembrane and cytoplasmic regions. The two conserved or constant TR domains ($C\alpha$ and $C\beta$; Fig. 1.8a, b) of both α and β chains [109, 110] are linked to the upper more diverse or variable domains ($V\alpha$ and $V\beta$; Fig. 1.8a, b), containing the CDR (complementarity determining region) 1, 2 and 3 loops (Fig. 1.8c) which recognize the pMHC at the TR/pMHC binding interface [111]. Interestingly, the overall structural assembly, domain organization and chain-fold of the TR proteins that recognize both pMHC-I and pMHC-II complexes are strikingly similar. The only significant difference that could contribute to pMHC-I or pMHC-II specificities of a given TR is the amino acid sequence variation of the $V\alpha$ and $V\beta$ CDR1, 2 and 3 loops.

The function of TR proteins is similar to certain cell surface receptors of B cell mediated immune responses, such as Fc receptors, found on the surface of macrophages or neutrophils, which bind to antigen-bound antibody, resulting in phagocytosis and lysis of the antigen or pathogen by macrophages or neutrophils [1, 112, 113]. However, the function of TR proteins differs from B cell mediated cell surface receptors in that, upon

TR/pMHC complex formation, the TR proteins do not actually cause the T cells to ingest and break down the pathogen. Instead, they trigger T cells to destroy the infected cells either directly (*via* CD8⁺ cytotoxic T cells) or indirectly (*via* CD4⁺ helper T cells).

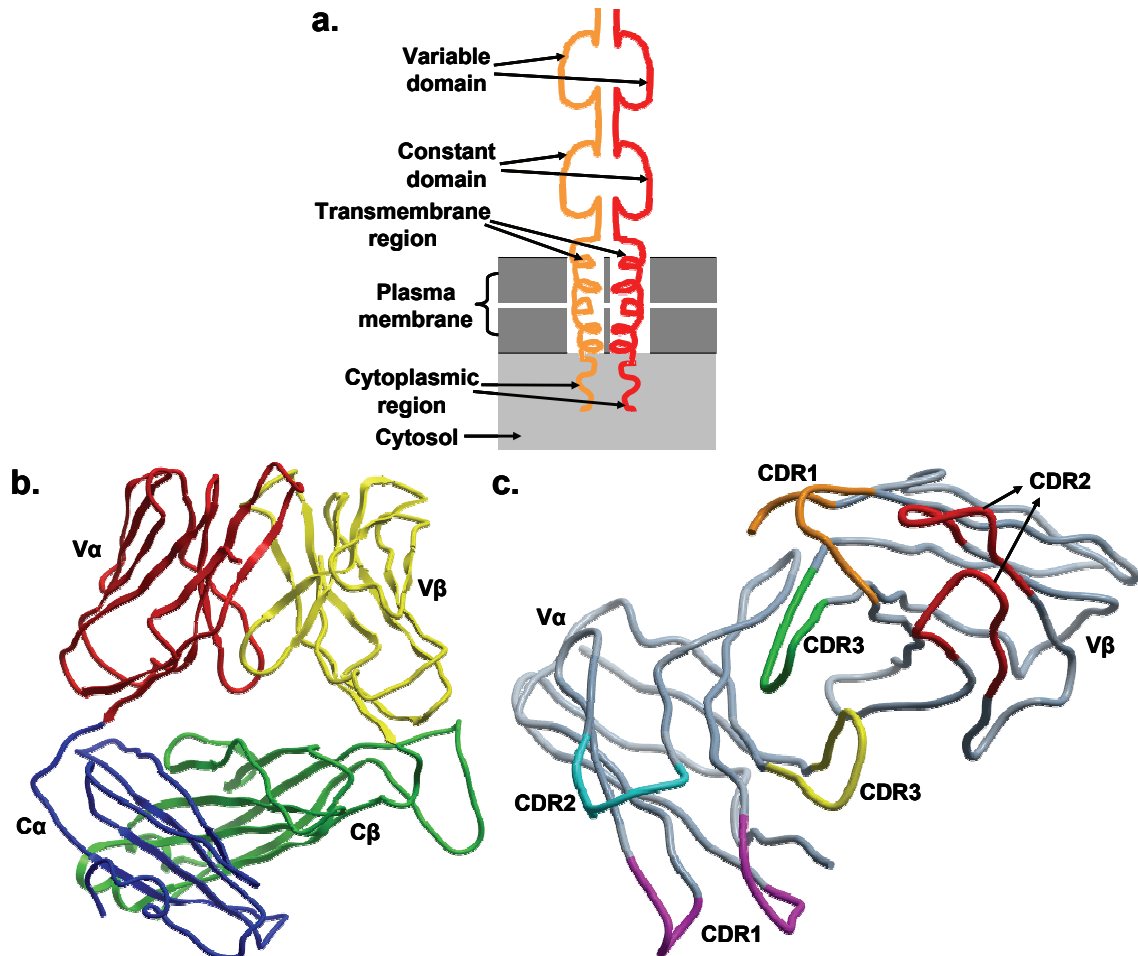


Figure 1.8: Domains in a TR. a. A cartoon depicting a typical $\alpha\beta$ TR, its various regions and domains. **b.** The variable and constant domains in a 21.30 TR from the TR/pMHC-II X-ray crystal structure 3mbe (PDB code; [114]). **c.** The $V\alpha$ and $V\beta$ domains in a M67 TR from the TR/pMHC-I X-ray crystal structure 3e2h (PDB code; [115]) rotated 180° along their interacting axis to show the CDR1, 2 and 3 loops. In **a.** the variable and constant domains along with the transmembrane and cytoplasmic regions within the TR α chain (orange) and TR β chain (red), the plasma membrane and the cytosol are labelled. In **b.** the $V\alpha$, $V\beta$, $C\alpha$ and $C\beta$ domains are labelled and coloured red, yellow, blue and green, respectively. In **c.** pMHC interacting CDR1, 2 and 3 loops from $V\alpha$ domain are labelled and coloured pink, turquoise and yellow, respectively. Similarly, the pMHC interacting CDR1, 2 and 3 loops from $V\beta$ domain are labelled and coloured orange, red and green, respectively. The $V\alpha$ and $V\beta$ domains are also labelled in **c.**

Another important thing to note about the function of TR proteins is that they are aided by co-receptors such as CD proteins [116, 117]. These CD proteins convey intracellular signals that are triggered when a TR engages with a pMHC [116-118]. Nevertheless, it is the recognition of pMHC complexes by the TR proteins that activates the T cells resulting

in the production or secretion of cytokines by the activated T cells [119]. The cytokines secreted by CD8⁺ T cells cause their differentiation into cytolytic or cytotoxic effector T cells (CTL), while the cytokines secreted by CD4⁺ helper T cells support their differentiation into effector helper T cells [120-122]. This phenomenon initiates the effector response and sheds light on the significance of a TR in the entire adaptive immune response mechanism.

1.7 First crystal structures of TR/pMHC complexes

Similar to the acceleration in research for identifying the structures of a MHC and a TR protein, the race to solve the crystal structure of a TR/pMHC complex began in the early 1990s. Diligent efforts were made in October 1996 by Garcia and co-workers [123], when they solved a crystal structure of a murine 2C TR (PDB code: 1tcr) and proposed its orientation or bound conformation to a pMHC-I complex from TR/pMHC crystals. Using this model, they were able to explain the positional orientations of CDR1, 2 and 3 loops of the V α and V β domains from the 2C TR. Shortly thereafter, in November 1996, the complete X-ray crystal structure of a TR/pMHC complex was solved by Garboczi *et al.* [7, 124], where they reported a TR/pMHC-I complex (PDB code: 1ao7) between the human A6 TR and a tax peptide from the human T cell lymphotropic virus HTLV-1 bound to HLA-A2*0201 allele.

Subsequently, the continual efforts of Garcia and co-workers [125] yielded results in 1998 when they crystallized a TR/pMHC-I structure between the murine 2C TR and the pMHC-I complex formed by the murine MHC-I H2-Kb allele and dEV8 peptide (PDB code 2ckb). The pioneering work from the Garcia and Garboczi's groups generated a lot interest among crystallographers and immunologists, who then began to work together to solve crystal structures of TR/pMHC complexes, in order to better understand the fundamental principles underlying TR recognition of pMHC complexes, TR/pMHC complex formation and hence explain T cell activation. Following in the footsteps of Garboczi and co-workers, Ding *et al.* [126] reported another crystal structure in 1998, between the pMHC-I complex containing the tax peptide and HLA-A*0201 allele and a different human TR known as B7 (PDB code 1bd2). This marked the beginning of an expansion in TR/pMHC structural data through the early 2000s. To date, there are 62 TR/pMHC structures reported for which crystal structures are available in the PDB. One of these (PDB code: 2icw; [127]) has a superantigen mediating the TR and pMHC binding and is thus not strictly a TR/pMHC complex.

1.8 TR/pMHC interaction

The mechanisms of combinatorial diversity and N-diversity of the variable domains of TR that create 10¹² TR per individual [3], the very high number of MHC alleles and most of all, a vast number of antigenic peptides together with the structural and functional complexities of these proteins (explained in the earlier sections) involved in this vital immunological synapse, make the underlying mechanism responsible for the specificity of TR/pMHC interactions, an elusive but extremely interesting area of research. As one would expect, right from the time the first crystal structures of TR/pMHC complexes were reported [7, 124-126], the elusive nature of this machinery's specificities have led researchers to propose various theories as probable concepts or reasons that direct and dictate these interactions. These theories range from the TR "germline bias," in which TR/pMHC binding is independent of the nature of the peptide and MHC restriction or TR specificity is based on specific conserved contacts between TR V (variable) domains and MHC proteins that co-evolve [128, 129], to the role of "diagonal" (below 70°; Fig. 1.9a) and "orthogonal" (above 70°; Fig. 1.9b) angle of TR binding or docking onto the pMHC in determining pMHC-I and pMHC-II specificities, respectively, for TR proteins [130, 131].

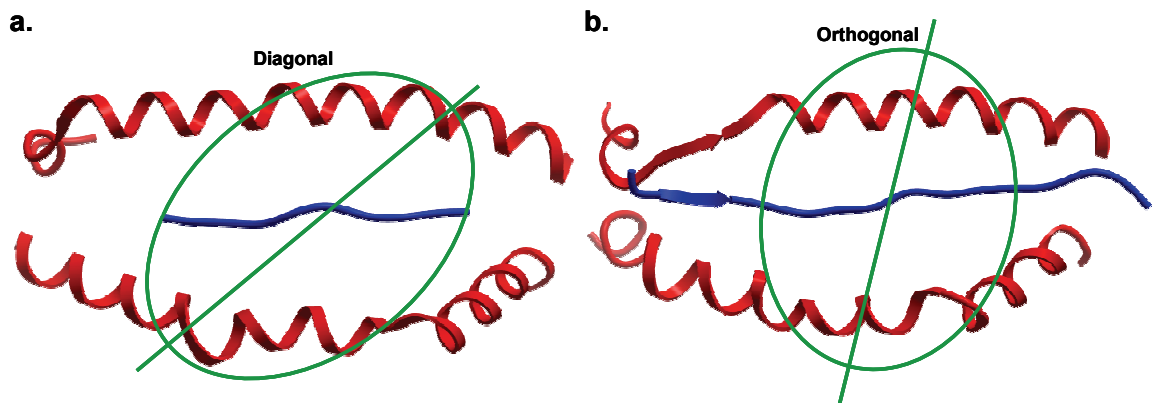


Figure 1.9: The TR docking angles for TR/pMHC structures. a. The interacting region of the pMHC from the TR/pMHC-I structure 2e7l (PDB code; [132]) showing the “diagonal” TR docking angle (44° in this case) seen in most TR/pMHC-I complexes [130, 131]. **b.** The interacting region of the pMHC from the TR/pMHC-II structure 1d9k (PDB code; [130]) showing the “orthogonal” TR docking angle (71° in this case) seen in most TR/pMHC-II complexes [130, 131]. The MHC-I G-ALPHA1 and G-ALPHA2 helices and MHC-II G-ALPHA and G-BETA helices are shown in red ribbon representation in **a.** and **b.** The cognate peptides are depicted in blue ribbon representation. Similarly, the green ellipses portray the orientation of the CDR1, 2 and 3 loops of the TR Vα and Vβ domains on the pMHC. The diagonal line (also in green) cutting across the ellipse (and hence through the centre of the mass of TR Vα and Vβ domains) shows the TR docking angle, with respect to the linear axis of the bound peptide, formed on the pMHC interface.

However, these explanations appear simplistic since, apart from the combinatorial issues described above, cross-reactivity of MHC and TR proteins [133-135] that effectively results in brief encounters between a TR protein and several pMHC complexes before the TR protein actually interacts with a specific pMHC complex, possibly explaining the feeble TR/pMHC interactions alluded to earlier, also adds to an impediment. Moreover, the significant role played by the peptide in determining the specificities of TR/pMHC interactions is widely accepted [133, 136-139]. Also, exceptional TR/pMHC-I [140] and TR/pMHC-II [99, 114] crystal structures have been reported with unusually “orthogonal” and “diagonal” TR docking angles, respectively, despite the fact that “diagonal” and “orthogonal” TR docking angles are the most common and conserved among TR/pMHC-I and TR/pMHC-II complexes, respectively [130, 131].

The increasing number of TR/pMHC X-ray crystal structures in PDB [62, 63] and in IMGT/3Dstructure-DB (<http://www.imgt.org/3Dstructure-DB/>) [57, 58], the reference database for immunoglobulins, T cell receptors, MHC and pMHC structures, has resulted in the identification of many structural characteristics that are common for most TR/pMHC interactions. Among these, two prominent characteristics are: (i) the common docking orientation or geometry of the TR proteins on pMHC complexes [53, 54, 139-141]; and (ii) the CDR3 loops of TR V α and V β domains, positioned in the center of TR/pMHC binding interface where they contact the peptide, whereas the CDR1 and CDR2 loops of TR V α and V β domains contact the tops of the MHC binding groove helices (G-ALPHA1 and G-ALPHA2 for pMHC-I and G-ALPHA and G-BETA for pMHC-II complexes), surrounding the central CDR3-peptide region like a “gasket” [8, 132, 142-146] (Fig. 1.10).

These characteristics are thus, on the whole, similar for both TR/pMHC-I and TR/pMHC-II interactions. Yet, there have been differences observed between most TR/pMHC-I and TR/pMHC-II interactions within these common characteristics. For example, the “diagonal” and “orthogonal” angle of TR docking observed for TR/pMHC-I and TR/pMHC-II structures [130, 131], respectively, although the orientations of the TR proteins on pMHC complexes are overall similar for both TR/pMHC-I and TR/pMHC-II complexes. Also, TR/pMHC structures [147, 148] have recently been identified where CDR1 and 3 loops from the TR V α and V β domains make extensive contacts with the peptide, which again, is an exception compared to most TR/pMHC complexes. Hence, the

above mentioned complexities are compounded with the overall commonalities found, suggest the increasing importance for an in-depth analysis over a broad spectrum of data to understand the minute physicochemical aspects of this vital binding.

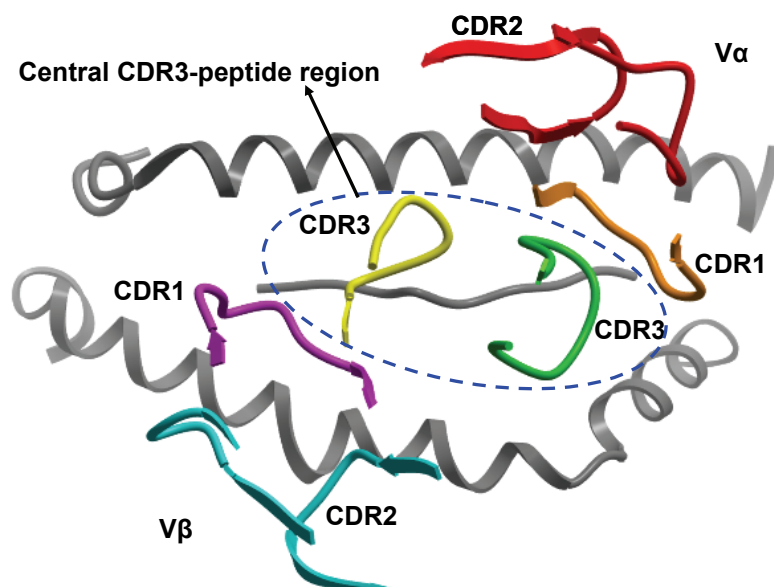


Figure 1.10: A pictorial representation of the central CDR3-peptide region surrounded by the CDR1 and 2 loops that interact with the MHC helices in the TR/pMHC-I structure 3h9s (PDB code; [149]). The V α and V β domains are labelled. The V α CDR1, 2 and 3 loops are labelled and coloured pink, turquoise and yellow, respectively. Similarly, the V β CDR1, 2 and 3 loops are labelled and coloured orange, red and green, respectively. The dotted blue ellipse represents the central CDR3-peptide region.

1.9 Issues with T cell epitope prediction

Apart from the complexities highlighted above, the identification of T cell epitopes is inundated with a number of intrinsic issues and difficulties. There occurs a great diversity in HLA genes among human population [150-155] with over 6000 known alleles or variants identified as of January 2011 (<http://www.ebi.ac.uk/imgt/hla/stats.html>) [156]. Peptide epitopes that bind strongly to MHC proteins are known to elicit T cell response, albeit with ~50% efficiency [16], forming the basis of T cell epitope prediction and hence T cell-based peptide vaccines. As mentioned earlier, the high polymorphism of HLA alleles along with allele specificity of candidate peptides [157], present the biggest obstacle in determining high-binders to a particular allele. Peptide binding studies have identified that each HLA allele possesses a unique spectrum of peptide binding specificities that limits them to choose from only a restricted set of peptides [158].

It has also been shown that strong and efficient pMHC binding is essential for immunogenicity or T cell activation [158]. At the same time, evidence indicating that efficient pMHC binding does not necessarily guarantee immunogenicity, also exists [16, 159]. Therefore, the binding of antigenic peptides to specific MHC alleles becomes a vital rate-limiting step in the process of T cell activation. Experimental identification of strong-binding peptides for every allele or T cell epitopes is a tedious, time consuming and forbiddingly expensive process, not suitable for application in studies involving large numbers of protein sequences or large-scale pathogen proteome studies [160-165]. Despite the recent increase in experimental data for HLA-binding peptides in databases such as IEDB (Immune Epitope Database; <http://www.immuneepitope.org/>) [166-169] and SYFPEITHI [170, 171] (<http://www.syfpeithi.de/>), there are a number of HLA variants for which experimental data is either unavailable or limited.

Robust computational approaches with tolerance for imprecision, errors, data bias, uncertainty, partial truth and limited amount of available data, are necessary and in particular demand to successfully accelerate the T cell epitope discovery process [17, 172], as imprecision, errors and biases prevail in currently available experimental data. Nevertheless, the accuracy of a T cell epitope prediction method or model is highly dependent on the quantity and quality of existing experimental data from biochemical and immuno-assays. Therefore, problems related to peptide data have significant implications on the selective ability and performance of the prediction model. A few major issues related to peptide data are described below.

1.9.1 Quantity of peptide data

The choice and quality of the prediction model is directly affected by the availability of considerable numbers of known peptide binders to specific alleles. As mentioned above, there is a vast need for experimental peptide binding data from biochemical studies for many HLA alleles. Structure-based predictive techniques (discussed later), especially docking methodologies, are usually preferred, due to their robustness, when little or no peptide data is available. For many years now, innate complexities involved with developing protocols for high-throughput screening of peptides, model building, data fitting and computational speed have severely hampered the development of computational tools for use in structure-based prediction methodologies. However recent advances in peptide docking protocols [10, 11, 49], can be harnessed for successful application of structure-based techniques in T cell epitope prediction.

Sequence-based predictive techniques (discussed later) however, are more useful as the number of available known peptide binders increases. Sequence-based methods using SVM (Support Vector Machines) outperform ANN (Artificial Neural Networks) based and decision tree based approaches when employed on a relatively small training dataset of 36 binders and 167 non-binders [173], although with increasing peptide data, ANN-dependent protocols are known to have a better predictive performance compared to that of methods using motifs, matrices and HMM (Hidden Markov Models) for T cell epitope prediction [17]. For MHC alleles with more than 100 known binders, ANN and HMM are the predictive methods of choice [17].

1.9.2 Quality of peptide data

The development of robust, generic, efficient and useful predictive models has always been impaired by the presence of noise and errors in accessible datasets. The role and impact of errors and noise in datasets on building predictive models using various sequence-based approaches, especially matrix-based methods, is well documented [174-177]. It has been shown that even a nominal 5% error in a dataset could potentially double the number of data points compared to a relatively ‘clean’ dataset, required to construct a reasonably accurate matrix-based models [174]. Ironically, the ability and performance of ANN and HMM based models to handle incomplete or inaccurate data is not significantly influenced by similar degrees of error [176-178]. These results again support the choice of ANN and HMM based methods to develop predictive models provided a high quantity of biochemical data is used.

1.9.3 Bias in peptide data

Another important factor that has a significant impact on the predictive ability of a model is the bias in the available data. This usually results in overfitting where a predictive model is extremely well adapted or overlapped to the training data. The general consequence of this is that the model includes random disturbances (noise) in the training set as being significant leading to inadequate performance of the machine learning technique on the given test dataset due to the fact that these disturbances mask the effect of the underlying distribution by not reflecting it [43]. The use of a regularizer [179-182] that replaces the observed amino acid distribution by its estimator, is the typical strategy adopted to overcome this particular problem. Structure-based protocols however, are usually less

affected by the above listed barriers, thereby resulting in strikingly accurate predictive models even for alleles with limited experimental data [11-15].

1.10 Databases and resources available

The role of bioinformatic databases, resources and tools in modeling the network of the immune system has been instrumental in advancing peptide vaccine discovery. Particular success has been reported in studies on anti-tumour vaccines [183], malaria [184], melanoma [185] and multiple sclerosis [186]. The development of various bioinformatic tools for in-depth analysis and prediction of pMHC and TR/pMHC interactions are greatly assisted by the availability of general and specialized databases that store, annotate, disseminate and depict pMHC and TR/pMHC binding information. The most important and commonly used general and specialized databases and resources for the study of pMHC and TR/pMHC interactions are discussed below along with some of their implications towards these studies.

1.10.1 General databases and resources

Several general databases contain useful information ranging from published literature to protein sequences to X-ray crystal structures of pMHC and TR/pMHC complexes. Table 1.1 gives a list of a few such important databases and resources that are used in day to day research on pMHC and TR/pMHC interactions. These databases are described below.

1.10.1.1 UniProt

UniProt [187-193] is a comprehensive, high-quality, annotated and freely available resource for information on protein sequences and their functions. It is a unified knowledgebase that combines databases such as Swiss-Prot [193-197], a computer-annotated supplement to Swiss-Prot called as TrEMBL (Translated EMBL) [195], UniRef [191-193, 198] (a database of protein sequence clusters, developed to speed up sequence similarity searches) and UniParc [191-193, 199] (an archive for protein sequences, used to keep track of protein sequence identifiers and changes in protein sequences). The TrEMBL database consists of all translation of European Molecular Biology Laboratory (EMBL) nucleotide sequences (from the EMBL Nucleotide Sequence Database [200-210]) that are not available in Swiss-Prot. As of January 2011, the combined number (including Swiss-Prot and TrEMBL records) of sequence entries within the UniProt knowledgebase is 13,593,921 which encompass 4,392,846,537 amino acids. The TrEMBL database contains 13,069,501 sequence entries, comprising 4,207,640,687 amino acids as of January 2011.

Table 1.1: List of generalized databases and resources used for the study of pMHC and TR/pMHC interactions

Title	Description	URL
UniProt [187-193]	A resource for protein sequence and functional information.	http://www.uniprot.org/
Swiss-Prot [193-197]	A curated and annotated protein sequence database.	http://au.expasy.org/sprot/
neXtProt	An innovative knowledge platform dedicated to human proteins.	http://beta.nextprot.org/
Protein Data Bank (PDB) [62, 63]	A resource for structural data of biological macromolecules.	http://www.rcsb.org/pdb/
PubMed [211-214]	A central repository for published scientific literature.	http://www.ncbi.nlm.nih.gov/pubmed/

1.10.1.2 Swiss-Prot

Swiss-Prot [193-197] is manually curated and annotated protein sequence database. Records within Swiss-Prot are deposited by biologists and are further validated by domain experts. Researchers at Swiss-Prot strive to minimize redundancy and thus furnish high quality annotation through manual curation. However, manual curation results in compromised data coverage within Swiss-Prot. It was due to this limitation that TrEMBL was created as a computer-annotated supplement to Swiss-Prot [195]. As of January 2011, the Swiss-Prot database within UniProt contains 524,420 sequence records that comprise 185,205,850 amino acids and are obtained from 194,602 published references.

1.10.1.3 neXtProt

neXtProt is a knowledge platform dedicated to human proteins. neXtProt is a new resource containing a wealth of high-quality data on all human proteins that are coded by the 20,000 protein-coding genes found in the human genome. This web-based interactive platform and repository has been developed to help understand the functionality and role of human proteins in health and diseases. The database's beta release incorporates 20,394 protein entries abstracted from 264,571 published articles as of January 2011.

1.10.1.4 Protein DataBank (PDB)

PDB [62, 63] is a one of a kind, up to date worldwide archive for primary (amino-acid sequence), secondary and tertiary structural data of biological macromolecules. It consists of protein, nucleic acids and carbohydrate structures. A four-letter identifier, referred to as the PDB-code or ID is assigned to each structure deposited in the PDB. The first character in a PDB-code is a number from 1–9. Often several entries correspond to one protein. These multiple entries could result from the structure being solved in different crystal forms, in different states of ligation, re-solved using more accurate data collection techniques or using better (higher resolution) crystals. PDB contains a total of 70,813 structures as of January 2011.

1.10.1.5 PubMed

PubMed is the central repository that comprises over 20 million citations as of January 2011. The citations are indexed for biomedical literature from MEDLINE, life science journals and online books. Among the fields included in PubMed citations and abstracts are medicine, nursing, dentistry, veterinary medicine, the health care system and preclinical sciences. Access to additional relevant and useful websites and links to the other National Center for Biotechnology Information (NCBI) molecular biology database and resources are also provided within PubMed. PubMed is a freely accessible resource, developed and maintained by NCBI, within the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH).

1.10.2 Specialized databases and resources

Besides the general databases described above, there are various specialized resources that focus primarily on pMHC and/or TR/pMHC interactions and contain valuable information pertaining to these significant interactions. A list of such databases, resources and tools is provided in Table 1.2. Among these, a few significant databases are today actively used in the study of both pMHC and TR/pMHC interactions. Described below are these pivotal resources that have contributed significantly over the years towards shaping the way research is currently pursued in the field of structural immunoinformatics.

Table 1.2: List of specialized databases, resources and tools used in the study of pMHC and TR/pMHC interactions.

Title	Description	URL
IMGT [215-231]	The international ImMunoGeneTics information system.	http://www.imgt.org/
IMGT/3Dstructure-DB [57, 58]	A database for immunoglobulin (IG), TR and MHC structural data.	http://www.imgt.org/3Dstructure-DB/
IMGT/HLA Database [150-156]	A specialist database for HLA sequences.	http://www.ebi.ac.uk/imgt/hla/
IEDB [166-169]	Immune Epitope Database.	http://www.immuneepitope.org/
SYFPEITHI [170, 171]	A database of MHC ligands and peptide motifs.	http://www.syfpeithi.de/
NCBI dbMHC [232]	A database for DNA and clinical data related to the human MHC.	http://www.ncbi.nlm.nih.gov/gv/mhc/
MHCBN [233, 234]	Comprehensive database of MHC-binding, non-binding peptides and T cell epitopes.	http://www.imtech.res.in/raghava/mhcbn/
Dana-Farber Repository for Machine Learning in Immunology	A repository containing data from MHCPEP [235-238] database and selected independent datasets of proteins, protein fragments, non-binding peptides and lists of T cell epitopes.	http://bio.dfci.harvard.edu/DFRMLI/
AntiJen [239]	A database containing experimentally determined quantitative binding data for MHC-binding, TAP-binding peptides and T cell epitopes.	http://www.darrenflower.info/AntiJen/
IMGT/LIGM-DB [240, 241]	A comprehensive database for IG and TR nucleotide sequences from human and other vertebrates.	http://www.imgt.org/cgi-bin/IMGTlect.jv/
IMGT/PRIMER-DB	A database for standardized information on oligonucleotides or primers of IG and TR.	http://www.imgt.org/IMGTPrimerDB/

Title	Description	URL
IMGT/GENE-DB [242]	A comprehensive database of IG and TR genes from human and mouse.	http://www.imgt.org/IMGT_GENE-DB/GENEselect
IPD-MHC Database [243-246]	A centralised repository for MHC sequences from different species.	http://www.ebi.ac.uk/ipd/mhc/
EPIMHC [247]	A curated database of MHC ligands.	http://bio.dfci.harvard.edu/epimhc/
Cancer Immunome Database [248]	A database focussing on human gene products against which an immune response is known in cancer.	http://ludwig-sun5.unil.ch/CancerImmunomeDB/
Epitome [249]	A database of structurally inferred antigenic epitopes in proteins	http://www.rostlab.org/services/epitome/
HLA Database	A database of HLA proteins.	http://bio.dfci.harvard.edu/Database/db_show_hla.html
HIV Molecular Immunology Database [250]	An annotated collection of HIV-1 cytotoxic, helper T cell epitopes and antibody binding sites.	http://www.hiv.lanl.gov/content/immunology/
IMGT Repertoire [220]	The global ImMunoGeneTics web resource for IG, TR, MHC and related proteins of the immune system (RPI).	http://www.imgt.org/textes/IMGTrepertoire/
IMGT-ONTOLOGY [251-256]	A resource for concise, non ambiguous and a formal specification of the terms to be used in the study of IG, TR and MHC proteins.	http://www.imgt.org/textes/IMGTindex/ontology.html
IMGT Scientific chart [218, 257-259]	A resource of standardized rules for sequence description, numbering and nomenclature of IG, TR, MHC and RPI from human and other vertebrate species, belonging to the immunoglobulin superfamily (IgSF) and MHC superfamily (MhcSF).	http://www.imgt.org/textes/IMGTindex/IMGTchart.html
IMGT/V-QUEST [260, 261]	A customized and integrated tool for IG and TR standardized V-J and V-D-J sequence analysis.	http://www.imgt.org/IMGT_vquest/share/textes/

Title	Description	URL
IMGT/JunctionAnalysis [262]	A program to analyse the junction of IG and TR nucleotide sequences.	http://www.imgt.org/IMGT_jcta/jcta
IMGT/HighV-QUEST [261, 263]	A tool to analyse large numbers of rearranged IG and TR sequences at once.	http://www.imgt.org/HighV-QUEST/
IMGT/Allele-Align	An alignment tool to identify nucleotide and amino acid differences by comparing two MHC alleles.	http://www.imgt.org/Allele-Align/
IMGT/PhyloGene [264]	An online software package to compute and draw phylogenetic trees for IG and TR V-REGION nucleotide sequences.	http://www.imgt.org/IMGTPhylogeny/
IMGT/DomainDisplay	A web-based tool to display amino acid sequences from the domains of the IgSF and MhcSF superfamilies.	http://www.imgt.org/3Dstructure-DB/cgi/DomainDisplay.cgi
IMGT/GeneView	A tool for visualization of a given gene in a locus for human MHC, IG, TR and mouse TR.	http://www.imgt.org/LocusView/
IMGT/LocusView	A program to view multiple genes in a locus for human MHC, IG, TR and mouse TR.	http://www.imgt.org/LocusView/
IMGT/GeneInfo [265, 266]	A tool to obtain information on data resulting from the mechanisms of V-J and V-D-J gene rearrangements in human and mouse TR loci.	http://www.imgt.org/GeneInfoServlets/htdocs/
IMGT/GeneFrequency [254]	A tool for graphical representation of rearranged IG and TR gene sequences.	http://www.imgt.org/IMGTGeneFrequency/
IMGT/DomainGap Align [58]	A web-based program for analysing amino acid sequences of IG, TR and MHC domains.	http://www.imgt.org/3Dstructure-DB/cgi/DomainGapAlign.cgi
IMGT/Collier-de-Perles [267-269]	An analysis tool for graphical representations of protein domains from their amino acid sequences.	http://www.imgt.org/3Dstructure-DB/cgi/Collier-de-Perles.cgi

Title	Description	URL
IMGT/DomainSuperimpose	A web-based tool to superimpose IG, TR and MHC domains.	http://www.imgt.org/3Dstructure-DB/cgi/DomainSuperimpose.cgi
IMGT/StructuralQuery [57]	A tool to retrieve and analyse IG, TR and MHC structural data from IMGT/3Dstructure-DB.	http://www.imgt.org/3Dstructure-DB/StructuralQuery

1.10.2.1 IMGT

Established in 1989 by Marie-Paule Lefranc, a pioneer in immunogenetics and immunoinformatics, the international ImMunoGeneTics information system (IMGT; [215-231]) is a global reference in immunogenetics and immunoinformatics that specializes and shares a wealth of extremely significant information on IG or antibodies, TR proteins, MHC proteins of human and other vertebrate species along with immunoglobulin superfamily (IgSF), MHC superfamily (MhcSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates. It is a unique centralized high-quality integrated knowledgebase that consists of sequence databases such as IMGT/LIGM-DB [240, 241] and IMGT/PRIMER-DB, a genome database called IMGT/GENE-DB [242], a structure database known as IMGT/3Dstructure-DB [57, 58] and a database of monoclonal antibodies (mAb) termed IMGT/mAb-DB [270]. IMGT also contains web resources such as IMGT Repertoire [220] and IMGT Scientific chart [218, 257-259] along with interactive analysis tools such as IMGT/V-QUEST [260, 261], IMGT/GeneInfo [265, 266] and IMGT/Collier-de-Perles [267-269]. The IMGT/HLA database, one of the most important databases used in the study of pMHC and TR/pMHC interactions, is also a part of the IMGT project and can be accessed through the “IMGT/MHC-DB” link on the IMGT website (<http://www.imgt.org/>).

1.10.2.2 IMGT/3Dstructure-DB and IMGT/2Dstructure-DB

IMGT/3Dstructure-DB [57, 58] is an exclusive and unique resource on IG, TR, MHC and RPI with known three-dimensional (3D) structures. The structural data is sourced from PDB [62, 63]. The high-quality standardized information within IMGT/3Dstructure-DB includes IMGT annotation on IG, TR, MHC and RPI sequences, their two-dimensional (2D) and 3D structures. IMGT/2Dstructure-DB consists of amino acid sequences, originally obtained from the World Health Organization’s (WHO) International Nonproprietary Names (INN; [271-274]) and the Kabat [275-277] database, for IG, mAb

and fusion proteins for immune applications (FPIA). IMGT/3Dstructure-DB portrays 3D structural information such as chain details and contact details at different levels such as at the domain/chain interface and the residue level. In particular, contact information relevant for immunological proteins like IG, TR and MHC is shed light upon since these proteins interact specifically with a great number of molecules making these interactions extremely vital for normal immune responses. As of January 2011, IMGT/3Dstructure-DB contains 2,416 entries out of which 1,987 are structural entries extracted from PDB along with IMGT/2Dstructure-DB containing 94 sequence entries obtained from INN and 335 are sequence entries sourced from the Kabat database.

1.10.2.3 IMGT/HLA Database

The IMGT/HLA sequence database [150-156] is a specialist database for HLA sequences and perhaps the most important resource used for the study of HLA proteins and their interactions apart from the IMGT knowledgebase itself [215-231]. It includes the official sequences for the WHO HLA Nomenclature Committee for Factors of the HLA System. In addition to the sequences of HLA proteins, the database contains detailed information concerning the source of the sequences and data on the validation of the sequences. Researchers at the IMGT/HLA database strive to avoid the problems associated with renaming already published sequences and the confusion of multiple names for the same sequence by officially naming a sequence, in compliance with the WHO HLA Nomenclature Committee's rules for naming HLA alleles, prior to the publication of a HLA sequence. The database also permits users to present complex queries about a particular HLA sequence, sequence features, references, contacts and allele designations. As of January 2011, the IMGT/HLA database contains 6,189 allele sequences out of which 6,074 are HLA alleles and 115 are non-HLA alleles.

1.10.2.4 IEDB

The Immune Epitope Database (IEDB; [166-169]) is another vital source of information for investigations pertaining to pMHC and TR/pMHC interactions. It contains functional MHC-binding and T cell response information for peptide epitopes, derived from published *in vitro* assays. The peptide data is related to antibody and T cell epitopes for humans, non-human primates, rodents and other species of animals. The available immune epitope and MHC-binding data are from a variety of different antigenic sources. The current peptide epitope data relating to all infectious diseases, including National Institute of Allergy and Infectious Diseases (NIAID) Category A, B, and C priority pathogens,

NIAID Emerging and Re-emerging infectious diseases, allergens, and autoimmune diseases, along with non-peptidic allergen epitope data is painstakingly yet completely curated. On-going curation includes that of peptide epitopes related to transplant/alloantigen epitopes and that of non-peptidic infectious disease and autoimmune epitope data. IEDB contains 79,348 confirmed peptide epitopes and 783 confirmed non-peptidic epitopes, sourced from 2,626 organisms, from 154,423 T cell assays and 201,071 MHC-binding assays, as of January 2011. The current peptide epitope and non-peptidic epitope data and information are extracted from 11,771 published references.

1.10.2.5 SYFPEITHI

The SYFPEITHI [170, 171] database comprises of more than 7000 peptide sequences (as of January 2011) that are known to bind MHC-I and MHC-II proteins, along with peptide motifs from various species such as human, non-human primates, cattle, chicken, and mouse. All currently available motifs can be accessed as individual entries in the database. All entries within this resource are obtained and compiled from published reports. It is possible to search the database for MHC alleles, peptide motifs, natural ligands, T cell epitopes, source proteins/organisms and references. The database also includes hyperlinks to data sources such EMBL and PubMed besides enabling the users of peptide binding predictions for a number of MHC alleles.

1.10.2.6 NCBI dbMHC

The NCBI dbMHC [232] is a semi-curated database for DNA and clinical data related to HLA proteins. Originally designed and built by NCBI as an open resource for registration and characterization of HLA DNA-typing kits and reagents [232], NCBI dbMHC continues to provide a platform where researcher around the globe can submit, edit, view, and exchange HLA data. The database currently hosts an online tool called Sequencing Based Typing (SBT; [278]) for typing highly polymorphic HLA sequences, sequence interpretation and evaluating the allelic composition in SBT results for complementary DNA (cDNA) or genomic sequences [278], a HLA microsatellite database [279, 280], a tool known as Microsatellite Markers to search descriptive information for some of the known short tandem repeats within the HLA gene region, a Primer/Probe database, an interactive Alignment Viewer for HLA and related genes, a Typing Kit Interface for HLA alleles, a program for viewing clustering trees for HLA alleles produced using Basic Local Alignment Search Tool (BLAST; [281]) pairwise alignments and a tool for graphical visualization of HLA genes, non-HLA genes and pseudogenes within chromosome 6. The

NCBI dbMHC database also provides links to the IMGT/HLA database for every allele and is fully integrated with other NCBI resources as well as with the International Histocompatibility Working Group (IHWG) website (<http://www.ihwg.org/>). Allele sequences, both curated and not-curated ones, housed in NCBI dbMHC are retrieved from IMGT/HLA and GenBank [282-291] databases, respectively.

1.10.2.7 MHCBN

MHCBN [233, 234] is a curated database of MHC-binding peptides, MHC non-binding peptides, TAP-binding peptides, TAP non-binding peptides and T cell epitopes compiled from published literature and existing databases. The database provides the scientific community with a number of web-based tools which allow the user to search for any information about MHC alleles and peptides, map experimentally determined MHC-binders, MHC non-binders and T cell epitopes in a given query protein sequence and conduct a BLAST search against related antigenic and MHC associated proteins. The resource contains other information such as sequence and structure data for source proteins of peptides and MHC molecules. MHCBN also provides hyperlinks to major databases including most NCBI resources, Swiss-Prot (for protein sequences and source information), PDB (for structural information), IMGT/HLA (for HLA allele sequences), PubMed (for published references), GenBank (for nucleotide sequences) and the Online Mendelian Inheritance in Man (OMIM; [292-299]; for MHC linked diseases) database. As of January 2011, the database contains 20,717 MHC-binders, 4,022 MHC non-binders, 1053 TAP-binding and non-binding peptides and 6,722 T cell epitopes for 450 MHC alleles extrapolated from 1,519 published articles.

1.10.2.8 Dana-Farber Repository for Machine Learning in Immunology

The Dana-Farber Repository for Machine Learning in Immunology contains all the data from the earlier MHCPEP [235-238] database and selected independent datasets of proteins, protein fragments, non-binding peptides and lists of T cell epitopes. This database has recommendations for scaling and comparison of performance for various sequence-based MHC-binding and T cell epitope prediction systems. There are also HLA binding peptide datasets for specific alleles such as HLA-A, HLA-B and HLA-DRB1 haplotypes, along with T cell epitope reference lists from tumour and viral antigens. This repository provides a unique resource that can be used in conjunction with IEDB datasets for the development of advanced machine learning and pattern recognition solutions which can be innovatively applied to develop T cell epitope prediction algorithms. MHCPEP is a

manually curated database that contains more than 13,000 experimentally validated MHC-binding peptide sequences [238]. Two sources, published reports and direct submissions of experimental data, are used to compile the entries in the MHCPEP database. Each entry or record consists of the peptide sequence, peptide's MHC specificity and where available, experimental method, observed activity, pMHC binding affinity, source protein, anchor positions or amino acid within the peptide and published citations.

1.10.2.9 AntiJen

AntiJen [239] is a database containing experimentally determined quantitative binding data for MHC-binding peptides, T cell epitopes, TAP-binding peptides and other significant proteins of the immune system. Archived in the database are continuous quantitative data on a variety of immunological molecular interactions including thermodynamic and kinetic measures of peptide interactions with TAP and MHC, pMHC complexes binding to TR proteins, antibodies binding to protein antigens and general immunological protein-protein interactions apart from functional and cellular data within the context of immunology and vaccinology. As of January 2011, the database contains over 24,000 entries [300]. The database is fully sourced from published literature. AntiJen also holds over 3,500 entries for linear and discontinuous B cell epitopes [300].

1.11 Methods available for T cell epitope prediction

Identification of T cell epitopes that activate both CD8⁺ cytotoxic T cells and CD4⁺ helper T cells is extremely important as it forms the basis for the development of peptide vaccines that are used in the treatment of allergic [301], autoimmune [302] and neoplastic diseases such as cancer [303, 304], besides combating infectious agents such as viruses [305]. Successful identification of T cell epitopes is also a significant means to understand disease pathogenesis [306]. Conventional means to identify T cell epitopes included the synthesis of overlapping peptides spanning the entire length of a protein, followed by experimental immuno-assays such as *in vitro* intracellular cytokine staining for each peptide [307], to determine T cell activation. Therefore, experimental detection of T cell epitopes has been doomed a tedious, time consuming and expensive process in the recent years [49].

It is well known that pMHC binding is a prerequisite for TR recognition of pMHC complexes and subsequent T cell activation. Moreover, it is widely regarded to be the event that most selectively defines immunogenic or T cell epitopes [308]. Therefore, T cell

epitope prediction relies primarily on predicting pMHC binding. Consequently, computational approaches have been developed as an alternative to traditional *in vitro* procedures, for the identification of T cell epitopes. Recently developed computational methods have proven to be vastly time and cost efficient in screening the vast oceans of peptides and MHC repertoires [17, 49], thereby, significantly decreasing the burden and cutting down the lead time associated with experimental identification of T cell epitopes. There are various criteria that have been applied to classify or categorize the available computational methods for T cell epitope prediction [18, 300, 309].

Nevertheless, two types of classification have stood the test of time. Tong *et al.* [18] have used the type of data employed for prediction to classify the methods into “sequence” and “structure-based” approaches. However, this type of classification groups the methods that employ both sequence-derived pMHC binding affinity data and 3D structural information to predict T cell epitopes [310, 311], under the structure-based approaches. On the other hand, Lafuente and Reche [309] have used the type of data and the technique employed for prediction to classify methods into “binding pattern recognition”, “quantitative binding affinity” and “modeling-based” models. Yet, this type of classification schema lists the above described methods that employ both sequence-derived pMHC binding affinity data and 3D structural information to predict T cell epitopes [310, 311], under the quantitative binding affinity models.

Therefore, it has now become important to add a third category namely, ‘sequence-structure-based’ approaches into the original classification by Tong *et al.* [18] to classify the methods that employ both sequence and structure-derived information to predict T cell epitopes. Hence, the currently available specialized computational methods for the prediction of T cell epitopes, can be broadly classified into three main categories: (i) methods based on identifying patterns in sequences of MHC-binding peptides (qualitative) along with those that attempt to quantify the actual pMHC binding affinity (quantitative), collectively called as sequence-based approaches; (ii) methods that employ 3D structures to model pMHC interactions termed structure-based approaches; and (iii) methods that employ both sequence-derived pMHC binding affinity data and 3D structural information to predict T cell epitopes, which can be referred to as sequence-structure-based approaches.

The first group includes protocols based on sequence motifs, motif matrices, quantitative matrices, decision trees, artificial neural networks (ANN), hidden Markov models (HMM)

and support vector machines (SVM). On the contrary, the second category represents techniques with distinct theoretical lineage and includes the use of 3D homology modeling, protein threading, docking and molecular dynamics (MD) techniques. The third category combines similarity matrices and structure-based techniques such as protein threading for T cell epitope prediction. Utilizing these algorithms and techniques, many web-based bioinformatics tools for T cell epitope prediction have been developed in the recent years. Table 1.3 provides a comprehensive list of such tools and web-servers that are widely used for the identification of strong-MHC-binding peptides. Described below are the above mentioned algorithms and techniques, their strengths and their weaknesses.

Table 1.3: List of available tools and web-servers for T cell epitope prediction

Title	Technique/Algorithm	MHC class	URL
Motif Scan	Sequence Motifs	I and II	http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan
SYFFPEITHI [170, 171]	Motif Matrices	I and II	http://www.syfpeithi.de/Scripts/MHCServer.dll/EpitopePrediction.htm
EPIMHC [247]	Position-Specific Scoring Matrix	I and II	http://imed.med.ucm.es/epimhc/
PEPVAC [312, 313]	Position-Specific Scoring Matrix	I	http://bio.dfci.harvard.edu/PEPVAC/
RANKPEP [314-316]	Position-Specific Scoring Matrix	I and II	http://bio.dfci.harvard.edu/RANKPEP/
BIMAS [317]	Quantitative Matrices	I	http://www-bimas.cit.nih.gov/molbio/hla_bind/
EpiJen [318]	Quantitative Matrices	I	http://www.darrenflower.info/EpiJen/
EpiMatrix [319]	Quantitative Matrices	I and II	http://www.epivax.com/immunogenicity-screening/epimatrix/
ProPred-I [320]	Quantitative Matrices	I	http://www.imtech.res.in/raghava/propred1/
ProPred [321]	Quantitative Matrices	II	http://www.imtech.res.in/raghava/propred/

Title	Technique/Algorithm	MHC class	URL
MAPPP [322]	Quantitative Matrices /Motif Matrices	I	http://www.mpiib-berlin.mpg.de/MAPPP/binding.html
IEDB [166-169]	Average Relative Binding-Quantitative Matrices/Stabilized Matrix Method-Quantitative Matrices/Artificial Neural Networks	I and II	http://tools.immuneepitope.org/analyze/html/mhc_binding.html http://tools.immuneepitope.org/analyze/html/mhc_II_binding.html
SMM [323]	Stabilized Matrix Method-Quantitative Matrices	I	http://zlab.bu.edu/SMM/
EpiTOP [324, 325]	Quantitative Structure-Activity Relationship	II	http://www.pharmfac.net/EpiTOP/
MHCPred [326-328]	Quantitative Structure-Activity Relationship	I and II	http://www.darrenflower.info/MHCPRED/
ANNPred [329]	Artificial Neural Networks	I	http://www.imtech.res.in/raghava/nhlaped/neural.html
NetMHCpan [330, 331]	Artificial Neural Networks	I	http://www.cbs.dtu.dk/services/NetMHCpan/
NetMHCIIpan [332, 333]	Artificial Neural Networks	II	http://www.cbs.dtu.dk/services/NetMHCIIpan/
MULTIPRED [334]	Artificial Neural Networks/Profile Hidden Markov model	I and II	http://antigen.i2r.a-star.edu.sg/multipred/
NetMHC [335-339]	Artificial Neural Networks-Weight Matrices	I	http://www.cbs.dtu.dk/services/NetMHC/
NetMHCII [340, 341]	Artificial Neural Networks/Stabilized Matrix Method-Quantitative Matrices	II	http://www.cbs.dtu.dk/services/NetMHCII/
KISS [342]	Support Vector Machines	I	http://cbio.enscm.fr/kiss/

Title	Technique/Algorithm	MHC class	URL
MHC2Pred [343]	Support Vector Machines	II	http://www.imtech.res.in/raghava/mhc2pred/
POPI [344]	Support Vector Machines	I	http://iclab.life.nctu.edu.tw/POP-I/
SVMHC [345, 346]	Support Vector Machines	I and II	http://www-apb.informatik.uni-tuebingen.de/Services/SVMHC/
SVRMHC [347-349]	Support Vector Machines Regression	I and II	http://svrmhc.biolead.org/
MHC-Thread [350]	Protein Threading	II	http://www.csd.abdn.ac.uk/~gjl k/MHC-Thread/
PREDEP [351]	Protein Threading	I	http://margalit.huji.ac.il/Teppred/mhc-bind/
HLABinding [311]	Adaptive Double Threading	I	http://atom.research.microsoft.com/hlabinding/hlabinding.aspx

1.11.1 Sequence-based approaches

The discoveries that peptides binding to specific MHC alleles are functionally related [352] and that they share residues with similar properties at various positions of their primary sequences [353] led to the earliest known attempts at predicting T cell epitopes [354, 355]. As alluded to earlier, MHC-I and MHC-II binding peptides are made up of residues with side-chains that fit into the cavities or ‘pockets’ made up of polymorphic complementary residues within the peptide binding cleft of the specific MHC proteins or alleles. These residues, referred to as the ‘anchor’ residues due to their role in anchoring the peptides firmly in the MHC binding cleft [352-354, 356-359], contribute the most towards pMHC binding by taking part in most of the pMHC binding interactions. This fact gave rise to the notion of “peptide motif” and subsequently helped researchers define peptide motifs [353, 354, 356] for an array of MHC-I and MHC-II alleles.

Following this, numerous research groups around the world began to develop computational tools that scan peptides fitting these motifs [317, 360-367]. Meanwhile, it was discovered that sequence motifs alone are inadequate to account for comprehensive binding ability of a candidate peptide and that residues along other positions (apart from

anchor residues) of a peptide also play a vital role in pMHC binding [368-370]. This resulted in a multitude of sequence-based techniques ranging from sequence motifs that use peptide motifs for prediction to SVM that use the entire length of the peptides for prediction, being employed by various researchers for large-scale screening of potential T cell epitopes from vast numbers of protein sequences. These sequence-based techniques and algorithms are highlighted below.

1.11.1.1 Sequence motifs

The simplest mode of representation of the peptide binding motif for a specific MHC allele is a sequence motif. Sequence motifs consist of a symbolic peptide string that lists the amino acid preferences of a given MHC protein for each residue position of the peptide. Although the general practice to obtain peptide binding motifs is to compare sets of peptide sequences that are known to bind to MHC proteins [170, 171], the first peptide binding motifs were identified by pool sequencing of peptide ligands eluted from MHC-I proteins [353, 356]. As said earlier, the SYFPEITHI database [170, 171] represents one of the largest collections of peptide binding motifs for MHC-I and MHC-II proteins. Peptide binding motifs specific for particular MHC alleles were the first models that enabled prediction of MHC-I restricted T cell epitopes [354, 355]. Although primitive, sequence motifs continue to be used for identification of T cell epitopes [371].

However, application of sequence motifs to the identification of T cell epitopes is today considered too simplistic, primarily because of the fact that peptide residues other than the anchor residues also contribute to binding [369, 372], as highlighted above. Moreover, immunodominant peptides without the required binding motifs have been identified [373] and it has also been shown that not all motif-conforming peptides bind to respective MHC alleles [374]. An investigation on the significance of the role played by peptide motifs in pMHC binding using *in vitro* binding assays on HLA-A*0201 binding peptides [369], has illustrated that only about 30% of motif-conforming peptides were actual MHC-binders. The extreme rigid nature of sequence motifs renders them unsuitable for T cell epitope prediction. Hence, use of simple motif models for T cell epitope prediction has proven to be both non-sensitive and non-specific [374]. Therefore, this approach fails to detect binders not pertaining to existing motifs and includes non-binding sequences that fit the required patterns, particularly yielding many false negatives [366]. Despite these limitations, this approach still presents a useful alternative to random guessing or using a set of overlapping peptides for the selection of candidate binders [17].

1.11.1.2 *Motif matrices*

Representing an enhancement of sequence motif models, motif matrices consist of tables whose coefficients quantify the contribution of position-specific amino acid frequencies found within candidate peptides that bind to a specific MHC allele [375, 376]. For a given peptide sequence, the consensus binding score is calculated by aligning a matrix with the target protein segments and computing (summing, multiplying or averaging) the relevant position specific and residue-matched matrix coefficients. These consensus peptide binding scores are generally continuous and thus, a binding threshold is usually put in place to distinguish the MHC-binders. First examples of motif matrices were developed by de Groot *et al.* [377] and Rammensee *et al.* [170].

First introduced by Gribskov and co-workers [378] in 1987, ‘profiles’, also known as position-specific scoring matrices (PSSM), are useful for detecting distantly related sequences and are similar to motif matrices. Fundamentally, PSSM consist of log-odds matrices with coefficients defined by the logarithmic ratios of observed amino acid frequencies with respect to the relevant background frequencies [309]. Later, it was Reche *et al.* [314-316] who first applied PSSM to the study of pMHC binding and developed profiles which were extracted from sets of aligned peptides that were known to bind specific MHC proteins. A significant improvement in identifying MHC-binders was achieved due to the use of PSSM, which can be attributed to the ability of profiles to tackle the problems of sequence redundancies (through sequence weights) and missing data (using pseudo-counts that are estimated from substitution matrices), unlike basic motif matrices.

Subsequently, using an expectation-maximization motif discovery program [379] and peptide binding scores obtained from MHC-II binding peptides, MHC-II-specific profiles were also generated [314-316] for successful identification of T cell epitopes presented by MHC-II proteins as well. This was followed by the creation of another type of motif-based matrices called as the ‘weight matrices’ by Nielsen *et al.* [337], who applied a Gibbs sampler to detect weak sequence motifs and characterize them in terms of weight matrices for MHC-I and MHC-II binding peptides. Weight matrices are almost indistinguishable from PSSM and perform virtually identically. Rajapakse *et al.* [380] later utilized a multi-objective evolutionary algorithm to identify a consensus motif for the murine MHC-II allele I-A(G7). Developed primarily from positive peptide data, all of the above described motif

matrices consisted of known MHC-binding peptides within the training sets and hence, lacked a control set (negative or non-binders) resulting in inability to accurately identify experimental negatives.

Mallios [381], realizing this issue, revolutionized the use of motif matrices for the prediction of peptides binding to MHC-II proteins by describing a motif matrix obtained and evaluated utilizing positive and negative peptide examples and a stepwise discriminating analysis (SDA) method. Contrasting to the methods utilizing sequence patterns, this method resulted in outputs with continuous peptide binding scores that discriminated peptides as binders and non-binders. Although advanced compared to the simple sequence motifs, the similarity in the underlying motif concept renders prediction of T cell epitopes using motif matrix-based predictive methods, susceptible to the same disadvantages as with utilizing sequence motifs. The basic limitation being the fact that, motif matrices also assume peptide residues to be contributing independently to pMHC binding. Despite being well supported by experimental data, such absolute assumptions are incorrect as there is evidence that supports the influence of neighboring residues on the contribution of peptide residues to pMHC binding [323], thereby, shedding light on the ignorance of the effect of the overall structure of peptide by motif matrices.

1.11.1.3 Quantitative matrices

In order to detect weak binding patterns and to account for noisy and collinear data, more complex forms of matrix-based predictive methods were developed in the following years. These matrices are termed the quantitative matrices and are the most widely used additive models in predicting pMHC binding. Although they resemble motif matrices, they are generated from actual peptide binding affinity data, unlike motif matrices, resulting in peptide binding scores that reflect actual pMHC binding affinity. The first implementation of quantitative matrices for the identification of MHC-I binding peptides was by Parker and co-workers [317]. Methods utilizing quantitative matrices that in turn use binding affinity data procured from positional scanning combinatorial peptide libraries (PSCPL), have also been developed [382, 383], where sets of sub-libraries represent all possible peptides of a particular length with one amino acid being fixed and the remaining residue positions containing mixtures of all amino acids in each sub-library. Characteristically, logarithmic peptide concentrations relative to a reference peptide library form the coefficients of quantitative matrices generated using PSCPL.

Many means that employ large sets of pMHC binding affinity data have since been used to construct quantitative matrices. These methods included use of average relative binding (ARB) [384] and stabilized matrix method (SMM) [385] to derive quantitative matrices for the prediction of pMHC-I and pMHC-II binding affinity. Although SMM was first applied to predict pMHC-I binding affinity [323], Nielsen and colleagues [340] applied an improvised SMM-align approach that focuses on the two most proximal (generally amino-terminal) peptide flanking residues (PFR) to compute the pMHC binding score, resulting in enhanced predictive performance for MHC-II proteins. Using peptide libraries to procure a quantitative representation of the amino acid interactions with pocket residues of the MHC-II HLA-DR alleles, Sturniolo *et al.* [386] were able to generate virtual quantitative matrices which are a close relative of quantitative matrices themselves. Their work highlighted the importance of selecting binding pocket profiles to compute pMHC binding affinity for MHC-II alleles, a principle that forms the basis of the TEPITOPE [387] prediction system.

A consequence of the similarities between motif matrices and quantitative matrices is that quantitative matrices also assume an independent contribution of peptide side chains to pMHC binding. To overcome this, Doytchinova *et al.* [161, 162] made use of a robust partial least squares (PLS) multivariate statistical approach to improve the predictive performance of their protocol by deriving quantitative structure-activity relationship (QSAR) matrices, where an additive equation was formulated to account for individual amino acid contributions at each position and interactions with neighbouring residues together as pMHC binding affinity. The matrices were subsequently solved employing PLS-regression. Later, Guan *et al.* [326-328] reassured the usefulness of PLS-QSAR-based quantitative matrix models and resultantly, incorporated this methodology to develop the web-server MHCPRED [326-328] for MHC-I and MHC-II restricted T cell epitope prediction. As good as it is, even the use of quantitative matrices has disadvantages such as heavy reliance on the availability of large comprehensive training sets of peptides rendering them inappropriate for accurate prediction of peptides in circumstances where the peptide data available is insufficient.

1.11.1.4 Decision trees

Decision trees are rule-based models that classify patterns using a sequence of well defined rules [388]. Due to their popularity as classification algorithms [389], they are also applicable for T cell epitope prediction. Embedded within the nodes of a decision tree are

rules extrapolated by converting position-specific binding motifs. Amino acid properties that correlate strongly to the physicochemical properties of binding peptides are thus indicated within the resulting tree structure. Subsequently, threading of peptide sequences occurs through a series of nodes. Finally, the outcome of prediction is determined by the result of all node-to-node transitions. Credit to its capability to elucidate both linear and non-linear problems, this approach has been adopted by several groups to identify higher-level rules for pMHC binding. Savoie *et al.* [390] were the first to construct a BONSAI decision tree to investigate TR preference and peptide epitope motifs for the peptides that bind to the human MHC-I allele HLA-A*0201. Segal *et al.* [391] adopted a similar tree-structured technique to predict peptides binding to murine MHC-I allele H2-Kb. Recently, Zhu *et al.* [392] used decision trees that were simultaneously trained on peptide binding data from different MHC-I alleles, to predict peptides that bind to a specific MHC-I proteins and achieved an enhancement in their prediction accuracy. An example of a decision tree network is shown below in Figure 1.11.

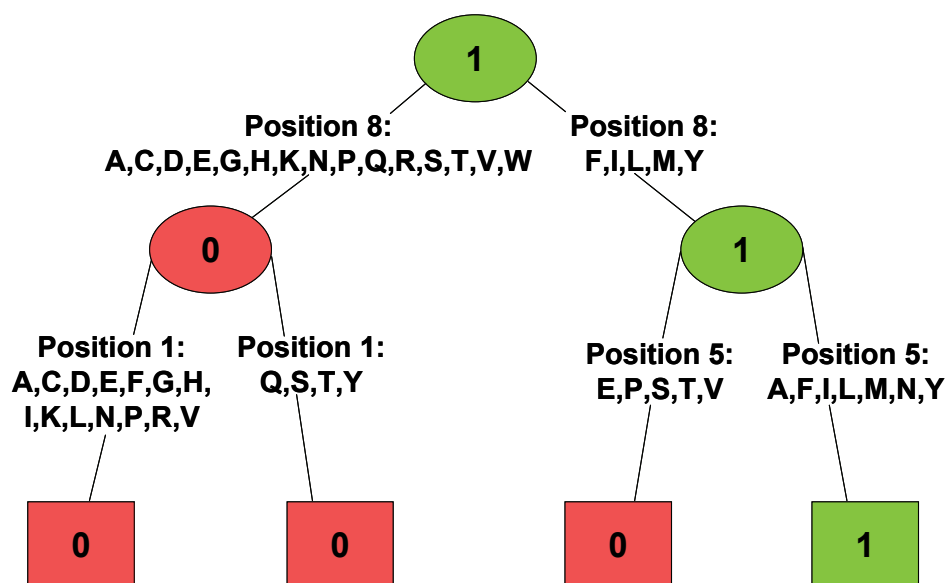


Figure 1.11: A pictorial representation of a subset of the decision tree network utilized by Segal *et al.* [391]. Represented as each node is the grouping of preferential or non-preferential amino acid residues at various positions for the peptides binding to the murine MHC-I allele H2-Kb. The ellipses denote internal nodes and the rectangles depict terminal nodes. The numbers 0 or 1 signify the predictions non-binding (bright red) or binding (bright green), respectively, at each node.

1.11.1.5 Artificial neural networks

Artificial Neural Networks (ANN) are connectionist models particularly well suited to perform classification and complex pattern recognition tasks [393]. Therefore they are one

of the most frequently and extensively used machine learning algorithms for recognizing pMHC binding patterns. ANN can even encode non-linear data and have been used for prediction of both MHC-I and MHC-II restricted T cell epitopes [175, 335, 336, 376, 394-396]. ANN were first employed to predict T cell epitopes restricted to MHC-I alleles, especially the human allele HLA-A*0201 [376, 394] and the murine allele H2-Kb [395]. They were later extended to MHC-II alleles, specifically applied to the human HLA-DR4 alleles [176, 396]. ANN work by representing peptide features through amino acid descriptors such as composition, hydrophobicity, volume and charge. The descriptors are used to train the ANN for classifying peptides into binders and non-binders. An example of the ANN architecture is illustrated in Figure 1.12.

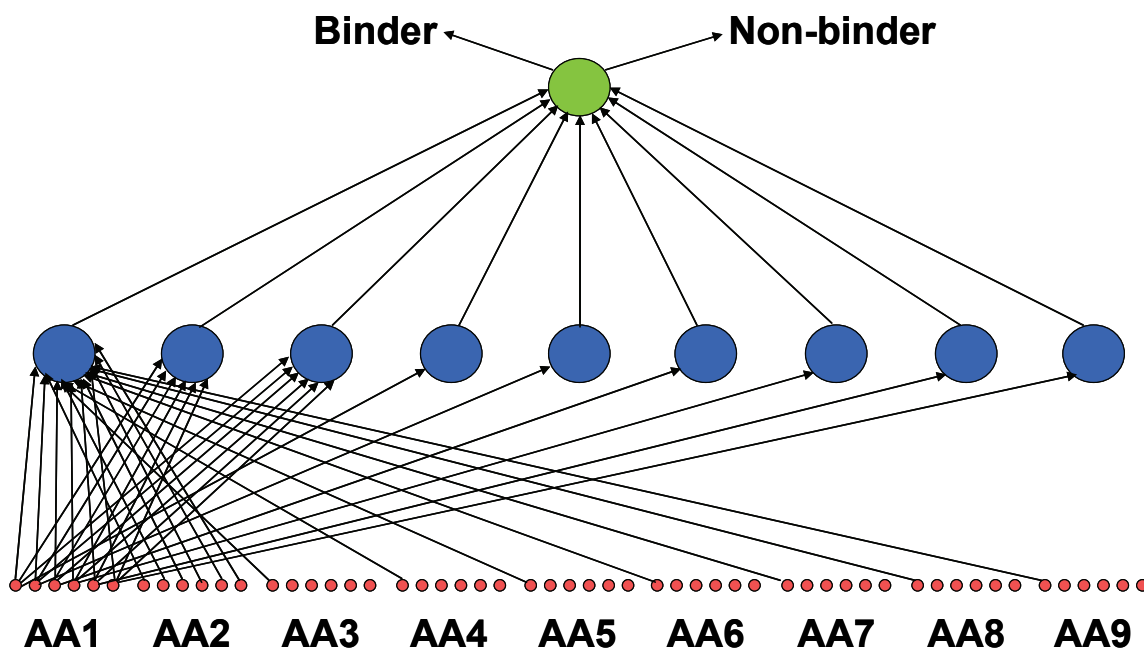


Figure 1.12: An example of the three-layer ANN derived by Brusica *et al.* [175] for predicting MHC-I restricted T cell epitopes. The first layer (small red circles) represents input nodes with the number of nodes corresponding to the length of the input peptide (in this case 9-mer; AA stands for amino acid). The number of nodes in the second (hidden; blue circles) layer equals the ideal length of the binding peptides (usually set to 9 residues) and a single output node (green circle) predicts binders and non-binders.

A comparative study [17] that investigated the predictive performance of ANN has revealed that, with a gradual increase in the training peptide data, ANN outperform motifs, motif matrices, quantitative matrices and even HMM, thereby, suggesting that ANN are better suited for T cell epitope prediction. ANN were also shown to be highly sensitive in their predictions of pMHC binding affinity for MHC-I proteins [335]. Many research

groups have since created hybrid versions of ANN to improvise pMHC binding prediction. For example, Nielsen *et al.* [336] trained series of ANN using a combination of novel input representations such as several sequence coding strategies including sparse encoding, blosum encoding and even HMM derived input to improve the predictive power of the system. In another example, Brusic *et al.* [176] successfully devised a system that automated the strength of matrix models and the efficiency of an evolutionary algorithm to identify the pMHC binding scores for MHC-II binding peptide dataset, which were subsequently utilized to train an ANN, resulting in accurate predictions.

The underlying concept that is commendable for the success of the approach integrating an evolutionary algorithm and ANN can be explained as follows. To begin with, the evolutionary algorithm selects new alignment matrices based on evolutionary principles. Two offspring matrices are produced by each parent matrix. One sibling matrix is an exact copy of the parent matrix and the other is a mutant copy. The child with higher fitness value is passed on to the next generation to improve accuracy and efficiency of the prediction. Finally, the ANN are trained by feeding them with the highest scoring alignments from the final generation matrices. As evident, a major limitation for the prediction of T cell epitopes is the availability of experimental peptide binding data. Yet again, to counter this short-coming, Nielsen *et al.* [330] developed a method where they combined the MHC-I peptide binding residues and pMHC binding affinity data for training the ANN, effecting the prediction of T cell epitopes even for MHC-I proteins with little binding data.

A similar procedure has also been followed in developing protocols which can be utilized to predict T cell epitopes restricted to uncharacterized HLA-DR alleles [332]. Recently, Soam *et al.* [397] have described a method where they have applied probability distribution functions to initialize the weights and biases of the ANN for HLA-A*0201 restricted T cell epitope prediction. Despite these recent advances in the use of ANN for high-throughput screening of peptides to predict T cell epitopes, the requirement of a fixed input length remains a major drawback of ANN-based methods [18]. The disability to predict peptide epitopes that are of a different length compared to those in the training dataset, is another back-drop of any given ANN model.

1.11.1.6 Hidden Markov models

Hidden Markov models (HMM) have a wide range of applications due to them being a type of probabilistic graphical models. They are the most widely used technique in speech, sequence and statistical pattern recognition and classification [398, 399]. Based on parametric statistical models, HMM work by assuming that the system that is being modeled is connected by a Markov chain of unknown hidden parameters extracted from data. Just like the previously described decision trees and ANN, HMM also possess the capacity to handle non-linear data and this ability renders them suitable for representing time-series sequences having flexible lengths. Each HMM has a series of discrete-state, time-homologous, first-order Markov chains associated with it. These Markov chains have an initial distribution and suitable transition probabilities between states. A discrete or continuous distribution over possible outputs is contained within each state.

Upon visiting a particular state or during transition from state to state, these outputs are generated. A set of transition and emission probability rules are followed for undergo transitions between states. The probability of moving from one state to another via a connected edge is called the transition probability and the probability of emitting a particular symbol at any particular state is known as the emission probability. The name ‘Hidden’ Markov model is derived from the sequences of states that are hidden from observance and underlie the Markov chains. By multiplying the emission and transition probabilities along the path, the overall probability of any given peptide sequence being a binder or a non-binder is computed. Using HMM has been known to be useful in surmounting the potential constraints of using ANN to predict T cell epitopes [177, 400].

The first instance of using HMM for T cell epitope prediction was reported Mamitsuka in 1998. The author had described two different HMM topologies known as the profile or pHMM (Figure 1.13a.) and the fully connected HMM (Figure 1.13b.). Recently, a new type of HMM topology has been devised for T cell epitope prediction namely structure-optimized HMM [401, 402]. Nevertheless, the first successful application of HMM to predict T cell epitopes restricted to HLA-A*0201 was through the use of fully connected HMM [400]. Figure 1.13b depicts the states that are pairwise connected such that the underlying digraph is complete within a fully connected HMM. With an exception of diagonal entries, which correspond to loops or self-transitions, the transition matrix of a fully connected HMM does not contain any zero entries. Another important aspect of the fully connected HMM models is the lack of any particular notable starting or terminating

state. This significant characteristic permits the representation of more than one peptide sequence pattern veiled within the binding peptide data used for training, due to no absolute constraints being imparted on the structure of a fully connected HMM. Therefore, fully connected HMM are very well suited to model nonlinear data as they are able to recognize different patterns in the binding peptides.

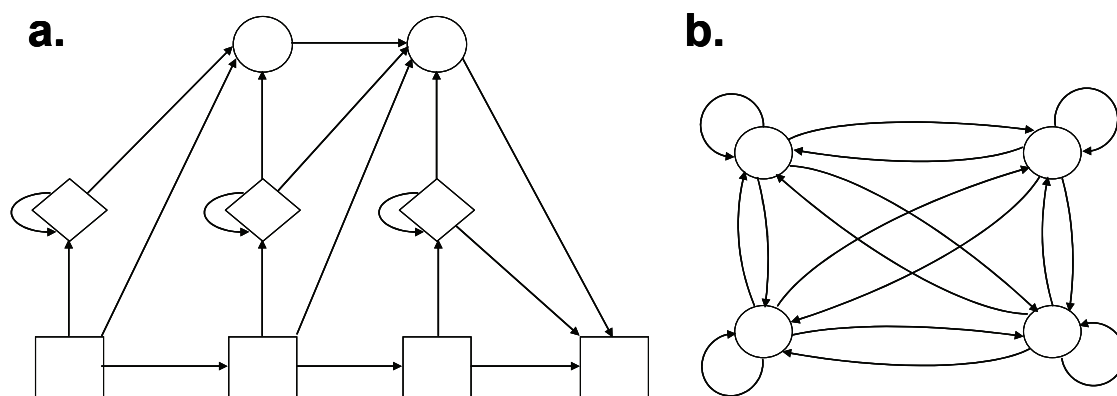


Figure 1.13: An illustration of the first HMM topologies implemented for T cell epitope prediction [400]. a. A pHMM and b. A fully connected HMM. The partial order of states and the lack of any given starting or terminating state in **a.** and **b.**, respectively, are evident.

By making use of tools such as the HMMER and SAM packages [403], pHMM are extrapolated from sets of aligned peptides. With the underlying directed graph being acyclic and an exception of loops, pHMM (Figure 1.13a) are linear left-right models. Therefore, they support a partial order of the states. Consisting of three classes of states known as the match state, the insert state and the delete state along with two sets of parameters namely the transition and emission probabilities [404], the pHMM architecture is unique. Amongst these states, always emitting a symbol are the match and insert states, while without emission probabilities the delete states act as the silent states. Requiring a drastically low computing power, pHMM are much weaker in modeling different patterns compared to fully connected HMM. pHMM derived from pMHC binding data are perhaps virtually identical to profile matrices or PSSM [404] due to almost ungapped alignments observed during pMHC binding [309]. Although structure-optimized HMM models have the capacity to model non-linear data, their connectivity compared to that of fully connected HMM is greatly reduced. Use of fully connected HMM is often associated with high computing costs. Furthermore, all HMM-based models can only be trained on known positive (binding) peptide data unlike other machine learning algorithms described above.

1.11.1.7 Support vector machines

Support vector machines (SVM) are a new type of machine learning algorithm where statistical learning methods are based on the structural risk minimization principle [405]. Due to their usefulness in identifying patterns, they are being extensively applied in life sciences [406-408]. Just like their predecessors (decision trees, ANN and HMM), SVM also have the skill to encode and work with both linear and non-linear data. Hence, SVM have also been utilized to predict T cell epitopes. Classification of data within SVM is done by separating the data optimally into categories which is carried out by constructing an N -dimensional hyperplane [309]. Representations of amino acid properties such as residue composition, solvent accessibility, charge, bulkiness, polarity and hydrophobicity are used to encode and assemble specific feature vectors that denote each peptide sequence that is being processed within a SVM. As noted above, the margin between the binders and non-binders is maximized by using an N -dimensional optimal separating hyperplane after mapping input vectors into a high dimensional feature space to train the parameters used for optimal classification.

Many researchers in the recent years have chosen to use SVM for T cell epitope prediction [343, 345, 346, 409] primarily due to their efficiency in the absence of large training datasets [173]. A consensus-based and combined prediction approach for T cell epitope prediction has also been embraced [410], where the authors have successfully integrated the strength of quantitative matrices, the robustness of ANN and the efficiency of SVM algorithms to create hybrid models. Although peptides are often represented in binary format, generally, different formats are used to encode the input peptide information or data that is used to train SVM or any given machine learning algorithm. However, that each peptide needs to be represented by a fixed length vector, presents a major limitation. To overcome this barrier, a kernel-based SVM trained on similarity scores of MHC-II binding allele-specific peptides, have recently been utilized to obtain better prediction results for MHC-II restricted T cell epitopes [411]. This approach was also able to model the influence of PFR on pMHC binding.

Subsequently, physicochemical properties of peptides known to bind to MHC-I proteins have been utilized for training SVM to improved prediction accuracy for MHC-I restricted T cell epitopes [344]. It is well known that unavailability of peptide binding data is a key limiting factor in the development of allele-specific T cell epitope prediction protocols. Recently, by combining the routine SVM formulation and a user-defined measure of

similarity between alleles, Jacob and Vert [342] have been able to predict T cell epitopes for MHC-I alleles with few known binders. Although sequence-based methods are well established and are frequently used to predict T cell epitopes, their use is still persistent with major limitations such as heavy reliance on the availability of large comprehensive training sets of peptides. Thus, in cases where the available data is insufficient, the sequence-based approaches are inappropriate for accurate prediction of peptides rendering their coverage only to subsets of binding peptides that belong to the most numerous groups. Therefore, for peptides that are least represented in the dataset [10, 49], these methods are unable to generate reliable data, implying that structural immunoinformatics is the only option for such peptides [11-14].

1.11.2 Structure-based approaches

Structure-based approaches are the methods that utilize three-dimensional data for detailed structural analysis of interactions between the MHC alleles and the respective bound antigenic segmental peptides [49]. These methods are not trained on peptide binding data and are exclusively based on the 3D structures of MHC proteins and pMHC complexes, hence are referred to as the structure-based methods. Due to the complexities and high developmental costs associated with structure-based methods, they have not been as extensively studied or used for T cell epitope prediction as the sequence-based methods. However, their applicability to any MHC allele means that these methods have the utmost potential to accurately predict T cell epitopes [309] provided their crystal structures are available. Therefore, diligent efforts have been made with great success in the recent years to both develop [10, 49, 412, 413] and apply [11-14, 414-416] various structure-based techniques for T cell epitope prediction. Described below are these structure-based techniques that have been successfully employed by researchers in the past few years for T cell epitope prediction.

1.11.2.1 Homology modeling

Originally developed in the early 1990s [413, 417, 418], homology modeling is arguably one of the most basic structure-based technique that was employed for T cell epitope prediction [415, 419]. It predicts the unknown structure of an amino acid sequence related to a homologous protein utilizing the available structure(s) of the homologous protein. As said, homology modeling has found extensive use in T cell epitope prediction. Here, given the 3D structures of bound peptides to homologous MHC proteins, homology modeling is primarily used for modeling the bound conformation of a peptide sequence with an

unknown structure. One of the first uses of homology modeling for T cell epitope prediction was reported when Hammer *et al.* [415] identified specific patterns of peptide binding from the pMHC crystallographic structure of a HA peptide bound to the human allele HLA-DRB1*0101 by constructing a series of synthetic pMHC models with designer peptides bound to other human rheumatoid arthritis associated (DRB1*0401 and DRB1*0404) and rheumatoid arthritis non-associated (HLA-DRB1*0402) alleles.

Utilizing their strategy, they were able to determine striking differences in pMHC binding for rheumatoid arthritis associated and non-associated alleles. Using homology models for different MHC-I proteins, Zhang *et al.* [420] were able to explain the structural principles that govern the development of peptide binding motifs for MHC-I alleles. Following this, construction of the bound conformation of peptides to a range of MHC-I alleles using a two-step approach was described by Rognan *et al.* [419]. The authors combined 3D models and a custom-built scoring function called “Fresno” to predict T cell epitopes restricted by the human MHC-I allele HLA-A*0204 (a close relative of the HLA-A*0201 allele) and the murine MHC-I allele H2-Kk. Crystal structures of many different proteins such as the 2C TR, the TR/pMHC complex of A6/Tax-HLA-A2, the 1934.4 TR V α chain, the 14.3.d TR V β chain and the pMHC complex of ovalbumin peptide-H2-Kb, were together used by Michielin *et al.* [421] to successfully develop a homology model of the TR/pMHC complex of T1/PbCS-H2-Kd.

Based on their previous success, Logean *et al.* [422] went on to compare and prove the superiority of their customized scoring function (Fresno) over other available scoring functions for T cell epitope prediction. They also applied a similar two-step prediction protocol to the one adopted by Rognan *et al.* [419] for HLA-B*2705 restricted T cell epitope prediction. Based on homology to the most similar MHC-bound peptide with available crystal structure, peptide termini are selected as the first step in their modeling protocol. Subsequently, by satisfaction of spatial restraints using a knowledge-based loop search procedure, the remaining residues were constructed as the second step in their technique. In 2002, identification of critical residues within the A6 TR interacting with peptide-HLA-A2 pMHC complex was presented by Michielin *et al.* [423] where they had applied their previously developed homology modeling-based methodology [421].

Recently, Kosmopoulou *et al.* have reported an improvised homology modeling-based approach to HLA-DQ2 and HLA-DQ7 restricted T cell epitope prediction. They have

subjected peptides that were initially identified using a combination of sequence patterns, quantitative matrices and ANN, to homology modeling using the crystal structure of the insulin-B peptide-HLA-DQ8 pMHC complex as a homologue. Finally, these structural models were placed into the structural models of the HLA-DQ2 and HLA-DQ7 MHC proteins, again built using the insulin-B peptide-HLA-DQ8 pMHC complex as a homologue and energy minimization was carried out to identify potential T cell epitopes. Although useful, use of this technique is basic, includes complexities in developing high-throughput T cell epitope prediction models and is not very accurate.

1.11.2.2 Protein threading

An improvised technique compared to homology modeling, protein threading [412] is the name given to the practice of computing an alignment between the spatial positions of a 3D structure and a target amino acid sequence. It is also generally referred to as side-chain conformational search [424]. Protein threading is made use of in T cell epitope prediction to replace the target peptide residues ($S_1, S_2 \dots S_n$) for the amino acids ($P_1, P_2 \dots P_n$) of a source peptide (by substituting P_i with S_i) bound to a MHC protein of interest. Usually, a scoring scheme for peptides is applied to discriminate the binders from non-binders after performing a search for the best side-chain conformations for the peptides. The first use of protein threading for T cell epitope prediction was documented by Altuvia *et al.* [414] when they introduced a reasonably quick and accurate (compared to homology modelling-based protocols) structure-based algorithm to predict HLA-A*0201 restricted T cell epitopes.

By adopting the protein threading technique and the knowledge-based potential matrix of Miyazawa and Jernigan [425], they were able to successfully detect binding peptides not conforming to HLA-A*0201 binding motifs. Subsequently, Altuvia *et al.* [426] extended this algorithm to predict T cell epitopes for a multitude of MHC-I alleles. The underlying feature that governs the predictive ability of their approach comprises of utilizing a form of protein threading called as peptide threading to fit the peptides within the peptide binding cleft of the MHC-I proteins and then assess the pairwise pMHC interactions by adopting the above mentioned statistical pairwise potential table or matrix of Miyazawa and Jernigan [425]. This approach was also exploited by Schueler-Furman *et al.* [427] to predict T cell epitopes for 23 MHC-I proteins. However, this approach could only identify T cell epitopes for MHC-I proteins with hydrophobic binding pockets and was

unsuccessful in identifying T cell epitopes for MHC-I alleles with pockets that are charged or hydrophilic in nature.

To counter this issue, by observing the number of solvent exposed hydrophobic residues on modeled peptides and the number of atomic clashes in pMHC binding, Kangueane *et al.* [428] ingeniously introduced the concept of knowledge-based rules for successful discrimination of binders and non-binders. Learning from the knowledge and useful information generated from researching the structures of pMHC complexes, Schueler-Furman *et al.* [351] were able to fruitfully overcome the hurdle of developing a generalized (applicable to most MHC-I proteins) protein threading-based T cell epitope prediction algorithm. They combined a different pairwise potential table, previously described by Betancourt and Thirumalai [429], with the peptide-threading approach to effectively come up with an algorithm that described hydrophilic interactions more appropriately, resulting in improved prediction for MHC-I alleles with hydrophilic binding pockets.

Following this, Zhao *et al.* [430] adopted a similar combined procedure and produced another novel knowledge-based potential matrix which, in combination with the peptide threading technique, allowed T cell epitope prediction for most MHC-I alleles. Recently, Singh and Mishra [431] have utilized the inverse folding approach by tethering protein threading, Miyazawa and Jernigan and Betancourt and Thirumalai pairwise potential tables together to predict T cell epitopes for MHC-I proteins. Peptide threading has also very recently been used to predict MHC-II restricted T cell epitopes [432]. Despite being superior to the basic homology modelling technique, this approach fails to account for both the appropriate inclusion of the flexibility of peptide side-chains and the flexibility of the MHC peptide binding groove (binding pockets) itself.

1.11.2.3 Docking

Among all other techniques used for T cell epitope prediction, docking is perhaps the most successful to date [11-14]. It is the name used to describe a very powerful and systematic computer-aided technique to investigate intermolecular interactions. It is also known as computer-simulated ligand binding. This is because this technique essentially performs *in silico* simulation or computer-based simulation of receptor and ligand binding. Usually this is carried out as a two-step protocol which involves: (i) rational determination of the most appropriate conformational, translational and rotational concurrence for a particular receptor-ligand pair and; (ii) evaluation of how well a ligand can bind to its receptor or the

relative goodness-of-fit, habitually estimated by calculating the binding energy (BE) value for receptor-ligand binding which can then be used for scoring purposes to identify the best fitting ligands among a group of ligands.

The advantages and robustness of docking have resulted in this technique being successfully applied to predict T cell epitopes for various MHC alleles in the recent years [11-14, 433]. So much so, that there has been a surge in both the development [10, 11, 49] and application [11-14, 433] of docking-based protocols to address the problems associated with T cell epitope prediction, in the past few years alone. Generally, methods pertaining to docking can be sub-divided into rigid, semi-flexible and flexible docking and typically, after the calculation of BE scores for all input peptides, they involve scoring of a series of peptide candidates using energy-scoring functions for T cell epitope prediction. Initially, rigid docking of the influenza matrix peptide to the human MHC-I allele HLA-A*0201 with the help of a Monte Carlo-based combinatorial build-up algorithm, was documented by Caflisch *et al.* [434] in 1992.

This triggered persistent efforts from various research groups around the globe to improve the quality and speed of the docking protocol for effective prediction of the structures of the bound peptides to MHC proteins as a first step for subsequent T cell epitope prediction. Resultantly, Rosenfeld *et al.* [435] presented a semi-flexible protocol for docking of peptides to MHC proteins where only the peptide backbone was rendered flexible and the MHC pocket residues were rendered rigid. They also applied this protocol to predict bound peptide structures for the human MHC-I allele HLA-A*0201 and the murine MHC-I allele H2-Kb. Later, Rosenfeld *et al.* [436] developed another protocol for semi-flexible docking of peptides utilizing a multiple copy algorithm to identify probable peptide termini conformations and constructing the structure for the middle residues of the peptide sequences using a loop closure algorithm.

In the very next year, Sezerman *et al.* [424] created a semi-flexible docking protocol by making use of the free energy maps of the MHC pocket residues to generate the docked conformations of the peptides for MHC-I proteins. However, all these methods were still treating the MHC binding site residues (pocket residues) as rigid entities and hence their accuracies varied. To conquer this obstacle, Desmet *et al.* [437] built a docking algorithm that treated both the peptide and the MHC binding cleft residues flexibly, thereby, giving birth to use of flexible docking protocols for the prediction of the bound conformation of

peptides to MHC proteins. Consequently, efforts were directed at developing new and improved flexible docking procedures that could cope with the complexities of pMHC binding and T cell epitope prediction. This resulted in the development of an accurate flexible docking protocol by Tong *et al.* [10, 11] by integrating the strength of Monte Carlo simulations and homology modeling to dock peptides to a number of MHC-I and MHC-II alleles.

They have also been successful in the implementation of their multi-step docking protocol [10, 11] to predict T cell epitopes (even for MHC alleles with limited peptide binding data) by making use of a custom-built *BE* scoring function [11-14]. Subsequently, Bordner and Abagyan [433] developed a Biased-Probability Monte Carlo procedure for flexible docking of peptides to MHC proteins and were able to successfully predict T cell epitopes for human MHC-I allele HLA-A*0201 and the murine MHC-I allele H2-Kb. However, the speed, accuracy and efficiency of these flexible docking protocols needed to be improved for high-throughput screening of peptides for fast and efficient T cell epitope prediction. To surmount this difficulty, Khan and Ranganathan [49] have recently built “pDOCK” (see publication 3 for details) by combining the strength of the Biased-Probability Monte Carlo procedure and the efficiency of grid-based fully flexible docking.

Their method (pDOCK) is a fully flexible docking protocol that renders full flexibility to the peptide backbone, peptide side-chains, the MHC binding cleft residue backbones and their side-chains. pDOCK is also a rapid, robust and efficient method for docking of peptides to both MHC-I and MHC-II alleles. It has been noted that accurate prediction of the appropriate geometry of peptides bound to both MHC-I and MHC-II proteins can drastically improve the accuracy of T cell epitope predictions, suggesting the usefulness and advantages of using fully flexible docking for T cell epitope prediction. However, like all structure based techniques, automation of fully flexible docking protocols for the development of web-servers to make them comparable with sequence-based methods in terms of high-throughput T cell epitope prediction, still poses a challenge mainly owing to the technical complexities and computational costs involved.

1.11.2.4 Molecular dynamics

Molecular dynamics (MD) is another form of generating computerized models of 3D protein structures through computer simulation of their physical movements based on statistical mechanics. It combines the abilities of molecular modeling and computer

simulation and hence, is a powerful and flexible tool to predict and/or analyse molecular and macromolecular systems [416]. Due to its qualities, MD has often been used in structure-based drug design. In the process of T cell epitope prediction, MD is generally used to sample the conformational spaces of the input peptides within the fixed environments of their respective MHC binding clefts [438]. Although MD has been in use for T cell epitope prediction since the early 1990s, progress in utilizing MD for T cell epitope prediction has been relatively slow mainly due to the fact that researchers have resented from the use of this technique owing to the complexities involved in simulating the pMHC and/or the TR/pMHC interactions.

The first use of MD for T cell epitope prediction was documented in 1994 by Rognan *et al.* [439]. They simulated the pMHC interactions for the human MHC-I allele HLA-B*2705 and six different peptides and found evidence supporting the importance of the role played by the peptide residues other than the anchor residues in pMHC binding. Similarly, structures of the human MHC-I allele HLA-A*0201 in complex with 9-mer peptides, were examined by Lim *et al.* [440] through the use of MD simulations. Their investigation led to conclude that the C-terminal residues of the non-binders are rotated away from the binding pockets by a conformational change within the non-binders resulting in subsequent release of the peptides from their respective MHC binding clefts. Following this, MD was used in conjunction with other techniques for T cell epitope prediction. A decade later however, Fagerberg *et al.* [438] used an advanced MD based technique known as simulated annealing (SA) to sample the conformational space of the peptides binding to MHC-I proteins.

A year later, Sieker *et al.* [441] performed a comparative analysis of tapasin-dependent pMHC binding and studied the conformational flexibility of the human MHC-I alleles HLA-B*4402 and HLA-B*4405 in the presence and absence of bound peptides using MD simulations. As noted above, up until now, the progress of using MD for T cell epitope prediction has been tentative, but emergence of high-performance computing and the development of coarse-grained simulation has now enabled researchers to exclusively use MD for not only simulating pMHC interactions but also TR/pMHC interactions along with the cell membranes and CD proteins as a whole to form the entire immunological synapse or the “immune complex” [416]. In this regard, Flower *et al.* [416] have very recently described the use of MD simulations to calculate the free energies of binding for pMHC

and TR/pMHC interaction in four distinct immune complexes and present the potentiality of MD as a possible T cell epitope prediction technique.

Therefore, the potential use of MD simulations for T cell epitope prediction presents an exciting prospect for the future due to its advantages in modeling the details of all dynamic behaviour involved during pMHC and TR/pMHC interactions including the details of the solvent and the ionic environments within which they occur [416]. Nevertheless, MD simulations have two prominent limitations: (i) the short time scale of MD simulations compared with those exhibited by biological processes reflected by the inadequate sampling of conformational space and; (ii) most, if not all, empirical force fields remain highly approximate despite increasingly sophisticated parameterization [416]. Hence, the intrinsic validity of such simulation exercises is still a cause of concern. Although, the emergence of supercomputing and widely accessible parallel MD code have addressed the first issue, the second problem still remains unaddressed despite many attempts with limited success as increasing the time scales results in more and more approximations [416]. Besides these prominent issues, the need for crystal structures of pMHC and TR/pMHC complexes and enduring limitations pertaining to molecular modeling and computer-aided simulation have also hampered the use of MD for T cell epitope prediction. Finally, their present success rate is very much lower than that of other structure-based techniques such as docking, implying that much work remains to be done on developing, refining and applying this technique for T cell epitope prediction.

1.11.3 Sequence-structure-based approaches

As referred to earlier, various research groups have used methods that combine the sequence-derived pMHC binding affinity data and 3D structural information of MHC proteins and pMHC complexes to predict T cell epitopes. Despite this, a few prominent examples of success by combining the two aspects have been documented. Doytchinova and Flower [160] were the first to introduce a discrimination schema by employing similarity indices and 3D quantitative structure-affinity relationship (QSAR) using the powerful comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) to predict T cell epitopes among 102 peptides known to bind the human MHC-I allele HLA-A*0201. Following this, Doytchinova and Flower [310] again utilized the power of molecular CoMSIA-based 3D-QSAR technique to predict T cell epitopes from a set of 266 peptides for the human MHC-I allele HLA-A*0201.

Computational predictive models or algorithms capable of extrapolating T cell epitope predictions for MHC proteins with very little binding data can also be generated using this combination of peptide binding data and 3D structural information. This was demonstrated recently by Jojic *et al.* [311] who implemented an adaptive double threading approach by making use of the peptide binding affinity data and the predictive ability of the protein threading technique to predict T cell epitopes for HIV related MHC-I alleles. Although relatively successful, these approaches inherit the coherent limitations of using both sequence based approaches such as similarity indices and structure-based approaches such as protein threading. These shortcomings greatly influence and/or limit the ability of these conjunctive approaches. Hence, successful use of this approach requires addressing several potential limitations that could emerge due to the use of sequence and structure data such as the need for significant amounts of peptide binding affinity data and the complexities with using structural data.

While a comprehensive overview of the field of structural immunoinformatics and its applications in peptide based vaccine design has been presented above, recent progress and previous work on successful prediction of T cell epitopes [11-14] using an accurate docking strategy [10, 11] has been reviewed in publication 1 below. Following this, an in-depth explanation pertaining to TR/pMHC complex formation, T cell activation, current knowledge of TR/pMHC interactions and their significance in clinical medicine is presented in publication 2 along with a preview of the newly characterized TR/pMHC interaction parameters applied to one TR/pMHC complex (PDB code: 1oga) as a primary understanding of the TR/pMHC binding.

Although we have used simplistic terminology such as ‘TR footprint’ to describe the residues on the pMHC interface that contact the TR in publication 1, we have adhered to standardized IMGT terminology such as ‘pMHC epitope’ and ‘TR paratope’ to describe residues on pMHC interface that contact the TR and residues on TR interface that contact the pMHC, respectively, from publication 2 onwards.

Pages 57-73 of this thesis have been removed as they contain published material under copyright. Removed contents published as:

Khan J.M., Tong J.C., Ranganathan S. (2009) Structural Immunoinformatics: Understanding MHC-Peptide-TR Binding. In: Flower D., Davies M., Ranganathan S. (eds) *Bioinformatics for Immunomics. Immunomics Reviews: (An Official Publication of the International Immunomics Society)*, vol 3. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-0540-6_7

TR recognition of MHC-peptide complexes

Javed Mohammed Khan, Department of Chemistry and Biomolecular Sciences
and ARC Center of Excellence in Bioinformatics, Macquarie University, NSW
2109, Australia, javed.khan@mq.edu.au

Shoba Ranganathan, Department of Chemistry and Biomolecular Sciences and
ARC Center of Excellence in Bioinformatics, Macquarie University, NSW 2109,
Australia and Department of Biochemistry, Yong Loo Lin School of Medicine,
National University of Singapore, 8 Medical Drive, Singapore 117597,
shoba.ranganathan@mq.edu.au

Synonyms

TCR recognition of MHC-peptide complexes, TR/pMHC interaction, TCR-pMHC binding.

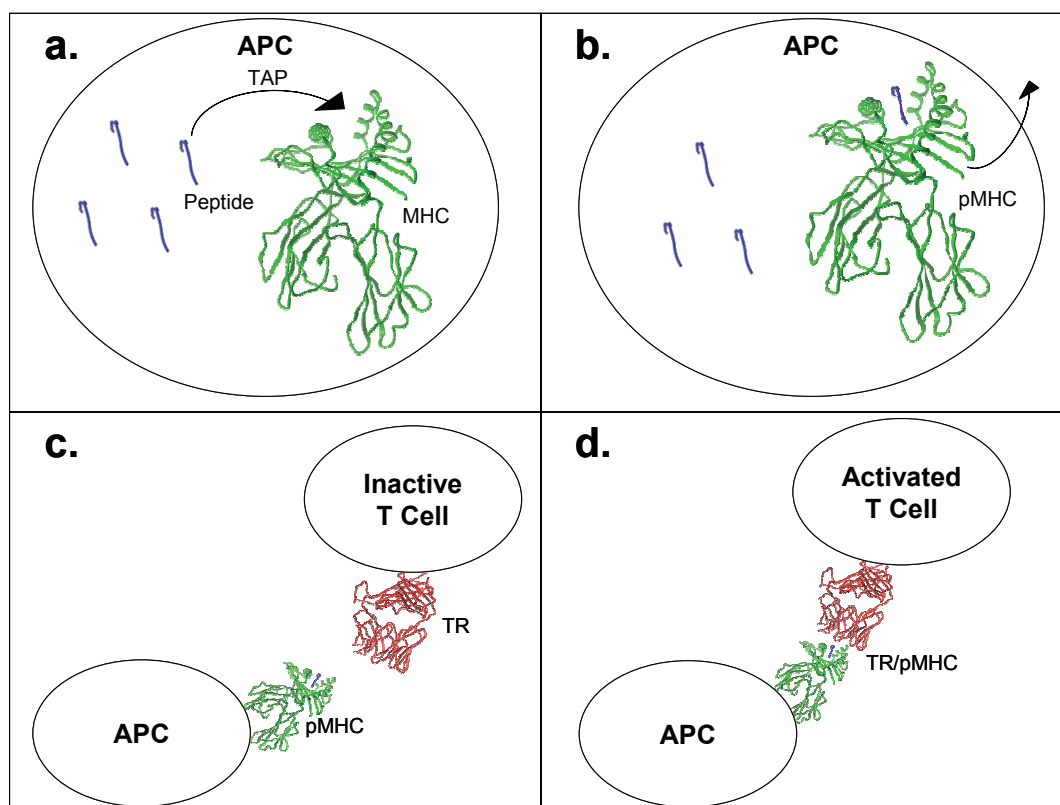


Figure 1. Steps leading to T cell activation in adaptive immune response. a. antigen processing and pMHC complex formation. b. transportation of pMHC to APC surface for presentation. c. pMHC presentation and TR surveillance. d. TR recognition of pMHC, TR/pMHC complex formation, T cell signaling and activation.

Definition

TR recognition of MHC-peptide complexes is the name given to the vital immunological synapse within the adaptive immune system of higher jawed vertebrates where, the antigenic peptide bound major histocompatibility complexes (pMHC) are recognized and bound by T cell receptor (TR) at the antigen presenting cell (APC) surface for T cell

signaling (Fig. 1) leading to an immediate immune response to either destroy infected cells directly (via CD8⁺ cytotoxic T cells) or activate (via CD4⁺ helper T cells) other immune system cells like B cells and macrophages to carry out the immune cascade.

Characteristics

The adaptive immune system plays an important role in defending higher jawed vertebrates against infectious, autoimmune, allergic and graft vs. host diseases. It is named “adaptive” due to its ability to adapt and respond to an ever changing variety of new pathogens thereby conferring long-lasting or protective immunity to the host. This significant phenomenon within the body’s defense mechanism works under the influence of a series of vital protein-protein interactions. These essential interactions are mediated by certain highly specific and selective proteins, similar to those involved in antibody or B cell mediated immune response (Alberts et al. 2002). Amongst these proteins, the most important from a clinician’s perspective are the ones involved in T cell activation namely, major histocompatibility complexes (MHC), antigenic or immunogenic peptides (p) derived from antigens and T cell receptors (TR) proteins. To maximize the immunological protection against a vast repertoire of pathogens, the adaptive immune response cascade causes MHC or human leukocyte antigens (HLA) in human, to bind to immunogenic peptides and present them as peptide-MHC (pMHC) complexes on the surface of antigen-presenting cells (APC), for recognition by TR (Fig. 1) which are bound to the surface of the T cells (Rudolph et al. 2006). Upon recognition by the TR, the TR and pMHC bind to form a ternary TR/pMHC complex which is called as the immunological synapse. This synapse activates the T cells leading to an immediate immune response to either destroy infected cells directly (via CD8⁺ cytotoxic T cells) or activate (via CD4⁺ helper T cells) other immune system cells like B cells and macrophages to carry out the immune cascade. Although it has been more than a decade since the first TR/pMHC structure was reported (Garboczi et al. 1996), this interaction still poses an intricate theoretically and structurally unscaled frontier in Structural Immunoinformatics. Therefore, it is extremely important to understand TR recognition of MHC-peptide complexes at the molecular level with a focus on its various physicochemical properties, beginning with an in-depth knowledge of the essential components involved, in order to gain insights into the likes and dislikes of a TR protein towards a specific pMHC complex and to comprehend the potential of a peptide epitope to elicit T cell response which, today serves as the first step in vaccine development through Reverse Vaccinology.

The key players:

MHC

MHC proteins are expressed within most cells and are arguably the most important element of T cell mediated immunity. They are structurally and functionally similar to antibodies secreted by B cells (Alberts et al. 2002). Typically, the MHC proteins are composed of two chains, α and β and are broadly classified into two types, MHC class I (MHC-I) proteins and MHC class II (MHC-II) proteins. MHC-I proteins are heterodimers, consisting of a heavy α chain (I-ALPHA) of about 45 kDa, and a light chain, β 2-microglobulin (B2M) of about 12 kDa with the α chain (I-ALPHA) consisting of α 1 (G-ALPHA1), α 2 (G-ALPHA2) and α 3 (C-LIKE) domains where G-ALPHA1 and G-ALPHA2 domains form the peptide binding groove or ‘cleft’ (Lefranc et al. 2005; Fig. 2a).

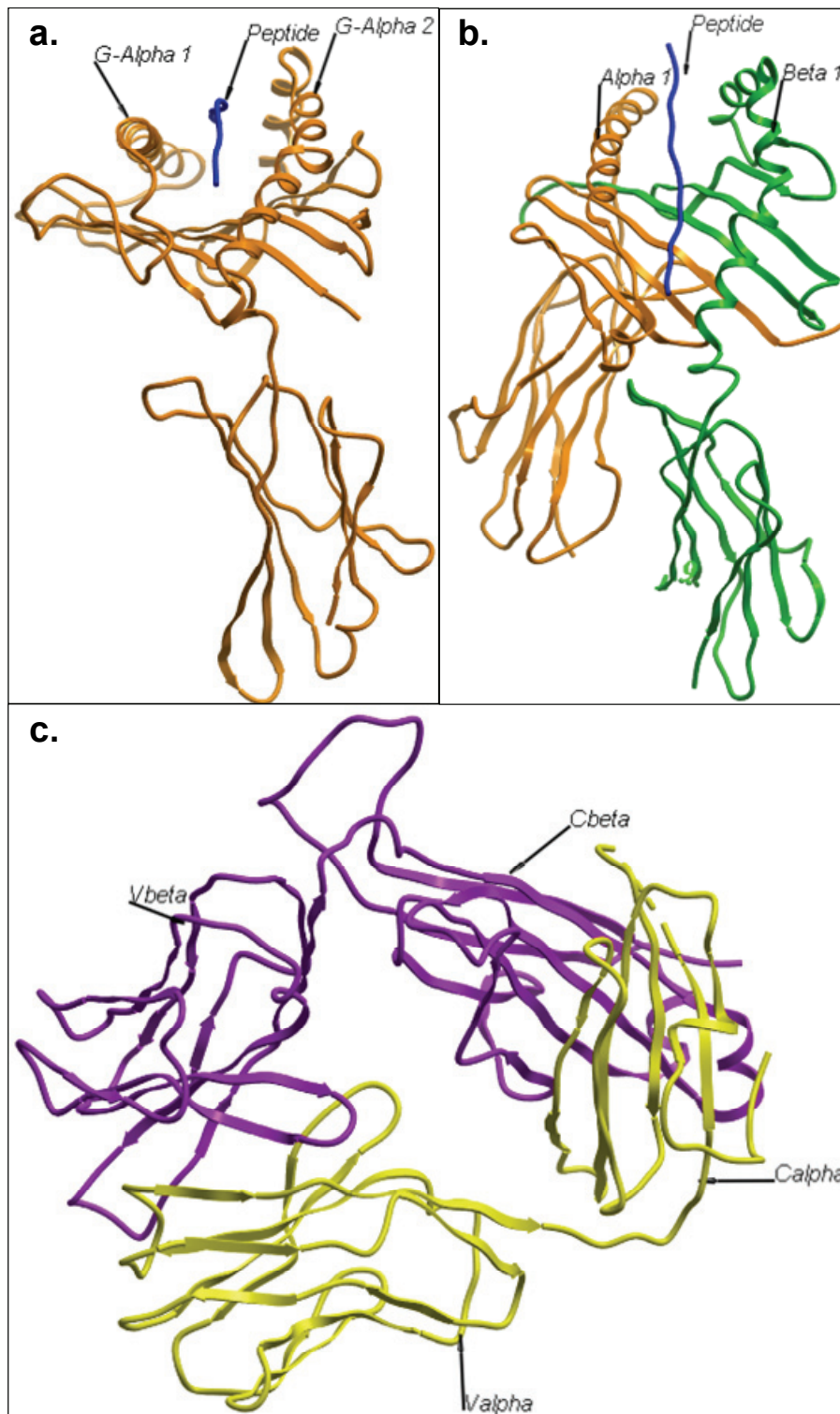


Figure 2. The key elements of T cell dependant immunity. a. MHC-I α chain (I-ALPHA; gold) from the PDB structure 1ao7 (Garboczi et al. 1996), with the bound peptide (blue) showing the peptide binding domains. b. MHC-II α chain (II-ALPHA; gold) and β chain (II-BETA; green) from the PDB structure 1fyt (Hennecke et al. 2000), depicting the peptide binding domains with the cognate peptide (blue). c. a typical TR protein from the TR/pMHC-II structure 1fyt (PDB code; Hennecke et al. 2000), with the two chains (α - yellow and β - purple) portraying the two constant and two variable domains on either chains.

MHC-II proteins are also heterodimeric proteins consisting of a α chain (II-ALPHA; 34 kDa) and a β chain (II-BETA; 29 kDa) with very similar overall quaternary structure to that of MHC-I proteins (Lefranc et al. 2005). However, their peptide binding groove is formed by the $\alpha 1$ and $\beta 1$ domains (Fig. 2b) of the two chains, α (II-ALPHA) and β (II-BETA). Generally, MHC-I complexes bind and present endogenous (processed within the cell) peptides whereas MHC-II complexes prefer exogenous (processed outside the cell) peptides.

Immunogenic peptide antigens

Immunogenic peptides or T cell epitopes are essential subunit peptides that are required to stimulate cellular immune responses, especially the adaptive immune responses. Peptide epitopes are presented for surveillance and recognition by the TR in an MHC allele (polymorphic MHC proteins) and supertype (groups of MHC proteins with similar peptide binding properties) dependent manner and can be of endogenous or exogenous origins. Usually, Peptides between 8-11 amino acids in length are presented by MHC-I. Cytosolic proteases within the cytosol of the cell 'chop' these peptides, which are then carried by "transporters associated with antigen processing" (TAP) proteins in an ATP-dependent manner to the MHC binding groove, for pMHC complex formation (Tong et al. 2004). This pMHC complex is then translocated to the cell surface and presented for recognition by the TR of CD8⁺ cytotoxic T cells or cytotoxic T lymphocytes (CTL). MHC-II presents peptides that are generally 12-20 amino acids in length. Endocytosed into the cell by the lysosomes, these peptides displace the native MHC-II ligand known as the 'CLIP' peptide, to form the pMHC complex (Tong et al. 2004). Just as with the pMHC-I complexes, pMHC-II complexes are then presented at the APC surface for recognition by the TR of CD4⁺ T helper cells.

TR

The TR proteins are another vital part of T cell dependant immune response. They function in a similar way as some cell surface receptors of the B cell mediated immunity such as, Fc receptors found on the surface of macrophages or neutrophils which bind to the antigen-bound antibody, resulting in phagocytosis and lysis of the antigen or pathogen by the macrophages or neutrophils (Alberts et al. 2002). The difference here is that, upon TR/pMHC complex formation, the TR proteins do not actually cause the T cell to ingest and break down the pathogen. Instead, they trigger T cells to destroy the infected cells either directly or indirectly as described above. A typical $\alpha\beta$ TR has two chains, α and β (Fig. 2c) which are divided into constant (encoded by the conserved constant (C) gene segment of the TR coding genes) and variable domains (encoded by rearranged variable (V), diversity (D) and joining (J) gene segments, V-J for α chain and V-D-J gene segments for β chain, respectively) which perform specific functions. The two conserved or constant domains ($C\alpha$ and $C\beta$; Fig. 2c) of the TR anchor it to the T cell surface through a transmembrane region. These constant domains are linked to the upper more diverse or variable domains ($V\alpha$ and $V\beta$; Fig. 2c) which recognize the pMHC at the TR/pMHC binding interface.

TR/pMHC interaction: what's understood

Many theories have been put forward as an answer to comprehend the rationale behind TR/pMHC interaction. An interesting one is the "TR germline bias" for MHC which

suggests that the basis of MHC restriction or TR specificity is a set of specific conserved and localized contacts between TR V gene (variable gene) products and MHC gene products that co-evolve (Jerne 2004). The combinatorial diversity problem due to a large number of antigenic peptides, the variety in the variable regions of TR proteins and many greater number of MHC alleles all complicate the issue further, contradicting the simplistic explanation provided by the TR germline bias theory. The cross-reactivity of MHC proteins implies the ability to the TR to briefly scan through several pMHC complexes before actually interacting with and binding to a specific one. Over the years, researchers have singled out many factors that could contribute to or influence the TR/pMHC binding.

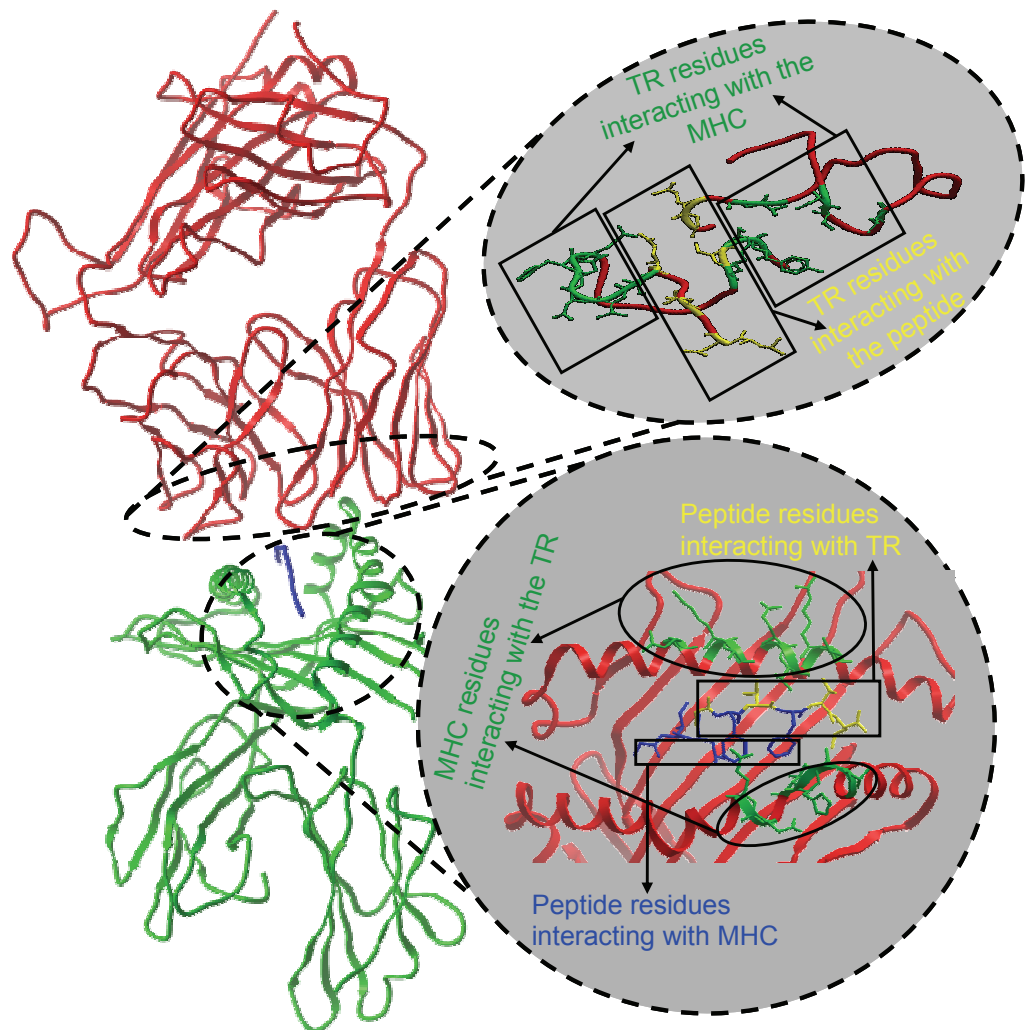


Figure 3. What does a TR “see”: TR/pMHC interaction zones in the structure 1oga (PDB code; Stewart-Jones et al. 2003). Inset - above: TR paratope and below: pMHC epitope (except the peptide residues that interact with the MHC – shown in blue). The MHC residues interacting with the TR and TR residues interacting with the MHC are in green. Similarly, the peptide residues interacting with the TR and TR residues interacting with the peptide are highlighted in yellow. All interacting residues are shown in the stick representation.

However, a thorough literature survey helped us identify four major factors: 1. Binding energy (BE) or binding free energy between the TR and the pMHC ligand; 2. Electrostatic

Potential computed and displayed on the TR and the pMHC interfaces (Rudolph et al. 2006); 3. The angle formed by docking of the TR onto the pMHC surface and; 4. Residues or certain broadly conserved structural determinants among pMHC and TR interacting sequences (Rudolph et al. 2006; Fig. 3) or in other words pMHC epitope (residues on pMHC interface that contact the TR; Kaas and Lefranc, 2005) and TR paratope (residues on TR interface that contact the pMHC; Kaas and Lefranc, 2005), that would constitute the “smoking gun” of “MHC bias” (Garcia et al. 2009).

A recent study on a limited subset (approximately 20) of TR/pMHC X-ray crystal structures has given some promising and favourable results to TR germline bias theory (Garcia et al. 2009). Nevertheless, vital interaction parameters like electrostatic interactions between the TR and the pMHC have not been taken into account by the authors, which could reveal the physicochemical basis of TR/pMHC interaction. Moreover, many more TR/pMHC crystal structures are now available. Hence, continuing from our preliminary analysis of the 1oga (PDB code; Stewart-Jones et al. 2003) complex (Khan et al. 2010), we have applied these new interaction parameters to the same structure (1oga; Stewart-Jones et al. 2003) for a primary understanding of the TR/pMHC binding. A BE value of -11.99 kcal/mol and a TR docking angle of 69° have been computed. The electrostatic potential for the pMHC interface of this TR/pMHC structure depicts a set of complementary charges on the TR and pMHC surfaces, which could serve as one of the underlying principles for TR recognition of pMHC complexes. However, a detailed and in-depth analysis of a larger subset of available TR/pMHC X-ray crystal structures using these factors as important interaction criteria, is needed in order to thoroughly understand the governing aspects of the pMHC recognition and TR signaling, perhaps a few conserved residues at the TR/pMHC binding interface.

Recent advances and implications in vaccine development

Considering the critical role that the peptide plays in determining TR specificity, identification of true T cell epitopes from repertoires of immunologically significant antigenic peptide sequences becomes a vital prerequisite in the process of conventional molecular vaccine design. Identifying T cell epitopes experimentally is a tedious, time consuming and expensive process, owing to the combinatorial diversity problem (mentioned earlier) and the extremely low chance of immunogenicity (1 in 2000 peptides; Khan et al. 2010). Recently, advanced computational methods have proven to be vastly time and cost efficient in screening the vast oceans of peptides and MHC repertoires. Current computational methods can be broadly classified into: sequence-based and structure-based approaches. The former generally require extensive sequence data for training, whereas the latter utilize three-dimensional structural analysis of interactions between the MHC and bound segmental antigenic peptides (Khan et al. 2010). Although sequence-based methods are well established and suitable for large-scale screening of potential T cell epitopes, a major limitation of such techniques is the heavy reliance on the availability of large comprehensive training sets of peptides (Khan et al. 2010). Hence, these approaches are inappropriate for accurate prediction of peptides in circumstances where the available data is limited. On the other hand, structure-based protocols work better for detailed analysis of short immunogenic regions of antigens and can generate reliable data for peptides that are least represented in a dataset (Khan et al. 2010), as they are computationally intensive and time consuming.

The development of new structural modeling and docking techniques and an increase in the number of protein structures is resulting in accurate structure-based flexible-docking approaches being more commonly used to predict potential T cell epitopes (Khan et al. 2010). Often producing modeled/docked structures of peptide ligands accurate to within 2.00Å root mean square deviation (RMSD) from the experimental crystal structure, these approaches provide a wealth of information for structural analysis and improvement of epitope prediction methods. An initial but accurate flexible-docking method (Tong et al. 2004) helped us accomplish quantitative predictions for both MHC-I and MHC-II alleles, with limited binding peptide data (Khan et al. 2010). However, this method is relatively slow. Therefore, there is an urgent need for a faster, more robust and accurate docking technique which, along with the results of a comprehensive analysis (mentioned earlier), could together form the basis of successful *in silico* identification of true T cell epitopes, from a large number of predicted MHC-binding peptides, for subsequent *in vitro* immune response assessment. Such an approach would significantly reduce the lead time involved in experimental vaccine development methods, resulting in swift production of effective, highly specific and efficient peptide vaccines.

Importance of supposedly insignificant molecules

The antigen and MHC allele-specific interaction between a TR on a T cell and a pMHC complex on an APC, appears to be governed largely by the composition and the electrostatic interaction on the TR and the pMHC interface regions (Rudolph et al. 2006). However, other proteins, which are usually considered insignificant, also play a significant role in this vital immunological synapse. For example, interactions between adhesion proteins called intercellular adhesion molecule–1 (ICAM-1) present on the APC and leukocyte function-associated antigen–1 (LFA-1) present on the T cell surface, bring the APC and T cell close to each other, leading to the formation of the immunological synapse (Alberts et al. 2002; Rudolph et al. 2006). Cluster of differentiation (CD) proteins bound to the T cell surface also contribute to the TR/pMHC binding. It is well known that the CD8 proteins specifically recognise MHC-I proteins and CD4 proteins are specific for MHC-II proteins. This could imply allele related specificity to the TR. The β -2 microglobulin chain found alongside the MHC α -chain in MHC-I structures, also has a stake in the synapse formation, by partially recognizing the CD8 proteins along with the MHC α -chain lower (constant) region.

These proteins that support the immunological synapse are collectively called the costimulatory proteins. Several structures have been reported highlighting the interaction of these proteins, which bear witness to the importance of their role in TR/pMHC binding and TR activation (Gao et al. 1997, Liu et al. 2003). It could also be inferred that once the TR recognises the pMHC complex, it is the interaction of the CD proteins that locks the TR/pMHC complex together, thereby giving the TR enough time to stabilize itself on the pMHC surface, resulting in T cell activation or immune response. Another important aspect of TR/pMHC binding is the presence of water molecules in and around the TR/pMHC complexes. These water molecules are usually considered to be water molecules of crystallization, but some of these lie within 4 Å from both the TR and the pMHC residues, forming hydrogen-bonded bridges between the pMHC and the TR residues and could be vital for the immunological synapse to occur. Thus, the activation of T cells depends not only on TR engagement with pMHC but also on the interaction of

costimulatory proteins. However, as these costimulatory proteins do not bind and/or present any antigenic/immunogenic peptide determinants, the primary players in the development of peptide-based vaccines through reverse vaccinology remain the immunogenic peptides, presented by MHC for recognition by TR proteins.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: *Molecular biology of the cell*, 4th edn. Garland Science, New York, pp 1363-1421.

Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* 24: 419-466.

Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* 384: 134-141.

Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriout A, Lefranc G (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* 29 (11): 917-938.

Hennecke J, Carfi A, Wiley DC (2000) Structure of a covalently stabilized complex of a human alphabeta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. *EMBO J* 19: 5611-5624.

Tong JC, Tan TW, Ranganathan S (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci* 13 (9): 2523-2532.

Jerne NK (2004) The somatic generation of immune recognition. 1971. *Eur J Immunol* 34: 1234-1242.

Kaas Q, Lefranc MP (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* 5: 505-528.

Garcia KC, Adams JJ, Feng D, Ely LK (2009) The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* 10: 143-147.

Stewart-Jones GB, McMichael AJ, Bell JI, Stuart DI, Jones EY (2003) A structural basis for immunodominant human T cell receptor recognition. *Nat Immunol* 4: 657-663.

Khan JM, Tong JC, Ranganathan S (2010) Structural Immunoinformatics: Understanding MHC-peptide-TR binding. In: Davies MN, Ranganathan S, Flower DR (eds) *Bioinformatics for Immunomics*, vol 3 (Immunomics Reviews Series). Springer, New York, pp 77-94.

Gao GF, Tormo J, Gerth UC, Wyer JR, McMichael AJ, Stuart DI, Bell JI, Jones EY, Jakobsen BK (1997) Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2. *Nature* 387: 630-634.

Liu Y, Xiong Y, Naidenko OV, Liu JH, Zhang R, Joachimiak A, Kronenberg M, Cheroutre H, Reinherz EL, Wang JH (2003) The crystal structure of a TL/CD8alphaalpha complex at 2.1 Å resolution: implications for modulation of T cell activation and memory. *Immunity* 18: 205-215.

Definitions

Adaptive Immune System

Synonyms

Adaptive immune response cascade, Adaptive immunity.

Definition

The adaptive immune system is a collective term given to a group of highly specialized, systematic cells and processes that prevent vertebrates from certain death by pathogenic infections (Alberts et al. 2002).

Reference

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: *Molecular biology of the cell*, 4th edn. Garland Science, New York, pp 1363-1421.

T cell signaling

Synonyms

T cell receptor signaling, TCR signaling, TR signaling.

Definition

A number of signaling cascades that occur after TR/pMHC binding and promote T cell activation through regulated production of cytokines to ultimately determine infected cell fate are together called as the T cell signaling process (Alberts et al. 2002; Rudolph et al. 2006).

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: *Molecular biology of the cell*, 4th edn. Garland Science, New York, pp 1363-1421.

Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* 24: 419-466.

B cell mediated immune response

Synonyms

B cell mediated immunity, Antibody dependant immune response, Antibody mediated immunity.

Definition

B cell mediated immune response is defined as the immune response cascade triggered by the binding of antibodies (produced by the B cells) to the antigens and subsequent identification by the cell surface receptors of macrophages, neutrophils or other cells of the B cell mediated immunity to destroy the antigens. It is a type of adaptive immunity in vertebrates (Alberts et al. 2002).

Reference

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: Molecular biology of the cell, 4th edn. Garland Science, New York, pp 1363-1421.

T cell activation

Synonyms

T cell receptor activation, TCR activation, TR activation.

Definition

Prior to target (infected) cell killing or activation of other immune system cells to do the same, by the cytotoxic or helper T cells, respectively, the T cells must be activated and this activation, called T cell activation, occurs via T cell signaling which is in turn caused by TR/pMHC binding or TR recognition of pMHC complexes (Alberts et al. 2002; Rudolph et al. 2006).

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: Molecular biology of the cell, 4th edn. Garland Science, New York, pp 1363-1421.

Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 24: 419-466.

Structural Immunoinformatics

Synonyms

Immunoinformatics, Structure-based Immunoinformatics.

Definition

Structural Immunoinformatics is the study of Immune system using computer-aided biotechnological (bioinformatics) tools and X-ray crystal structures of immune system components (Khan et al. 2010).

Reference

Khan JM, Tong JC, Ranganathan S (2010) Structural Immunoinformatics: Understanding MHC-peptide-TR binding. In: Davies MN, Ranganathan S, Flower DR (eds) Bioinformatics for Immunomics, vol 3 (Immunomics Reviews Series). Springer, New York, pp 77-94.

Reverse Vaccinology

Synonyms

Computer-based vaccine development, Computer-aided Vaccinology.

Definition

Reverse Vaccinology is a quick and efficient method of determining potential vaccine targets by screening entire pathogenic genomes using bioinformatics approaches, which later undergo normal wet-lab testing for immunological responses.

T cell epitopes

Synonyms

Peptide epitopes, Immunogenic peptides, Peptide antigens.

Definition

T cell epitopes are endogenous or exogenous immunogenic peptide antigens that are bound to and presented by the MHC proteins for recognition by the TR at the APC surface leading to T cell signaling and activation (Tong et al. 2004; Khan et al. 2010).

References

Tong JC, Tan TW, Ranganathan S (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci 13 (9): 2523-2532.

Khan JM, Tong JC, Ranganathan S (2010) Structural Immunoinformatics: Understanding MHC-peptide-TR binding. In: Davies MN, Ranganathan S, Flower DR (eds) Bioinformatics for Immunomics, vol 3 (Immunomics Reviews Series). Springer, New York, pp 77-94.

TR germline bias

Synonyms

MHC bias, TCR germline bias.

Definition

TR germline bias is the name given to a theory which suggests that the basis of MHC restriction or TR specificity are certain specific conserved constellations of contacts between TR V gene (variable gene) products and MHC gene products that co-evolve (Jerne 2004).

Reference

Jerne NK (2004) The somatic generation of immune recognition. 1971. Eur J Immunol 34: 1234-1242.

pMHC epitope

Synonyms

TCR footprint, TCR footprint on the pMHC.

Definition

The residues on the pMHC binding interface that contact and/or bind to corresponding residues on the TR interface are collectively called as the pMHC epitope on the pMHC surface (Kaas and Lefranc, 2005).

Reference

Kaas Q, Lefranc MP (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. In Silico Biol 5: 505-528.

TR paratope

Synonyms

MHC imprint, pMHC imprint on the TR.

Definition

The residues on the TR binding interface that contact and/or bind to corresponding residues on the pMHC interface are collectively called as the TR paratope on the TR surface (Kaas and Lefranc, 2005).

Reference

Kaas Q, Lefranc MP (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. In Silico Biol 5: 505-528.

1.12 Objectives

Currently, there is an immense need to develop a means by which characterization and identification of disease-implicated immunogenic antigens or T cell epitopes can be performed quickly and efficiently so as to cut down the lead time involved with traditional (experimental) vaccine design protocols and to gain and share knowledge and information for a global perspective in peptide-based vaccine development. Since experimental determination of T cell epitopes for every single disease-implicated MHC allele is prohibitively expensive, recently developed computer-aided techniques, especially structure-based techniques such as docking [10, 11], have contributed immensely to the problem of efficient T cell epitope identification, thereby, assisting in the planning of critical experiments leading to peptide vaccine design. This is particularly true for alleles with insufficient biochemical peptide binding data [11-14], a case where most commonly used sequence-based predictive techniques underperform.

Although accurate, the available docking protocol had multiple steps and hence was relatively slow compared to the sequence-based methods, posing a limitation for high-throughput structure-based T cell epitope identification. Until now most prediction models including docking-based predictive approaches [11-14] have neglected the influence and significance of TR/pMHC interactions on T cell activation. This was primarily due to the relatively small number of crystal structures available for TR/pMHC complexes. Nevertheless, if not large, a substantial number of TR/pMHC complexes are now available for analysis to characterize TR/pMHC interactions and identify the TR/pMHC interaction parameters. Also, no work has been done on predicting how well a pMHC complex can bind to TR. Therefore, three overall aims for this thesis are described below and addressed in detail in the five specific aims presented thereafter:

- I. To develop a new fast, efficient and robust protocol for docking of peptides to MHC proteins to improve the speed and efficiency of pMHC docking and make structure-based methods comparable with sequence-based methods for high-throughput screening of peptide epitopes.
- II. To analyze all available TR/pMHC crystal structures to characterize the protein-protein interactions in TR/pMHC complexes and extrapolate the TR/pMHC interaction parameters.

III. To apply the new docking protocol and the derived TR/pMHC interaction parameters to predict immunogenic peptides with high TR avidity for an example MHC allele.

Consequently, specific project goals were developed to address the main aims set out above. These are

1. To optimize the new pMHC docking technique, benchmark it with the previous method and validate it against previously published studies (publication 3).
2. To develop a new database for sequence-structure-function information of pMHC and TR/pMHC complexes with crystal structures, augmenting it with advanced features and new parameters for analysis of pMHC and TR/pMHC structures (publication 4).
3. To identify common structural characteristics of TR/pMHC complexes using existing crystallographic data (publication 5) and use these to develop methods for accurate prediction of T cell epitopes (publication 6).
4. To enhance the strategies for effective discrimination of MHC-binding peptides from the background (publication 6).
5. To use the combined predictive technique to develop a prediction model for identifying peptides that can lead to pMHC complexes with improved TR recognition and thus understand which peptides are actually responsible for T cell activation in HLA-DQ8 associated diseases (publication 6).

Chapter 2: Methods and Applications

Methods and applications that were developed and used in this study are summarised in Table 2.1. The ensuing publications have also been listed and included in the relevant chapter.

Table 2.1: Methods, applications and publications

Methods/Applications	Chapter	Refer to Publication
Structural Immunoinformatics: Understanding MHC-peptide-TR binding.	1	1
TR recognition of MHC-peptide complexes.	1	2
pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes.	3	3
MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures.	4	4
Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one.	5	5
<i>In silico</i> prediction of immunogenic T cell epitopes for HLA-DQ8.	6	6

Chapter 3: pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes

3.1 Summary

Immunogenic peptides or T cell epitopes are an integral part of the vital immunological synapse between the pMHC complexes and the TR proteins resulting in TR/pMHC complex formation which activate the T cells leading to the initiation of the adaptive immune response cascade [1-4, 9]. Identification of these antigenic peptide epitopes is an essential prerequisite in T cell-based molecular vaccine design. Experimental identification of T cell epitopes is a tedious, time consuming and expensive process. Recently developed computational methods, especially structure-based protocols such as docking that are even suited to alleles with limited epitope data [11-14], have proven to be vastly inexpensive and efficient compared to experimental approaches in screening numerous peptides against their cognate MHC alleles [10, 11, 17, 18]. The first step in these structure-based docking techniques is to identify strong MHC-binding peptides. These docking techniques need improvement in speed and efficiency to be useful in large-scale screening studies.

Therefore, this publication 3 discusses “pDOCK” which is a new computational technique for rapid and accurate fully flexible docking of peptides to MHC proteins which has been primarily applied on a non-redundant dataset of 186 pMHC (149 pMHC-I and 37 pMHC-II) complexes with X-ray crystal structures. 159 out of 186 peptides had a C α RMSD of less than 1.00 Å with a mean of 0.56 Å from initial testing of pDOCK for re-docking of peptides into their respective MHC grooves. 23 out of 25 peptides used for single and variant template docking had their C α RMSD values below 1.00 Å. pDOCK shows upto 2.5 fold improvement in the accuracy and is ~60% faster compared to our earlier docking methodology [10, 11]. A seven-fold increase in pDOCK accuracy has been recorded by validation against previously published studies [419, 424, 433, 435-437, 442, 443].



PROCEEDINGS

Open Access

pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes

Javed Mohammed Khan¹, Shoba Ranganathan^{1,2*}

From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics (InCoB2010)

Tokyo, Japan. 26-28 September 2010

Abstract

Background: Identification of antigenic peptide epitopes is an essential prerequisite in T cell-based molecular vaccine design. Computational (sequence-based and structure-based) methods are inexpensive and efficient compared to experimental approaches in screening numerous peptides against their cognate MHC alleles. In structure-based protocols, suited to alleles with limited epitope data, the first step is to identify high-binding peptides using docking techniques, which need improvement in speed and efficiency to be useful in large-scale screening studies. We present pDOCK: a new computational technique for rapid and accurate docking of flexible peptides to MHC receptors and primarily apply it on a non-redundant dataset of 186 pMHC (MHC-I and MHC-II) complexes with X-ray crystal structures.

Results: We have compared our docked structures with experimental crystallographic structures for the immunologically relevant nonameric core of the bound peptide for MHC-I and MHC-II complexes. Primary testing for re-docking of peptides into their respective MHC grooves generated 159 out of 186 peptides with $\text{C}\alpha$ RMSD of less than 1.00 Å, with a mean of 0.56 Å. Amongst the 25 peptides used for single and variant template docking, the $\text{C}\alpha$ RMSD values were below 1.00 Å for 23 peptides. Compared to our earlier docking methodology, pDOCK shows upto 2.5 fold improvement in the accuracy and is ~60% faster. Results of validation against previously published studies represent a seven-fold increase in pDOCK accuracy.

Conclusions: The limitations of our previous methodology have been addressed in the new docking protocol making it a rapid and accurate method to evaluate pMHC binding. pDOCK is a generic method and although benchmarks against experimental structures, it can be applied to alleles with no structural data using sequence information. Our outcomes establish the efficacy of our procedure to predict highly accurate peptide structures permitting conformational sampling of the peptide in MHC binding groove. Our results also support the applicability of pDOCK for *in silico* identification of promiscuous peptide epitopes that are relevant to higher proportions of human population with greater propensity to activate T cells making them key targets for the design of vaccines and immunotherapies.

* Correspondence: shoba.ranganathan@mq.edu.au

¹Department of Chemistry and Biomolecular Sciences and ARC Center of Excellence in Bioinformatics, Macquarie University, NSW 2109, Australia
Full list of author information is available at the end of the article

Background

The molecular machinery by which an antigen presenting cell (APC) presents T cell epitopes for recognition by T cell receptors (TR) and subsequent activation of T cells followed by the immune response cascade is fascinating. T cell epitopes are short antigenic peptide sequences (p) that are bound to and presented by the major histocompatibility complexes (MHC) for recognition by the TR [1]. These epitopes are essential subunit peptides that are required in order to stimulate cellular immune responses, especially the adaptive immune responses. Peptide epitopes can be of endogenous (processed within the cell) or exogenous (processed outside the cell) origins, which are presented for surveillance and recognition by the TR in an MHC allele and supertype dependant manner. Broadly classified into two types, MHC class I (MHC-I) complexes bind and present endogenous peptides whereas MHC class II (MHC-II) complexes prefer exogenous peptides. Typically, MHC-I proteins are heterodimers, consisting of a heavy α chain (I-ALPHA) of about 45 kDa, and a light chain, β 2-microglobulin (B2M) of about 12 kDa [2,3]. The α chain consists of α 1 (G-ALPHA1), α 2 (G-ALPHA2) and α 3 (C-LIKE) domains where G-ALPHA1 and G-ALPHA2 domains form the peptide binding groove or 'cleft' [4]. MHC-II proteins are also heterodimeric proteins consisting of an α chain (II-APLHA; 34 kDa) and a β chain (II-BETA; 29 kDa) with very similar overall quaternary structure to that of MHC-I proteins [5-10]. However, their peptide binding groove is formed by the α 1 and β 1 domains of the two chains.

Peptides presented by MHC-I are generally between 8-11 amino acids in length. These peptides are 'chopped' within the cytosol of the cell by cytosolic proteases and are transported to the MHC binding groove within the endoplasmic reticulum by the transporters associated with antigen processing (TAP) proteins in an ATP dependant manner. Following which, the peptides bind to the MHC to form the peptide-MHC (pMHC) complex which is then transported to the APC cell surface and presented for recognition by the TR of CD8⁺ cytotoxic T cells (CTLs). Similarly, the peptides presented by MHC-II are usually 12-25 amino acids in length and are endocytosed into the cell by the lysosomes where they bind the MHC-II proteins by displacing the original MHC-II ligand known as the 'CLIP' peptide to form the pMHC complex. And again, they are transported to the APC cell surface for recognition by the TR of the CD4⁺ T helper cells. Identification of true T cell epitopes from the repertoires of immunologically significant antigenic peptide sequences is a vital prerequisite in the process of conventional molecular vaccine design for prevention and treatment of infectious, autoimmune, allergic and graft vs. host diseases. The key step in TR-mediated immune response is thus the binding and presentation

of antigenic endogenous or exogenous peptide epitopes, which can be reasonably well predicted using sequence-based methods for alleles with large datasets of known binding peptides, as reviewed earlier [11,12].

Experimental identification of T cell epitopes is a tedious, time consuming and expensive process owing to the large number and diversity of both MHC alleles and the antigenic peptides. Not to mention, is the extremely low chance of immunogenicity (1 in 2000 peptides) even amongst the peptides that bind strongly to the MHC (50%) [13]. Recently developed computational methods have proven to be vastly time and cost efficient in screening the vast oceans of peptides and MHC repertoires [14]. Current computational methods can be broadly classified into: 1. Sequence-based approaches which use sequence motifs [15], matrix models [16,17], Artificial Neural Network [18-20], Hidden Markov Model [21] and Support Vector Machine [22-24] for large-scale screening of potential T cell epitopes from protein sequence databanks and 2. Structure-based approaches such as protein threading [25,26], homology modeling [27,28], rigid docking [29] and flexible docking [2,3] which utilize three-dimensional data for detailed structural analysis of interactions between the MHC and bound segmental antigenic peptides. The former are more suitable for large-scale screening of potential T cell epitopes, while the latter work better for detailed analysis of short immunogenic regions of antigens [2]. Although sequence-based methods are well established, a major limitation of such techniques is the heavy reliance on the availability of large comprehensive training sets of peptides. Thus, these approaches are not appropriate for accurate prediction of peptides in circumstances where the data available is insufficient. Therefore, the coverage of sequence-based techniques is limited to subsets of binding peptides that belong to the most numerous groups and cannot generate reliable data for peptides that are least represented in the dataset [2], leaving structural immunoinformatics as the only option for such peptides [3,5-7].

Antigenic peptides that bind strongly to MHC alleles are known to elicit T cell responses [1-3,5-7,11]. Hence, their identification is a vital first step in the process of structure-based immune epitope prediction. The usual approach adopted to address this important issue is to utilize a powerful concept, based on the principle of structure-based drug design called "docking", where peptides are computationally placed in MHC grooves in the best orientation, reflecting steric and electrostatic complementarity, using structure-based docking techniques. The accuracy with which the peptides are docked is measured in terms of Root Mean Square Deviation (RMSD) values obtained by comparing the docked

conformations of the peptides to their original bound conformations in the respective X-ray crystal structures. With the development of new structural modeling and docking techniques and an increase in the number of protein structures deposited in the Protein Data Bank (PDB) [30] and the IMGT/3Dstructure-DB [31,32], structure-based approaches are being more commonly used to predict potential T cell epitopes [33], often producing modeled structures accurate to within 2.00 Å RMSD from the experimental crystal structure, providing a wealth of information for structural analysis and the development of prediction methods.

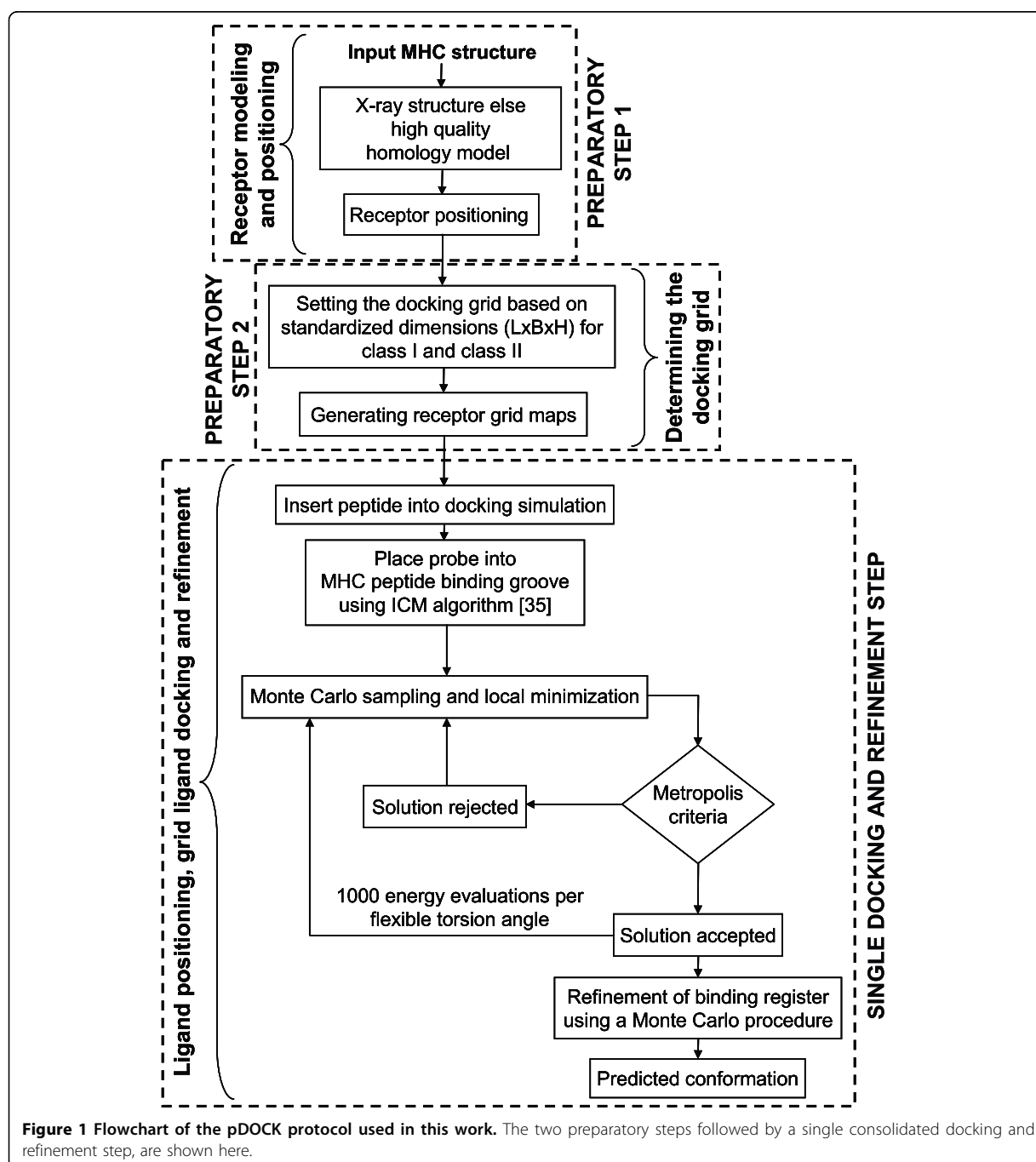
The development of an accurate protocol for flexible docking has helped us to successfully carry out quantitative predictions for both MHC-I and MHC-II alleles even with limited binding peptide data [3,5-7]. Our earlier docking protocol consisted of three steps (extended to four for pMHC-II complexes for incorporating the flanking residues on either side of the nonameric core, which is the 9-mer anchored to the MHC molecule): (1) rigid docking of the peptide nonamer termini into the MHC binding groove; (2) loop closure of central residues by satisfaction of spatial constraints; (3) followed by iterative *ab initio* refinements of ligand backbone and; (4) extension of flanking peptide residues by satisfaction of spatial constraints [2,3] (only for MHC-II related peptides). While accurate, this approach has multiple steps, resulting in suboptimal computational speeds. Therefore, the efficiency of this protocol for peptide docking to MHC needs to be improved for large-scale screening of T cell epitopes. A grid-based docking methodology has earlier been reported [34] to be highly accurate in pMHC docking over a limited MHC-I data. Hence, we have developed a grid-based peptide docking method (pDOCK) and have extensively tested it on both MHC-I and MHC-II peptides. The motivation behind the development of a faster and more accurate peptide docking methodology was to eventually improve the qualitative and quantitative efficacy of structure-based T cell epitope prediction.

In this study, we present pDOCK: a new computational technique for rapid and accurate docking of flexible peptides to the MHC receptors and primarily apply it to re-dock a non-redundant dataset of 186 (149 MHC-I and 37 MHC-II related) peptides, from MPID-T2 (<http://biolinfo.org/mpid-t2>) database for which X-ray crystal structures are available in the PDB and the IMGT/3Dstructure-DB, back into their respective MHC grooves. pDOCK comprises of two input preparatory steps followed by a single consolidated docking and refinement step as depicted in Figure 1. The pDOCK protocol involves: Preparatory step 1: receptor modeling and positioning; Preparatory step 2: determining the docking grid by defining the grid dimensions (length x

breadth x height) for ligand placement and grid map generation within the vicinity of the receptor's binding site and; Final docking and refinement step: ligand positioning within the grid, flexible docking of the peptide into the peptide binding groove and refinement of all ligand and binding site residues using the Internal Coordinate Mechanics (ICM) global optimization, docking algorithm [35] and a biased Monte Carlo procedure (see Methods section for more details). Our preliminary analysis of all pMHC complexes from the MPID-T2 database has provided us with standardized dimensions for the 3-D docking grids for both class I and class II pMHC structures. These standardized values were used to set the dimensions of the docking grids in all our experiments. Unlike the previously reported grid-based docking method [34], homology model building for MHC receptors has not been used in the development of pDOCK, instead using only experimentally determined X-ray crystallographic structures. The pDOCK method, however, is generic and is applicable to high quality homology models of alleles when experimental structures are not available. Here, the receptor modeling sub-step mentioned in the preparatory step 1 (Figure 1) can be used in the absence of structural data for the MHC proteins. Thus, the direct use of X-ray crystal structures in our docking simulations ensures accurate results.

The first experiment that we conducted was to ensure that an extended peptide bound to its cognate MHC receptor preferentially selecting the same nonameric core peptides as in the crystal structure and then to evaluate the accuracy of the docked peptide. Hence, we performed re-docking of 186 peptides back to their cognate MHC receptors to check for conformational accuracy of the predicted binding registers and their Cα RMSD against their respective crystal structures. We have then benchmarked pDOCK with our earlier docking protocol [2,3] for a dataset of 50 selected (35 MHC-I and 15 MHC-II) pMHC complexes to verify the speed and accuracy of pDOCK against our earlier method. This was followed by validation and accuracy checks for pDOCK against available flexible peptide docking results obtained from the literature for a dataset of 15 peptides.

In the process of selecting immunogenic peptides for vaccine design, the two main aspects are to determine: (1) multiple peptides that bind to the same allele or MHC molecule and; (2) promiscuous or same peptides that bind multiple alleles. Therefore, as a secondary experiment, we have pursued to test the efficacy and robustness of our docking protocol in modeling the bound conformations of novel peptides to specific MHC alleles by carrying out docking of multiple peptides to a single MHC template structure (same MHC allele), suitable for immune epitope prediction from an antigenic



protein, using a moving window of 9-mers along the entire sequence [3,5-7]. Our third experiment was to dock a single peptide from particular PDB structures onto multiple MHC templates (multiple alleles) from other crystal structures, suitable for determining promiscuous peptides capable of binding to a set of related alleles and therefore, important for vaccine design.

The $C\alpha$ RMSD values have been calculated only for the nonameric core of the peptide (for both MHC-I and MHC-II related peptides) which is a contiguous immunogenic segment that forms the “binding register” within the MHC peptide binding cleft, as reported earlier by our group [3]. For the peptides with nine and less number of amino acid residues the entire peptide

was used for $\text{C}\alpha$ RMSD calculation. pDOCK accurately detected all 186 binding registers, i.e., the nonameric cores of the peptides are identical to their respective crystal structures. pDOCK generated 85.5% of all the peptides with $\text{C}\alpha$ RMSD of less than 1.00 Å compared to their respective X-ray crystal structures. Our benchmarking results imply up to 2.5 fold improvement in the accuracy of the new peptide docking methodology. The validation results represent a sevenfold improvement in the accuracy of our technique compared to that of the existing methodologies in flexible docking and modeling of peptides into MHC grooves. Amongst the 21 peptides docked in the second experiment, the $\text{C}\alpha$ RMSD values for docked peptides compared to their respective crystal structures were below 1.00 Å for 20 peptides (details in Results and discussion section). The third experiment accounted for all 4 peptides docked with less than 1.00 Å $\text{C}\alpha$ RMSD compared to the same peptides from the corresponding template crystal structures (details in Results and discussion section). Overall, pDOCK is up to 60% faster than our earlier protocol and hence provides a rapid and accurate docking method to evaluate pMHC binding for large scale immune-epitope prediction.

Results and discussion

The fact that our earlier method was comparatively slower and that it involved rigid-docking of the peptide termini, acted as the platform for us to 'revisit' our pMHC docking methodology. Based on these requirements, we have developed a single step pMHC docking protocol (details in Methods section) as shown in Figure 1, which allows flexibility over the entire length of the peptide antigen and can be used as a generic method to obtain the conformations of bound peptide ligands to MHC binding grooves of both class I and class II MHC proteins. A systematic evaluation of pDOCK is performed as three separate tests: (1) exhaustive re-docking of all non-redundant peptides to their respective MHC grooves as a test case, benchmarking and validation; we then address two very significant practical problems faced by immunologists during the process of allele-specific peptide vaccine design: (2) the docking of multiple peptides that bind to same MHC allele, for immunogenic epitope scanning of antigenic sequences and; (3) docking of promiscuous peptides or same peptides binding to multiple MHC alleles for vaccine design, based on groups of disease-implicated alleles. A correctly docked structure is defined as the peptide with at most 2.50 Å $\text{C}\alpha$ RMSD from the respective experimental X-ray crystal structure [2]. pDOCK has also been benchmarked against our previous docking protocol and validated on published peptide modeling and docking results from the literature. Bordner and Abagyan [34] suggested that

while grid-based docking could be applied for pMHC-II, it was a more difficult problem. pDOCK has been successfully applied for MHC-II peptide docking as well with excellent results.

Experiment 1

Re-docking bound peptides to their cognate MHC grooves

pDOCK has been applied on a non-redundant dataset of 186 (149 MHC-I and 37 MHC-II) pMHC complexes from the MPID-T2 database (details in Methods section, data and docking results in Additional File 1 – Table S1). Initially, the peptides were extracted from the experimental pMHC complexes, randomized and set to extended conformations. This was followed by optimization of the peptide ligands and re-docking of the separated peptides back to their respective MHC grooves. As depicted in Figure 2, our technique generated 159 out of 186 peptides with $\text{C}\alpha$ RMSD values less than 1.00 Å: 124 out of 149 peptides (83%) and 35 out of 37 peptides (~95%) for class I and class II MHC proteins,

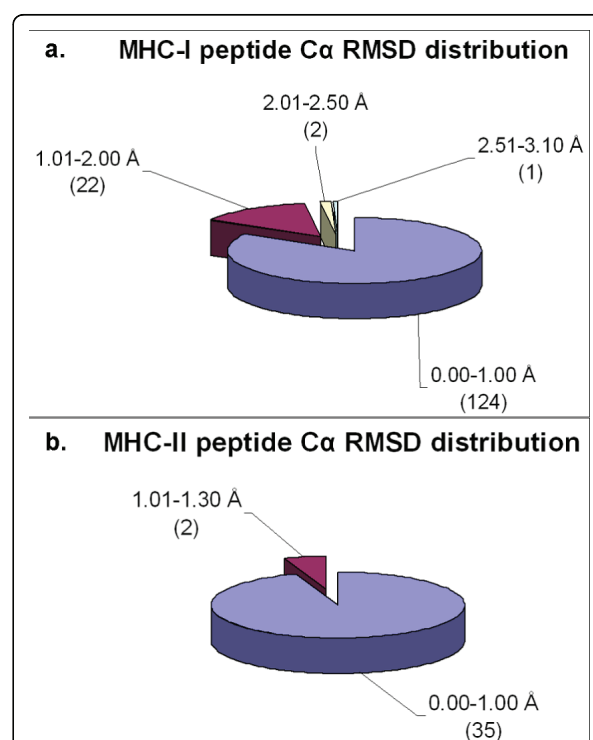


Figure 2 Distribution of $\text{C}\alpha$ RMSD of the docked peptides and their respective crystal structures across the non-redundant MPID-T2 dataset for peptides for a. MHC-I complexes and b. MHC-II complexes. Most of the peptides from both MHC-I (124/149; 83%) and MHC-II (35/37; ~95%) datasets have their $\text{C}\alpha$ RMSD values below 1.00 Å, highlighting the accuracy of our docking protocol. The number of peptides in each $\text{C}\alpha$ RMSD range is given in parentheses.

respectively. ~15% (22/149) and ~1% (2/149) of the peptides have their C α RMSD values within the ranges 1.01-2.00 Å and 2.01-2.50 Å, respectively amongst the MHC-I peptides docked (Figure 2a). Similarly, ~5% of the peptides have their C α RMSD values within a range of 1.01-1.30 Å amongst the MHC-II related peptides that were docked using pDOCK (Figure 2b). On an average, pDOCK resulted in a C α RMSD value of about 0.56 Å for re-docking of peptides into their respective MHC grooves over the entire dataset of 186 pMHC complexes.

Our best results are shown in Figure 3, with structural comparison between the lowest energy docked conformation and the native conformation of the bound peptides for MHC-I (PDB code 1s7q) and MHC-II (PDB code 1d5x) structures. These docked conformations of peptide structures have the best C α RMSD values of 0.09 Å and 0.11 Å respectively, obtained over the entire dataset. The MHC-II peptide in Figure 3b has 5 out of 6 amino acid residues replaced by amino acid analogues (chemical mimics) in the crystal structure. Nonetheless, it has the best C α RMSD value among all the MHC-II related peptides used in this study, supporting pDOCK's applicability to peptide or peptide analogues (containing amino acid mimics in structure-based drug design). pDOCK also generated the least energy docked orientations for all the peptides with accurate determination of their respective binding registers, i.e. having the exact nonameric core in the binding grooves, with respect to their native bound conformations in the X-ray crystal structures. All peptides except one from the class I pMHC crystal structure (PDB code 2gtw; C α RMSD of 3.08 Å) were within the acceptable 2.50 Å C α RMSD from their respective native conformations (Figure 2a). Also, none of the MHC-II related peptides showed any deviation from the acceptable 2.50 Å C α RMSD threshold (Figure 2b).

We carefully examined the re-docked conformation of the peptide LAGIGILTV in the MHC groove of the complex 2gtw, with the X-ray structure. In 2gtw, peptide residues 1 to 5 interact with a formic acid molecule, which was not explicitly introduced into the docking simulation. When the formic acid molecule was included in the docking simulation, the predicted orientation of the peptide using pDOCK is energetically more favourable for pMHC complex formation than the predicted conformation when the formic acid molecule is omitted. The improvement in accuracy by the inclusion of the formic acid molecule is ~13 folds. This is portrayed in Figure 4 which clearly indicates that the peptide residues Leu 1, Ala 2, Gly 3, Ile 4 and Gly 5 that are not correctly predicted in the absence of the formic acid molecule (Figure 4a), are accurately docked when the formic acid molecule is introduced into the docking simulation (Figure 4b), resulting in an improvement in the C α RMSD value from 3.08 Å to 0.24 Å.

Although water molecules and other common biological ions such as phosphate and chloride may mediate pMHC interactions in some cases, they were omitted from our experiments because the significance and contributions of these molecules towards pMHC binding vary immensely between different peptides and specific alleles over a large dataset like the one used in this study (186 complexes). Our previous protocol achieves a C α RMSD of 1.53 Å for the bound structure of the peptide from pMHC complex 1jf1, due to the presence of a water molecule positioned around the peptide residues 5 to 7 in the crystal structure leading to erroneous prediction of the loop formed, which resulted in incorrect positioning of interacting residues [2]. However, pDOCK successfully overcomes this restriction to accurately predict the least energy bound conformation of this peptide with a C α RMSD value of 0.30 Å. The enhancement in accuracy of docking is a direct

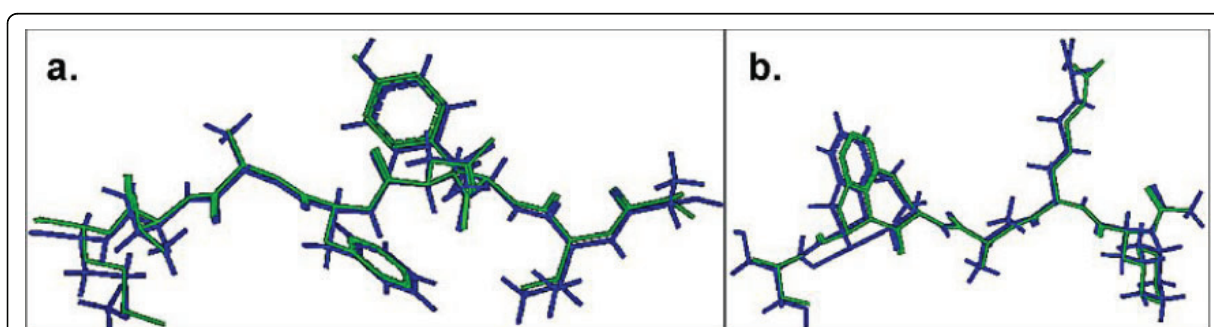
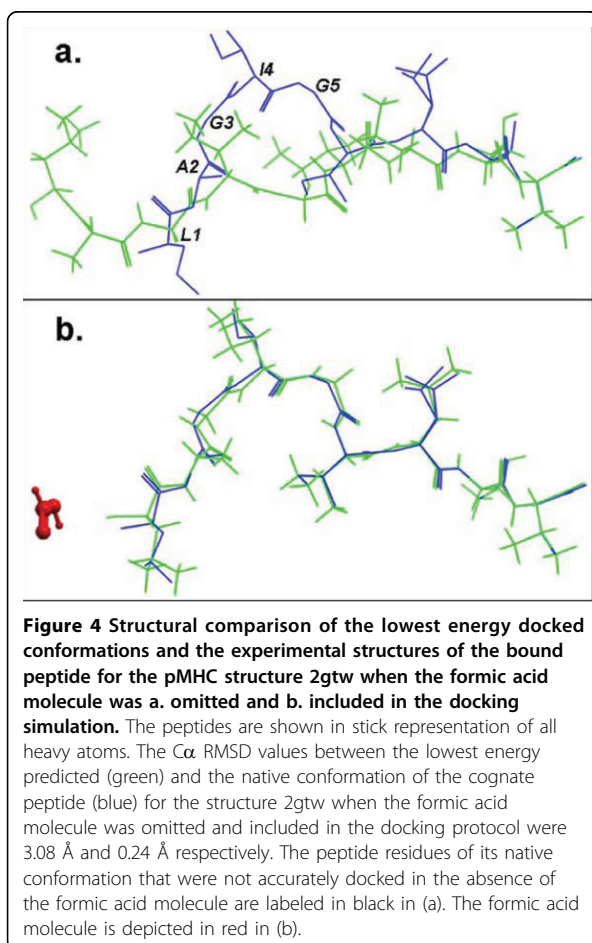


Figure 3 Comparison of the lowest energy predicted and the experimental structures of the cognate peptides with the least RMSD values across the pDOCK test set. **a.** KAVYNFATM peptide in the MHC-I complex 1s7q (PDB code). **b.** XXRXXX peptide in the MHC-II complex 1d5x (PDB code). The peptides are shown in stick representation of all heavy atoms. The C α RMSD values between the lowest energy docked conformation (green) and the native conformation of the bound peptides (blue) for the MHC-I structure 1s7q (PDB code) and the MHC-II structure 1d5x (PDB code) are 0.09 Å and 0.11 Å, respectively. X: Amino acid analogues (chemical mimics).



consequence of the improved sampling of available conformational space in pDOCK. This preliminary experiment is a critical first step as it establishes the validity of our approach and helps us test the ability of our technique to accurately dock cognate peptides into their respective MHC receptors, using the proposed single-step docking procedure.

Benchmarking with our previous methodology

In order to ascertain the improvement in speed and accuracy of pDOCK compared to the old technique, we have benchmarked our peptide docking methodology with our earlier pMHC docking protocol [2,3] over a subset of 50 pMHC complexes (35 MHC-I and 15 MHC-II) from the complete non-redundant dataset (listed in Additional File 1 – Table S1) and the results are presented in Table 1. pDOCK results are consistently better than our earlier docking methodology in terms of accuracy (α RMSD) of the modelled or docked peptide compared to their X-ray crystal structures after docking back into their respective MHC grooves. The new protocol also generates the least energy docked conformations for all 50

peptides with α RMSD values less than 1.00 Å, compared to eight peptides, docked using the earlier method, having α RMSD values above 1.00 Å (graphically shown in Additional File 2 – Figure S1). The new procedure outmatches the old protocol particularly well for complexes 1s9y, 1hhh, 1jfl, 1e27, 1jpf, 1qo3, 1wbz and 1g7p (Table 1) amongst the MHC-I structures and for structures 1uvq and 1aqd (Table 1) amongst the MHC-II structures (highlighted in yellow in Additional File 2 – Figure S1).

These results suggest that some of the conformational limitations of our previous methodology, such as the presence of water molecules in and around the peptide and within the peptide binding groove in the original PDB structure, have been addressed in our new docking protocol making it highly accurate. Besides an improvement in the accuracy, pDOCK is also able to accurately model docked conformations for some peptides especially for MHC-II related peptides with more than 9 amino acid residues, thereby improving the coverage over the entire length of the peptides. Peptides from the pMHC complexes 1uvq and 2iam were among the highest coverage (20 and 15 residues respectively) obtained in this experiment with α RMSD values 0.42 Å and 0.46 Å respectively over the length of the entire peptide (results not shown). The reliability for the accurate prediction of flanking residues (especially for MHC-II peptides) depends upon their interactions with the MHC residues outside the peptide binding groove and therefore, have not been included in the calculation of α RMSD values reported.

In terms of the computational time to complete a single docking experiment, pDOCK is up to 60% faster (on an average) than the earlier method as summarized in Table 2. The average time taken by pDOCK is approximately 10 min. (the preparatory receptor positioning step of ~3 sec. {0.50%}, determining the docking grid taking ~42.6 sec. {7.10%} and the single docking and refinement step of ~9.24 min. {92.4%}), compared to 23.50 to 24.50 min (Step 1 taking ~5 min., Step 2 of ~30 sec., Step 3 taking ~18 min. and Step 4, which was only applicable to MHC-II related peptides, of ~1 min.) using the old protocol on a 2 CPU 3.20 GHz 3 GB RAM workstation. The average time taken for each of the steps using either of the methodologies is calculated over the entire non-redundant dataset of 186 pMHC complexes catalogued in additional file 1 – table S1. The mean α RMSD value for the least energy docked conformations of peptides, from the dataset of 50 peptides used for benchmarking, was 0.27 Å using pDOCK compared to 0.65 Å for the old procedure. This denotes almost two and a half fold improvement in the accuracy of our novel docking strategy over a larger dataset (50 peptides) than that used previously (40 peptides) [2].

Table 1 Benchmarking pDOCK with our earlier methodology

S. No.	Allele	PDB	Peptide Length	Peptide Sequence	Cα RMSD (Å)	
					Previous method	pDOCK
MHC-I						
1	HLA-A*1101	1qvo	10	QVPLRPMTYK	0.53	0.24
2	HLA-A*0201	1qr1	9	IISAVVGIL	0.46	0.29
3	HLA-A*0201	1akj	9	ILKEPVHGV	0.87	0.39
4	HLA-A*0201	1i1y	9	YLKEPVHGV	0.70	0.66
5	HLA-A*0201	1i7r	9	FAPGFFPYL	0.59	0.47
6	HLA-A*0201	1i7u	9	ALWGFVPVL	0.32	0.29
7	HLA-A*0201	1oga	9	GILGFVFTL	0.32	0.16
8	HLA-A*0201	1qsf	9	LLFGYPVAV	0.54	0.34
9	HLA-A*0201	1lp9	9	ALWGFPPVL	0.58	0.26
10	HLA-A*0201	1s9y	9	SLLMWITQS	1.09	0.39
11	HLA-A*0201	1hhh	10	FLPSDFFPSV	1.10	0.49
12	HLA-A*0201	1jf1	10	ELAGIGILTV	1.53	0.30
13	HLA-B*0801	1agc	8	GGKKKYQL	0.28	0.23
14	HLA-B*0801	1mi5	9	FLRGRAYGL	0.42	0.37
15	HLA-B*2705	1ogt	9	RRKWRRWHL	0.51	0.18
16	HLA-B*2705	2a83	9	RRRWHRWRL	0.55	0.18
17	HLA-B*3501	2cik	9	KPIVLHGY	0.74	0.26
18	HLA-B*3508	3bwa	8	FPTKDVAL	0.56	0.26
19	HLA-B*5101	1e27	9	LPPWAKEI	1.27	0.18
20	HLA-B*5301	1a1m	9	TPYDINQML	0.59	0.28
21	HLA-Cw*0401	1im9	9	QYDDAVYKL	0.49	0.34
22	HLA-G*0101	2dyp	9	RIIPRHLQL	0.43	0.16
23	H2-Db	1fg2	9	KAVYNFATC	0.25	0.19
24	H2-Db	3buy	9	LSLRNPILV	0.63	0.23
25	H2-Db	1yn7	10	SLENFAAYV	0.62	0.14
26	H2-Db	1jpf	11	SGVENPGGYCL	1.14	0.36
27	H2-Dd	1qo3	10	RGPGRAFVTI	1.49	0.17
28	H2-Kb	1t0m	8	SSIEFARL	0.66	0.21
29	H2-Kb	1vac	8	SIINFEKL	0.32	0.22
30	H2-Kb	1wbz	9	SSYRRPVGI	0.89	0.19
31	H2-Kb	1s7q	9	KAVYNFATM	0.20	0.09
32	H2-Kb	1g7p	9	SRDHSRTPM	0.97	0.17
33	H2-Kd	1vgk	9	SYVNTNMGL	0.86	0.25
34	H2-Kk	1zt1	8	FEANGNLI	0.57	0.45
35	H2-Ld	2e7l	9	QLSPFPFDL	0.37	0.35
MHC-II						
36	HLA-DQB1*0602	1uvq	20	MNL PSTKVS WAAVGGGSLV	1.09	0.23
37	HLA-DRB1*0301	1a6a	15	PVSK MRMATPLLMQA	0.38	0.30
38	HLA-DRB1*0101	1aqd	14	GSD WRFLRGYHQA	1.08	0.28
39	HLA-DRB1*0101	1fyt	13	PK YVKQNTLKLAT	0.68	0.23
40	HLA-DRB1*0101	2iam	15	GEL IGILNAAKVPAD	0.56	0.24
41	HLA-DRB1*0401	1d5x	6	XXRXXX	0.23	0.11
42	HLA-DRB1*0401	1d5z	7	XXRAXSX	0.33	0.22
43	HLA-DRB1*0401	1d6e	8	XXRXMASX	0.32	0.14
44	HLA-DRB1*0401	1j8h	13	PK YVKQNTLKLAT	0.59	0.20
45	HLA-DRB3*0101	2q6w	11	AWRSEALPLG	0.54	0.30
46	HLA-DRB5*0101	1fv1	20	NPVVHFF KNIVTPRT PPPSQ	0.88	0.59
47	I-Ad	1iao	14	RG ISQAVHAAHAEI	0.81	0.27

Table 1 Benchmarking pDOCK with our earlier methodology (Continued)

48	I-Ak	1iak	13	STDYGILQINSRW	0.42	0.23
49	I-Au	2pxy	11	RGGASQYRPSQ	0.78	0.28
50	I-Ek	1r5v	13	ADLIAYPKAATKF	0.82	0.28

α RMSD values are calculated only for the nonamer binding cores (shown in bold) for peptides with more than 9 residues in the X-ray crystal structures. X: Amino acid analogues (chemical mimics). A graphical representation of the results is available in Additional File 2 – Figure S1.

Validation against previously published studies

Keeping in mind the essence of improving the accuracy and robustness of the proposed strategy, we have validated pDOCK with seven studies involving MHC-I peptide docking/modeling and one study involving MHC-II peptide docking, covering 15 pMHC structures and compared the results by re-running our earlier method. The results of our validation experiments are compiled into Table 3. Peptides 1, 2, 3, 4 and 15 (Table 3) are new in this study and are collated from recent publications [34,36,37], whereas the remaining 10 were from the validation studies reported for our earlier methodology [2]. To the best of our knowledge, these results represent a sevenfold increase in the accuracy of pDOCK compared to available flexible docking techniques in the remodeling of pMHC complexes. Interestingly, the validation criteria for almost all of the previously published studies [34,36-40] involved either docking or remodeling of peptides back into their original crystal structure. Although the α RMSD values (0.29 Å and 0.30 Å, respectively) for peptides 2 and 3 (Table 3) were slightly higher, they are still comparable with the α RMSD values reported earlier (0.23 Å and 0.22 Å, respectively) [34]. Peptide 1 (Table 3) however, was generated with a better α RMSD (0.31 Å) compared to the α RMSD (0.76 Å) reported in the same earlier grid-based docking study [34]. The enhancement in the accuracy for peptide 1 could be a direct implication of more conformational sampling space in a flexible environment resulting from a relatively larger docking grid (35.36 Å x 35.52 Å x 35.79 Å) for MHC-I peptides and a lower temperature (300 K) used in pDOCK compared to the grid dimensions (34 Å

x 34 Å x 25 Å) and temperature (700 K) used in the previous grid-based docking study [34]. Thus, pDOCK is not only comparable to but also surpasses the available techniques in flexible docking and remodeling of peptides with regards to the accuracy (α RMSD) with which it predicts the bound structure of a peptide to its respective MHC groove. By and large, our results illustrate the advantages of using grid-based flexible docking over conventional docking protocols.

Figure 5 provides a pictorial representation of an example of the above discussed accuracy. This structural comparison between the least energy docked conformation generated using pDOCK and that of the native conformation of the cognate peptide in the complex 1duz portrays not only the highly accurate predicted conformation of the peptide, α RMSD of 0.33 Å compared to that of 3.01 Å reported earlier [37], but also highlights the fact that the peptide's N-terminal residues (Leu 1, Leu 2 and Phe 3) were better modeled and structurally well aligned to that of its native conformation when compared to the lowest energy docked conformation reported earlier [37]. Notably, the least energy docked conformations generated for a common murine MHC (H2-Kb) related Sendai virus nucleocapsid peptide FAPGNYPAL and a very familiar human HLA (A*0201) related Influenza A virus matrix peptide GILGFVFTL have significantly lower α RMSD values of 0.25 Å and 0.16 Å respectively (Table 3) than those reported in earlier studies (2.70 Å and 0.46 Å, 1.60 Å, 1.40 Å respectively) [38,40-42] and those obtained using our previous protocol (0.40 Å and 0.32 Å). These observations establish the efficacy of pDOCK to dock highly accurate multi-species related peptide structures permitting conformational sampling of the peptide in the binding groove during flexible docking.

Table 2 Comparison of computational time of pDOCK with our earlier docking method

Previous method	pDOCK
Step 1: ~ 5 min	
Step 2: ~ 30 s	Preparatory Step 1: ~ 3*s
Step 3: ~ 18 min	Preparatory Step 2: ~ 42.6 s
Step 4 [#] : ~ 1 min	Single docking and refinement step: ~ 9.24 min
Total: ~ 23.50 – 24.50 min	Total: ~ 10 min

Both methodologies were applied using a 2 CPU 3.20 GHz 3 GB RAM workstation. *Only for X-ray crystal structures of MHC proteins. The time taken for this step would increase if homology modeling needs to be carried out.

[#]Applicable only to MHC-II related peptides.

Experiment 2

Docking of multiple peptides onto a single template

We applied pDOCK to a subset of 25 non-redundant pMHC complexes (obtained from the pDOCK test set of 186 pMHC complexes), with either a common allele or a common peptide core. The dataset of 18 MHC-I and seven MHC-II complexes comprises 21 (15 MHC-I and six MHC-II related) novel peptides which were known to bind to a single template (same allele) and four (three MHC-I and one MHC-II related) promiscuous peptides that were known to bind variant templates

Table 3 Comparison of pDOCK with published MHC-peptide modeling and flexible docking methods

S.No	Technique	Peptide Sequence	MHC class	PDB	RMSD (Å)		
					Published	Previousmethod	pDOCK
1	Grid-based Flexible docking [34]	RGVYQGL	I	1kpu [#]	0.76	0.59	0.31
2	Grid-based Flexible docking [34]	ALWGFVPVL	I	1i7u	0.23	0.32	0.29
3	Grid-based Flexible docking [34]	ELAGILTV	I	1jfi	0.22	1.53	0.30
4	Monte Carlo annealing [37]	LLFGYPVYV	I	1duz [#]	3.01	0.33	0.33
5	Simulated annealing [38]	FLPSDFFPSV	I	1hhh	1.59	1.10	0.48
6	Simulated annealing [38]	GILGFVFTL	I	1hhi [#]	0.46	0.32	0.16
7	Simulated annealing [38]	ILKEPVHGV	I	1hhj [#]	0.87	0.87	0.55
8	Simulated annealing [38]	LLFGYPVYV	I	1hkh [#]	0.78	0.33	0.33
9	Combinatorial buildup algorithm [39]	RGVYQGL	I	2vaa [#]	0.56	0.32	0.22
10	Combinatorial buildup algorithm [40]	LLFGYPVYV	I	1hkh [#]	1.40	0.33	0.33
11	Combinatorial buildup algorithm [40]	ILKEPVHGV	I	1hhj [#]	1.30	0.87	0.55
12	Combinatorial buildup algorithm [40]	GILGFVFTL	I	1hhi [#]	1.60	0.32	0.16
13	Multiple copy algorithm [41]	FAPGNYPAL	I	2vab [#]	2.70	0.40	0.25
14	Multiple copy algorithm [42]	GILGFVFTL	I	1hhi [#]	1.40	0.32	0.16
15	GOLD/GLIDE Flexible docking [36]	XXRXMASX	II	1d6e	1.24/3.06	0.32	0.14

X: Amino acid analogues (chemical mimics). [#]These structures are not listed in Additional File 1 - Table S1 due to redundancy in MPID-T2.

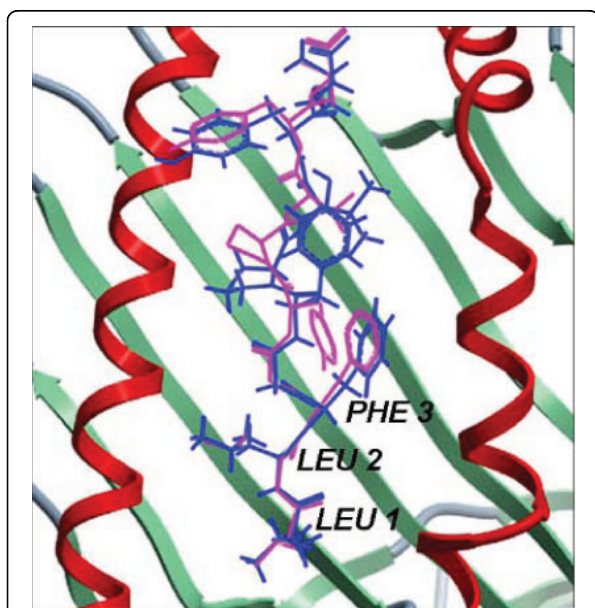


Figure 5 Structural comparison between the native conformation and the lowest energy docked conformation of the cognate peptide in MHC-I complex 1duz. The peptide is shown in stick representation wherein the native conformation is in pink and the docked conformation is in blue. The MHC peptide binding 'groove' is shown as ribbons. The C α RMSD between the native and the lowest energy docked conformation of the bound peptide from our work is 0.33 Å which is up to three and a half times better than an earlier reported C α RMSD of 3.01 Å [37]. The peptide residues of our lowest energy docked conformation that were better modeled and aligned to that of its native conformation when compared to the lowest energy docked conformation reported earlier [37] are labeled in black. This structure is not listed in Additional File 1 - Table S1 since it was a redundant structure in MPID-T2.

(multiple alleles). Due to lack of sufficient promiscuous peptides available in the PDB, only four peptides are currently tested. The results obtained from the docking of peptides onto single templates are tabulated in Table 4. 20 out of 21 peptides were docked onto a single template with C α RMSD values less than 1.00 Å compared to their respective experimental structures. Amongst the results from single template docking experiments, the most accurate docked conformation of the least energy peptide, with a C α RMSD of 0.06 Å compared to its relevant PDB peptide structure (Table 4), was achieved for the peptide from the structure 1kbg docked onto the MHC from the structure 1nam having the same murine MHC allele (H2-Kb) as the complex 1kbg.

Experiment 3

Docking of same peptides onto variant templates

Results from variant template docking experiments are listed in Table 5. It is worth noting that the C α RMSD values for the peptides docked onto variant templates were calculated in comparison to the same peptides present in the respective template structures. This was done due to the fact that although the peptides may be similar, the environments encountered by the same peptides are different in the binding grooves of different MHC alleles. All four promiscuous peptides were docked onto variant templates with C α RMSD values below 1.00 Å (Table 5). This observation suggests the robustness of pDOCK in docking promiscuous peptides onto multiple MHC alleles and its adaptability in ordering the binding registers or conformations of the peptides according to the changed environments, due to changes in the amino acid sequences, of the MHC

Table 4 Docking novel peptides onto a single template: pDOCK compared to our previous method

MHC class	Peptide PDB Allele	Peptide PDB	MHC Template Structure	Template Allele	Peptide Length	Peptide Sequence	C α RMSD (Å)	
							Previous method	pDOCK
I	HLA-A*0201	2v2w	1qrm	HLA-A*0201	9	SLYNTVATL	0.63	0.38
I	HLA-A*0201	1hhh	1qrm	HLA-A*0201	10	FLPSDFFPSV	0.58	0.25
I	HLA-A*0201	1qse	1qrm	HLA-A*0201	9	LLFGYPRYV	0.62	0.30
I	HLA-A*0201	2bnq	1qrm	HLA-A*0201	9	SLLMWITQV	0.97	0.77
I	HLA-A*0201	2gj6	1qrm	HLA-A*0201	9	LLFGKPVYV	0.56	0.24
I	HLA-A*0201	1qr1	1qrm	HLA-A*0201	9	IISAVGIL	0.87	0.36
I	HLA-A*0201	1qsf	1qrm	HLA-A*0201	9	LLFGYPVAV	0.94	0.41
I	HLA-A*0201	1bd2	1qrm	HLA-A*0201	9	LLFGYPYV	0.68	0.46
I	HLA-A*0201	1hhg	1i4f	HLA-A*0201	9	TLTSCNTSV	0.58	0.56
I	HLA-A*0201	1hhh	1i4f	HLA-A*0201	10	FLPSDFFPSV	1.48	0.57
I	H2-Kb	1osz	1nam	H2-Kb	8	RGYLYQGL	0.85	0.47
I	H2-Kb	1fo0	1nam	H2-Kb	8	INFDENTI	0.62	0.35
I	H2-Kb	1g6r	1nam	H2-Kb	8	SIYRYGL	0.66	0.11
I	H2-Kb	1kbg	1nam	H2-Kb	8	RGYVYXGL	0.40	0.06
I	H2-Kb	1g7p	1nam	H2-Kb	9	SRDHSRTPM	1.41	0.82
II	HLA-DRB1*0101	1fyt	2iam	HLA-DRB1*0101	13	PKYVKQNTLKLAT	0.69	0.35
II	HLA-DRB1*0101	1klu	2iam	HLA-DRB1*0101	15	GELIGTLNAAKVPAD	0.85	0.59
II	HLA-DRB1*0101	1t5w	2iam	HLA-DRB1*0101	13	AAYSDAQATPLLS	0.99	0.65
II	HLA-DRB1*0101	1pyw	2iam	HLA-DRB1*0101	9	FVKQNXAL	0.40	0.32
II	HLA-DRB1*0101	1sje	2iam	HLA-DRB1*0101	15	PEVIPMFSALSEGAT	0.70	0.37
II	HLA-DRB1*0101	1aqd	2iam	HLA-DRB1*0101	14	GSDWRFLRGYHQYA	1.68	1.01

C α RMSD values are calculated only for the nonamer binding core (shown in bold) for peptides with more than 9 residues in the X-ray crystal structures. X: Amino acid analogues (chemical mimics).

grooves in different MHC alleles. Out of the 4 promiscuous peptides, the peptide FAPGNYPAL from the pMHC structure 2vaa having the murine MHC allele H2-Kb, when docked onto the MHC from the structure 1ce6 with the murine MHC allele H2-Db, was generated with the best C α RMSD of 0.21 Å (Table 5) compared to the same peptide from 1ce6. The highest C α RMSD value (0.79 Å) obtained using pDOCK during this experiment was when the peptide from the structure 1zsd was docked onto the MHC from the structure 2ak4 (Table 5). This value is still well within the acceptable value of 2.50 Å.

In all, only one peptide generated using pDOCK from the single template docking experiments has the C α RMSD value above 1.00 Å (Table 4) compared to 5 peptides (three from single template docking and two from variant template docking) with C α RMSD values above 1.00 Å using our previous methodology (Table 4 and Table 5). It is thus clear that pDOCK accurately predicts the structure of cognate peptides in both single and variant template docking cases. These evaluation steps are also vital to establish the efficiency with which our new method can dock and subsequently predict novel peptides onto given MHC proteins.

Table 5 Docking promiscuous peptides onto variant templates: comparison of pDOCK with our previous method

MHC class	Peptide PDB Allele	Peptide PDB	MHC Template Structure	Template Allele	Peptide Length	Peptide Sequence	C α RMSD compared to template peptides (Å)	
							Previous method	pDOCK
I	HLA-B*3501	1zhk [#]	1zhl [#]	HLA-B*3508	13	LPEP L PQGQLTAY	0.62	0.44
I	HLA-B*3501	1zsd [#]	2ak4	HLA-B*3508	11	EPL P QGQLTAY	1.15	0.79
I	H2-Kb	2vaa [#]	1ce6	H2-Db	9	FAPGNYPAL	0.73	0.21
II	HLA-DRB1*1501	1bx2 [#]	1fv1	HLA-DRB5*0101	14	ENPV V HFFKNIVTP	1.01	0.22

C α RMSD values are calculated only for the nonamer binding core (shown in bold) for peptides with more than 9 residues in the X-ray crystal structures. [#] The structures are not listed in Additional file – Table S1 due to redundancy in MPID-T2.

Conclusions

We have developed pDOCK as a fast, accurate and robust method for flexible docking of peptides to MHC-I and MHC-II proteins. Our results provide evidence of overcoming limitations pertaining to the application of our previous methodology, such as the presence of water molecules in and around the peptide and within the peptide binding groove in the template and relatively longer computational time required. Benchmarking with our previous method for a dataset of 50 non-redundant pMHC complexes consistently produced least energy docked conformations of peptides below 1.00 Å Cα RMSD from their respective native orientations for all 50 peptides. The Cα RMSD range for the same dataset was 0.09 Å (1s7q) to 0.66 Å (1i1y) using pDOCK compared to a Cα RMSD range from 0.20 Å (1s7q) to 1.53 Å (1jfl) applying our previous protocol. These observations imply an improvement in the accuracy by upto two and a half folds compared to our previous protocol. The outcomes of our validation experiments suggest a seven-fold improvement in the accuracy of the pDOCK docking protocol. pDOCK can therefore be successfully applied as a generalized, efficient protocol for docking of peptides to MHC-I and MHC-II receptors with improved accuracy, greater coverage of peptide residues and vastly reduced computational time (up to 60% compared to our earlier method).

The average time taken to perform each step using pDOCK has also improved drastically compared to our old technique on a 2 CPU 3.20 GHz 3 GB RAM workstation. This is mainly due to the consolidation of the docking and refinement protocols into a single step docking and refinement procedure. Our results establish the efficacy of pDOCK to model highly accurate pMHC complex structures permitting conformational sampling of the peptide in MHC binding groove. The current study thus presents one of the most accurate pMHC docking protocols developed to date. pDOCK targets a more generic approach to generation of docked conformations of peptides using a single template for each allele. For some pMHC complexes however, appropriate addition of mediating molecules or considerations of solvent effects may lead to a possible improvement in docking accuracy. Rapid and large scale docking and scanning for identification of potential candidates for immunogenicity from repertoires of immunologically significant antigenic peptide sequences is possible by automating all the steps. No requirement for experimental data to be trained and the need of only a suitable template for a particular allele give pDOCK a prominent edge over other sequence-based techniques such as Artificial Neural Networks, Support Vector Machines, and Hidden Markov Models.

pDOCK is also highly efficient in accurately predicting the docked conformations of amino acid analogues or chemical components within the peptide ligand suggesting its possible use as a docking and evaluation tool in structure-based drug design protocols and chemoinformatics. The single and variant template docking experiments along with the validation experiments also serve as strong benchmarks for pDOCK against our old method. pDOCK can correctly predict the conformation of residues that extend into the MHC binding cleft and therefore could help identify essential contacts with the MHC receptor, responsible for reducing the half life of the pMHC complex such that the peptide is held long enough within the MHC groove for presentation at the APC cell surface leading to surveillance and recognition by the TR molecules which in turn results in the activation of T cells and triggers the adaptive immune response cascade. Another significant improvement in this study is that the peptide ligand is allowed full flexibility within the peptide binding groove of the MHC proteins, unlike our previous method where the peptide termini were docked rigidly to the MHC groove. This aspect of pDOCK has helped us carry out fully flexible peptide docking to the MHC proteins. Our results also indicate the successful application of this protocol for easy *in silico* identification of promiscuous peptide epitopes that are applicable to higher proportions of human population with greater propensity to bind to MHC proteins and consequently activate T cells making them key targets for the design of vaccines and immunotherapies.

Methods

Data

pDOCK was tested on a non-redundant dataset of 186 (149 MHC-I and 37 MHC-II) pMHC complexes from the MPID-T2 (<http://biolinfo.org/mpid-t2>) database for which X-ray crystal structures are available in the PDB and the IMGT/3Dstructure-DB. When there is more than one complex with the same bound peptide and the same allele, the structure with the highest resolution is selected to avoid redundancy. When more than one bound peptide is available in the selected crystal structure, all bound peptides in that crystal structure are analyzed. TR/pMHC structures in MPID-T2 database are treated as non-redundant entries unless they have the same peptide, allele and TR type. In which case, the structure with the best resolution is considered non-redundant. Similarly, a dataset of 25 (18 MHC-I and 7 MHC-II) pMHC complexes was selected from the pDOCK test set for single and variant template docking. When more than one allele is available as template for docking of peptides into a single or variant template,

the allele with the highest resolution was selected. Redundancy in MPID-T2 data is primarily decided from the similarities in peptides, MHC alleles and TR types (in case of TR/pMHC structures). Since one publication can refer to crystal structures of many complexes, redundancy in the literature is not considered as a criterion for redundancy. Some redundant structures were used for variant template docking (Table 5) due to limited number of crystal structures with promiscuous peptides bound to different alleles in the PDB. Although the MPID-T2 database contains 294 pMHC complexes (273 classical and 21 non classical), the 21 non-classical and 87 redundant structures were discarded from this study in order to avoid any biasness in our results.

pMHC complexes for benchmarking and validation

A non-redundant dataset of 50 high quality (35 MHC-I and 15 MHC-II) pMHC complexes, with maximum 3.00 Å resolutions, was selected from the 186 pMHC complexes in the pDOCK test set for benchmarking with the previous methodology. 15 pMHC complexes were chosen for validation experiments depending on the ones used in the corresponding reference studies [34,36-42].

The pDOCK protocol

Unlike our earlier method [2,3], the new technique incorporates flexibility into the entire length of the peptide ligand. We have now incorporated a receptor modeling sub-step at the beginning of our novel schema (Figure 1), which involves rigorous homology modeling of MHC proteins from available MHC sequences by satisfaction of spatial restraints using MODELLER [43] followed by structure optimization and stringent structural quality assessment protocols to affirm the generation of high quality homology models of MHC proteins to be subsequently used in the pMHC docking strategy. Thereby, accounting for the validity of our methodology even in the absence of experimental structures for the MHC proteins and when only MHC sequences are available. However, this sub-step was not used in the current study as testing, benchmarking, validation, single template and variant template docking experiments are performed only on X-ray crystal structures of pMHC complexes.

The current pMHC docking technique is applied on MHC-I and MHC-II related peptides in two preparatory steps and a single consolidated docking and receptor step as follows: Preparatory step 1: receptor positioning using the Internal Coordinate Mechanics (ICM) global optimization algorithm [35]; Preparatory step 2: determining the docking grid using standardized values for MHC supertypes (MHC-I and MHC-II) from our preliminary studies and; A single docking and refinement

step involving: ligand positioning, grid ligand docking followed by iterative ab initio refinements of backbone and ligand interacting side-chain dihedral angles of the MHC binding site residues to eliminate or minimize atomic clash regions at the pMHC interface using a Biased Monte Carlo procedure. The preparatory steps were together used to generate the receptor maps and the final single docking and refinement step was used to carry out ligand docking, generate the final least energy conformation and further refine the product.

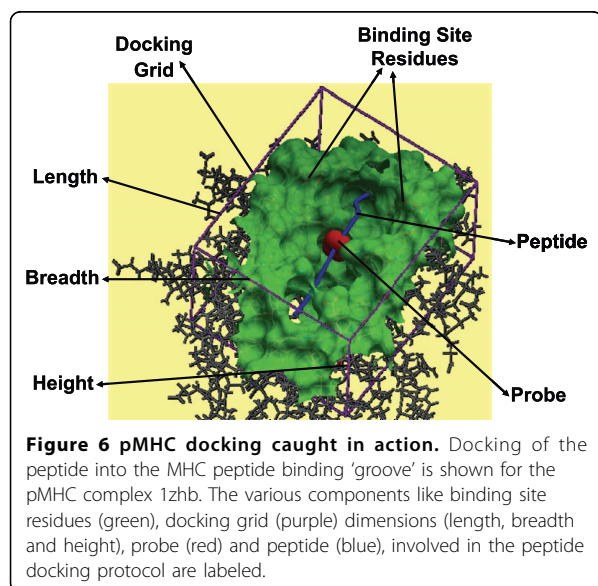
Preparatory steps

Receptor modeling and positioning

Positioning of the MHC receptor is a major requirement in the pMHC docking simulation to ensure a best fit of the flexible peptide in the MHC groove. This first preparatory step (receptor modeling and positioning) is the least time consuming (only applicable to sub-step 'b') in the pDOCK docking protocol and involves two vital sub-steps: (a) homology model building by satisfaction of spatial restraints for MHC sequences where no structural data is available or inserting the MHC crystal structure into the docking simulation and; (b) positioning of the receptor within the docking simulation. Although not used in this study, high quality homology models can be generated, using our previously described three-step homology modeling procedure [44], for alleles with no structural data. Receptor positioning using the ICM global optimization algorithm assures the addition of any important missing residues in the template besides optimizing the zero occupancy side chains and any polar hydrogen atoms.

Determining the docking grid

The second, relatively small preparatory step of our docking procedure is to determining the docking grid which constitutes two major sub-steps: (a) defining the dimensions (*length x breadth x height*) of the 3-D docking grid and; (b) grid map generation for the receptor using the ICM stochastic global optimization algorithm. The ICM algorithm generates a three-dimensional docking grid (purple box in Figure 6), which encloses all MHC binding site residue atoms along with peptides residue atoms, soon after the previous step for generation of receptor maps. This ensures the localization of the peptide ligand for docking within the vicinity of the MHC peptide binding site residues and thereby limits the flexibility of the allowed peptide side chain torsion angles to be randomly sampled within the MHC groove. The dimensions of this 3-D docking box are set to standardized values derived from our preliminary analysis of all available pMHC complexes from the MPID-T2 database, for both MHC-I (35.36 Å x 35.52 Å x 35.79 Å) and MHC-II (58.32 Å x 56.36 Å x 48.87 Å) complexes used in testing, benchmarking, validation, single and



variant template docking simulations. The ICM algorithm then selects all the binding site residues within the MHC groove by creating three-dimensional spheres from and around the centre of the MHC groove with 5.00 Å radii and selecting all the atoms of the MHC binding groove residues falling in and on the spheres (shown in green in Figure 6). 5.00 Å is set as the default radii to select the binding site residues amongst the residues forming the MHC groove as these are the MHC residues that are most likely to form hydrogen bonds (maximum allowed distance – 3.65 Å) and van der Waals contacts (maximum allowed distance – 4.50 Å) with the peptide residues, resulting in strong enough interactions to hold the bound peptide for presentation at the APC cell surface leading to surveillance and recognition by the TR molecules. The stochastic global optimization in internal coordinates with pseudo-Brownian and collective “probability-biased” random moves allow flexibility to the peptide ligand interface side chains and generate a grid potential map of the receptor energy localized to small cubic regions of side 1.00 Å from the carbon-alpha backbone of the peptide.

Single step docking and refinement

Ligand positioning, grid ligand docking and refinement

As with receptor positioning, ligand positioning is also equally important in achieving the best docked conformations, with the lowest energy values for flexible peptides using pDOCK. The final, most exhaustive (in terms of computational time required compared to the other two steps) single step docking and refinement part deals with ligand positioning, grid ligand docking and

refinement, comprising three very important sub-steps: (a) positioning of the peptide ligand either by using the original crystal structure or by inserting a peptide model into the docking simulation using the peptide sequence; (b) placing and positioning of the probe into the peptide binding groove using the Internal Coordinate Mechanics global optimization algorithm and; (c) flexible docking of the peptide into the MHC groove and refinement of all ligand and binding site residues using a Biased Monte Carlo procedure. Ligand positioning was carried out either by using ICM algorithm for existing peptides within the X-ray crystal structures of pMHC complexes or by manually inserting a peptide model into the docking simulation for each of the available peptide sequences (docking of novel peptides to a single template and docking of promiscuous peptides to variant templates). This was followed by placing a probe (red in Figure 6) in the MHC groove which provides an initial position for conformational sampling and docking simulations using the ICM algorithm.

ICM docking algorithm [35] runs flexible docking of peptide ligands to MHC peptide binding clefts. During the docking simulation, the ligand side-chain torsions that have been previously stored within the grid receptor maps (preparatory step 2) are changed in each random step using a Biased Monte Carlo procedure, which begins by pseudo-randomly selecting a set of torsion angles in the ligand and consequently finding the local energy minima about those angles. Upon satisfaction of the Metropolis criteria, novel conformations are adopted with a probability $\min(1, \exp[-\Delta G/RT])$, where R is the universal gas constant and T is the absolute temperature of the simulation. The temperature was set to 300 K for the current study. To keep the ligand molecule close to the starting conformation, loose restraints are imposed on its positional variables. The internal energy function adopted for our simulations integrates internal van der Waals interactions energy (calculated using an extension of ECEPP/3 with force field parameters) [45], hydrogen bonding energy, torsion energy, electrostatic energy with a distance-dependent dielectric constant ($\epsilon = 4r$; where ϵ is the distance-dependent dielectric constant and r is the distance) [46] and hydrophobic potential between the atoms of peptide residues and atoms of the binding site residues. The final energy incorporates configurational entropy of side chains and the surface-based solvation energy to select the best-iterated orientations. In brief, the complete optimal energy function, E , is made up of the internal energy of the ligand and the intermolecular energy of the optimized receptor potential maps and can be summarized as:

$$E = E_{vw} + E_{en} + 2.16 E_{el}^{Solv} + 2.53 E_{hb} + 4.35 E_{hp} + 0.20 E_{solv}$$

where E_{vw} is the internal van der Waals interaction energy, E_{en} is the configurational/conformational entropy, E_{el}^{Solv} is the electrostatic energy of solvation, E_{hb} is the hydrogen bonding energy, E_{hp} is the hydrophobic potential and E_{solv} is the surface-based solvation energy.

Finally, to improve the accuracy of the initial predicted conformation, refinement of the ligand as well as binding site residues backbone and side chains was performed as described in our previous methodology [2,3] to overcome any atomic clashes detected at the pMHC binding interface, using ICM Biased Monte Carlo procedure. Again, restraints are imposed upon the positional variables of the C α atoms of the peptide residues. The early stages of the refinement efforts try to trounce the consequences of docking fully flexible ligands to rigid receptors by introducing partial flexibility to the backbone of MHC peptide binding residues. Refinements of ligand and receptor side-chain torsions in the vicinity of 4.00 Å from the receptor were executed upon the final backbone structure of the peptides to keep the docked peptides closest to their starting conformations. The energy function, E , utilized for this refinement sub-step, is the sum of energy terms arising from the optimal energy function described above:

$$E = E_{vw} + E_{hbonds} + E_{tors} + E_{elec} + E_{solv} + E_{en}$$

where E_{tors} is the torsion energy, E_{elec} is the electrostatic energy and E_{en} is the entropic term.

Additional File 1: Table S1. Application of pDOCK to the 186 (149 MHC-I and 37 MHC-II) non-redundant structures from MPID-T2 database. (*.pdf) Application of pDOCK to the 186 (149 MHC-I and 37 MHC-II) non-redundant structures from MPID-T2 database.

Additional File 2: Figure S1. Comparison of C α RMSD values obtained using pDOCK and our previous method across the benchmarking dataset (*.pdf) Comparison of C α RMSD values obtained using pDOCK and our previous method across the benchmarking dataset

Acknowledgements

JMK gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship and a Macquarie University Postgraduate Research Fund. We also thank Dr. J.C. Tong, Institute for Infocomm Research, Singapore for useful discussions on grid docking. Open access publication charges were borne by Macquarie University. This article has been published as part of *Immunome Research* Volume 6 Supplement 1, 2010: Ninth International Conference on Bioinformatics (InCoB2010): Immunome Research. The full contents of the supplement are available online at <http://www.immunome-research.com/supplements/6/S1>.

Author details

¹Department of Chemistry and Biomolecular Sciences and ARC Center of Excellence in Bioinformatics, Macquarie University, NSW 2109, Australia.

²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

Authors' contributions

JMK developed the methodology, carried out the computational simulation studies and drafted the manuscript. JMK and SR participated in the design of the study and interpretation of data. SR conceived the project and

finalized the manuscript. Both authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 27 September 2010

References

- Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, Peters B: Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* 2008, **4**:2.
- Tong JC, Tan TW, Ranganathan S: Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci* 2004, **13**(9):2523-2532.
- Tong JC, Zhang GL, Tan TW, August JT, Brusic V, Ranganathan S: Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics* 2006, **22**(10):1232-1238.
- Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G: IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhCSF) G-LIKE-DOMAIN. *Dev Comp Immunol* 2005, **29**(11):917-938.
- Tong JC, Bramson J, Kanduc D, Chow S, Sinha AA, Ranganathan S: Modeling the bound conformation of Pemphigus vulgaris-associated peptides to MHC class II DR and DQ alleles. *Immunome Res* 2006, **2**:1.
- Tong JC, Tan TW, Sinha AA, Ranganathan S: Prediction of desmoglein-3 peptides reveals multiple shared T-cell epitopes in HLA DR4- and DR6-associated Pemphigus vulgaris. *BMC Bioinformatics* 2006, **7**(Suppl 5):S7.
- Tong JC, Zhang ZH, August JT, Brusic V, Tan TW, Ranganathan S: In silico characterization of immunogenic epitopes presented by HLA-Cw*0401. *Immunome Res* 2007, **3**:7.
- Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC: Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 1993, **364**(6432):33-39.
- Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, Wiley DC: Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 1994, **368**(6468):215-221.
- Stern LJ, Wiley DC: Antigenic peptide binding by class I and class II histocompatibility proteins. *Structure* 1994, **2**(4):245-251.
- Khan JM, Tong JC, Ranganathan S: Structural Immunoinformatics: Understanding MHC-peptide-TR binding. *Bioinformatics for Immunomics* Springer, Immunomics Reviews Series Davies MN, Ranganathan S, Flower DR 2009, **3**:77-94.
- Tong JC, Tan TW, Ranganathan S: Methods and protocols for prediction of immunogenic epitopes. *Brief Bioinform* 2007, **8**(2):96-108.
- Yewdell JW, Bennink JR: Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol* 1999, **17**:51-88.
- Yu K, Petrovsky N, Schonbach C, Koh JY, Brusic V: Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol Med* 2002, **8**(3):137-148.
- Nielsen M, Lundegaard C, Wornig P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004, **20**(9):1388-1397.
- Vordermeier M, Whelan AO, Hewinson RG: Recognition of mycobacterial epitopes by T cells across mammalian species and use of a program that predicts human HLA-DR binding peptides to predict bovine epitopes. *Infect Immun* 2003, **71**(4):1860-1867.
- Doytchinova IA, Flower DR: class I T-cell epitope prediction: improvements using a combination of proteasome cleavage, TAP affinity, and MHC binding. *Mol Immunol* 2006, **43**(13):2037-2044.
- Schonbach C, Kun Y, Brusic V: Large-scale computational identification of HIV T-cell epitopes. *Immunol Cell Biol* 2002, **80**(3):300-306.
- Bhasin M, Raghava GP: A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci* 2007, **32**(1):31-42.
- Zhang GL, Khan AM, Srinivasan KN, Heiny A, Lee K, Kwok CK, August JT, Brusic V: Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. *BMC Bioinformatics* 2008, **9**(Suppl 1):S19.

21. Srinivasan KN, Zhang GL, Khan AM, August JT, Brusic V: Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. *Bioinformatics* 2004, **20**(Suppl 1):i297-302.
22. Donnes P, Elofsson A: Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 2002, **3**:25.
23. Li S, Yao X, Liu H, Li J, Fan B: Prediction of T-cell epitopes based on least squares support vector machines and amino acid properties. *Anal Chim Acta* 2007, **584**(1):37-42.
24. Liu W, Wan J, Meng X, Flower DR, Li T: In silico prediction of peptide-MHC binding affinity using SVMHC. *Methods Mol Biol* 2007, **409**:283-291.
25. Singh SP, Mishra BN: Ranking of binding and nonbinding peptides to MHC class I molecules using inverse folding approach: Implications for vaccine design. *Bioinformation* 2008, **3**(2):72-82.
26. Knapp B, Omasits U, Frantal S, Schreiner W: A critical cross-validation of high throughput structural binding prediction methods for pMHC. *J Comput Aided Mol Des* 2009, **23**(5):301-307.
27. Logean A, Rognan D: Recovery of known T-cell epitopes by computational scanning of a viral genome. *J Comput Aided Mol Des* 2002, **16**(4):229-243.
28. Kosmopoulou A, Vlasi M, Stavrakoudis A, Sakarellos C, Sakarellos-Daitsiotis M: T-cell epitopes of the La/SSB autoantigen: prediction based on the homology modeling of HLA-DQ2/DQ7 with the insulin-B peptide/HLA-DQ8 complex. *J Comput Chem* 2006, **27**(9):1033-1044.
29. Sauton N, Lagorce D, Villoutreix BO, Miteva MA: MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics* 2008, **9**:184.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**(1):235-242.
31. Kaas Q, Ruiz M, Lefranc MP: IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* 2004, **32**(Database issue):D208-210.
32. Ehrenmann F, Kaas Q, Lefranc MP: IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* 2010, **38**(Database issue):D301-307.
33. Ranganathan S, Tong JC, Tan TW: Structural immunoinformatics. *Immunoinformatics* Springer, Immunomics Reviews Series Schonbach C, Ranganathan S, Brusic V 2008, 51-61.
34. Bordner AJ, Abagyan R: Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* 2006, **63**(3):512-526.
35. Abagyan RA, Totrov MM: Ab initio folding of peptides by the Optimal-Bias Monte Carlo Minimization Procedure. *J Comput Phys* 1999, **151**:402-421.
36. Wei HY, Tsai KC, Lin TH: Modeling ligand-receptor interaction for some MHC class II HLA-DR4 peptide mimetic inhibitors using several molecular docking and 3D QSAR techniques. *J Chem Inf Model* 2005, **45**(5):1343-1351.
37. Liu Z, Dominy BN, Shakhnovich EI: Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. *J Am Chem Soc* 2004, **126**(27):8515-8528.
38. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V: Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 1999, **42**(22):4650-4658.
39. Desmet J, De Maeyer M, Spriet J, Lasters I: Flexible docking of peptide ligands to proteins. *Methods Mol Biol* 2000, **143**:359-376.
40. Sezerman U, Vajda S, DeLisi C: Free energy mapping of class I MHC molecules and structural determination of bound peptides. *Protein Sci* 1996, **5**(7):1272-1281.
41. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: Computing the structure of bound peptides. Application to antigen recognition by class I major histocompatibility complex receptors. *J Mol Biol* 1993, **234**(3):515-521.
42. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: Flexible docking of peptides to class I major-histocompatibility-complex receptors. *Genet Anal* 1995, **12**(1):1-21.
43. Sali A, Blundell TL: Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993, **234**(3):779-815.
44. Khan JM, Ranganathan S: A multi-species comparative structural bioinformatics analysis of inherited mutations in alpha-D-mannosidase

reveals strong genotype-phenotype correlation. *BMC Genomics* 2009, **10**(Suppl 3):S33.

45. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA: Energy parameters in polypeptides, 10: Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J Phys Chem* 1992, **96**(15):6472-6484.
46. Fernandez-Recio J, Totrov M, Abagyan R: Soft protein-protein docking in internal coordinates. *Protein Sci* 2002, **11**(2):280-291.

doi:10.1186/1745-7580-6-S1-S2

Cite this article as: Khan and Ranganathan: pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Research* 2010 **6**(Suppl 1):S2.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Additional File 1

pDOCK: A new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes

Javed M. Khan and Shoba Ranganathan

Table S1 – Application of pDOCK to the 186 (149 MHC-I and 37 MHC-II) non-redundant structures from MPID-T2 database.

C α RMSD values are calculated only for the nonameric core (shown in bold) forming the MHC binding register for peptides with more than 9 residues in the X-ray crystal structures. X: Chemical mimics.

Allele	PDB code	Peptide Sequence	Peptide Length	Res. (Å)	C α RMSD (Å)
MHC-I					
HLA-A*0101	1w72	EADPTGHSY	9	2.15	0.42
HLA-A*1101	1qvo	QVPLRPMTYK	10	2.22	0.24
HLA-A*1101	1x7q	KTFPPTEPK	9	1.45	0.36
HLA-A*1101	2hn7	AIMPARFYPK	10	1.60	1.40
HLA-A*1101	1q94	AIFQSSMTK	9	2.40	1.52
HLA-A*0201	1oga	GILGFVFTL	9	1.40	0.16
HLA-A*0201	1t1y	SLYNVVATL	9	2.00	0.78
HLA-A*0201	1s8d	SLANTVATL	9	2.20	0.65
HLA-A*0201	2gtz	ALGIGILTV	9	1.70	0.95
HLA-A*0201	1duy	LFGYPVYV	8	2.15	1.12
HLA-A*0201	2p5w	SLLMWITQC	9	2.20	0.35
HLA-A*0201	2v2w	SLYNTVATL	9	1.60	0.55
HLA-A*0201	2v2x	SLFNTVATL	9	1.60	0.78
HLA-A*0201	1qr1	IISAVVGIL	9	2.40	0.29
HLA-A*0201	1qrn	LLFGYAVYV	9	2.80	0.91
HLA-A*0201	1qse	LLFGYPRYV	9	2.80	0.31
HLA-A*0201	1qsf	LLFGYPVAV	9	2.80	0.34
HLA-A*0201	1hhg	TLTSCNTSV	9	2.60	1.18
HLA-A*0201	1lp9	ALWGFFPVL	9	2.00	0.26
HLA-A*0201	1s9x	SLLMWITQA	9	2.50	0.44
HLA-A*0201	1s9y	SLLMWITQS	9	2.30	0.39
HLA-A*0201	1tvb	ITDQVPFSV	9	1.80	0.38
HLA-A*0201	1tvh	IMDQVPFSV	9	1.80	0.43
HLA-A*0201	1akj	ILKEPVHGV	9	2.65	0.39
HLA-A*0201	1eey	ILSALVGIV	9	2.25	0.92
HLA-A*0201	2guo	AAGIGILTV	9	1.90	2.14
HLA-A*0201	1t20	SLYNTIATL	9	2.20	0.82
HLA-A*0201	1ao7	LLFGYPVYV	9	2.60	0.58
HLA-A*0201	1eez	ILSALVGIL	9	2.30	0.71

Allele	PDB code	Peptide Sequence	Peptide Length	Res. (Å)	Cα RMSD (Å)
HLA-A*0201	2gj6	LLFGKPVYV	9	2.56	0.97
HLA-A*0201	1hhh	FLPSDFFPSV	10	3.00	0.49
HLA-A*0201	1i1f	FLKEPVHGV	9	2.80	0.64
HLA-A*0201	2clr	MLLSVPLLIG	10	2.00	1.01
HLA-A*0201	2gt9	EAAGIGILTV	10	1.75	0.26
HLA-A*0201	1i7r	FAPGFFPYL	9	2.20	0.47
HLA-A*0201	1i7t	ALWGVFPVL	9	2.80	0.42
HLA-A*0201	1i7u	ALWGFVPVL	9	1.80	0.29
HLA-A*0201	1jf1	ELAGIGILTV	10	1.85	0.30
HLA-A*0201	1t1x	SLYLTVATL	9	2.20	1.72
HLA-A*0201	1i1y	YLKEPVHGV	9	2.20	0.66
HLA-A*0201	1t1z	ALYNTAAAL	9	1.90	1.15
HLA-A*0201	2bnq	SLLMWITQV	9	1.70	0.55
HLA-A*0201	2c7u	SLFNTIAVL	9	2.38	0.86
HLA-A*0201	1i4f	GVYDGREHTV	10	1.40	1.25
HLA-A*0201	1bd2	LLFGYPVYV	9	2.50	0.33
HLA-A*0201	2gtw	LAGIGILTV	9	1.55	3.08
HLA-A*2402	2bck	VYGFVRACL	9	2.80	0.79
HLA-A*6801	1tmc	EVAPPEYHRK	10	2.30	0.54
HLA-B*0801	1agc	GGKKKYQL	8	2.10	0.23
HLA-B*0801	1agd	GGKKKYKL	8	2.05	0.45
HLA-B*0801	1agb	GGRKKYKL	8	2.20	0.28
HLA-B*0801	1mi5	FLRGRAYGL	9	2.50	0.37
HLA-B*1501	1xr8	LEKARGSTY	9	2.30	1.15
HLA-B*1501	3c9n	VQQESSFVM	9	1.87	0.80
HLA-B*2101	3bev	GHAEEYGAETL	11	2.10	0.90
HLA-B*2101	3bew	REVDEQLLSV	10	2.60	0.33
HLA-B*2705	1uxs	RRRWRLTV	9	1.55	0.75
HLA-B*2705	1ogt	RRKWRRWHL	9	1.47	0.18
HLA-B*2705	2bsr	RRIYDLIEL	9	2.30	1.72
HLA-B*2705	2bst	SRYWAIRTR	9	2.10	1.36
HLA-B*2705	2a83	RRRWHRWRL	9	1.40	0.18
HLA-B*2705	1w0v	RRLPIFSRL	9	2.27	1.41
HLA-B*2709	1w0w	RRLPIFSRL	9	2.10	0.64
HLA-B*2709	1k5n	GRFAAAIAK	9	1.09	0.75
HLA-B*2709	1uxw	RRRWRLTV	9	1.71	0.47
HLA-B*2709	1jgd	RLLRGHNQY	10	1.90	1.42
HLA-B*3501	2cik	KPIVVLHGY	9	1.75	0.26
HLA-B*3501	2axg	APQPAPENAY	10	2.00	1.14
HLA-B*3501	1a9b	LPPLDITPY	9	3.20	0.39
HLA-B*3501	1qew	FLWGPRALV	9	2.20	0.53
HLA-B*3501	1a1n	VPLRPMTY	8	2.00	0.39
HLA-B*3508	3bw9	CPSQEPMSIYVY	12	1.75	0.27
HLA-B*3508	3bwA	FPTKDVAL	8	1.30	0.26
HLA-B*3508	2ak4	LPEPLPQGQLTAY	13	2.50	1.09

Allele	PDB code	Peptide Sequence	Peptide Length	Res. (Å)	Ca RMSD (Å)
HLA-B*3508	2axf	APQPAPENAY	10	1.80	0.51
HLA-B*4402	1m6o	EEFGRAFSF	9	1.60	1.33
HLA-B*4403	1sys	EEPTVIKKY	9	2.40	0.65
HLA-B*4403	1n2r	EEFGRAFSF	9	1.70	0.99
HLA-B*4405	1syv	EEFGRAFSF	9	1.70	0.77
HLA-B*5101	1e27	LPPVVAKEI	9	2.20	0.18
HLA-B*5101	1e28	TAFTIPSI	8	3.00	0.26
HLA-B*5301	1a1m	TPYDINQML	9	2.30	0.28
HLA-B*5301	1a1o	KPIVQYDNF	9	2.30	0.84
HLA-B*5703	2bvq	KAFSPEVIP	9	2.00	0.50
HLA-B*5703	2bvo	KAFSPEVIPMF	11	1.65	2.17
HLA-B*5703	2bvp	ISPRTLDAW	9	1.35	0.65
HLA-Cw*0304	1efx	GAVDPLLAL	9	3.00	0.35
HLA-Cw*0401	1im9	QYDDAVYKL	9	2.80	0.34
HLA-E*0101	2esv	VMAPRTLIL	9	2.60	0.66
HLA-E*0103	1kpr	VMAPRTVLL	9	2.80	1.02
HLA-E*0103	1kti	VTAPRTLLL	9	3.10	0.45
HLA-E*0103	3cdg	VMAPRTLFL	9	3.40	1.97
HLA-G*0101	2dyp	RIIPRHLQL	9	2.50	0.16
H2-Db	1jpf	SGVENPGGYCL	11	2.18	0.36
H2-Db	1jpg	FQPQNGQFI	9	2.20	0.38
H2-Db	1juf	SSVIGVWYL	9	2.00	0.29
H2-Db	1bz9	FAPGVFPYM	9	2.80	1.19
H2-Db	1ce6	FAPGNYPAL	9	2.90	0.24
H2-Db	1hoc	ASNENMETM	9	2.40	0.46
H2-Db	1ffo	AAVYNFATM	9	2.65	0.25
H2-Db	1ffp	SAVYNFATM	9	2.60	0.33
H2-Db	1fg2	KAVYNFATC	9	2.75	0.19
H2-Db	1inq	SSVVGWVYL	9	2.20	0.42
H2-Db	1n3n	SNLQNAASIA	10	3.00	1.34
H2-Db	1qlf	FAPSNYPAL	9	2.65	0.26
H2-Db	1s7v	KAVYNLATM	9	2.20	0.27
H2-Db	1s7w	KALYNFATM	9	2.40	0.62
H2-Db	1s7x	KAVFNFATM	9	2.41	0.47
H2-Db	1wbx	SQLKNNAKEI	10	1.90	0.38
H2-Db	1wby	SSLENFRAYV	10	2.30	0.31
H2-Db	1yn7	SSLENFAAYV	10	2.20	0.14
H2-Db	2f74	KAVYNFATM	9	2.70	0.27
H2-Db	3buy	LSLRNPILV	9	2.60	0.23
H2-Dd	1qo3	RGPGRFVTI	10	2.30	0.17
H2-Kb	1g6r	SIYRYYGL	8	2.80	0.34
H2-Kb	1s7q	KAVYNFATM	9	1.99	0.09
H2-Kb	1s7r	KAVYNLATM	9	2.95	1.26
H2-Kb	1s7s	KALYNFATM	9	1.99	0.28
H2-Kb	1s7t	KAVFNFATM	9	2.30	0.19

Allele	PDB code	Peptide Sequence	Peptide Length	Res. (Å)	Cα RMSD (Å)
H2-Kb	1g7p	SRDHSRTPM	9	1.50	0.17
H2-Kb	1g7q	SAPDTRPA	8	1.60	0.36
H2-Kb	1kgg	RGYVYXGL	8	2.20	0.47
H2-Kb	1t0m	SSIEFARL	8	2.00	0.21
H2-Kb	1vac	SIINFEKL	8	2.50	0.22
H2-Kb	1wbz	SSYRRPVGI	9	2.00	0.19
H2-Kb	1rjz	SEIEFARL	8	2.60	0.48
H2-Kb	1kj2	KVITFIDL	8	2.71	0.38
H2-Kb	1lk2	GNYSFYAL	8	1.35	0.53
H2-Kb	1zhh	KALYNYAPI	9	2.70	0.24
H2-Kb	1mwa	EQYKFYSV	8	2.40	0.27
H2-Kb	2fo4	SAPDFRPL	8	2.70	0.60
H2-Kb	1n59	AVYNFATM	8	2.95	0.44
H2-Kb	2ol3	SQYYNSL	8	2.90	0.30
H2-Kb	1nam	RGYVYQGL	8	2.70	0.38
H2-Kb	1fo0	INFDNTI	8	2.50	0.34
H2-Kb	1osz	RGYLYQGL	8	2.10	0.28
H2-Kd	1vgk	SYVNTNMGL	9	2.06	0.25
H2-Kd	2fwo	TYQRTRALV	9	2.60	0.26
H2-Kk	1zt1	FEANGNLI	8	2.50	0.45
H2-Kk	1zt7	SEFLLEKRI	9	3.00	0.45
H2-Ld	1ldp	APAAAAAAM	9	3.10	0.59
H2-Ld	1ld9	YPNVNIHNF	9	2.40	0.56
H2-Ld	2e7l	QLSPFPFDL	9	2.50	0.35
H2-Ld	2oi9	QLSPFPFDL	9	2.35	0.55
H2-M3	1mhc	MYFINILT	9	2.10	1.16
H2-Qa-2	1k8d	ILMEHIHKL	9	2.30	0.55
Mamu-A*01	1zvs	TTPESANL	8	2.80	0.65
RT1.Aa	1kjm	AQFSASASR	9	2.35	0.49
RT1-A1C	1kju	NPRAMQALL	9	1.48	0.33
MHC-II					
HLA-DQB1*0201	1s9v	LQFPQPPELPY	11	2.22	0.33
HLA-DQB1*0302	1jk8	LVEALYLVCGERGG	14	2.40	0.31
HLA-DQB1*0302	2nna	SGEGSFQPSQENP	13	2.10	0.22
HLA-DQB1*0602	1uvq	MNLPSTKVSAAVGGGGSLV	20	1.80	0.23
HLA-DRA*0101	1zgl	VHHFKNIVTPRTPG	14	2.80	1.27
HLA-DRB1*0101	1aqd	GSDWRFLRGYHQYA	14	2.45	0.28
HLA-DRB1*0101	1fyt	PKYVKQNTLKLAT	13	2.60	0.23
HLA-DRB1*0101	1klu	GELIGTLNAAKVPAD	15	1.93	0.20
HLA-DRB1*0101	1pyw	FVKQNAXAL	9	2.10	0.81
HLA-DRB1*0101	1sje	PEVIPMFSALESEGAT	15	2.45	0.46
HLA-DRB1*0101	1sjh	PEVIPMFSALESEG	13	2.25	0.22
HLA-DRB1*0101	1t5w	AAYSDAQATPLLLS	13	2.40	0.25
HLA-DRB1*0101	2fse	AGFKGEQGPKGEPG	14	3.10	0.64

Allele	PDB code	Peptide Sequence	Peptide Length	Res. (Å)	Ca RMSD (Å)
HLA-DRB1*0101	2iam	GELIGILNAAKVPAD	15	2.80	0.24
HLA-DRB1*0301	1a6a	PVSKMRMATPLLMQA	15	2.75	0.30
HLA-DRB1*0401	1d5m	XXRAMXSX	8	2.00	0.13
HLA-DRB1*0401	1d5x	XXRXXX	6	2.45	0.11
HLA-DRB1*0401	1d5z	XXRAXSX	7	2.00	0.22
HLA-DRB1*0401	1d6e	XXRXMASX	8	2.45	0.14
HLA-DRB1*0401	1j8h	PKYVKQNTLKLAT	13	2.40	0.20
HLA-DRB1*0401	2seb	AYMRADAAAGGA	12	2.50	0.31
HLA-DRB1*1501	1ymm	ENPVVHFFKNIVTP	14	3.50	0.28
HLA-DRB3*0101	2q6w	AWRSDEALPLG	11	2.25	0.30
HLA-DRB5*0101	1fv1	NPVVHFFKNIVTPRTPPPSQ	20	1.90	0.59
HLA-DRB5*0101	1h15	GGVYHFVKKHVHES	14	3.10	0.22
HLA-DRB5*0101	1hqr	VHFFKNIVTP	10	3.20	0.56
I-Ab	1muj	PVSKMRMATPLLMQA	15	2.15	0.15
I-Ad	1iao	RGISQAVHAAHAEI	14	2.60	0.27
I-Ad	2iad	GHATQGVTAASSHE	14	2.40	0.56
I-A(G7)	1es0	YEIAPVFLLEYVT	14	2.60	0.38
I-Ak	1f3j	AMKRHGLDNYRGYS	14	3.10	0.28
I-Ak	1iak	STDYGILQINSRW	13	1.90	0.23
I-Ak	1jl4	GNSHRGAIEWEGIESG	16	4.30	0.35
I-Au	1u3h	SRGGASQYRPSQ	12	2.42	0.95
I-Au	2pxy	RGGASQYRPSQ	11	2.23	0.28
I-Ek	1r5v	ADLIAYPKAATKF	13	2.50	0.28
I-Ek	1r5w	ADLIAYFKAATKF	13	2.90	1.26

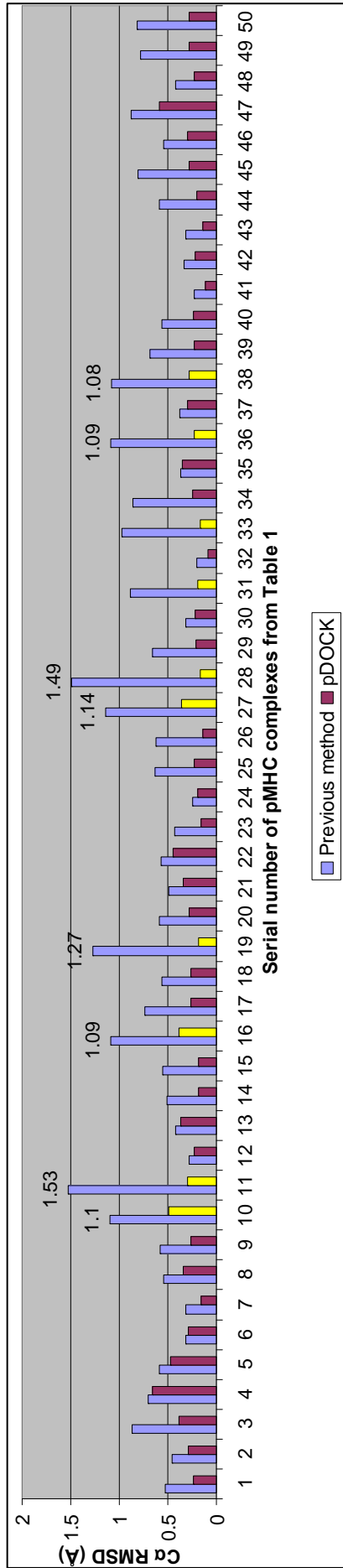
Additional File 2

pDOCK: A new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes

Javed M. Khan and Shoba Ranganathan

Figure S1 – Comparison of C α RMSD values obtained using pDOCK and our previous method across the benchmarking dataset.

The pMHC complex number refers to the serial number of the complexes listed in Table 1. The most significant pDOCK results are highlighted in yellow.



3.2 Conclusions

pDOCK is a fast, accurate and robust method for flexible docking of peptides to MHC-I and MHC-II proteins. The limitations pertaining to the application of our previous methodology [10, 11] have been addressed in pDOCK. Consolidation of the docking and refinement protocols into a single step docking and refinement procedure has resulted in a decreased average time required to perform each docking using pDOCK. Although pDOCK benchmarks against experimental structures, it can be applied to alleles with no structural data using sequence information and a previously described homology modeling procedure [444] to build structural models that can be subsequently used for docking. pDOCK can be automated to perform rapid and large scale docking and scanning for identification of potential candidates for immunogenicity from repertoires of immunologically significant antigenic peptide sequences. pDOCK can therefore be successfully applied as a generalized, efficient protocol for docking of peptides to MHC proteins with improved accuracy, greater coverage of peptide residues and vastly reduced computational time (up to 60% compared to our earlier method [10, 11]).

pDOCK gets a prominent edge over other sequence-based techniques such as ANN, SVM, and HMM owing to no requirement for large numbers of experimental data for training and the need of only a suitable template for a particular allele. Conformation of residues that extend into the MHC binding cleft can also be correctly predicted using pDOCK. This suggests that essential pMHC contacts responsible for reducing the half life of the pMHC complexes could possibly be identified using pDOCK. The full flexibility allowed for the peptide residues and the MHC pocket residues within the peptide binding groove of the MHC proteins, unlike our previous method where the peptide termini were docked rigidly to the MHC groove, highlights a significant improvement in the pDOCK peptide docking procedure.

Chapter 4: MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures

4.1 Summary

Understanding the mechanisms underlying pMHC and TR/pMHC binding and recognition relies mainly of the sequence-structure-function information of these vital immune system interactions [84]. The knowledge of the physicochemical basis for the selection of certain specific peptide epitopes by MHC alleles and the consequent recognition of pMHC ligands by TR proteins is critical for the design of T cell based peptide vaccines [15]. With recent rise in TR/pMHC structural data in the PDB [62, 63] and in IMGT/3Dstructure-DB [57, 58], and newly recognized interaction parameters [51], there is an increasing demand for more effective and efficient computational protocols to predict T cell epitopes. Thus, publication 4 describes a new database for sequence-structure-function information on pMHC and TR/pMHC interactions known as “MHC-Peptide Interaction Database-TR version 2 (MPID-T2)”, that has been developed and augmented with latest PDB and IMGT/3Dstructure-DB data, advanced features and new parameters for analysis of pMHC and TR/pMHC structures, to gain an in-depth understanding of structural determinants underlying TR/pMHC binding and recognition.

4.2 Data

MPID-T2 contains interaction information on all available experimental X-ray crystal structures of pMHC and TR/pMHC complexes extracted from PDB. It is a semi-automatically curated structure-derived MySQL database hosted on a Linux server. The November 2010 update of the database comprises 415 entries from five MHC sources (human: 282, murine: 127, rat: 3, chicken: 2 and monkey: 1), spanning 56 alleles. The 415 entries have 353 pMHC structures (Table 4.1) and 62 TR/pMHC structures (Table 4.2) from 352 MHC-I complexes and 63 MHC-II complexes, comprising 327 non-redundant entries (MHC-I: 279 and MHC-II: 48). The database includes non-classical structures (with T cell receptor like antibodies, CD proteins and natural killer cell immunoglobulin like receptors or KIR associated to the pMHC) and complexes with non-standard residues. Within the database, the most accurate and complete structure is stored for PDB structures with multiple molecular assemblies. Manual verification, classification and analysis for pMHC and TR/pMHC interactions is carried out on each structure and the results are stored in the database. MPID-T2 is available online at: <http://biolinfo.org/mpid-t2>.

Table 4.1: List of pMHC structures in MPID-T2

PL: peptide length; MCI: MHC chain identifier(s); PCI: peptide chain identifier; Res.: Resolution

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0135	1W72	A*0101	9	Melanoma-associated antigen 1	EADPTGHSY	A	C	2.15	2004
I	Human	MHCA0255	3BO8	A*0101	9	Melanoma-associated antigen 1	EADPTGHSY	A	C	1.80	2008
I	Human	MHCA0070	1DUY	A*0201	8	Tax peptide	LFGYPPVYV	A	C	2.15	2000
I	Human	MHCA0002	1AKJ	A*0201	9	HIV-1 RT	ILKEPVHGV	A	C	2.65	1997
I	Human	MHCA0004	1AO7	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.60	1997
I	Human	MHCA0006	1B0G	A*0201	9	Human-peptide P0149	ALWGFFPVL	A	C	2.60	1998
I	Human	MHCA0010	1B0R	A*0201	9	Influenza matrix	GILGFVFTL	A	C	2.90	1999
I	Human	MHCA0005	1BD2	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.50	1998
I	Human	MHCA0040	1DUZ	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	1.80	2000
I	Human	MHCA0071	1EEY	A*0201	9	Gp2 Peptide Variant	ILSALVGIV	A	C	2.25	2003
I	Human	MHCA0072	1EEZ	A*0201	9	Gp2 Peptide Variant	ILSALVGIL	A	C	2.30	2003
I	Human	MHCA0008	1HHG	A*0201	9	HIV-1 gp 120	TLTSCNTSV	A	C	2.60	1993
I	Human	MHCA0009	1HHI	A*0201	9	Synthetic	GILGFVFTL	A	C	2.50	1993

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0001	1HHJ	A*0201	9	Synthetic	ILKEPVHGV	A	C	2.50	1993
I	Human	MHCA0003	1HHK	A*0201	9	Synthetic	LLFGYPVYV	A	C	2.50	1993
I	Human	MHCA0039	111F	A*0201	9	HIV-RT	FLKEPVHGV	A	C	2.80	2000
I	Human	MHCA0038	111Y	A*0201	9	HIV-1RT	YLKEPVHGV	A	C	2.20	2000
I	Human	MHCA0060	117R	A*0201	9	Synthetic	FAPGFFPYL	A	C	2.20	2001
I	Human	MHCA0059	117T	A*0201	9	Synthetic	ALWGVFPVL	A	C	2.80	2001
I	Human	MHCA0061	117U	A*0201	9	Synthetic	ALWGFVPVL	A	C	1.80	2001
I	Human	MHCA0051	1IM3	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	2.20	2001
I	Human	MHCA0058	1JHT	A*0201	9	Mart-1	ALGIGILTV	A	C	2.15	2001
I	Human	MHCA0094	1LP9	A*0201	9	Self peptide	ALWGFFPVL	A	C	2.00	2003
I	Human	MHCA0108	1OGA	A*0201	9	Synthetic	GILGFVFTL	A	C	1.40	2003
I	Human	MHCA0110	1P7Q	A*0201	9	Pol Polyprotein	ILKEPVHGV	A	C	3.40	2003
I	Human	MHCA0114	1QR1	A*0201	9	Gp2 Peptide	IISAVVGIL	A	C	2.40	2000
I	Human	MHCA0057	1QRN	A*0201	9	Tax peptide P6A	LLFGYAVYV	A	C	2.80	1999
I	Human	MHCA0056	1QSE	A*0201	9	Tax peptide	LLFGYPRYV	A	C	2.80	1999
I	Human	MHCA0055	1QSF	A*0201	9	Tax peptide	LLFGYPVAV	A	C	2.80	1999
I	Human	MHCA0146	1S8D	A*0201	9	Gag peptide	SLANTVATL	A	C	2.20	2005
I	Human	MHCA0123	1S9W	A*0201	9	Ny-Eso-1 Peptide	SLLMWITQC	A	C	2.20	2004

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0124	1S9X	A*0201	9	Ny-Eso-1 Peptide Analogue S9A	SLLMWITQA	A	C	2.50	2004
I	Human	MHCA0125	1S9Y	A*0201	9	Ny-Eso-1 Peptide Analogue S9S	SLLMWITQS	A	C	2.30	2004
I	Human	MHCA0147	1T1W	A*0201	9	Gag peptide	SLFNTIAVL	A	C	2.20	2005
I	Human	MHCA0148	1T1X	A*0201	9	Gag peptide	SLYLTVAATL	A	C	2.20	2005
I	Human	MHCA0149	1T1Y	A*0201	9	Gag peptide	SLYNNVATL	A	C	2.00	2005
I	Human	MHCA0150	1T1Z	A*0201	9	Gag peptide	ALYNTAAAL	A	C	1.90	2005
I	Human	MHCA0151	1T20	A*0201	9	Gag peptide	SLYNTIATL	A	C	2.20	2005
I	Human	MHCA0152	1T21	A*0201	9	Gag peptide	SLYNTVAATL	A	C	2.19	2005
I	Human	MHCA0153	1T22	A*0201	9	Gag peptide	SLYNTVAATL	A	C	2.20	2005
I	Human	MHCA0130	1TVB	A*0201	9	Epitope of melanocyte protein Pmel 17	ITDQVPFSV	A	C	1.80	2005
I	Human	MHCA0131	1TVH	A*0201	9	Epitope of melanocyte protein Pmel 17	IMDQVPFSV	A	C	1.80	2005
I	Human	MHCA0170	2AV1	A*0201	9	HTLV-1 TAX peptide	LLFGYPVYV	A	C	2.05	2005
I	Human	MHCA0169	2AV7	A*0201	9	HTLV-1 TAX peptide	LLFGYPVYV	A	C	1.95	2005
I	Human	MHCA0175	2BNQ	A*0201	9	Synthetic peptide	SLLMWITQV	A	C	1.70	2005

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0176	2BNR	A*0201	9	Synthetic peptide	SLLMWITQC	A	C	1.90	2005
I	Human	MHCA0177	2C7U	A*0201	9	Gag peptide	SLFNTIAVL	A	C	2.38	2006
I	Human	MHCA0160	2F53	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	2.10	2006
I	Human	MHCA0161	2F54	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	2.70	2006
I	Human	MHCA0162	2GIT	A*0201	9	HTLV-1 TAX peptide	LLFGKPVYV	A	C	1.70	2006
I	Human	MHCA0163	2GJ6	A*0201	9	Modified HTLV-1 TAX (Y5K-IBA) peptide	LLFGKPVYV	A	C	2.56	2006
I	Human	MHCA0165	2GTW	A*0201	9	Octapeptide from Melan-A/MART-1	LAGIGILTV	A	C	1.55	2007
I	Human	MHCA0166	2GTZ	A*0201	9	Octapeptide from Melan-A/MART-1	ALGIGILTV	A	C	1.70	2007
I	Human	MHCA0167	2GUO	A*0201	9	Melan-A/MART-1(27-35) peptide	AAGIGILTV	A	C	1.90	2007
I	Human	MHCA0171	2J8U	A*0201	9	Self peptide P1049	ALWGFFPVL	A	C	2.88	2007
I	Human	MHCA0172	2JCC	A*0201	9	Self peptide P1049	ALWGFFPVL	A	C	2.50	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0178	2P5E	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	1.89	2007
I	Human	MHCA0179	2P5W	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	2.20	2007
I	Human	MHCA0180	2PYE	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	2.30	2007
I	Human	MHCA0168	2UWE	A*0201	9	Self peptide from C15ORF24 uncharacterized protein	ALWGFFPVL	A	C	2.40	2007
I	Human	MHCA0173	2V2W	A*0201	9	HIV-1 P17 peptide	SLYNTVATL	A	C	1.60	2006
I	Human	MHCA0174	2V2X	A*0201	9	HIV-1 P17 peptide	SLFNTVATL	A	C	1.60	2006
I	Human	MHCA0181	2VLJ	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	2.40	2008
I	Human	MHCA0182	2VLK	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	2.50	2008
I	Human	MHCA0183	2VLL	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	1.60	2008

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0184	2VLR	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	2.30	2008
I	Human	MHCA0313	2X4N	A*0201	9	Restricted influenza matrix epitope	KILGGVFXV	A	C	2.34	2010
I	Human	MHCA0312	2X4O	A*0201	9	HIV envelope glycoprotein GP 160	KLTPLCVTL	A	C	2.30	2010
I	Human	MHCA0311	2X4P	A*0201	9	Restricted influenza matrix epitope	MILGGVFXV	A	C	2.30	2010
I	Human	MHCA0310	2X4Q	A*0201	9	Restricted influenza matrix epitope	MILGGVFXV	A	C	1.90	2010
I	Human	MHCA0309	2X4R	A*0201	9	Cytomegalovirus (CMV) PP65 epitope	NLVPMVATV	A	C	2.30	2010
I	Human	MHCA0308	2X4S	A*0201	9	Epitope of the H5N1 (avian flu) nucleoprotein	AMDSNTLEL	A	C	2.55	2010
I	Human	MHCA0307	2X4T	A*0201	9	Peiodate-cleavable peptide 65 KDA phosphoprotein	NLVXMMVATV	A	C	2.30	2010

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0306	2X4U	A*0201	9	HIV reverse transcriptase/ ribonuclease H	ILKEPVHGV	A	C	2.10	2010
I	Human	MHCA0248	3BGM	A*0201	9	Nonameric peptide from Serine/threonine-protein kinase D2	RQASLSISV	A	C	1.60	2008
I	Human	MHCA0261	3D25	A*0201	9	Nonameric peptide from HA-1	VLHDDLLEA	A	C	1.30	2009
I	Human	MHCA0338	3D39	A*0201	9	Modified HTLV-1 TAX (Y5(4fluoro)F) peptide	LLFGFPVYV	A	C	2.81	2009
I	Human	MHCA0337	3D3V	A*0201	9	Modified HTLV-1 TAX (Y5(3,4-difluoro)F) peptide	LLFGFPVYV	A	C	2.80	2009
I	Human	MHCA0265	3FQT	A*0201	9	Peptide 38-46 from cell division cycle 25b (CDC25b)	GLLGSPVRA	A	C	1.80	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0264	3FQU	A*0201	9	Phospho-peptide 38-46 from cell division cycle 25b (CDC25b)	GLLGSPVRA	A	C	1.80	2009
I	Human	MHCA0263	3FQW	A*0201	9	Peptide 1097-1105 from insulin receptor substrate 2 (IRS2)	RVASPTSGV	A	C	1.93	2009
I	Human	MHCA0262	3FQX	A*0201	9	Phospho-peptide 1097-1105 from insulin receptor substrate 2 (IRS2)	RVASPTSGV	A	C	1.70	2009
I	Human	MHCA0275	3FT2	A*0201	9	Citrulline variant HA-1 peptide	VLRDDLLEA	A	P	1.80	2009
I	Human	MHCA0274	3FT3	A*0201	9	Histidine variant HA-1 peptide	VLHDDLLEA	A	P	1.95	2009
I	Human	MHCA0273	3FT4	A*0201	9	Arginine variant HA-1 peptide	VLRDDLLEA	A	P	1.90	2009
I	Human	MHCA0272	3GJF	A*0201	9	NYESO-1 peptide	SLLMWITQV	A	C	1.90	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0341	3GSN	A*0201	9	HCMV pp65 fragment 495-503	NLVPMVATV	H	P	2.80	2009
I	Human	MHCA0288	3GSO	A*0201	9	HCMV pp65 fragment 495-503	NLVPMVATV	A	P	1.60	2009
I	Human	MHCA0287	3GSQ	A*0201	9	HCMV pp65 fragment 495-503, variant M5S	NLVPSVATV	A	P	2.12	2009
I	Human	MHCA0286	3GSR	A*0201	9	HCMV pp65 fragment 495-503, variant M5V	NLVPPVATV	A	P	1.95	2009
I	Human	MHCA0285	3GSU	A*0201	9	HCMV pp65 fragment 495-503, variant M5T	NLVPTVATV	A	P	1.80	2009
I	Human	MHCA0284	3GSV	A*0201	9	HCMV pp65 fragment 495-503, variant M5Q	NLVPPQVATV	A	P	1.90	2009
I	Human	MHCA0283	3GSW	A*0201	9	HCMV pp65 fragment 495-503, variant T8A	NLVPMVAAV	A	P	1.81	2009
I	Human	MHCA0282	3GSX	A*0201	9	HCMV pp65 fragment 495-503, variant T8V	NLVPMVAVV	A	P	2.10	2009
I	Human	MHCA0299	3H7B	A*0201	9	Tellp peptide	MLWGYLQYV	A	C	1.88	2010
I	Human	MHCA0298	3H9H	A*0201	9	Tellp peptide	MLWGYLQYV	A	C	2.00	2010

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0344	3H9S	A*0201	9	Tellp peptide	MLWGYLEQYV	A	C	2.70	2010
I	Human	MHCA0278	3HAE	A*0201	9	NYESO-1 Peptide	SLLMWITQV	A	C	2.90	2009
I	Human	MHCA025	3HPJ	A*0201	9	Wilms tumor antigen WT126 peptide	RMFPNAPYL	A	C	2.00	2010
I	Human	MHCA0324	3I6G	A*0201	9	SARS-CoV membrane glycoprotein	GLMWLSYFV	A	C	2.20	2010
I	Human	MHCA0297	3IXA	A*0201	9	Tax peptide	LLFGYPVYV	A	C	2.10	2010
I	Human	MHCA0303	3KLA	A*0201	9	NYESO-1 peptide analogue	SLLMWITQL	A	C	1.65	2010
I	Human	MHCA0349	3MYJ	A*0201	9	Wilms tumor protein	YMFPNAPYL	A	C	1.89	2010
I	Human	MHCA0012	1HHH	A*0201	10	HBV nucleocapsid	FLPSDFFPSV	A	C	3.00	1993
I	Human	MHCA0064	1I4F	A*0201	10	MAGE-4 Antigen	GVYDGREHTV	A	C	1.40	2001
I	Human	MHCA0027	1JF1	A*0201	10	Mart-1	ELAGIGILTV	A	C	1.85	2001
I	Human	MHCA0011	2CLR	A*0201	10	Synthetic	MLLSVPLLIG	A	C	2.00	1998
I	Human	MHCA0164	2GT9	A*0201	10	Melan-A/MART-1(26-35) peptide	EAAIGILTV	A	C	1.75	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0246	3BH8	A*0201	10	Decameric peptide from lymphocyte-specific protein 1	RQASIELPSM	A	C	1.65	2008
I	Human	MHCA0247	3BH9	A*0201	10	Decameric peptide from protein POF1B	RTYSGPMNKV	A	C	1.70	2008
I	Human	MHCA0245	3BHB	A*0201	10	NEDD4-binding protein 2	KMDSFLDMQL	A	C	2.20	2008
I	Human	MHCA0267	3FQN	A*0201	10	Peptide 30-39 from β -catenin	YLDSGIHSGA	A	C	1.65	2009
I	Human	MHCA0266	3FQR	A*0201	10	Phospho-peptide 30-39 from β -catenin	YLDSGIHSGA	A	C	1.70	2009
I	Human	MHCA0280	3GIV	A*0201	10	HIV-1 peptide	SLFNTVATLY	A	C	2.00	2009
I	Human	MHCA0340	3HG1	A*0201	10	Cancer/MART-1	ELAGIGILTV	A	C	3.00	2009
I	Human	MHCA0323	3I6K	A*0201	10	SARS-CoV membrane glycoprotein	TLACFVLAAN	A	C	2.80	2010
I	Human	MHCA0321	3MGO	A*0201	10	10-meric peptide from hemagglutinin	RLYQNPTTYI	A	C	2.30	2010

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0320	3MGT	A*0201	10	10-meric peptide from hemagglutinin	KLYQNPTTYI	A	C	2.20	2010
I	Human	MHCA0143	1Q94	A*1101	9	POL polyprotein	AIFQSSMTK	A	C	2.40	2004
I	Human	MHCA0145	1X7Q	A*1101	9	SARS nucleocapsid peptide	KTFPPTEPK	A	C	1.45	2005
I	Human	MHCA0144	1QVO	A*1101	10	Negative factor	QVPLRPMTYK	A	C	2.22	2004
I	Human	MHCA0159	2HN7	A*1101	10	Synthetic DNA polymerase peptide homologue	AIMPARFYPK	A	C	1.60	2006
I	Human	MHCA0185	2BCK	A*2402	9	Telomerase reverse transcriptase	VYGFVRACL	A	C	2.80	2006
I	Human	MHCA0332	3I6L	A*2402	9	Epitope of SARS-CoV nucleoprotein peptide	QFKDNVILL	D	F	2.40	2010
I	Human	MHCA0013	1TMC	A*6801	10	Synthetic	EVAPPEYHRK	A	C	2.30	1995
I	Human	MHCA0014	1AGB	B*0801	8	HIV-1 gag	GGRKKYKL	A	C	2.20	1997
I	Human	MHCA0015	1AGC	B*0801	8	HIV-1 gag	GGKKKYQL	A	C	2.10	1997
I	Human	MHCA0016	1AGD	B*0801	8	HIV-1 gag	GGKKKYKL	A	C	2.05	1997
I	Human	MHCA0017	1AGE	B*0801	8	HIV-1 gag	GGRKKYKL	A	C	2.30	1997

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0018	1AGF	B*0801	8	HIV-1 gag	GGKKRYKL	A	C	2.20	1997
I	Human	MHCA0095	1M05	B*0801	9	Ebna-3 nuclear protein	FLRGRAYGL	A	E	1.90	2003
I	Human	MHCA0099	1MI5	B*0801	9	Epstein Barr virus peptide	FLRGRAYGL	A	C	2.50	2003
I	Human	MHCA0335	3FFC	B*0801	9	FLRGRAYGL peptide from an EBV protein	FLRGRAYGL	A	C	2.80	2009
I	Human	MHCA0260	3BVN	B*1402	9	Latent membrane protein 2 (LMP2) of EBV	RRRWRRLTV	A	C	2.55	2009
I	Human	MHCA0259	3BXN	B*1402	9	Cathepsin A signal sequence octapeptide	IRAAPPLF	A	C	1.86	2009
I	Human	MHCA0140	1XR8	B*1501	9	Ebna-3 nuclear protein	LEKARGSTY	A	C	2.30	2005
I	Human	MHCA0141	1XR9	B*1501	9	Ubiquitin-conjugating enzyme E2 E1	ILGPPGSVY	A	C	1.79	2005
I	Human	MHCA0206	3C9N	B*1501	9	SARS corona virus derived peptide	VQQESSFVM	A	C	1.87	2008
I	Human	MHCA0019	1HSA	B*2705	9	N.A.	ARAAAAAAA	A	C	2.10	1992
I	Human	MHCA0080	1JGE	B*2705	9	Peptide M9	GRFAAAIAK	A	C	2.10	2002

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0107	1OF2	B*2705	9	Vasoactive intestinal polypeptide receptor	RRKWRRWHL	A	C	2.20	2004
I	Human	MHCA0109	1OGT	B*2705	9	Vasoactive intestinal polypeptide receptor 1	RRKWRRWHL	A	C	1.47	2004
I	Human	MHCA0154	1UXS	B*2705	9	EBV gene terminal protein LMP2	RRRWRLTV	A	C	1.55	2004
I	Human	MHCA0133	1W0V	B*2705	9	Butyrate response factor 2	RRLPIFSRL	A	C	2.27	2005
I	Human	MHCA0187	2A83	B*2705	9	Glucagon receptor (GR) peptide	RRRWHRWRL	A	C	1.40	2005
I	Human	MHCA0188	2BSR	B*2705	9	Epstein-Barr nuclear antigen-6	RRIYDLIEL	A	C	2.30	2005
I	Human	MHCA0186	2BST	B*2705	9	Influenza nucleoprotein	SRYWAIATR	A	C	2.10	2005
I	Human	MHCA0242	3B6S	B*2705	9	Vasoactive intestinal polypeptide receptor 1	RRKWRRWHL	A	C	1.80	2008

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0253	3BP4	B*2705	9	Cathepsin A signal sequence peptide, pCatA	IRAAPPPLF	A	C	1.85	2008
I	Human	MHCA0276	3DTX	B*2705	9	Intestinal polypeptide receptor	RRKWRRRWHL	A	C	2.10	2009
I	Human	MHCA0189	2BSS	B*2705	10	HIV peptide	KRWIILGLNK	A	C	2.00	2005
I	Human	MHCA0082	1K5N	B*2709	9	Nonameric model peptide M9	GRFAAAIAK	A	C	1.09	2002
I	Human	MHCA0132	1UXW	B*2709	9	Gene terminal protein (membrane protein Lmp-2A/Lmp-2B)	RRRWRRRLTV	A	C	1.71	2004
I	Human	MHCA0134	1W0W	B*2709	9	Butyrate response factor 2	RRLPIFSRL	A	C	2.10	2005
I	Human	MHCA0243	3B3I	B*2709	9	Vasoactive intestinal polypeptide receptor 1	RRKWRRRWHL	A	C	1.86	2008
I	Human	MHCA0254	3BP7	B*2709	9	Cathepsin A signal sequence peptide, pCatA	IRAAPPPLF	A	C	1.80	2008

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0271	3CZF	B*2709	9	Glucagon receptor peptide	RRRWHRWRL	A	C	1.20	2009
I	Human	MHCA0277	3D18	B*2709	9	Latent membrane protein 2 (LMP2) of EBV	RRRWRRRLTL	A	C	1.74	2009
I	Human	MHCA0079	1JGD	B*2709	10	Peptide S10R	RRLLRGHNQY	A	C	1.90	2003
I	Human	MHCA0279	3HCV	B*2709	9	Double citrullinated vasoactive intestinal polypeptide receptor	RRKWRRRWHL	A	C	1.95	2009
I	Human	MHCA0020	1A1N	B*3501	8	HIV-1 Nef	VPLRPMTY	A	C	2.00	1998
I	Human	MHCA0022	1A9B	B*3501	9	EBNA-3C	LPPLDITPY	A	C	3.20	1998
I	Human	MHCA0021	1A9E	B*3501	9	EBV-Ebna3c	LPPLDITPY	A	C	2.50	1998
I	Human	MHCA0155	1CG9	B*3501	9	EBV EBNA-3C peptide	LPPLDITPY	A	C	2.70	2003
I	Human	MHCA0111	1QEW	B*3501	9	Melanoma-associated Antigen 3	FLWGPRLV	A	C	2.20	2003
I	Human	MHCA0192	2CIK	B*3501	9	cytochrome p450 peptide	KPIVV LHGY	A	C	1.75	2006

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0191	2H6P	B*3501	9	9-mer peptide from cytochrome P450	KPIVVVLHGY	A	C	1.90	2006
I	Human	MHCA0331	3LKN	B*3501	9	NP418 epitope from 1918 influenza strain	LPFERATIM	A	C	2.00	2010
I	Human	MHCA0330	3LKO	B*3501	9	NP418 epitope from 1934 influenza strain	LPFDRTTIM	A	C	1.80	2010
I	Human	MHCA0329	3LKP	B*3501	9	NP418 epitope from 1972 influenza strain	LPFDKSTIM	A	C	1.80	2010
I	Human	MHCA0328	3LKQ	B*3501	9	NP418 epitope from 1977 influenza strain	LPFDKTTIM	A	C	1.80	2010
I	Human	MHCA0327	3LKR	B*3501	9	NP418 epitope from 2009 swine-influenza strain	LPFERATVM	A	C	2.00	2010
I	Human	MHCA0326	3LKS	B*3501	9	NP418 epitope from 1980 influenza strain	LPFEKSTVM	A	C	1.90	2010
I	Human	MHCA0190	2AXG	B*3501	10	10-mer peptide from BZLF1 trans-activator protein	APQPAPENAY	A	C	2.00	2005

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0157	1ZSD	B*3501	11	BZLF1 trans-activator protein	EPLPQGQLTAY	A	C	1.70	2005
I	Human	MHCA0193	2NX5	B*3501	11	EBV peptide	EPLPQGQLTAY	A	C	2.70	2007
I	Human	MHCA0347	3MV7	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFEY	A	C	2.00	2010
I	Human	MHCA0346	3MV8	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFEY	A	C	2.10	2010
I	Human	MHCA0345	3MV9	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFEY	A	C	2.70	2010
I	Human	MHCA0156	1ZHK	B*3501	13	EBV peptide	LPEPLPQGQLTAY	A	C	1.60	2005
I	Human	MHCA0139	1XH3	B*3501	14	Aa 4-17 of Alternative Reading Frame of M-Csf	LPAVVGLSPGEQE _Y	A	C	1.48	2004
I	Human	MHCA0208	3BWA	B*3508	8	FPT peptide	FPTKDVAL	A	C	1.30	2008
I	Human	MHCA0194	2NW3	B*3508	11	EBV peptide	EPLPQGQLTAY	A	C	1.70	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0196	2AXF	B*3508	10	10-mer peptide from BZLF1 trans-activator protein	APQPAPENAY	A	C	1.80	2005
I	Human	MHCA0350	2FYY	B*3508	11	Epstein-Barr nuclear antigen 1	HPVGEADYFEY	A	C	1.50	2006
I	Human	MHCA0197	2FZ3	B*3508	11	11-mer peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFEY	A	C	1.90	2006
I	Human	MHCA0207	3BW9	B*3508	12	CPS peptide	CPSQEPMSIYVY	A	C	1.75	2008
I	Human	MHCA0158	1ZHL	B*3508	13	EBV peptide	LPEPLPQQQLTAY	A	C	1.50	2005
I	Human	MHCA0195	2AK4	B*3508	13	EBV peptide	LPEPLPQQQLTAY	A	C	2.50	2005
I	Human	MHCA0322	3KWW	B*3508	13	Trans-activator protein BZLF1	LPEPLPQQQLTAY	A	C	2.18	2010
I	Human	MHCA0348	3KXF	B*3508	13	Trans-activator protein BZLF1	LPEPLPQQQLTAY	A	Q	3.10	2010
I	Human	MHCA0096	1M6O	B*4402	9	Hla Dpa*0201 Peptide	EEFGRAFSF	A	C	1.60	2003
I	Human	MHCA0294	3KPM	B*4402	9	Mimotope peptide	EEYLKAWTF	A	C	1.60	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0295	3KPL	B*4402	9	Self peptide from the ATP binding cassette protein ABCD3	EEYLQAFY	A	C	1.96	2009
I	Human	MHCA0319	3L3D	B*4402	9	Self-peptide derived from DPA*0201	EEAGRAFSF	A	C	1.80	2010
I	Human	MHCA0318	3L3G	B*4402	9	Self-peptide derived from DPA*0201	EEFGAAFSF	A	C	2.10	2010
I	Human	MHCA0316	3L3I	B*4402	9	Self-peptide derived from DPA*0201	EEFGRAASF	A	C	1.70	2010
I	Human	MHCA0315	3L3J	B*4402	9	Self-peptide derived from DPA*0201	EEAGAAFSF	A	C	2.40	2010
I	Human	MHCA0314	3L3K	B*4402	9	Self-peptide derived from DPA*0201	EEFGAAASF	A	C	2.60	2010
I	Human	MHCA0258	3DX6	B*4402	10	EBV decapeptide epitope	EENLLDFVRF	A	C	1.70	2009
I	Human	MHCA0101	1N2R	B*4403	9	Hla Dpa*0201 Peptide	EEFGRAFSF	A	C	1.70	2004
I	Human	MHCA0126	1SYS	B*4403	9	Sorting Nexin 5	EEPTVIKKY	A	C	2.40	2004

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0293	3KPN	B*4403	9	Self peptide from the ATP binding cassette protein ABCD3	EEYLQAFTY	A	C	2.00	2009
I	Human	MHCA0292	3KPO	B*4403	9	Mimotope peptide	EEYLKAWTF	A	C	2.30	2009
I	Human	MHCA0257	3DX7	B*4403	10	EBV decapeptide epitope	EENLLDFVRF	A	C	1.60	2009
I	Human	MHCA0127	1SYV	B*4405	9	Self Ligand	EEFGRAFSF	A	C	1.70	2004
I	Human	MHCA0336	3DXA	B*4405	9	EBV decapeptide epitope	EENLLDFVRF	A	C	3.50	2009
I	Human	MHCA0291	3KPP	B*4405	9	Self peptide from the ATP binding cassette protein ABCD3	EEYLQAFTY	A	C	1.90	2009
I	Human	MHCA0290	3KPQ	B*4405	9	Mimotope peptide	EEYLKAWTF	A	C	1.84	2009
I	Human	MHCA0343	3KPR	B*4405	9	Mimotope peptide	EEYLKAWTF	A	C	2.60	2009
I	Human	MHCA0342	3KPS	B*4405	9	Self peptide from the ABCD3 protein	EEYLQAFTY	A	C	2.70	2009
I	Human	MHCA0256	3DX8	B*4405	10	EBV decapeptide epitope	EENLLDFVRF	A	C	2.10	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0043	1E28	B*5101	8	HIV-1 Km2	TAFTIPSI	A	C	3.00	2000
I	Human	MHCA0042	1E27	B*5101	9	HIV-1 Kml	LPPVVAKEI	A	C	2.20	2000
I	Human	MHCA0023	1A1M	B*5301	9	HIV-2 gag	TPYDINQML	A	C	2.30	1998
I	Human	MHCA0024	1A1O	B*5301	9	HIV-1 Nef	KPIVQYDNF	A	C	2.30	1998
I	Human	MHCA0241	2RFX	B*5701	9	Self peptide	LSSPVTKSF	A	C	2.50	2008
I	Human	MHCA0199	2BVP	B*5703	9	Gag peptide HIV-P24	ISPRTLDAW	A	C	1.35	2005
I	Human	MHCA0200	2BVQ	B*5703	9	Gag peptide HIV-P24	KAFSPEVIP	A	C	2.00	2005
I	Human	MHCA0198	2BVO	B*5703	11	Gag peptide	KAFSPEVIPMF	A	C	1.65	2005
I	Human	MHCA0201	2HJK	B*5703	11	Gag peptide HIV-1	KGFNPEVIPMF	A	C	1.85	2006
I	Human	MHCA0202	2HJL	B*5703	11	Gag peptide HIV-1	KAFNPEIIPMF	A	C	1.50	2006
I	Human	MHCA0054	1EFX	Cw*0304	9	Importin a 2	GAVDPLLAL	A	C	3.00	2000
I	Human	MHCA0113	1QQD	Cw*0401	9	Synthetic	QYDDAVYKL	A	C	2.70	1999
I	Human	MHCA0053	1IM9	Cw*0401	9	Synthetic	QYDDAVYKL	A	C	2.80	2001
I	Human	MHCA0098	1MHE	E*0101	9	Synthetic	VMAPRTVLL	A	P	2.85	1999
I	Human	MHCA0203	2ESV	E*0101	9	Peptide from CMV gpUL40	VMAPRTLIL	A	P	2.60	2006

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0237	3BZE	E*0101	9	Leader peptide of HLA class I histocompatibility antigen, Cw-7 α chain	VMAPRALLL	A	P	2.50	2008
I	Human	MHCA0112	3BZF	E*0101	9	Leader peptide of HLA class I histocompatibility antigen, Cw-7 α chain	VMAPRALLL	A	P	2.50	2008
I	Human	MHCA0238	3CII	E*0101	9	UNP residues 3-11	VMAPRTLFL	A	C	4.41	2008
I	Human	MHCA0087	1KPR	E*0103	9	Synthetic	VMAPRTVLL	A	P	2.80	2003
I	Human	MHCA0090	1KTL	E*0103	9	Peptide B27	VTAPRTLTL	A	P	3.10	2003
I	Human	MHCA0209	3CDG	E*0103	9	HLA class I leader peptide	VMAPRTLFL	A	P	3.40	2008
I	Human	MHCA0142	1YDP	G*0101	9	Histone 2A	RIIPRHLQL	A	P	1.90	2005
I	Human	MHCA0205	2D31	G*0101	9	9-mer peptide from histone H2A	RIIPRHLQL	A	C	3.20	2006
I	Human	MHCA0204	2DYP	G*0101	9	9-mer peptide from histone H2A.x	RIIPRHLQL	A	C	2.50	2006

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Human	MHCA0305	3KYN	G*0101	9	Self peptide	KGPPAALTL	A	P	2.40	2010
I	Human	MHCA0304	3KYO	G*0101	9	Self peptide	KLPAQFYIL	A	P	1.70	2010
I	Chicken	MHCA0211	3BEW	B*2101	10	10-mer from Tubulin β -6 chain	REVDEQLLSV	A	C	2.60	2008
I	Chicken	MHCA0210	3BEV	B*2101	11	Hemoglobin subunit α -A	GHAEYGAETL	A	C	2.10	2008
I	Monkey	MHCA0212	1ZVS	Mamu-A*01	8	Tat-T18 peptide	TTPEANL	A	C	2.80	2006
I	Murine	MHCA0031	1BZ9	H2-Db	9	Peptide P1027	FAPGVFPYM	A	P	2.80	1998
I	Murine	MHCA0032	1CE6	H2-Db	9	SV nucleoprotein	FAPGNYPAL	A	C	2.90	1999
I	Murine	MHCA0073	1FFN	H2-Db	9	Gp33 Peptide	KAVYNFATM	A	C	2.70	2002
I	Murine	MHCA0074	1FFO	H2-Db	9	Gp33 Peptide	AAVYNFATM	A	C	2.65	2002
I	Murine	MHCA0075	1FFP	H2-Db	9	Gp33 Peptide	SAVYNFATM	A	C	2.60	2002
I	Murine	MHCA0044	1FG2	H2-Db	9	Gp33 Peptide	KAVYNFATC	A	C	2.75	2000
I	Murine	MHCA0077	1HOC	H2-Db	9	Influenza virus nucleoprotein	ASNENMETM	A	C	2.40	1994
I	Murine	MHCA0078	1INQ	H2-Db	9	H13A	SSVVGVWYL	A	C	2.20	2002
I	Murine	MHCA0063	1JPG	H2-Db	9	LCMV peptide	FQPQNGQFI	A	C	2.20	2001

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0081	1JUF	H2-Db	9	H13B	SSVIGVWYL	A	C	2.00	2002
I	Murine	MHCA0103	1N5A	H2-Db	9	Gp33 Derived From Lymphocytic Choriomeningitis Virus	KAVYNFATM	A	C	2.85	2003
I	Murine	MHCA0033	1QLF	H2-Db	9	SV-nucleoprotein	FAPSNYPAL	A	C	2.65	1999
I	Murine	MHCA0119	1S7U	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVYNFATM	A	C	2.20	2004
I	Murine	MHCA0120	1S7V	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVYNLATM	A	C	2.20	2004
I	Murine	MHCA0121	1S7W	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KALYNFATM	A	C	2.40	2004
I	Murine	MHCA0122	1S7X	H2-Db	9	Lcmv- Derived Gp33 Index Peptide	KAVFNFATM	A	C	2.41	2004
I	Murine	MHCA0224	2F74	H2-Db	9	LCMV-derived immunodominant peptide gp33	KAVYNFATM	A	C	2.70	2006
I	Murine	MHCA0240	2ZOK	H2-Db	9	Spike glycoprotein	ASLWNGPHL	A	I	2.10	2008

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0239	2ZOL	H2-Db	9	Spike glycoprotein	ASLSNGPHL	A	E	2.70	2008
I	Murine	MHCA0236	3BUY	H2-Db	9	Peptide of PB1-F2	LSLRNPILV	A	C	2.60	2008
I	Murine	MHCA0268	3CC5	H2-Db	9	Nonameric peptide from Melanocyte protein Pmel 17	KVPRNQDWL	A	C	1.91	2009
I	Murine	MHCA0269	3CCH	H2-Db	9	Nonameric peptide murine gp100	EGSRNQDWL	A	C	2.60	2009
I	Murine	MHCA0270	3CH1	H2-Db	9	Nonameric peptide chimeric gp100	EGPRNQDWL	A	C	2.30	2009
I	Murine	MHCA0296	3FTG	H2-Db	9	NP366-N3A variant peptide from influenza virus	ASAEENMETM	A	C	2.60	2009
I	Murine	MHCA0036	1DDH	H2-Db	10	HIV-1 gp120	RGPGRAFVTI	A	P	3.10	1999
I	Murine	MHCA0102	1N3N	H2-Db	10	Mycobacterial Hsp60 Decameric Epitope	SNLQNAASIA	A	I	3.00	2003
I	Murine	MHCA0136	1WBX	H2-Db	10	Influenza A Peptide	SQLKNNAKEI	A	C	1.90	2005
I	Murine	MHCA0137	1WBY	H2-Db	10	Influenza A Peptide	SSLENFRA YV	A	C	2.30	2005

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0213	1YN6	H2-Db	10	10-mer peptide from RNA-directed RNA polymerase subunit P2	SSLENFRAYV	A	C	2.20	2005
I	Murine	MHCA0214	1YN7	H2-Db	10	10-mer peptide from RNA-directed RNA polymerase subunit P2	SSLENFAAYV	A	C	2.20	2005
I	Murine	MHCA0318	3L3H	H2-Db	10	Influenza A acid polymerase	SSLENARAYV	A	C	2.70	2010
I	Murine	MHCA0062	1JPF	H2-Db	11	LCMV peptide	SGVENPGGYCL	A	C	2.18	2001
I	Murine	MHCA0352	3E6F	H2-Dd	9	Envelope glycoprotein	IGPGRAFYA	A	P	2.41	2009
I	Murine	MHCA0289	3E6H	H2-Dd	9	HIV (BaL) envelope glycoprotein 120	IGPGRAFYTI	A	P	2.10	2009
I	Murine	MHCA0034	1BII	H2-Dd	10	HIV-1 P18-100	RGPGRAFVTI	A	P	2.40	1998
I	Murine	MHCA0007	1QO3	H2-Dd	10	HIV	RGPGRAFVTI	A	P	2.30	2000
I	Murine	MHCA0339	3DMM	H2-Dd	10	Synthetic Peptide	RGPGRAFVTI	A	P	2.60	2009
I	Murine	MHCA0281	3ECB	H2-Dd	10	Peptide P18-I10 from HIV gp160	RGPGRAFVTI	A	C	1.70	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0069	1BQH	H2-Kb	8	Vesicular stomatitis virus	RGYVYQGL	A	C	2.80	1998
I	Murine	MHCA0045	1FO0	H2-Kb	8	Natural peptide	INFDFNTI	H	P	2.50	2000
I	Murine	MHCA0047	1FZJ	H2-Kb	8	VSV nucleoprotein	RGYVYQGL	A	P	1.90	2001
I	Murine	MHCA0048	1FZM	H2-Kb	8	VSV nucleoprotein	RGYVYQGL	A	P	1.80	2001
I	Murine	MHCA0050	1G6R	H2-Kb	8	Syir protein	SIYRYYGL	H	P	2.80	2000
I	Murine	MHCA0068	1G7Q	H2-Kb	8	mucin1,transmembrane	SAPDTRPA	A	P	1.60	2002
I	Murine	MHCA0084	1KBG	H2-Kb	8	VSV nucleoprotein	RGYVYXGL	H	P	2.20	1999
I	Murine	MHCA0065	1KJ2	H2-Kb	8	Naturally processed	KVITFIDL	H	P	2.71	2002
I	Murine	MHCA0066	1KJ3	H2-Kb	8	Naturally processed	KVITFIDL	H	P	2.30	2002
I	Murine	MHCA0088	1KPU	H2-Kb	8	Vsv8, Nucleocapsid Fragment	RGYVYQGL	A	P	1.50	2003
I	Murine	MHCA0091	1LEG	H2-Kb	8	Dev 8	EQYKFYSV	A	P	1.75	2002
I	Murine	MHCA0092	1LEK	H2-Kb	8	Dev 8	EQYKFYSV	A	P	2.15	2002
I	Murine	MHCA0093	1LK2	H2-Kb	8	Synthetic	GNYSFYAL	A	P	1.35	2003
I	Murine	MHCA0100	1MWA	H2-Kb	8	Dev 8	EQYKFYSV	H	P	2.40	2002

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0104	1N59	H2-Kb	8	Gp33 Derived From Lcmv	AVYNFATM	A	P	2.95	2003
I	Murine	MHCA0105	1NAM	H2-Kb	8	Vesicular Stomatitis Virus Nucleoprotein Fragment	RGYVYQGL	H	P	2.70	2003
I	Murine	MHCA0106	1NAN	H2-Kb	8	Riken Cdna 2410004N11	INFDNTI	H	M	2.30	2003
I	Murine	MHCA0025	1OSZ	H2-Kb	8	VSV nucleoprotein	RGYLYQGL	A	C	2.10	1999
I	Murine	MHCA0351	1P1Z	H2-Kb	8	Ovalbumin peptide	SIINFEKL	A	P	3.26	2003
I	Murine	MHCA0218	1P4L	H2-Kb	8	Ovalbumin peptide	SIINFEKL	A	P	2.90	2003
I	Murine	MHCA0215	1RJY	H2-Kb	8	Glycoprotein B peptide	SSIEFARL	A	P	1.90	2004
I	Murine	MHCA0216	1RJZ	H2-Kb	8	Glycoprotein B peptide	SEIEFARL	A	P	2.60	2004
I	Murine	MHCA0219	1RK0	H2-Kb	8	Glycoprotein B peptide	SSIEFARL	A	P	2.61	2004
I	Murine	MHCA0220	1RK1	H2-Kb	8	Glycoprotein B peptide	SEIEFARL	A	P	2.10	2004

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0128	1T0M	H2-Kb	8	Glycoprotein B	SSIEFARL	A	P	2.00	2004
I	Murine	MHCA0129	1T0N	H2-Kb	8	Glycoprotein B	SSIEFARL	A	P	1.80	2004
I	Murine	MHCA0028	1VAC	H2-Kb	8	Ovalbumin	SIINFEKL	A	P	2.50	1996
I	Murine	MHCA0052	2CKB	H2-Kb	8	Dev 8	EQYKFYSV	H	P	3.20	1998
I	Murine	MHCA0225	2CLV	H2-Kb	8	PBM8 peptide from RBM5 protein	SQYYYNLS	A	C	1.90	2006
I	Murine	MHCA0226	2CLZ	H2-Kb	8	PBM1 peptide from RBM5 protein	INFDFNTI	A	C	1.90	2006
I	Murine	MHCA0232	2FO4	H2-Kb	8	8-mer peptide from Mucin-1	SAPDFRPL	A	P	2.70	2006
I	Murine	MHCA0076	2MHA	H2-Kb	8	Vesicular stomatitis virus	RGYVYQGL	A	E	2.80	1993
I	Murine	MHCA0227	2OL3	H2-Kb	8	Naturally processed octapeptide PBM8	SQYYYNLS	H	P	2.90	2007
I	Murine	MHCA0228	2QRI	H2-Kb	8	Ovalbumin-derived peptide (linked to MHC)	SIINFEKL	A	A	2.00	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0230	2QRS	H2-Kb	8	Ovalbumin-derived peptide (linked to MHC)	SIINFEKL	A	A	2.00	2007
I	Murine	MHCA0229	2QRT	H2-Kb	8	Ovalbumin-derived peptide (linked to MHC)	SIINFEKL	A	A	1.80	2007
I	Murine	MHCA0030	2VAA	H2-Kb	8	VSV nucleoprotein	RGYVYQGL	A	P	2.30	1996
I	Murine	MHCA0250	2ZSV	H2-Kb	8	8-mer peptide from spike glycoprotein	RAQIFANI	A	E	1.80	2008
I	Murine	MHCA0249	2ZSW	H2-Kb	8	8-mer peptide from spike glycoprotein	RAYIFANI	A	M	2.80	2008
I	Murine	MHCA0235	3C8K	H2-Kb	8	Ovalbumin peptide	SIINFEKL	A	P	2.90	2008
I	Murine	MHCA0244	3CVH	H2-Kb	8	Ovalbumin	SIINFEKL	A	C	2.90	2008
I	Murine	MHCA0049	1FZK	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	1.70	2001
I	Murine	MHCA0046	1FZO	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	1.80	2001
I	Murine	MHCA0067	1G7P	H2-Kb	9	α -glucosidase p1	SRDHSRTPM	A	P	1.50	2002
I	Murine	MHCA0089	1KPV	H2-Kb	9	Sev9, Nucleoprotein Fragment	FAPGNYPAL	A	P	1.71	2003

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0115	1S7Q	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVYNFATM	A	C	1.99	2004
I	Murine	MHCA0116	1S7R	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVYNLATM	A	C	2.95	2004
I	Murine	MHCA0117	1S7S	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KALYNFATM	A	C	1.99	2004
I	Murine	MHCA0118	1S7T	H2-Kb	9	Lcmv- Derived Gp33 Index Peptide	KAVNFATM	A	C	2.30	2004
I	Murine	MHCA0029	1VAD	H2-Kb	9	Yeast α glucosid	SRDHSRTPM	A	P	2.50	1996
I	Murine	MHCA0138	1WBZ	H2-Kb	9	Influenza A Peptide	SSYRRPVGI	A	P	2.00	2005
I	Murine	MHCA0221	1ZHB	H2-Kb	9	9-mer peptide from Dopamine β -monooxygenase	KALYNYAPI	A	C	2.70	2005
I	Murine	MHCA0026	2VAB	H2-Kb	9	SV nucleoprotein	FAPGNYPAL	A	P	2.50	1996
I	Murine	MHCA0251	3CPL	H2-Kb	9	NP366 peptide	ASNENAETM	A	E	2.50	2008
I	Murine	MHCA0217	1VGK	H2-Kd	9	HBV immunodominant peptide	SYVNTNMGL	A	C	2.06	2005

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0231	2FWO	H2-Kd	9	Peptide derived from Influenza nucleoprotein	TYQRTALV	A	P	2.60	2006
I	Murine	MHCA0222	1ZT1	H2-Kk	8	Influenza virus peptide	FEANGNLI	A	P	2.50	2005
I	Murine	MHCA0223	1ZT7	H2-Kk	9	SV40 peptide	SEFLLEKRI	A	P	3.00	2005
I	Murine	MHCA0302	3FOL	H2-Kwm7	8	Self peptide	VNDIFERI	A	P	2.50	2010
I	Murine	MHCA0301	3FOM	H2-Kwm7	8	Self peptide	IQQSIERL	A	P	2.10	2010
I	Murine	MHCA0300	3FON	H2-Kwm7	8	Self peptide	VNDIFEAI	A	E	2.03	2010
I	Murine	MHCA0037	1LD9	H2-Ld	9	Synthetic	YPNVNIHNF	A	C	2.40	1998
I	Murine	MHCA0035	1LDP	H2-Ld	9	Natural peptide	APAAAAAAM	H	P	3.10	1998
I	Murine	MHCA0233	2E7L	H2-Ld	9	Synthetic peptide	QLSPFPFDL	E	P	2.50	2007
I	Murine	MHCA0234	2OI9	H2-Ld	9	Synthetic peptide	QLSPFPFDL	A	Q	2.35	2007
I	Murine	MHCA0252	3ERY	H2-Ld	9	2-oxoglutarate dehydrogenase E1	QLSPFPFDL	A	P	1.95	2008
I	Murine	MHCA0334	3E2H	H2-Ld	9	QL9 peptide	QLSPFPFDL	A	Q	3.80	2008
I	Murine	MHCA0333	3E3Q	H2-Ld	9	QL9 Peptide	QLSPFPFDL	A	G	2.95	2008
I	Murine	MHCA0097	1MHC	H2-M3	9	Rat Nadh Dehydrogenase	MYFINILTL	A	C	2.10	1996

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
I	Murine	MHCA0083	1K8D	H2-Qa-2	9	60S Ribosomal Protein	ILMEHIHKL	A	P	2.30	2001
I	Rat	MHCA0085	1KJM	RT1.Aa	9	B6 Peptide	AQFSASASR	A	P	2.35	2002
I	Rat	MHCA0041	1ED3	RT1.Aa	13	Rat atapase	ILFPSSERLISNR	A	C	2.55	2000
I	Rat	MHCA0086	1KJV	RT1.A1c	9	Peptide Npr	NPRAMQALL	A	P	1.48	2002
II	Human	MHCB0041	1S9V	DQB1*0201	11	A-I Gliadin	LQFPQPPELPY	A, B	C	2.22	2004
II	Human	MHCB0056	2NNA	DQB1*0302	13	Gluten peptide	SGEGSFQPSQENP	A, B	C	2.10	2007
II	Human	MHCB0018	1JK8	DQB1*0302	14	Insulin B Peptide	LVEALYLVCGERG G	A, B	C	2.40	2001
II	Human	MHCB0042	1UVQ	DQB1*0602	20	Hypocretin Peptide	MNLPSTKVSAAV GGGGS LV	A, B	C	1.80	2004
II	Human	MHCB0047	1ZGL	DRA*0101	14	Myelin basic protein	VHHFKNIVTPRTPG	A, B	C	2.80	2005
II	Human	MHCB0037	1PYW	DRB1*0101	9	Influenza Virus Hemagglutinin Related Peptide	FVKQNAXAL	A, B	C	2.10	2003

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Human	MHCB0002	1DLH	DRB1*0101	13	Influenza Virus Peptide	PKYVKQNTLKLAT	A, B	C	2.80	1994
II	Human	MHCB0009	1FYT	DRB1*0101	13	Hemagglutinin Ha1 Peptide Chain	PKYVKQNTLKLAT	A, B	C	2.60	2000
II	Human	MHCB0016	1HXY	DRB1*0101	13	Hemagglutinin	PKYVKQNTLKLAT	A, B	C	2.60	2001
II	Human	MHCB0028	1JWM	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.70	2003
II	Human	MHCB0029	1JWS	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.60	2003
II	Human	MHCB0030	1JWU	DRB1*0101	13	Ha Peptide	PKYVKQNTLKLAT	A, B	C	2.30	2003
II	Human	MHCB0020	1KG0	DRB1*0101	13	Hemagglutinin Ha Peptide	PKYVKQNTLKLAT	A, B	D	2.65	2002
II	Human	MHCB0034	1LO5	DRB1*0101	13	Hemagglutinin Peptide	PKYVKQNTLKLAT	A, B	C	3.20	2002
II	Human	MHCB0038	1R5I	DRB1*0101	13	Hemagglutinin Peptide	PKYVKQNTLKLAT	A, B	C	2.60	2004

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Human	MHCB0003	1SEB	DRB1*0101	13	Endogeneous Peptide	AAAAAAAAAAAAA A	A, B	C	2.70	1996
II	Human	MHCB0044	1SJH	DRB1*0101	13	Gag Polyprotein	PEVIPMFSALSEG	A, B	C	2.25	2004
II	Human	MHCB0045	1T5W	DRB1*0101	13	Fragment Of Regulatory Protein Mig1	AAYSDDQATPLLS	A, B	C	2.40	2004
II	Human	MHCB0046	1T5X	DRB1*0101	13	Fragment Of Regulatory Protein Mig1	AAYSDDQATPLLS	A, B	C	2.50	2004
II	Human	MHCB0049	2G9H	DRB1*0101	13	Haemagglutinin peptide	PKYVKQNTLKLAT	A, B	C	2.00	2006
II	Human	MHCB0050	2OJE	DRB1*0101	13	Haemagglutinin peptide	PKYVKQNTLKLAT	A, B	C	3.00	2007
II	Human	MHCB0001	1AQD	DRB1*0101	14	Endogeneous Peptide	GSDWRFLRGYHQY A	A, B	C	2.45	1998
II	Human	MHCB0054	2FSE	DRB1*0101	14	Collagen α -1(II)	AGFKGEQGPKEP G	A, B	E	3.10	2006

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Human	MHCB0005	1A6A	DRB1*0301	15	Clip	PVSKMRMATPLLM QA	A, B	C	2.75	1998
II	Human	MHCB0021	1KLG	DRB1*0101	15	Triosephosphate Isomerase Peptide	GELIGILNAAKVPA D	A, B	C	2.40	2002
II	Human	MHCB0022	1KLU	DRB1*0101	15	Triosephosphate Isomerase Peptide	GELIGTLNAAKV AD	A, B	C	1.93	2002
II	Human	MHCB0043	1SJE	DRB1*0101	15	Gag Polyprotein	PEVIPMFSALSEGA T	A, B	C	2.45	2004
II	Human	MHCB0051	2IAM	DRB1*0101	15	15-mer peptide from Triosephosphate isomerase	GELIGILNAAKVPA D	A, B	P	2.80	2007
II	Human	MHCB0052	2IAN	DRB1*0101	15	15-mer peptide from Triosephosphate isomerase	GELIGTLNAAKV AD	A, B	C	2.80	2007
II	Human	MHCB0024	1D5X	DRB1*0401	6	Dipeptide Mimetic Inhibitor	XXRXXX	A, B	D	2.45	2000
II	Human	MHCB0025	1D5Z	DRB1*0401	7	Peptidomimetic Inhibitor	XXRAXSX	A, B	D	2.00	2000

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Human	MHCB0023	1D5M	DRB1*0401	8	Peptide Inhibitor	XXRAMXSX	A, B	D	2.00	2000
II	Human	MHCB0026	1D6E	DRB1*0401	8	Peptidomimetic Inhibitor	XXRXMASX	A, B	D	2.45	2000
II	Human	MHCB0006	2SEB	DRB1*0401	12	Collagen II Peptide	AYMRADAAAAGGA	A, B	E	2.50	1998
II	Human	MHCB0019	1J8H	DRB1*0401	13	Hemagglutinin Ha1 Peptide Chain	PKYVKQNTLKLAT	A, B	C	2.40	2002
II	Human	MHCB0004	1BX2	DRB1*1501	14	Human Myelin Basic Protein	ENPVVHFFKNIVTP	A, B	C	2.60	1998
II	Human	MHCB0048	1YMM	DRB1*1501	14	MBP peptide	ENPVVHFFKNIVTP	A, B	C	3.50	2005
II	Human	MHCB0055	2Q6W	DRB3*0101	11	Integrin β -3	AWRSDEALPLG	A, B	C	2.25	2007
II	Human	MHCB0061	3C5J	DRB3*0301	13	Elongation factor 1- α ₂	QVIILNHPGQISA	A, B	C	1.80	2008
II	Human	MHCB0010	1HQR	DRB5*0101	10	Myelin Basic Protein	VHFFKNIVTP	A, B	C	3.20	2001

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Human	MHCB0027	1H15	DRB5*0101	14	Epstein Barr Virus (Ebv) DNA Polymerase	GGVYHFVKKHVVHES	A, B	C	3.10	2002
II	Human	MHCB0014	1FV1	DRB5*0101	20	Myelin Basic Protein	NPVVHFFKKNIVTPRTPPPSQ	A, B	C	1.90	2000
II	Human	MHCB0060	3L6F	HLA-DR1	15	Melanoma antigen recognized by T-cells 1	APPAYEKLSAEQSP	A, B	C	2.10	2010
II	Murine	MHCB0013	1ES0	I-A(G7)	14	Glutamic Acid Decarboxylase Peptide	YEIAPVFFVLLLEYVT	A, B	B	2.60	2000
II	Murine	MHCB0063	3CUP	I-A(G7)	15	GAD221-235 peptide	KKMREIIGWPGGS	A, B	P	3.09	2009
II	Murine	MHCB0062	3MBE	I-A(G7)	18	Polypeptide	GAMKRRHGLDNYRGYSLGN	A, B	P	2.89	2010
II	Murine	MHCB0035	1LNU	I-Ab	13	Eα3K Peptide	FEAQKAKANKAVD	A, B	B	2.50	2002
II	Murine	MHCB0036	1MUJ	I-Ab	15	Clip Peptide	PVSKMRMATPLLMQA	A, B	C	2.15	2003

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Murine	MHCB0007	1IAO	I-Ad	14	Ovalbumin Peptide	RGISQAVHAAHAEI	A, B	B	2.60	1998
II	Murine	MHCB0008	2IAD	I-Ad	14	Influenza Hemagglutinin Peptide	GHATQGVTAASSH E	A, B	B	2.40	1998
II	Murine	MHCB0015	1IAK	I-Ak	13	Hen Eggwhite Lysozyme Peptide	STDYGILQINSRW	A, B	P	1.90	1998
II	Murine	MHCB0011	1F3J	I-Ak	14	Gallus Gallus	AMKRHGLDNYRG YS	A, B	P	3.10	2000
II	Murine	MHCB0012	1D9K	I- Ak	16	Conalbumin Peptide	GNSHRGAIEWEGIE SG	C, D	P	3.20	1999
II	Murine	MHCB0017	1JL4	I-Ak	16	Ovotransferrin	GNSHRGAIEWEGIE SG	A, B	C	4.30	2001
II	Murine	MHCB0059	2PXY	I-Au	11	Myelin basic protein	RGGASQYRPSQ	C, D	P	2.23	2007
II	Murine	MHCB0058	2Z31	I-Au	11	Myelin basic protein	RGGASQYRPSQ	C, D	P	2.70	2007
II	Murine	MHCB0031	1K2D	I-Au	12	Myelin Basic Protein Peptide	SRGGASQYRPSQ	A, B	P	2.20	2003
II	Murine	MHCB0057	1U3H	I-Au	12	Myelin basic protein	SRGGASQYRPSQ	C, D	P	2.42	2005
II	Murine	MHCB0032	1KT2	I-Ek	12	Moth Cytochrome C Peptide	ADLIAYLKQATK	A, B	B	2.80	2002

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	Res. (Å)	Release Year
II	Murine	MHCB0039	1R5V	I-Ek	13	Artificial Peptide	ADLIAYPKAAATKF	A, B	E	2.50	2004
II	Murine	MHCB0040	1R5W	I-Ek	13	Artificial Peptide	ADLIAYFKAAATKF	A, B	E	2.90	2004
II	Murine	MHCB0033	1KTD	I-Ek	14	Pigeon Cytochrome C Peptide	AADLIAYLKQASA K	A, B	B	2.40	2002

Table 4.2: List of TR-pMHC structures in MPID-T2

PL: peptide length; MCI: MHC chain identifier(s); PCI: peptide chain identifier; TRCI: TR chain identifier; Res.: Resolution

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Human	MHCA0135	1W72	A*0101	9	Melanoma-associated antigen 1	EADPTGHSY	A	C	H, L	FAB-HYB3	2.15	2004
I	Human	MHCA0004	1AO7	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	D, E	A6	2.60	1997
I	Human	MHCA0005	1BD2	A*0201	9	HTLV-1 Tax	LLFGYPVYV	A	C	D, E	B7	2.50	1998
I	Human	MHCA0094	1LP9	A*0201	9	Self peptide	ALWGFFPVL	A	C	E, F	mAH31 2.2	2.00	2003
I	Human	MHCA0108	1OGA	A*0201	9	Synthetic	GILGFVFTL	A	C	D, E	Vb17V a10.2	1.40	2003
I	Human	MHCA0057	1QRN	A*0201	9	Tax peptide P6A	LLFGYAVYV	A	C	D, E	A6	2.80	1999
I	Human	MHCA0056	1QSE	A*0201	9	Tax peptide	LLFGYPRYV	A	C	D, E	A6	2.80	1999
I	Human	MHCA0055	1QSF	A*0201	9	Tax peptide	LLFGYPVAV	A	C	D, E	A6	2.80	1999
I	Human	MHCA0175	2BNQ	A*0201	9	Synthetic peptide	SLLMWITQV	A	C	D, E	1G4	1.70	2005
I	Human	MHCA0176	2BNR	A*0201	9	Synthetic peptide	SLLMWITQC	A	C	D, E	1G4	1.90	2005
I	Human	MHCA0160	2F53	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	D, E	1G4	2.10	2006

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Human	MHCA0161	2F54	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	D, E	1G4	2.70	2006
I	Human	MHCA0163	2GJ6	A*0201	9	Modified HTLV-1 TAX (Y5K-IBA) peptide	LLFGKPVYV	A	C	D, E	A6	2.56	2006
I	Human	MHCA0171	2J8U	A*0201	9	Self peptide P1049	ALWGFPPVL	A	C	E, F	AH3	2.88	2007
I	Human	MHCA0172	2JCC	A*0201	9	Self peptide P1049	ALWGFPPVL	A	C	E, F	AH3	2.50	2007
I	Human	MHCA0168	2UWE	A*0201	9	Self peptide from C15ORF24 uncharacterized protein	ALWGFPPVL	A	C	E, F	AH3	2.40	2007
I	Human	MHCA0178	2P5E	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	D, E	1G4	1.89	2007
I	Human	MHCA0179	2P5W	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	D, E	1G4	2.20	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Human	MHCA0180	2PYE	A*0201	9	Cancer/testis antigen 1B	SLLMWITQC	A	C	D, E	1G4	2.30	2007
I	Human	MHCA0181	2VLJ	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	D, E	Vb17V a10.2	2.40	2008
I	Human	MHCA0182	2VLK	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	D, E	Vb17V a10.2	2.50	2008
I	Human	MHCA0184	2VLR	A*0201	9	Influenza matrix peptide	GILGFVFTL	A	C	D, E	Vb17V a10.2	2.30	2008
I	Human	MHCA0338	3D39	A*0201	9	Modified HTLV-1 TAX (Y5(4-fluoro)F) peptide	LLFGFPVYV	A	C	D, E	A6	2.81	2009
I	Human	MHCA0337	3D3V	A*0201	9	Modified HTLV-1 TAX (Y5(3,4-di-fluoro)F) peptide	LLFGFPVYV	A	C	D, E	A6	2.80	2009
I	Human	MHCA0341	3GSN	A*0201	9	HCMV pp65 fragment 495-503	NLVPMVATV	H	P	A, B	RA14	2.80	2009

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Human	MHCA0344	3H9S	A*0201	9	Tellp peptide	MLWGYLQYV	A	C	D, E	A6	2.70	2010
I	Human	MHCA0340	3HG1	A*0201	10	Cancer/MART-1	ELAGIGILTV	A	C	D, E	MEL5	3.00	2009
I	Human	MHCA0099	1MI5	B*0801	9	Epstein Barr Virus Peptide	FLRGRAYGL	A	C	D, E	LC13	2.50	2003
I	Human	MHCA0335	3FFC	B*0801	9	FLRGRAYGL peptide from an EBV protein	FLRGRAYGL	A	C	D, E	CF34	2.80	2009
I	Human	MHCA0193	2NX5	B*3501	11	EBV peptide	EPLPQGQLTAY	A	C	D, E	ELS4	2.70	2007
I	Human	MHCA0347	3MV7	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFE _Y	A	C	D, E	TK3	2.00	2010
I	Human	MHCA0346	3MV8	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFE _Y	A	C	D, E	TK3	2.10	2010
I	Human	MHCA0345	3MV9	B*3501	11	HPVG peptide from Epstein-Barr nuclear antigen 1	HPVGEADYFE _Y	A	C	D, E	TK3	2.70	2010

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Human	MHCA0195	2AK4	B*3508	13	EBV peptide	LPEPLPQGQLTAY	A	C	D, E	SB27	2.50	2005
I	Human	MHCA0348	3KXF	B*3508	13	Trans-activator protein BZLF1	LPEPLPQGQLTAY	A	Q	D, E	SB27	3.10	2010
I	Human	MHCA0343	3KPR	B*4405	9	Mimotope peptide	EEYLKAWTF	A	C	D, E	LC13	2.60	2009
I	Human	MHCA0342	3KPS	B*4405	9	Self peptide from ABCD3 protein	EEYLQAFTY	A	C	D, E	LC13	2.70	2009
I	Human	MHCA0203	2ESV	E*0101	9	Peptide from CMV gpUL40	VMAPRTLIL	A	P	D, E	KK50.4	2.60	2006
I	Murine	MHCA0045	1FO0	H2-Kb	8	Natural peptide	INFDFNTI	H	P	A, B	BM3.3	2.50	2000
I	Murine	MHCA0050	1G6R	H2-Kb	8	Syir protein	SIYRYYYGL	H	P	A, B	2C	2.80	2000
I	Murine	MHCA0065	1KJ2	H2-Kb	8	Naturally processed	KVITFIDL	H	P	A, B	KB5-C20	2.71	2002
I	Murine	MHCA0105	1NAM	H2-Kb	8	Vesicular stomatitis virus nucleoprotein fragment	RGYVYQGL	H	P	A, B	BM3.3	2.70	2003
I	Murine	MHCA0100	1MWA	H2-Kb	8	Dev 8	EQYKFYSV	H	P	A, B	2C	2.40	2002

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
I	Murine	MHCA0052	2CKB	H2-Kb	8	Dev 8	EQYKFYSV	H	P	A, B	2C	3.20	1998
I	Murine	MHCA0227	2OL3	H2-Kb	8	Naturally processed octa-peptide PBM8	SQYYYNLSL	H	P	A, B	BM3.3	2.90	2007
I	Murine	MHCA0233	2E7L	H2-Ld	9	Synthetic peptide	QLSPFPFDL	E	P	A, C	M6	2.50	2007
I	Murine	MHCA0234	2OI9	H2-Ld	9	Synthetic peptide	QLSPFPFDL	A	Q	B, C	2C	2.35	2007
I	Murine	MHCA0334	3E2H	H2-Ld	9	QL9 peptide	QLSPFPFDL	A	Q	B, C	M67	3.80	2008
I	Murine	MHCA0333	3E3Q	H2-Ld	9	QL9 Peptide	QLSPFPFDL	A	G	C, E	M13	2.95	2008
II	Human	MHCB0047	1ZGL	DRA*0101	14	Myelin basic protein	VHHFKNIVTPR TPG	A, B	C	M, P	3A6	2.80	2005
II	Human	MHCB0009	1FYT	DRB1*0101	13	Hemagglutinin Ha1 Peptide	PKYVKQNTLK LAT	A, B	C	D, E	HA1.7	2.60	2000
II	Human	MHCB0051	2IAM	DRB1*0101	15	15-mer peptide from triosephosphate isomerase	GELIGILNAAK VPAD	A, B	P	C, D	E8	2.80	2007
II	Human	MHCB0052	2IAN	DRB1*0101	15	15-mer peptide from triosephosphate isomerase	GELIGTLNAAK VPAD	A, B	C	D, E	E8	2.80	2007

MHC Class	MHC Source	MPID-T2 ID	PDB ID	Allele	PL	Peptide Source	Peptide Sequence	MCI	PCI	TRCI	TR type	Res. (Å)	Release Year
II	Human	MHCB0053	2ICW	DRB1*0101	13	Haemagglutinin peptide	PKYVKQNTLK LAT	A, B	C	I, J	T7	2.41	2007
II	Human	MHCB0019	1J8H	DRB1*0401	13	Hemagglutinin Ha1 Peptide Chain	PKYVKQNTLK LAT	A, B	C	D, E	HA1.7	2.40	2002
II	Human	MHCB0048	1YMM	DRB1*1501	14	MBP peptide	ENPVVHFFKNI VTP	A, B	C	D, E	Ob.1A1 ₂	3.50	2005
II	Murine	MHCB0012	1D9K	I-Ak	16	Conalbumin Peptide	GNSHRGAIEWE GIESG	C, D	P	C, D	D10	3.20	1999
II	Murine	MHCB0062	3MBE	I-A(G7)	18	Polypeptide	GAMKRHGLDN YRGYSLGN	A, B	P	C, D	21.3	2.89	2010
II	Murine	MHCB0059	2PXY	I-Au	11	Myelin basic protein	RGGASQYRPSQ	C, D	P	A, B	1934.4 Vb8.2	2.23	2007
II	Murine	MHCB0058	2Z31	I-Au	11	Myelin basic protein	RGGASQYRPSQ	C, D	P	A, B	cl19Vb _{8.2}	2.70	2007
II	Murine	MHCB0057	1U3H	I-Au	12	Myelin basic protein	SRGGASQYRPS Q	C, D	P	A, B	172.10 Vb8.2	2.42	2005

MPID-T2: a database for sequence–structure–function analyses of pMHC and TR/pMHC structures

Javed Mohammed Khan¹, Harish Reddy Cheruku¹, Joo Chuan Tong^{2,3}
and Shoba Ranganathan^{1,2,*}

¹Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia, ²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore and ³Data Mining Department, Institute for Infocomm Research, Singapore, Singapore

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Sequence–structure–function information is critical in understanding the mechanism of pMHC and TR/pMHC binding and recognition. A database for sequence–structure–function information on pMHC and TR/pMHC interactions, MHC–Peptide Interaction Database-TR version 2 (MPID-T2), is now available augmented with the latest PDB and IMGT/3Dstructure-DB data, advanced features and new parameters for the analysis of pMHC and TR/pMHC structures.

Availability: <http://biolinfo.org/mpid-t2>.

Contact: shoba.ranganathan@mq.edu.au

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on November 26, 2010; revised on February 15, 2011; accepted on February 18, 2011

1 INTRODUCTION

Major histocompatibility complexes (MHC) or human leukocyte antigens (HLAs) in human are important elements of T cell-mediated immunity. They are cell surface glycoproteins among which MHC-I proteins are ubiquitously expressed by most cells and MHC-II proteins are expressed by antigen-presenting cells (APC; Lefranc *et al.*, 2005). MHC proteins bind immunogenic peptide epitopes (p) derived from antigens and present them as peptide–MHC (pMHC) complexes on the cell surface, for subsequent recognition by T-cell receptors (TR), leading to TR/pMHC complexes, which are responsible for T-cell activation and triggering the adaptive immune response cascade (Khan *et al.*, 2010). Understanding the physicochemical basis for the selection of certain specific peptide epitopes by MHC alleles and the consequent recognition of pMHC ligands by TR proteins is critical for the design of T cell-based peptide vaccines (Khan *et al.*, 2010).

An early collection of pMHC X-ray crystal structures in the Protein Data Bank (PDB; Berman *et al.*, 2000) led to the development of MPID (Govindarajan *et al.*, 2003), comprising 86 classical pMHC structures, reporting pMHC interaction parameters. With increasing pMHC and TR/pMHC structures in the PDB and in the IMGT/3Dstructure-DB (Kaas *et al.*, 2004) and reports of new interaction parameters (Kaas and Lefranc, 2005), MPID-T (Tong *et al.*, 2006) was developed, with 187 pMHC and 16 TR/pMHC

structures along with interaction parameters for the analysis of pMHC structures alone.

With recent rise in TR/pMHC structural data in the PDB and in IMGT/3Dstructure-DB (Ehrenmann *et al.*, 2010), and updated interaction parameters (Kaas *et al.*, 2008), there is an increasing demand for more effective and efficient computational protocols to predict T-cell epitopes. Hence, we have updated MPID-T, augmenting it with advanced features and new parameters for the analysis of both pMHC and TR/pMHC structures, to gain an in-depth understanding of the structural determinants underlying TR/pMHC binding and recognition.

2 RESOURCE DESCRIPTION

MPID-T2 is a semiautomatically curated structure-derived MySQL database hosted on a Linux server, containing interaction information on all available experimental X-ray crystal structures of pMHC and TR/pMHC complexes extracted from PDB. MPID-T2 (November 2010 update) comprises 415 entries from five MHC sources (human: 282, murine: 127, rat: 3, chicken: 2 and monkey: 1), spanning 56 alleles; 353 pMHC structures, 62 TR/pMHC complexes; 352 MHC class I (MHC-I) complexes and 63 MHC class II (MHC-II) structures. Overall, 327 entries are non-redundant (MHC-I: 279 and MHC-II: 48). MPID-T2 includes non-classical structures (structures with T-cell receptor like antibodies, cluster of differentiation {CD} molecules and natural killer cell immunoglobulin like receptors {KIR} associated to the pMHC) and complexes with non-standard residues. For PDB structures with multiple molecular assemblies, the most accurate and complete structure is stored. Each structure is manually verified, classified and analyzed for pMHC and TR/pMHC interactions.

2.1 Definitions of interaction parameters

2.1.1 Predefined interaction parameters Existing MPID-T interaction parameters namely (i) intermolecular hydrogen bonds; (ii) gap volume; (iii) gap index; and (iv) interface area have been applied to all new pMHC complexes and extended to all TR/pMHC structures (Tong *et al.*, 2006, 2007).

2.1.2 New interaction parameters Specific new interaction parameters in MPID-T2, vital for characterizing pMHC and/or TR/pMHC binding, computed from the 3D coordinates of the crystal structures, are listed below.

*To whom correspondence should be addressed.

Binding energy: binding energy (BE) is a measure of the strength of the interaction between the ligand and the receptor in terms of binding free energy (ΔG). Values for BE between peptide and MHC for all structures and between pMHC and TR for TR/pMHC structures were calculated using DCOMPLEX (Liu *et al.*, 2004).

Molecular surface electrostatic potential: interactions between TR and pMHC depend vastly on charges displayed by TR and pMHC binding interfaces. Hence, we used webPIPSA (Richter *et al.*, 2008) to calculate and ICM (Internal Coordinate Mechanics; Abagyan *et al.*, 1994) to visualize molecular surface electrostatic potential (MSEP) at the binding interfaces (Supplementary Fig. S1a, b).

TR docking angle: TR docking angle is the angle formed by the TR interface (paratope) on the pMHC interface (epitope) with respect to the linear axis of the cognate peptide within the MHC groove. This value ' θ ' (Supplementary Fig. S1a) was calculated by matching the respective pMHC and TR interface MSEP for complementarity of charges, augmented by TR/pMHC interacting residues from the literature. The charged residues at the pMHC interface form an ellipse. The angle between the major axis of the ellipse and the α backbone axis of the peptide was measured using ICM.

Contact area: contact area (CA) is the area enclosed by the interacting residues of the two molecules (Supplementary Fig. S1c), as compared to interface area, which is the interaction area at the molecular level. We have used ICM to compute CA values between peptide and MHC for all structures and between pMHC and TR for TR/pMHC structures.

3 IMPLEMENTATION

Each entry in MPID-T2 is given a unique identifier for ease of identification, comparison, characterization and rapid visualization. Information for each pMHC and TR/pMHC structure in MPID-T2 is classified into five major categories: (i) MHC (chain-id, allele, class and source); (ii) peptide (chain-id, sequence, source and length); (iii) computed pMHC interaction parameters (intermolecular hydrogen bonds, gap volume, gap index, interface area, BE, CA and MSEP); (iv) structural information (structure determination method, resolution, PDB release year and publication reference); and (v) hyperlinks to related external databases like PDB (for sequence-structure information), SYFPEITHI (Rammensee *et al.*, 1999; for MHC ligands and peptide motifs), IMGT/HLA (Robinson *et al.*, 2001; for HLA sequences) and IMGT/3Dstructure-DB (for annotations on pMHC and TR/pMHC sequences with 3D structures; Ehrenmann *et al.*, 2010). However, TR/pMHC structures in MPID-T2 have additional TR/pMHC interaction parameters (BE, MSEP, TR docking angle, CA, gap volume, gap index and interface area). Search page of the database presents a web interface that allows searching for pMHC and TR/pMHC complexes based on different categories (MHC class, allele, source organism, peptide length, user-defined output required and TR type) or PDB information (PDB-ID, resolution and release year; Supplementary Fig. S2a). The search output (Supplementary Fig. S2b) shows various fields; noticeably, TR/pMHC, pMHC, MHC, peptide and TR 3-D coordinates are downloadable for structural visualization. The alignment page illustrates pMHC and TR/pMHC structural alignments based on species, MHC allele, peptide length and TR type. To portray vital pMHC and TR/pMHC interactions, precomputed schematic diagrams, generated using LIGPLOT (Wallace *et al.*, 1995), are provided. Also available in the patterns page of MPID-T2 are consensus patterns, obtained

using WebLogo (Crooks *et al.*, 2004), showing the conservation of residues among peptides with same lengths and alleles. MPID-T2 help page lists database usability details, definitions for interaction parameters and other useful resources.

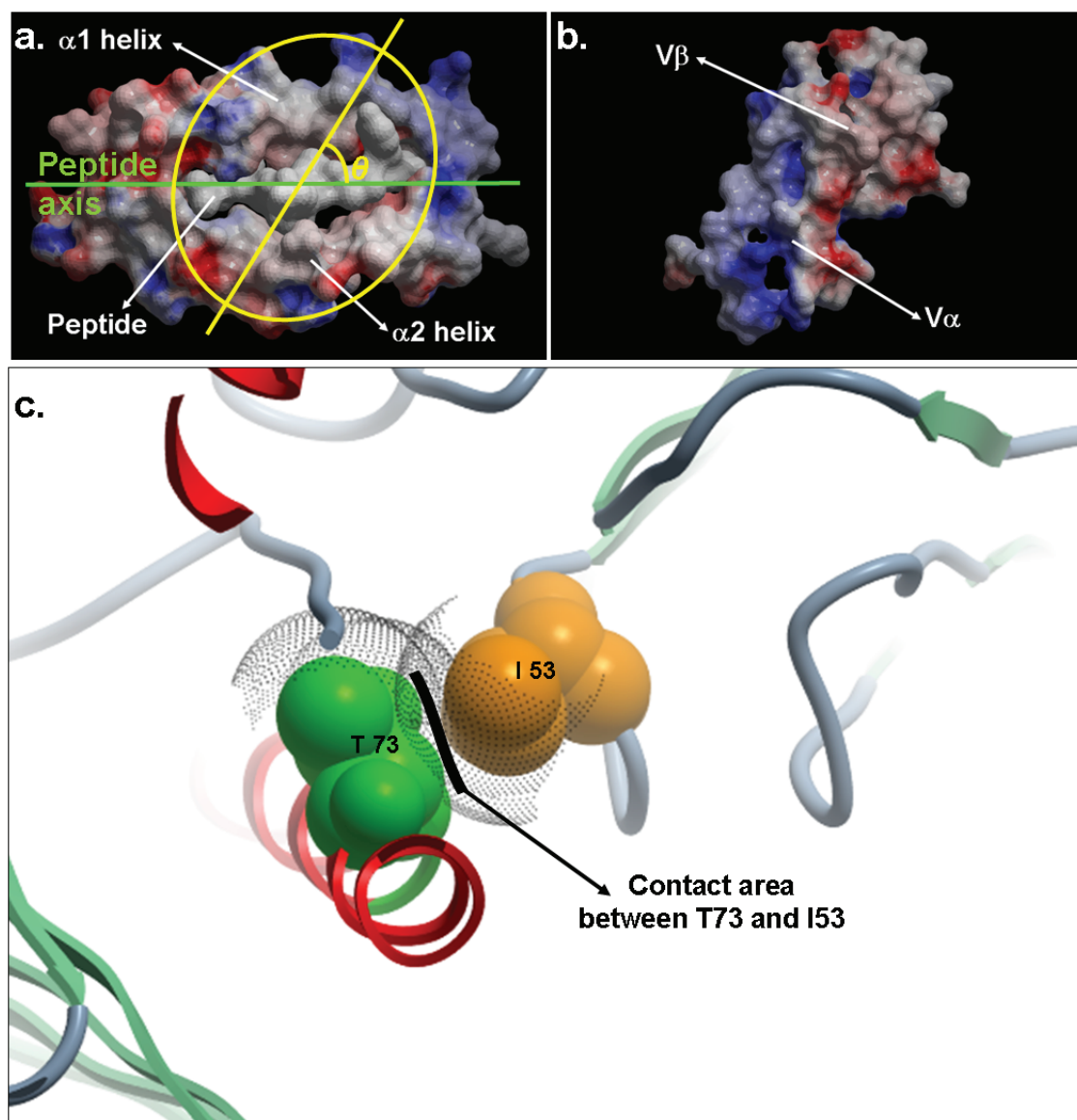
4 DISCUSSION

MPID-T2 aims to facilitate mining of fundamental relationships and structural descriptors hidden within TR/pMHC and pMHC interactions for in-depth characterization. Inclusion of structural descriptors like BE, MSEP, TR docking angle and CA have facilitated in understanding the principles underlying TR/pMHC binding (Khan and Ranganathan, unpublished results). These descriptors can be used as parameters defining pMHC and TR/pMHC interactions, thereby facilitating rational development of methods to identify strong MHC binding T-cell epitopes with greater propensity to activate T cells. This highlights the utility of MPID-T2 in vaccine research. We have now enabled TR-specific searches by classifying TR/pMHC structures based on TR types. Future enhancements will include listing post-translational modifications (PTM) for peptides to help understand the effect of PTM on TR/pMHC binding and interaction. MPID-T2 will be updated on a quarterly basis.

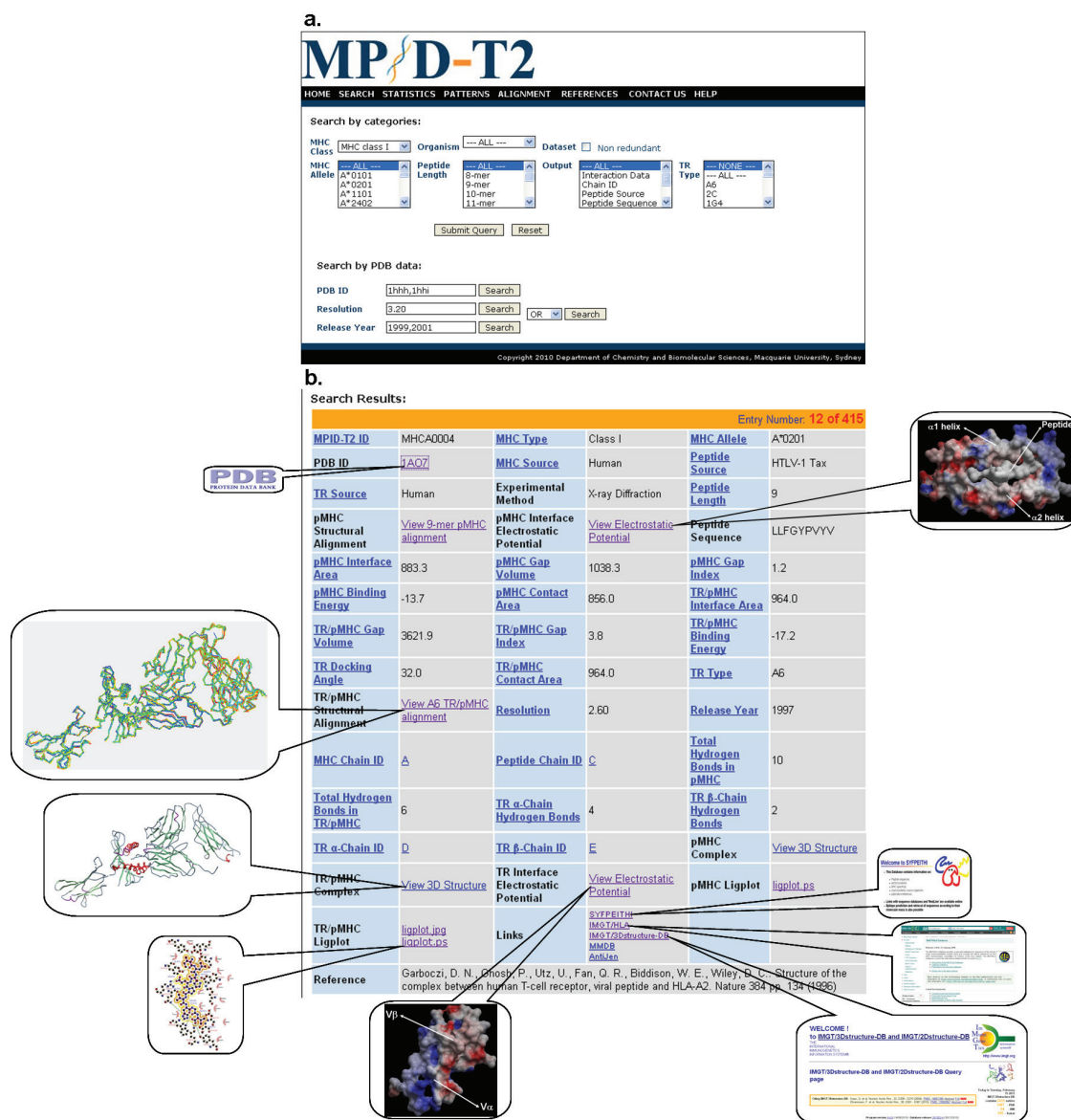
Conflict of Interest: none declared.

REFERENCES

- Abagyan, R.A. *et al.* (1994) ICM: a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, **15**, 488–506.
- Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Crooks, G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
- Ehrenmann, F. *et al.* (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.*, **38**, D301–D307.
- Govindarajan, K.R. *et al.* (2003) MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics*, **19**, 309–310.
- Kaas, Q. and Lefranc, M.P. (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol.*, **5**, 505–528.
- Kaas, Q. *et al.* (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.*, **32**, D208–D210.
- Kaas, Q. *et al.* (2008) IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In Schoenbach, C. *et al.* (eds) *Immunoinformatics, Immunomics Reviews Series*. Springer, New York, pp. 19–49.
- Khan, J.M. *et al.* (2010) Structural immunoinformatics: understanding MHC-peptide-TR binding. In Davies, M.N. *et al.* (eds) *Bioinformatics for Immunomics*, Vol. 3, Springer, New York, pp. 77–94.
- Lefranc, M.P. *et al.* (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev. Comput. Immunol.*, **29**, 917–938.
- Liu, S. *et al.* (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.
- Rammensee, H. *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
- Richter, S. *et al.* (2008) webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res.*, **36**, W276–W280.
- Robinson, J. *et al.* (2001) IMGT/HLA Database—a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.*, **29**, 210–213.
- Tong, J.C. *et al.* (2006) MPID-T: database for sequence-structure-function information on T-cell receptor/peptide/MHC interactions. *Appl. Bioinformatics*, **5**, 111–114.
- Tong, J.C. *et al.* (2007) *In silico* grouping of peptide/HLA class I complexes using structural interaction characteristics. *Bioinformatics*, **23**, 177–183.
- Wallace, A.C. *et al.* (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **8**, 127–134.



Supplementary Figure S1. Examples of Molecular Surface Electrostatic Potential (MSEP) and Contact Area (CA) as TR/pMHC interaction parameters. a. MSEP for a Tax-HLA-A2 pMHC interface (PDB code: 1A07). **b.** MSEP for A6 TR interface (1A07). **c.** CA between two interacting residues T73 and I53 from the $\alpha 1$ -helix (G-ALPHA1) of an MHC-I (HLA-A2) allele and the β -chain of a $V\beta 17V\alpha 10.2$ TR, respectively, in the TR/pMHC complex 1OGA (PDB code). The component parts/domains of both pMHC and TR interfaces are labeled in a. and b. TR interface in b. has been rotated 180° with respect to the pMHC interface in a. $V\alpha$ domain of TR interface interacts with $\alpha 2$ (G-ALPHA2) helix of the MHC and N-terminal half of the peptide, whereas, $V\beta$ domain interacts with $\alpha 1$ (G-ALPHA1) helix of the MHC and C-terminal half of the peptide. The ellipse (in yellow, with major axis marked diagonally) represents the paratope of the TR on the pMHC surface, while the green line represents the peptide axis. The TR docking angle, θ , is the angle between the peptide axis and the major axis of the ellipse.



Supplementary Figure S2. Screenshots of the search page and the search result for a TR/pMHC-I structure (1AO7) from the MPID-T2 database. a. The web interface for searching with user defined input parameters (including TR type). b. Search result for 1AO7 depicting various fields of pMHC and TR/pMHC information. Values for new pMHC and/or TR/pMHC interaction parameters: BA, CA and TR docking angle can be noted while MSEP images for both pMHC and TR interfaces can be accessed by clicking on the “View Electrostatic Potential” links provided as shown in the callout boxes. Structural alignment for TR/pMHC complexes based on TR types can also be visualized.

Table 4.3: Computed pMHC interaction parameters for pMHC-I structures in MPID-T2

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
8	A*0201	MHCA0070	1DUY	717.70	1116.10	1.60	-11.60	693.00	7
8	B*0801	MHCA0014	1AGB	820.80	881.70	1.10	-14.90	805.00	15
8	B*0801	MHCA0015	1AGC	812.50	688.10	0.90	-14.20	799.00	18
8	B*0801	MHCA0016	1AGD	819.70	816.10	1.00	-14.30	802.00	16
8	B*0801	MHCA0017	1AGE	812.30	920.60	1.10	-14.50	792.00	15
8	B*0801	MHCA0018	1AGF	860.00	765.40	0.90	-15.00	883.00	14
8	B*3501	MHCA0020	1A1N	857.90	670.20	0.80	-16.70	837.00	11
8	B*3508	MHCA0208	3BWA	837.30	628.70	0.80	-13.80	837.00	14
8	B*5101	MHCA0043	1E28	804.90	724.00	0.90	-15.10	783.00	7
8	H2-Kb	MHCA0069	1BQH	882.60	658.30	0.80	-12.30	866.00	12
8	H2-Kb	MHCA0045	1FO0	902.80	793.10	0.90	-12.90	891.00	18
8	H2-Kb	MHCA0047	1FZJ	834.90	857.60	1.00	-15.50	813.00	14
8	H2-Kb	MHCA0048	1FZM	933.20	732.70	0.80	-14.50	915.00	16
8	H2-Kb	MHCA0050	1G6R	865.20	1034.80	1.20	-11.10	855.00	13
8	H2-Kb	MHCA0068	1G7Q	752.10	891.10	1.20	-10.70	726.00	15
8	H2-Kb	MHCA0084	1KBG	938.20	753.80	0.80	-13.50	923.00	14
8	H2-Kb	MHCA0065	1KJ2	913.40	695.20	0.80	-11.90	902.00	16

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
8	H2-Kb	MHCA0066	1KJ3	905.30	555.20	0.60	-12.80	889.00	14
8	H2-Kb	MHCA0088	1KPU	896.80	732.60	0.80	-15.00	891.00	16
8	H2-Kb	MHCA0091	1LEG	880.60	815.00	0.90	-14.90	867.00	15
8	H2-Kb	MHCA0092	1LEK	870.20	1147.50	1.30	15.30	870.00	14
8	H2-Kb	MHCA0093	1LK2	776.20	835.40	1.10	14.50	788.00	15
8	H2-Kb	MHCA0100	1MWA	897.80	845.60	0.90	-13.10	887.00	19
8	H2-Kb	MHCA0104	1N59	864.70	767.60	0.90	-12.70	865.00	9
8	H2-Kb	MHCA0105	1NAM	928.30	613.80	0.70	-12.10	915.00	14
8	H2-Kb	MHCA0106	1NAN	892.20	800.40	0.90	-12.40	882.00	14
8	H2-Kb	MHCA0025	1OSZ	909.50	756.20	0.80	-14.90	912.00	18
8	H2-Kb	MHCA0351	1P1Z	812.00	998.90	1.20	-9.50	812.00	6
8	H2-Kb	MHCA0218	1P4L	795.30	921.10	1.20	-10.90	795.00	12
8	H2-Kb	MHCA0215	1RJY	858.70	774.10	0.90	-11.90	859.00	13
8	H2-Kb	MHCA0216	1RJZ	860.40	716.30	0.80	-12.50	860.00	15
8	H2-Kb	MHCA0219	1RK0	811.00	734.70	0.90	-14.80	811.00	14
8	H2-Kb	MHCA0220	1RK1	824.30	712.70	0.90	-14.80	824.00	12
8	H2-Kb	MHCA0128	1T0M	844.30	807.00	1.00	-12.20	839.00	17
8	H2-Kb	MHCA0129	1T0N	855.00	769.10	0.90	-11.70	820.00	15

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
8	H2-Kb	MHCA0028	1VAC	876.90	691.20	0.80	-15.40	865.00	14
8	H2-Kb	MHCA0052	2CKB	882.00	877.10	1.00	-13.00	866.00	10
8	H2-Kb	MHCA0225	2CLV	907.90	410.10	0.40	-13.50	908.00	18
8	H2-Kb	MHCA0226	2CLZ	917.50	590.30	0.60	-12.30	918.00	14
8	H2-Kb	MHCA0232	2FO4	791.60	812.00	1.00	-14.70	792.00	13
8	H2-Kb	MHCA0076	2MHA	874.20	1182.40	1.40	-11.00	851.00	6
8	H2-Kb	MHCA0227	2OL3	885.30	909.80	1.00	-15.30	885.00	19
8	H2-Kb	MHCA0228	2QRI	851.50	648.70	0.80	-12.60	865.00	15
8	H2-Kb	MHCA0230	2QRS	840.60	675.30	0.80	-12.70	850.00	13
8	H2-Kb	MHCA0229	2QRT	829.20	910.30	1.10	-12.40	819.00	15
8	H2-Kb	MHCA0030	2VAA	909.70	738.30	0.80	-14.80	895.00	16
8	H2-Kb	MHCA0250	2ZSV	823.90	805.90	1.00	-11.20	822.00	16
8	H2-Kb	MHCA0249	2ZSW	835.80	712.70	0.80	-11.50	836.00	14
8	H2-Kb	MHCA0235	3C8K	842.60	593.90	0.70	-12.40	843.00	12
8	H2-Kb	MHCA0244	3CVH	814.40	1011.70	1.20	-12.20	814.00	12
8	H2-Kk	MHCA0222	1ZT1	826.80	983.00	1.20	-13.80	827.00	19
8	H2-Kwm7	MHCA0302	3FOL	852.60	859.60	1.00	-10.80	853.00	14
8	H2-Kwm7	MHCA0301	3FOM	878.20	817.70	0.90	-11.90	878.00	18

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
8	H2-Kwm7	MHCA0300	3FON	814.80	712.20	0.90	-10.40	815.00	20
8	Mamu-A*01	MHCA0212	1ZVS	724.30	910.80	1.30	-9.70	724.00	14
9	A*0201	MHCA0002	1AKJ	857.20	746.50	0.90	-12.80	839.00	13
9	A*0201	MHCA0004	1AO7	883.30	1038.30	1.20	-13.70	856.00	10
9	A*0201	MHCA0006	1B0G	860.20	441.30	0.50	-14.90	848.00	12
9	A*0201	MHCA0010	1B0R	642.40	724.00	1.10	-9.80	712.00	7
9	A*0201	MHCA0005	1BD2	874.20	813.60	0.90	-13.40	852.00	11
9	A*0201	MHCA0040	1DUZ	863.00	1069.60	1.20	-14.20	834.00	11
9	A*0201	MHCA0071	1EEY	787.10	900.30	1.10	-10.20	753.00	11
9	A*0201	MHCA0072	1EEZ	829.80	704.20	0.90	-11.00	793.00	9
9	A*0201	MHCA0008	1HHG	765.80	1039.90	1.40	-10.50	743.00	12
9	A*0201	MHCA0009	1HHI	842.00	455.70	0.50	-13.70	833.00	9
9	A*0201	MHCA0001	1HHJ	847.90	827.40	1.00	-12.20	820.00	14
9	A*0201	MHCA0003	1HHK	865.50	1083.40	1.30	-14.10	838.00	10
9	A*0201	MHCA0039	1I1F	850.90	800.80	0.90	-11.00	821.00	11
9	A*0201	MHCA0038	1I1Y	877.90	745.00	0.90	-12.40	855.00	13
9	A*0201	MHCA0060	1I7R	902.50	805.40	0.90	-12.00	878.00	11

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*0201	MHCA0059	1I7T	847.90	591.40	0.70	-12.90	823.00	9
9	A*0201	MHCA0061	1I7U	845.20	545.80	0.70	-13.90	827.00	11
9	A*0201	MHCA0051	1IM3	855.00	954.90	1.10	-13.80	837.00	11
9	A*0201	MHCA0058	1JHT	781.00	588.80	0.80	-13.60	763.00	12
9	A*0201	MHCA0094	1LP9	855.70	549.70	0.60	-14.70	840.00	12
9	A*0201	MHCA0108	1OGA	856.00	539.50	0.60	-14.10	843.00	12
9	A*0201	MHCA0110	1P7Q	810.80	648.10	0.80	-11.90	794.00	7
9	A*0201	MHCA0114	1QR1	827.10	722.40	0.90	-9.90	802.00	9
9	A*0201	MHCA0057	1QRN	871.50	921.60	1.10	-12.80	847.00	11
9	A*0201	MHCA0056	1QSE	873.00	1097.70	1.30	-12.00	839.00	11
9	A*0201	MHCA0055	1QSF	828.60	959.50	1.20	-12.80	808.00	10
9	A*0201	MHCA0146	1S8D	751.70	853.00	1.10	-12.30	752.00	11
9	A*0201	MHCA0123	1S9W	904.90	933.90	1.00	-15.20	899.00	13
9	A*0201	MHCA0124	1S9X	872.30	933.10	1.10	-14.30	854.00	12
9	A*0201	MHCA0125	1S9Y	883.70	979.00	1.10	-14.30	871.00	11
9	A*0201	MHCA0147	1T1W	829.80	731.10	0.90	-14.60	830.00	14
9	A*0201	MHCA0148	1T1X	816.30	788.00	1.00	-13.10	816.00	13
9	A*0201	MHCA0149	1T1Y	818.70	754.20	0.90	-15.10	819.00	13

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*0201	MHCA0150	1T1Z	787.00	748.00	1.00	-12.30	787.00	10
9	A*0201	MHCA0151	1T20	836.60	735.70	0.90	-13.50	837.00	12
9	A*0201	MHCA0152	1T21	806.40	759.80	0.90	-14.60	806.00	13
9	A*0201	MHCA0153	1T22	828.80	814.60	1.00	-13.60	829.00	14
9	A*0201	MHCA0130	1TVB	829.20	866.80	1.00	-11.60	807.00	12
9	A*0201	MHCA0131	1TVH	865.10	950.60	1.10	-12.70	840.00	10
9	A*0101	MHCA0135	1W72	805.80	1003.80	1.20	-9.50	789.00	16
9	A*0201	MHCA0170	2AV1	784.80	1151.50	1.50	-14.60	785.00	10
9	A*0201	MHCA0169	2AV7	790.30	1132.50	1.40	-14.20	790.00	10
9	A*0201	MHCA0175	2BNQ	779.20	1199.10	1.50	-11.30	779.00	13
9	A*0201	MHCA0176	2BNR	809.30	1113.10	1.40	-12.90	809.00	13
9	A*0201	MHCA0177	2C7U	806.80	918.00	1.10	-12.20	807.00	8
9	A*0201	MHCA0160	2F53	802.00	1127.90	1.40	-12.80	802.00	13
9	A*0201	MHCA0161	2F54	764.20	1201.20	1.60	-11.30	764.00	12
9	A*0201	MHCA0162	2GIT	813.00	952.80	1.20	-13.30	813.00	11
9	A*0201	MHCA0163	2GJ6	857.20	710.10	0.80	-12.70	857.00	11
9	A*0201	MHCA0165	2GTW	740.20	615.40	0.80	-10.90	740.00	8
9	A*0201	MHCA0166	2GTZ	766.00	669.20	0.90	-12.00	766.00	10

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*0201	MHCA0167	2GUO	726.00	694.30	1.00	-10.10	726.00	11
9	A*0201	MHCA0171	2J8U	835.50	670.20	0.80	-14.40	836.00	10
9	A*0201	MHCA0172	2JCC	845.90	504.30	0.60	-14.30	846.00	12
9	A*0201	MHCA0178	2P5E	808.10	1134.60	1.40	-10.80	808.00	15
9	A*0201	MHCA0179	2P5W	796.20	1171.00	1.50	-1.60	796.00	14
9	A*0201	MHCA0180	2PYE	797.80	1277.50	1.60	-11.30	798.00	14
9	A*0201	MHCA0168	2UWE	844.70	459.30	0.50	-14.70	845.00	12
9	A*0201	MHCA0173	2V2W	807.60	759.80	0.90	-11.90	808.00	13
9	A*0201	MHCA0174	2V2X	830.00	791.50	1.00	-12.50	830.00	16
9	A*0201	MHCA0181	2VLJ	843.60	588.30	0.70	-13.80	844.00	11
9	A*0201	MHCA0182	2VLK	824.90	593.40	0.70	-14.30	825.00	9
9	A*0201	MHCA0183	2VLL	835.20	465.90	0.60	-13.90	835.00	12
9	A*0201	MHCA0184	2VLR	853.00	692.20	0.80	-13.90	853.00	11
9	A*0201	MHCA0313	2X4N	622.50	662.00	1.10	-8.10	602.00	8
9	A*0201	MHCA0312	2X4O	831.50	945.20	1.10	-12.50	831.00	12
9	A*0201	MHCA0311	2X4P	830.30	1129.00	1.40	-8.10	830.00	7
9	A*0201	MHCA0310	2X4Q	819.90	918.50	1.10	-9.10	820.00	11
9	A*0201	MHCA0309	2X4R	774.70	842.70	1.10	-10.10	775.00	10

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*0201	MHCA0308	2X4S	817.70	877.00	1.10	-11.40	818.00	10
9	A*0201	MHCA0307	2X4T	776.80	867.80	1.10	-9.70	777.00	14
9	A*0201	MHCA0306	2X4U	855.90	810.00	1.00	-12.10	856.00	13
9	A*0201	MHCA0248	3BGM	810.60	1238.00	1.50	-12.30	811.00	15
9	A*0101	MHCA0255	3BO8	813.70	767.50	0.90	-11.90	814.00	17
9	A*0201	MHCA0261	3D25	804.80	895.00	1.10	-12.70	805.00	12
9	A*0201	MHCA0338	3D39	843.30	936.40	1.10	-12.10	843.00	10
9	A*0201	MHCA0337	3D3V	843.20	1006.10	1.20	-12.00	843.00	10
9	A*0201	MHCA0265	3FQT	715.10	740.30	1.00	-10.40	715.00	11
9	A*0201	MHCA0264	3FQU	702.50	860.20	1.20	-12.30	702.00	11
9	A*0201	MHCA0263	3FQW	798.30	727.50	0.90	-11.40	798.00	13
9	A*0201	MHCA0262	3FQX	841.50	867.80	1.00	-10.60	841.00	15
9	A*0201	MHCA0275	3FT2	844.60	1097.70	1.30	-11.60	845.00	15
9	A*0201	MHCA0274	3FT3	865.60	818.70	1.00	-11.60	866.00	14
9	A*0201	MHCA0273	3FT4	868.80	839.70	1.00	-13.50	869.00	15
9	A*0201	MHCA0272	3GJF	834.80	932.90	1.10	-13.00	835.00	11
9	A*0201	MHCA0341	3GSN	763.40	1029.10	1.40	-10.30	763.00	11
9	A*0201	MHCA0288	3GSO	794.30	929.80	1.20	-12.40	794.00	15

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*0201	MHCA0287	3GSQ	763.20	699.90	0.90	-13.30	763.00	13
9	A*0201	MHCA0286	3GSR	789.00	705.50	0.90	-12.60	789.00	14
9	A*0201	MHCA0285	3GSU	788.70	732.20	0.90	-13.60	789.00	14
9	A*0201	MHCA0284	3GSV	803.30	661.50	0.80	-13.70	803.00	14
9	A*0201	MHCA0283	3GSW	776.00	779.80	1.00	-12.00	776.00	14
9	A*0201	MHCA0282	3GSX	776.40	782.80	1.00	-13.90	776.00	13
9	A*0201	MHCA0299	3H7B	885.00	797.70	0.90	-13.30	885.00	11
9	A*0201	MHCA0298	3H9H	879.20	893.40	1.00	-14.10	879.00	11
9	A*0201	MHCA0344	3H9S	894.30	704.00	0.80	-13.40	894.00	11
9	A*0201	MHCA0278	3HAE	757.00	1293.30	1.70	-11.10	757.00	10
9	A*0201	MHCA025	3HPJ	927.00	818.20	0.90	-13.70	927.00	14
9	A*0201	MHCA0324	3I6G	868.00	818.20	0.90	-15.20	869.00	9
9	A*0201	MHCA0297	3IXA	876.30	726.50	0.80	-13.90	876.00	11
9	A*0201	MHCA0303	3KLA	845.40	1109.00	1.30	-13.40	845.00	11
9	A*0201	MHCA0349	3MYJ	881.50	853.50	1.00	-14.20	881.00	12
9	A*1101	MHCA0143	1Q94	876.80	802.30	0.90	-10.10	877.00	12
9	A*1101	MHCA0145	1X7Q	912.50	596.50	0.60	-13.60	912.00	16
9	A*2402	MHCA0185	2BCK	862.50	1181.20	1.40	-13.60	862.00	11

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	A*2402	MHCA0332	3I6L	904.50	905.70	1.00	-14.40	905.00	12
9	B*0801	MHCA0095	1M05	1000.60	897.40	0.90	-14.20	977.00	18
9	B*0801	MHCA0099	1M15	1028.00	850.00	0.80	-14.30	1011.00	17
9	B*0801	MHCA0335	3FFC	956.70	881.10	0.90	-14.30	957.00	23
9	B*1402	MHCA0260	3BVN	986.90	1035.30	1.10	-15.20	987.00	13
9	B*1402	MHCA0259	3BXN	917.00	636.90	0.70	-15.90	917.00	12
9	B*1501	MHCA0140	1XR8	860.70	968.10	1.10	-13.10	852.00	16
9	B*1501	MHCA0141	1XR9	883.00	414.10	0.50	-15.30	923.00	18
9	B*1501	MHCA0206	3C9N	899.40	752.60	0.80	-13.10	899.00	19
9	B*2705	MHCA0019	1HSA	691.80	1148.40	1.70	-8.40	658.00	14
9	B*2705	MHCA0080	1JGE	815.40	994.40	1.20	-12.90	791.00	15
9	B*2705	MHCA0107	1OF2	1087.70	1015.50	0.90	17.00	1053.00	17
9	B*2705	MHCA0109	1OGT	1096.00	849.30	0.80	-16.60	1162.00	21
9	B*2705	MHCA0154	1UXS	1072.60	935.40	0.90	-15.10	1073.00	23
9	B*2705	MHCA0133	1W0V	1007.20	898.90	0.90	-16.00	978.00	16
9	B*2705	MHCA0187	2A83	1131.90	908.30	0.80	-17.10	1132.00	22
9	B*2705	MHCA0188	2BSR	953.30	1288.20	1.40	-14.70	953.00	15
9	B*2705	MHCA0186	2BST	943.30	1073.70	1.10	-14.90	943.00	17

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	B*2705	MHCA0242	3B6S	1092.70	842.80	0.80	-16.00	1093.00	19
9	B*2705	MHCA0253	3BP4	903.20	755.20	0.80	-13.40	903.00	17
9	B*2705	MHCA0276	3DTX	1093.90	804.40	0.70	-14.50	1094.00	20
9	B*2709	MHCA0082	1K5N	780.70	879.90	1.10	-12.70	777.00	19
9	B*2709	MHCA0132	1UXW	1071.90	1116.80	1.00	-15.60	1031.00	18
9	B*2709	MHCA0134	1W0W	999.90	968.10	1.00	-16.60	979.00	18
9	B*2709	MHCA0243	3B3I	1047.00	856.60	0.80	-14.80	1047.00	19
9	B*2709	MHCA0254	3BP7	914.50	758.30	0.80	-13.50	915.00	18
9	B*2709	MHCA0271	3CZF	1098.70	962.00	0.90	-16.60	1099.00	21
9	B*2709	MHCA0277	3D18	1032.20	1274.90	1.20	-15.70	1032.00	19
9	B*2709	MHCA0279	3HCV	1030.40	909.80	0.90	-14.80	1030.00	19
9	B*3501	MHCA0022	1A9B	855.40	847.40	1.00	-14.30	832.00	12
9	B*3501	MHCA0021	1A9E	882.60	779.30	0.90	-18.00	863.00	12
9	B*3501	MHCA0155	1CG9	823.40	794.10	1.00	-18.40	823.00	13
9	B*3501	MHCA0111	1QEW	843.20	855.90	1.00	-15.10	822.00	12
9	B*3501	MHCA0192	2CIK	880.60	846.80	1.00	-17.80	881.00	10
9	B*3501	MHCA0191	2H6P	886.30	756.20	0.80	-17.60	886.00	10
9	B*3501	MHCA0331	3LKN	887.80	776.20	0.90	-16.60	888.00	10

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	B*3501	MHCA0330	3LKO	894.50	922.60	1.00	-16.90	894.00	13
9	B*3501	MHCA0329	3LKP	865.20	828.40	1.00	-15.80	865.00	12
9	B*3501	MHCA0328	3LKQ	874.10	816.60	0.90	-16.10	874.00	11
9	B*3501	MHCA0327	3LKR	890.40	879.60	1.00	-16.50	890.00	10
9	B*3501	MHCA0326	3LKS	886.00	656.90	0.70	-16.40	886.00	11
9	B*4402	MHCA0096	1M6O	937.30	903.20	1.00	-12.60	928.00	16
9	B*4402	MHCA0295	3KPL	979.60	677.90	0.70	-14.90	980.00	17
9	B*4402	MHCA0294	3KPM	973.50	817.70	0.80	-13.10	974.00	18
9	B*4402	MHCA0319	3L3D	898.40	826.90	0.90	-13.50	898.00	17
9	B*4402	MHCA0317	3L3G	905.10	608.30	0.70	-12.50	905.00	16
9	B*4402	MHCA0316	3L3I	892.00	862.20	1.00	-12.70	892.00	17
9	B*4402	MHCA0315	3L3J	859.50	758.80	0.90	-13.00	860.00	17
9	B*4402	MHCA0314	3L3K	839.90	880.10	1.10	-11.30	840.00	16
9	B*4403	MHCA0101	1N2R	901.00	949.90	1.10	-13.70	882.00	15
9	B*4403	MHCA0126	1SYS	941.90	1076.00	1.10	-14.40	927.00	15
9	B*4403	MHCA0293	3KPN	980.00	640.00	0.60	-15.30	980.00	17
9	B*4403	MHCA0292	3KPO	958.70	766.50	0.80	-15.90	959.00	17
9	B*4405	MHCA0127	1SYV	901.80	915.60	1.00	-13.20	879.00	15

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	B*4405	MHCA0336	3DXA	935.90	1147.40	1.20	-12.20	936.00	12
9	B*4405	MHCA0291	3KPP	985.30	595.50	0.60	-15.50	985.00	16
9	B*4405	MHCA0290	3KPQ	957.80	768.50	0.80	-14.80	958.00	18
9	B*4405	MHCA0343	3KPR	1037.60	723.50	0.70	-13.80	1038.00	17
9	B*4405	MHCA0342	3KPS	1015.60	699.40	0.70	-13.50	1016.00	16
9	B*5101	MHCA0042	1E27	865.20	809.00	0.90	-15.20	841.00	8
9	B*5301	MHCA0023	1A1M	823.90	971.30	1.20	-15.40	803.00	12
9	B*5301	MHCA0024	1A1O	965.20	778.80	0.80	-15.50	947.00	10
9	B*5701	MHCA0241	2RFX	805.10	1022.00	1.30	-14.00	805.00	13
9	B*5703	MHCA0199	2BVP	839.10	695.30	0.80	-15.60	839.00	14
9	B*5703	MHCA0200	2BVQ	785.80	705.00	0.90	-12.80	786.00	11
9	Cw*0304	MHCA0054	1EFX	826.20	831.50	1.00	-11.20	798.00	10
9	Cw*0401	MHCA0113	1QQD	959.10	1116.10	1.20	-16.60	945.00	15
9	Cw*0401	MHCA0053	1IM9	949.10	1272.90	1.30	-14.30	937.00	13
9	E*0101	MHCA0098	1MHE	905.50	857.30	1.00	-12.90	900.00	8
9	E*0101	MHCA0203	2ESV	866.00	1103.40	1.30	-12.90	866.00	9
9	E*0101	MHCA0237	3BZE	911.40	989.70	1.10	-13.70	911.00	9
9	E*0101	MHCA0112	3BZF	875.40	898.00	1.00	-13.70	975.00	9

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	E*0101	MHCA0238	3CII	905.50	1110.50	1.20	-12.90	905.00	10
9	E*0103	MHCA0087	1KPR	895.40	906.20	1.00	-12.90	864.00	10
9	E*0103	MHCA0090	1KTL	890.40	985.60	1.10	12.10	869.00	10
9	E*0103	MHCA0209	3CDG	905.00	1168.40	1.30	-13.60	905.00	10
9	G*0101	MHCA0142	1YDP	991.70	859.50	0.90	-17.40	975.00	15
9	G*0101	MHCA0205	2D31	960.10	860.20	0.90	-13.40	960.00	16
9	G*0101	MHCA0204	2DYP	960.90	1074.70	1.10	-15.00	961.00	14
9	G*0101	MHCA0305	3KYN	762.80	945.20	1.20	-14.30	763.00	13
9	G*0101	MHCA0304	3KYO	868.10	884.70	1.00	-12.50	868.00	13
9	H2-Db	MHCA0031	1BZ9	1164.20	897.00	0.80	-13.50	855.00	10
9	H2-Db	MHCA0032	1CE6	867.70	787.70	0.90	-13.90	841.00	15
9	H2-Db	MHCA0073	1FFN	946.80	672.10	0.70	-11.40	934.00	18
9	H2-Db	MHCA0074	1FFO	878.40	548.90	0.60	-11.20	859.00	17
9	H2-Db	MHCA0075	1FFP	905.20	373.30	0.40	-11.20	885.00	21
9	H2-Db	MHCA0044	1FG2	928.40	519.10	0.60	-11.70	920.00	17
9	H2-Db	MHCA0077	1HOC	895.00	721.70	0.80	-13.30	885.00	20
9	H2-Db	MHCA0078	1INQ	870.20	532.20	0.60	-13.40	848.00	16
9	H2-Db	MHCA0063	1JPG	947.80	643.60	0.70	-15.00	929.00	19

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	H2-Db	MHCA0081	1JUF	865.10	583.90	0.70	-13.60	852.00	16
9	H2-Db	MHCA0103	1N5A	953.60	540.90	0.60	-11.60	944.00	16
9	H2-Db	MHCA0033	1QLF	879.50	567.30	0.70	-15.20	861.00	13
9	H2-Db	MHCA0119	1S7U	959.90	600.00	0.60	-11.70	948.00	20
9	H2-Db	MHCA0120	1S7V	914.80	774.90	0.90	-11.90	903.00	15
9	H2-Db	MHCA0121	1S7W	993.00	519.10	0.50	-12.30	975.00	18
9	H2-Db	MHCA0122	1S7X	950.00	484.80	0.50	-11.30	936.00	18
9	H2-Db	MHCA0224	2F74	930.90	636.90	0.70	-12.50	931.00	17
9	H2-Db	MHCA0240	2ZOK	892.30	449.50	0.50	-11.40	892.00	19
9	H2-Db	MHCA0239	2ZOL	854.00	460.80	0.50	-13.60	854.00	19
9	H2-Db	MHCA0236	3BUY	951.20	620.50	0.60	-14.80	951.00	15
9	H2-Db	MHCA0268	3CC5	999.40	600.60	0.60	-12.40	999.00	20
9	H2-Db	MHCA0269	3CCH	882.90	802.80	0.90	-10.10	883.00	18
9	H2-Db	MHCA0270	3CH1	939.20	634.40	0.70	-11.30	939.00	18
9	H2-Dd	MHCA0352	3E6F	828.50	919.00	1.10	-12.90	828.00	13
9	H2-Dd	MHCA0289	3E6H	866.90	869.40	1.00	-14.50	867.00	12
9	H2-Db	MHCA0296	3FTG	838.50	827.40	1.00	-12.90	838.00	16
9	H2-Kb	MHCA0049	1FZK	834.00	1067.50	1.30	-14.30	817.00	10

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	H2-Kb	MHCA0046	1FZO	818.60	814.60	1.00	-14.00	795.00	10
9	H2-Kb	MHCA0067	1G7P	860.70	993.70	1.20	-12.80	833.00	20
9	H2-Kb	MHCA0089	1KPV	814.20	1135.10	1.40	-13.90	797.00	10
9	H2-Kb	MHCA0115	1S7Q	842.50	671.40	0.80	-14.40	828.00	11
9	H2-Kb	MHCA0116	1S7R	950.40	831.80	0.90	-12.30	931.00	13
9	H2-Kb	MHCA0117	1S7S	877.50	730.50	0.80	-14.80	858.00	10
9	H2-Kb	MHCA0118	1S7T	852.30	672.90	0.80	-12.50	841.00	11
9	H2-Kb	MHCA0029	1VAD	861.00	939.50	1.10	-12.90	847.00	21
9	H2-Kb	MHCA0138	1WBZ	830.70	720.30	0.90	-11.30	807.00	15
9	H2-Kb	MHCA0221	1ZHB	950.50	529.90	0.60	-13.20	950.00	15
9	H2-Kb	MHCA0026	2VAB	817.90	1301.00	1.60	-14.30	802.00	12
9	H2-Kb	MHCA0251	3CPL	843.90	492.00	0.60	-9.90	844.00	15
9	H2-Kd	MHCA0217	1VGK	899.70	338.40	0.40	-15.40	900.00	17
9	H2-Kd	MHCA0231	2FWO	987.50	584.20	0.60	-16.60	988.00	19
9	H2-Kk	MHCA0223	1ZT7	951.70	886.80	0.90	-13.70	952.00	15
9	H2-Ld	MHCA0037	1LD9	952.70	746.50	0.80	-12.60	933.00	7
9	H2-Ld	MHCA0035	1LDP	748.60	889.30	1.20	10.10	720.00	9
9	H2-Ld	MHCA0233	2E7L	918.30	672.80	0.70	-13.70	918.00	10

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
9	H2-Ld	MHCA0234	2O19	900.60	710.10	0.80	-15.90	901.00	9
9	H2-Ld	MHCA0252	3ERY	892.50	581.10	0.60	-12.60	892.00	9
9	H2-Ld	MHCA0334	3E2H	905.80	772.10	0.80	-13.00	906.00	7
9	H2-Ld	MHCA0333	3E3Q	918.60	596.00	0.60	-13.80	919.00	11
9	H2-M3	MHCA0097	1MHC	848.10	1370.50	1.60	-17.80	839.00	9
9	H2-Qa-2	MHCA0083	1K8D	912.10	1185.40	1.30	-17.20	886.00	11
9	RT1.Aa	MHCA0085	1KJM	842.80	715.90	0.90	13.50	825.00	17
9	RT1-A1C	MHCA0086	1KJV	804.00	620.40	0.80	-15.40	823.00	18
10	A*0201	MHCA0012	1HHH	918.40	530.40	0.60	-18.10	908.00	11
10	A*0201	MHCA0064	1I4F	820.10	877.10	1.10	-13.00	804.00	15
10	A*0201	MHCA0027	1JF1	870.40	492.50	0.60	-14.60	885.00	11
10	A*0201	MHCA0011	2CLR	896.50	911.40	1.00	-15.30	876.00	10
10	A*0201	MHCA0164	2GT9	816.50	621.60	0.80	-10.10	816.00	11
10	A*0201	MHCA0246	3BH8	923.20	1087.00	1.20	-13.60	923.00	13
10	A*0201	MHCA0247	3BH9	880.30	860.70	1.00	-11.90	880.00	17
10	A*0201	MHCA0245	3BHB	949.10	1046.50	1.10	-15.00	949.00	12
10	A*0201	MHCA0267	3FQN	739.30	1039.90	1.40	-11.00	740.00	8
10	A*0201	MHCA0266	3FQR	828.70	1167.40	1.40	-13.30	829.00	9

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
10	A*0201	MHCA0280	3GIV	848.90	848.90	1.00	-14.50	849.00	12
10	A*0201	MHCA0340	3HG1	866.90	590.30	0.70	-11.40	867.00	12
10	A*0201	MHCA0323	3I6K	821.90	839.70	1.00	-12.10	822.00	8
10	A*0201	MHCA0321	3MGO	918.60	846.80	0.90	-13.40	919.00	7
10	A*0201	MHCA0320	3MGT	915.20	972.80	1.10	-13.60	915.00	11
10	A*1101	MHCA0144	1QVO	996.20	959.00	1.00	-13.40	996.00	12
10	A*1101	MHCA0159	2HN7	897.20	1130.00	1.30	-14.10	897.00	13
10	A*6801	MHCA0013	1TMC	918.10	926.20	1.00	-12.80	900.00	14
10	B*2101	MHCA0211	3BEW	952.30	910.80	1.00	-12.20	952.00	13
10	B*2705	MHCA0189	2BSS	937.20	1165.80	1.20	-14.70	937.00	18
10	B*2709	MHCA0079	1JGD	979.70	994.40	1.00	-14.30	979.00	23
10	B*3501	MHCA0190	2AXG	818.40	839.20	1.00	-13.50	818.00	12
10	B*3508	MHCA0196	2AXF	849.10	716.30	0.80	-16.00	849.00	12
10	B*4402	MHCA0258	3DX6	945.70	1224.70	1.30	-12.80	946.00	17
10	B*4403	MHCA0257	3DX7	937.90	1152.00	1.20	-13.80	938.00	17
10	B*4405	MHCA0256	3DX8	876.20	1313.80	1.50	-13.80	876.00	15
10	H2-Db	MHCA0036	1DDH	949.70	966.70	1.00	-13.00	903.00	9
10	H2-Db	MHCA0102	1N3N	816.60	626.90	0.80	-11.30	799.00	17

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
10	H2-Db	MHCA0136	1WBX	959.50	708.60	0.70	-14.20	941.00	20
10	H2-Db	MHCA0137	1WBY	884.10	793.90	0.90	-15.50	864.00	20
10	H2-Db	MHCA0213	1YN6	894.60	701.90	0.80	-13.60	895.00	23
10	H2-Db	MHCA0214	1YN7	895.90	645.60	0.70	-14.50	896.00	22
10	H2-Db	MHCA0318	3L3H	866.80	844.30	1.00	-13.70	867.00	13
10	H2-Dd	MHCA0034	1BII	923.50	792.10	0.90	-13.90	896.00	14
10	H2-Dd	MHCA0007	1QO3	934.50	999.80	1.10	-14.70	912.00	14
10	H2-Dd	MHCA0339	3DMM	899.60	870.40	1.00	-12.30	900.00	14
10	H2-Dd	MHCA0281	3ECB	901.40	1111.60	1.20	-14.20	901.00	12
11	B*2101	MHCA0210	3BEV	889.60	1134.60	1.30	-15.80	890.00	16
11	B*3501	MHCA0157	1ZSD	886.90	860.70	1.00	-16.10	887.00	16
11	B*3501	MHCA0193	2NX5	902.40	925.20	4.00	-13.30	902.00	11
11	B*3501	MHCA0345	3MV9	855.10	1199.60	1.40	-13.30	855.00	12
11	B*3501	MHCA0346	3MV8	872.70	1279.50	1.50	-13.00	873.00	13
11	B*3501	MHCA0347	3MV7	878.50	1148.40	1.30	-13.00	878.00	14
11	B*3508	MHCA0350	2FYV	715.90	875.50	1.20	-12.00	716.00	12
11	B*3508	MHCA0197	2FZ3	830.50	1069.00	1.30	-15.10	830.00	15
11	B*3508	MHCA0194	2NW3	901.40	849.90	0.90	-15.80	901.00	17

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
11	B*5703	MHCA0201	2HJK	848.30	1398.30	1.60	-14.80	848.00	9
11	B*5703	MHCA0202	2HJL	844.30	1288.20	1.50	-13.40	844.00	9
11	B*5703	MHCA0198	2BVO	821.60	1260.50	1.50	-14.10	822.00	11
11	H2-Db	MHCA0062	1JPF	854.40	645.90	0.80	-15.10	836.00	19
12	B*3508	MHCA0207	3BW9	969.80	764.40	0.80	-18.00	970.00	19
13	B*3501	MHCA0156	1ZHK	898.40	1753.60	2.00	-17.10	898.00	13
13	B*3508	MHCA0158	1ZHL	912.00	1743.30	1.90	-17.60	912.00	13
13	B*3508	MHCA0195	2AK4	919.80	1563.10	1.70	-14.00	920.00	12
13	B*3508	MHCA0322	3KWW	917.10	1479.20	1.60	-16.40	917.00	13
13	B*3508	MHCA0348	3KXF	889.50	1667.00	1.90	-13.40	893.00	12
13	RT1.Aa	MHCA0041	1ED3	1089.70	721.90	0.70	-17.80	1071.00	17
14	B*3501	MHCA0139	1XH3	927.00	1198.50	1.30	-18.10	890.00	13

Table 4.4: Computed pMHC interaction parameters for pMHC-II structures in MPID-T2

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (\AA^2)	Gap volume (\AA^3)	Gap index (\AA)	Binding energy (kcal/mol)	Contact Area (\AA^2)	H-bonds
6	DRB1*0401	MHCB0024	1D5X	530.40	760.40	1.40	-5.00	761.00	10
7	DRB1*0401	MHCB0025	1D5Z	666.10	640.80	1.00	-6.40	882.00	12
8	DRB1*0401	MHCB0023	1D5M	799.20	723.90	0.90	-7.90	860.00	13
8	DRB1*0401	MHCB0026	1D6E	737.20	680.20	0.90	-7.00	925.00	10
9	DRB1*0101	MHCB0037	1PYW	892.40	861.70	1.00	-13.30	1044.00	14
10	DRB5*0101	MHCB0010	1HQR	854.90	1087.00	1.30	-16.20	937.00	8
11	DQB1*0201	MHCB0041	1S9V	871.30	1313.70	1.50	-14.30	966.00	10
11	DRB3*0101	MHCB0055	2Q6W	955.90	623.10	0.60	-19.60	1136.00	18
11	I-Au	MHCB0059	2PXY	925.40	912.90	1.00	-15.60	1062.00	12
11	I-Au	MHCB0058	2Z31	902.50	1060.30	1.20	-15.80	1025.00	15
12	DRB1*0401	MHCB0006	2SEB	933.40	836.10	0.90	-15.00	1054.00	14
12	I-Au	MHCB0031	1K2D	965.20	968.80	1.00	-15.20	1083.00	12
12	I-Au	MHCB0057	1U3H	975.00	1034.80	1.10	-15.20	1122.00	11
12	I-Ek	MHCB0032	1KT2	1051.00	1177.30	1.10	-17.60	1205.00	14
13	DQB1*0302	MHCB0056	2NNA	1062.60	1203.70	1.10	-10.30	1249.00	18
13	DRB1*0101	MHCB0002	1DLH	1139.10	1081.70	1.00	-19.20	1318.00	17
13	DRB1*0101	MHCB0009	1FYT	1125.60	979.50	0.90	-19.30	1303.00	18

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
13	DRB1*0101	MHCB0016	1HXY	1110.60	1246.70	1.10	-18.80	1274.00	15
13	DRB1*0101	MHCB0028	1JWM	1129.60	1117.60	1.00	-19.10	1303.00	16
13	DRB1*0101	MHCB0029	1JWS	1122.90	1095.70	1.00	-18.60	1282.00	16
13	DRB1*0101	MHCB0030	1JWU	1114.00	1203.60	1.10	-18.90	1328.00	18
13	DRB1*0101	MHCB0020	1KG0	1120.80	1238.10	1.10	-19.30	1293.00	15
13	DRB1*0101	MHCB0034	1LO5	1119.10	978.30	0.90	-18.40	1272.00	13
13	DRB1*0101	MHCB0038	1R5I	1120.40	1054.90	0.90	-19.00	1301.00	19
13	DRB1*0101	MHCB0003	1SEB	806.70	964.10	1.20	-4.70	881.00	12
13	DRB1*0101	MHCB0044	1SJH	993.00	1127.80	1.10	-17.00	1100.00	16
13	DRB1*0101	MHCB0045	1T5W	1067.70	792.40	0.70	-18.00	1244.00	13
13	DRB1*0101	MHCB0046	1T5X	1056.90	906.10	0.90	-17.80	1216.00	13
13	DRB1*0101	MHCB0049	2G9H	1067.60	1383.90	1.30	-19.50	1247.00	20
13	DRB1*0101	MHCB0053	2ICW	1051.70	1304.10	1.20	-18.90	1234.00	22
13	DRB1*0101	MHCB0050	2OJE	1109.20	1079.80	1.00	-19.60	1308.00	16
13	DRB1*0401	MHCB0019	1J8H	1146.30	841.40	0.70	-19.90	1337.00	22
13	DRB3*0301	MHCB0061	3C5J	947.00	1450.50	1.50	-18.50	1088.00	18
13	I-Ab	MHCB0035	1LNU	1147.00	1127.00	1.00	-13.30	1278.00	18
13	I-Ak	MHCB0015	1IAK	1169.80	1180.30	1.00	-20.50	1350.00	16

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
13	I-Ek	MHCB0039	1R5V	1071.40	1087.70	1.00	-17.00	1177.00	15
13	I-Ek	MHCB0040	1R5W	1026.30	1235.70	1.20	-18.40	1132.00	7
14	DQB1*0302	MHCB0018	1JK8	1081.10	1472.70	1.40	-20.20	1256.00	20
14	DRA*0101	MHCB0047	1ZGL	968.00	1993.70	2.10	-16.80	1060.00	9
14	DRB1*0101	MHCB0001	1AQD	1190.30	1182.70	1.00	-22.00	1342.00	18
14	DRB1*0101	MHCB0054	2FSE	954.10	1673.20	1.80	-14.50	1070.00	12
14	DRB1*1501	MHCB0004	1BX2	985.70	1269.30	1.30	-17.10	1052.00	15
14	DRB1*1501	MHCB0048	1YMM	1086.20	1332.70	1.20	-18.50	1218.00	18
14	DRB5*0101	MHCB0027	1H15	1089.00	1468.20	1.40	-20.10	1218.00	11
14	I-A(G7)	MHCB0013	1ES0	1152.60	1109.70	1.00	-18.10	1296.00	14
14	I-Ad	MHCB0007	1IAO	1056.50	1449.50	1.40	-14.80	1142.00	14
14	I-Ad	MHCB0008	2IAD	1190.40	1475.10	1.20	-13.30	1030.00	13
14	I-Ak	MHCB0011	1F3J	1258.40	1462.30	1.20	-16.90	1405.00	18
14	I-Ek	MHCB0033	1KTD	1130.30	1182.40	1.10	-18.50	1282.00	12
15	DRB1*0101	MHCB0021	1KLG	1124.80	1130.00	1.00	-15.60	1227.00	15
15	DRB1*0101	MHCB0022	1KLU	1093.40	1217.80	1.10	-18.40	1212.00	16
15	DRB1*0101	MHCB0043	1SJE	1017.30	1084.00	1.10	-16.50	1127.00	16
15	DRB1*0101	MHCB0051	2IAM	1050.20	1194.50	1.10	-18.60	1194.00	16

Peptide length	Allele	MPID-T2 ID	PDB ID	Interface area (Å ²)	Gap volume (Å ³)	Gap index (Å)	Binding energy (kcal/mol)	Contact Area (Å ²)	H-bonds
15	DRB1*0101	MHCB0052	2IAN	1013.50	1225.70	1.20	-18.30	1156.00	15
15	DRB1*0301	MHCB0005	1A6A	1143.70	1204.70	1.10	-19.00	1301.00	19
15	HLA-DR1	MHCB0060	3L6F	1060.00	1331.70	1.30	-16.70	1232.00	13
15	I-A(G7)	MHCB0063	3CUP	1051.40	1402.40	1.30	-16.80	1205.00	17
15	I-Ab	MHCB0036	1MUJ	1220.10	1197.80	1.00	-19.80	1384.00	12
16	I- Ak	MHCB0012	1D9K	1159.00	837.30	1.60	-15.60	1312.00	19
16	I-Ak	MHCB0017	1JL4	1119.00	1968.10	1.80	-13.50	1298.00	13
18	I-A(G7)	MHCB0062	3MBE	1164.90	1917.50	1.60	-18.40	1321.00	14
20	DQB1*0602	MHCB0042	1UVQ	1234.10	1892.50	1.50	-18.40	1373.00	19
20	DRB5*0101	MHCB0014	1FV1	1065.60	1778.70	2.10	-20.40	1159.00	12

Table 4.5: Computed TR/pMHC interaction parameters for TR-pMHC-I structures in MPID-T2

Peptide Length	Allele	MPID-T2 ID	PDB ID	Interface Area (Å ²)	Gap Volume (Å ³)	Gap Index (Å)	Binding Energy (kcal/mol)	Contact Area (Å ²)	Total TR H-bonds	TR Docking Angle (°)
8	H2-Kb	MHCA0045	1FO0	646.80	5023.40	7.80	-9.60	647.00	5	72.00
8	H2-Kb	MHCA0050	1G6R	868.40	4083.00	4.70	-12.60	868.00	3	52.00
8	H2-Kb	MHCA0065	1KJ2	850.70	3722.10	4.40	-15.30	851.00	3	37.00
8	H2-Kb	MHCA0100	1MWA	975.20	3488.20	3.60	-14.10	975.00	4	43.00
8	H2-Kb	MHCA0105	1NAM	750.00	4361.60	5.80	-12.00	750.00	0	67.00
8	H2-Kb	MHCA0052	2CKB	936.70	3701.60	4.00	-11.80	937.00	2	62.00
8	H2-Kb	MHCA0227	2OL3	706.00	4477.30	6.30	-11.50	706.00	3	71.00
9	A*0201	MHCA0004	1AO7	964.00	3621.90	3.80	-17.20	964.00	6	32.00
9	A*0201	MHCA0005	1BD2	889.50	4371.30	4.90	-14.30	890.00	3	45.00
9	A*0201	MHCA0094	1LP9	968.80	3995.70	4.10	-23.00	969.00	4	30.00
9	A*0201	MHCA0108	1OGA	733.50	3797.50	5.20	-12.00	734.00	4	69.00
9	A*0201	MHCA0057	1QRN	950.60	4290.40	4.50	-16.60	951.00	6	35.00
9	A*0201	MHCA0056	1QSE	962.90	3604.50	3.70	-17.80	963.00	6	31.00
9	A*0201	MHCA0055	1QSF	876.00	3412.40	3.90	-15.00	876.00	3	39.00
9	A*0101	MHCA0135	1W72	884.20	4111.80	4.70	-9.30	884.00	2	73.00
9	A*0201	MHCA0175	2BNQ	1019.40	2783.80	2.70	-15.20	1019.00	4	38.00

Peptide Length	Allele	MPID-T2 ID	PDB ID	Interface Area (Å ²)	Gap Volume (Å ³)	Gap Index (Å)	Binding Energy (kcal/mol)	Contact Area (Å ²)	Total TR H-bonds	TR Docking Angle (°)
9	A*0201	MHCA0176	2BNR	1008.50	2906.10	2.90	-14.90	1008.00	5	39.00
9	A*0201	MHCA0160	2F53	965.90	3393.00	3.50	-16.10	954.00	6	36.00
9	A*0201	MHCA0161	2F54	1016.00	2906.10	2.90	-15.60	1016.00	6	36.00
9	A*0201	MHCA0163	2GJ6	912.50	3788.80	4.20	-15.50	913.00	7	37.00
9	A*0201	MHCA0171	2J8U	971.00	2914.80	3.00	-21.00	971.00	1	25.00
9	A*0201	MHCA0172	2JCC	948.40	3626.40	3.80	-22.70	948.00	4	20.00
9	A*0201	MHCA0178	2P5E	1007.50	3447.30	3.40	-16.10	1007.00	6	36.00
9	A*0201	MHCA0179	2P5W	1026.40	2799.10	2.70	-16.70	1026.00	6	34.00
9	A*0201	MHCA0180	2PYE	934.10	3186.60	3.40	-15.20	934.00	6	38.00
9	A*0201	MHCA0168	2UWE	967.00	3928.40	4.10	-22.40	967.00	3	21.00
9	A*0201	MHCA0181	2VLJ	710.00	3605.90	5.10	-11.80	710.00	4	70.00
9	A*0201	MHCA0182	2VLK	735.50	3728.30	5.10	-11.50	736.00	4	70.50
9	A*0201	MHCA0184	2VLR	749.20	3700.20	4.90	-11.90	749.00	3	70.00
9	A*0201	MHCA0338	3D39	923.20	3653.70	4.00	-15.20	923.00	5	38.50
9	A*0201	MHCA0337	3D3V	944.30	4109.20	4.30	-15.40	930.00	4	37.00
9	A*0201	MHCA0341	3GSN	912.90	4149.20	4.50	-16.70	913.00	3	34.50
9	A*0201	MHCA0344	3H9S	889.60	4024.70	4.50	-14.90	890.00	5	39.00

Peptide Length	Allele	MPID-T2 ID	PDB ID	Interface Area (Å ²)	Gap Volume (Å ³)	Gap Index (Å)	Binding Energy (kcal/mol)	Contact Area (Å ²)	Total TR H-bonds	TR Docking Angle (°)
9	B*0801	MHCA0099	1MI5	1078.30	3327.40	3.10	-17.20	1078.00	2	32.00
9	B*0801	MHCA0335	3FFC	1024.00	3864.50	3.80	-18.60	1024.00	2	30.00
9	B*4405	MHCA0336	3DXA	987.90	4031.40	4.50	-18.10	988.00	1	31.00
9	B*4405	MHCA0343	3KPR	1139.40	3490.30	3.10	-21.90	1139.00	2	22.00
9	B*4405	MHCA0342	3KPS	1116.40	3434.50	3.10	-21.50	1116.00	2	23.00
9	E*0101	MHCA0203	2ESV	961.50	3348.90	3.50	-12.80	962.00	5	50.00
9	H2-Ld	MHCA0233	2E7L	895.10	3711.40	4.20	-13.70	895.00	3	44.00
9	H2-Ld	MHCA0234	2OI9	832.00	4220.80	5.10	-13.10	832.00	2	47.00
9	H2-Ld	MHCA0334	3E2H	844.10	3562.50	4.20	-12.20	843.00	1	60.00
9	H2-Ld	MHCA0333	3E3Q	800.80	3265.00	4.80	-12.80	801.00	2	50.00
10	A*0201	MHCA0340	3HG1	950.40	3700.60	3.90	-17.60	950.00	2	30.00
11	B*3501	MHCA0193	2NX5	1016.20	4053.40	4.00	-14.80	1016.00	4	40.00
11	B*3501	MHCA0347	3MV7	958.20	3369.50	3.50	-17.00	890.00	8	33.00
11	B*3501	MHCA0346	3MV8	945.90	3492.80	3.70	-16.70	906.00	10	34.00
11	B*3501	MHCA0345	3MV9	898.30	3579.30	4.00	-16.40	898.00	7	35.00
13	B*3508	MHCA0195	2AK4	885.60	3850.70	4.30	-14.80	886.00	9	40.00
13	B*3508	MHCA0348	3KXF	849.00	6787.40	8.00	-18.70	849.00	8	28.00

Table 4.6: Computed TR/pMHC interaction parameters for TR-pMHC-II structures in MPID-T2

Peptide Length	Allele	MPID-T2 ID	PDB ID	Interface Area (Å ²)	Gap Volume (Å ³)	Gap Index (Å)	Binding Energy (kcal/mol)	Contact Area (Å ²)	Total TR H-bonds	TR Docking Angle (°)
11	I-Au	MHCB0059	2PXY	972.00	3304.90	3.40	-14.50	972.00	1	77.00
11	I-Au	MHCB0058	2Z31	955.10	3445.70	3.60	-13.00	955.00	1	87.00
12	I-Au	MHCB0057	1U3H	892.80	3358.70	3.80	-15.10	893.00	2	73.00
13	DRB1*0101	MHCB0009	1FYT	998.90	3569.10	3.60	-15.90	999.00	6	60.00
13	DRB1*0101	MHCB0053	2ICW	-	-	-	-	-	-	-
13	DRB1*0401	MHCB0019	1J8H	1014.20	3998.70	3.90	-16.80	1014.00	6	56.00
14	DRB1*1501	MHCB0048	1YMM	958.90	3445.20	3.60	-11.80	959.00	2	112.00
14	DRA*0101	MHCB0047	1ZGL	1069.90	3888.10	3.60	-15.60	1070.00	4	61.00
15	DRB1*0101	MHCB0051	2IAM	1057.80	3244.50	3.10	-13.70	1058.00	4	79.00
15	DRB1*0101	MHCB0052	2IAN	1075.20	3367.40	3.10	-15.00	1075.00	4	75.00
16	I-Ak	MHCB0012	1D9K	957.60	4281.80	4.50	-15.40	958.00	1	71.00
18	I-A(G7)	MHCB0062	3MBE	1297.20	3124.20	2.40	-22.40	1297.00	6	38.00

4.3 Data analysis

This section deals with the comparison of all predefined and new interaction parameters for all pMHC-I, pMHC-II, TR/pMHC-I and TR/pMHC-II structures listed in MPID-T2, to understand the correlation between the structural characteristics and dependencies of the interaction parameters upon each other and similarities in structural characteristics across the pMHC and TR/pMHC datasets. These computed interaction parameters for the pMHC-I, pMHC-II, TR/pMHC-I and TR/pMHC-II structures are listed in Table 4.3, Table 4.4, Table 4.5 and Table 4.6, respectively. Similarly, Figure 4.1 shows the graphs for the correlation between different computed interaction parameters for all pMHC-I complexes in MPID-T2.

From the figure, it is clearly evident that extremely poor correlations are obtained between sets of two interaction parameters such as pMHC-I interface area and pMHC-I gap volume (Figure 4.1a; $r^2=3E-05$), pMHC-I interface area and pMHC-I gap index (Figure 4.1b; $r^2=0.1048$), pMHC-I gap index and pMHC-I H-bonds (Figure 4.1f; $r^2=0.0848$), pMHC-I gap index and pMHC-I BE (Figure 4.1g; $r^2=0.01$), pMHC-I gap volume and pMHC-I H-bonds (Figure 4.1h; $r^2=0.0271$), pMHC-I gap volume and pMHC-I BE (Figure 4.1i; $r^2=0.0064$), pMHC-I H-bonds and pMHC-I BE (Figure 4.1j; $r^2=0.028$), pMHC-I gap index and pMHC-I contact area (Figure 4.1l; $r^2=0.1059$) and, pMHC-I gap volume and pMHC-I contact area (Figure 4.1m; $r^2=0.0003$). Same is the case with a few other sets where slightly better yet poor correlations are seen. These sets include pMHC-I interface area and pMHC-I BE (Figure 4.1c; $r^2=0.2362$), pMHC-I interface area and pMHC-I H-bonds (Figure 4.1d; $r^2=0.2069$), pMHC-I H-bonds and pMHC-I contact area (Figure 4.1n; $r^2=0.229$) and, pMHC-I BE and pMHC-I contact area (Figure 4.1o; $r^2=0.2326$). However, two distinct sets of pMHC-I structural descriptors portray excellent correlations, these are pMHC-I gap index and pMHC-I gap volume (Figure 4.1e; $r^2=0.8797$) and, pMHC-I interface area and pMHC-I contact area (Figure 4.1k; $r^2=0.9223$).

The above observations imply that for all pMHC-I complexes investigated in this study, the gap index, which measures geometric and electrostatic complementarity between the bound peptide and MHC protein, is inversely correlated with interface area and contact area (Figures 4.1b and 4.1l). This suggests that complexes with larger interface area and contact area have better geometric and electrostatic complementarity (i.e. smaller gap index) which indirectly results in the formation of more intermolecular hydrogen bonds (Figures 4.1d and 4.1n) contributing to the stability of the pMHC-I complexes. This could

be the reason behind the observed dependency of BE on interface area and contact area (Figures 4.1c and 4.1o). On the other hand, gap volume has almost no correlation with contact area and interface area (Figure 4.1m and 4.1a). A change in gap index or gap volume is also likely to have very little direct effect on the formation of H-bonds between peptides and MHC-I proteins (Figures 4.1f and 4.1h), hinting at their almost nil contribution towards pMHC-I BE (Figures 4.1g and 4.1i). Surprisingly, their BE seems to be independent of the number of H-bonds (Figure 4.1j). However, as expected, their gap volumes and gap indices are directly related just like their interface areas and contact areas (Figures 4.1e and 4.1k).

The average interface area for pMHC-I complexes is 874 \AA^2 and the average gap volume is 852.7 \AA^3 . Their almost similar values of average interface area and average gap volume have resulted in a gap index of 1 \AA on an average. Due to their relatively low averages for interface area and contact area (867 \AA^2), their mean BE (-13.5 kcal/mol) remains low despite the average number of H-bonds (13) formed between MHC-I binding peptides and their respective MHC-I alleles being relatively higher as indicated by the correlations portrayed in Figure 4.1.

Figure 4.2 depicts the graphs showing the correlation between different computed interaction parameters for all pMHC-II complexes in MPID-T2. Evidently, extremely poor correlations are observed between sets of two interaction parameters such as pMHC-II interface area and pMHC-II gap index (Figure 4.2b; $r^2=0.0013$), pMHC-II gap index and pMHC-II H-bonds (Figure 4.2f; $r^2=0.0535$), pMHC-II gap index and pMHC-II BE (Figure 4.2g; $r^2=0.0009$), pMHC-II gap volume and pMHC-II H-bonds (Figure 4.2h; $r^2=0.0002$), pMHC-II gap volume and pMHC-II BE (Figure 4.2i; $r^2=0.1036$), pMHC-II gap index and pMHC-II contact area (Figure 4.2l; $r^2=0.0132$) and, pMHC-II gap volume and pMHC-II contact area (Figure 4.2m; $r^2=0.0968$). Slightly better yet poor correlations are seen with a few other sets, these are pMHC-II interface area and pMHC-II gap volume (Figure 4.2a; $r^2=0.1797$), pMHC-II interface area and pMHC-II H-bonds (Figure 4.2d; $r^2=0.2442$), pMHC-II H-bonds and pMHC-II BE (Figure 4.2j; $r^2=0.1641$) and, pMHC-II H-bonds and pMHC-II contact area (Figure 4.2n; $r^2=0.3313$). Unlike pMHC-I complexes, the interaction parameters for pMHC-II complexes have a few good correlations such as pMHC-II interface area and pMHC-II BE (Figure 4.2c; $r^2=0.548$), pMHC-II gap index and pMHC-II gap volume (Figure 4.2e; $r^2=0.6237$) and, pMHC-II BE and pMHC-II contact area (Figure 4.2o; $r^2=0.5598$). However, only one set of structural descriptors portrays excellent

correlations among pMHC-II complexes. This set includes pMHC-II interface area and pMHC-II contact area (Figure 4.2k; $r^2=0.8607$).

These findings suggest that for all pMHC-II complexes studied here, unlike pMHC-I complexes, the geometric complementarity (gap index) plays no direct or indirect part in the either the increase or decrease of both interface area and contact area of the complexes (Figures 4.2b and 4.2l). Yet, the complexes with larger interface area and contact area have more intermolecular hydrogen bonds (Figures 4.2d and 4.2n) which assist in the stability of the pMHC-II complexes. This again could underlie the observed dependency of BE on interface area and contact area (Figures 4.2c and 4.2o). Contrary to observations in pMHC-I complexes, the gap volumes of pMHC-II complexes seem to be directly correlated to their contact areas and interface areas (Figures 4.2m and 4.2a). Similar to pMHC-I complexes, a change in gap index or gap volume is unlikely to have any effect on the formation of H-bonds between peptides and MHC-II proteins (Figures 4.2f and 4.2h). However, while their gap indices may portray no relationship with their BE values (Figure 4.2g), their gap volumes have a slight contribution towards their BE values (Figure 4.2i). As expected, although unlike pMHC-I structures, their BE values depend on the number of pMHC-II H-bonds (Figure 4.2j). The strong interdependencies of gap volume and gap index and, interface area and contact area are apparent in pMHC-II complexes, though not as much as in the case of pMHC-I structures (Figures 4.2e and 4.2k).

For pMHC-II structures, the mean interface area and gap volume are 1040.4 \AA^2 and 1187.6 \AA^3 . Although both these averages are greater than those of pMHC-I structures (874 \AA^2 and 852.7 \AA^3 , respectively), the systematic increase in both these values for pMHC-II structures, can be attributed to the mean gap index of 1.2 \AA making it comparable with the mean gap index of pMHC-I complexes (1 \AA). However, an increase in the means of their interface area, contact area (1181.7 \AA^2) and a relatively larger average number of H-bonds (15) formed between MHC-II binding peptides and their respective MHC-II alleles, have resulted in a significant methodical increase in their mean BE (-16.6 kcal/mol) as suggested by the correlations illustrated in Figure 4.2. Considering pMHC binding is a vital first step in T cell based immunity, these results indicate the suitability of analysing disease-implicated MHC-II alleles and their corresponding peptide antigens for T cell epitope prediction (chapter 6) and the design of MHC-II binding peptide vaccines to combat various diseases.

The graphs showing the correlation between various computed interaction parameters for all TR/pMHC-I complexes in MPID-T2 are exhibited in Figure 4.3. It is obvious that there occur extremely poor correlations between sets of two interaction parameters. These sets include TR/pMHC-I interface area and TR/pMHC-I H-bonds (Figure 4.3d; $r^2=0.0039$), TR/pMHC-I gap index and TR/pMHC-I H-bonds (Figure 4.3f; $r^2=0.0012$), TR/pMHC-I gap volume and TR/pMHC-I H-bonds (Figure 4.3h; $r^2=0.0021$), TR/pMHC-I gap volume and TR/pMHC-I BE (Figure 4.3i; $r^2=0.0114$), TR/pMHC-I H-bonds and TR/pMHC-I BE (Figure 4.3j; $r^2=0.0055$), TR/pMHC-I gap volume and TR docking angle (Figure 4.3m; $r^2=0.0313$), TR/pMHC-I H-bonds and TR docking angle (Figure 4.3n; $r^2=0.0501$) and, TR/pMHC-I H-bonds and TR/pMHC-I contact area (Figure 4.3r; $r^2=0.0003$). Poor correlations occur among a few other sets such as TR/pMHC-I interface area and TR/pMHC-I gap volume (Figure 4.3a; $r^2=0.1854$), TR/pMHC-I gap index and TR/pMHC-I BE (Figure 4.3g; $r^2=0.1772$), TR/pMHC-I gap index and TR docking angle (Figure 4.3l; $r^2=0.2825$) and, TR/pMHC-I gap volume and TR/pMHC-I contact area (Figure 4.3q; $r^2=0.1769$).

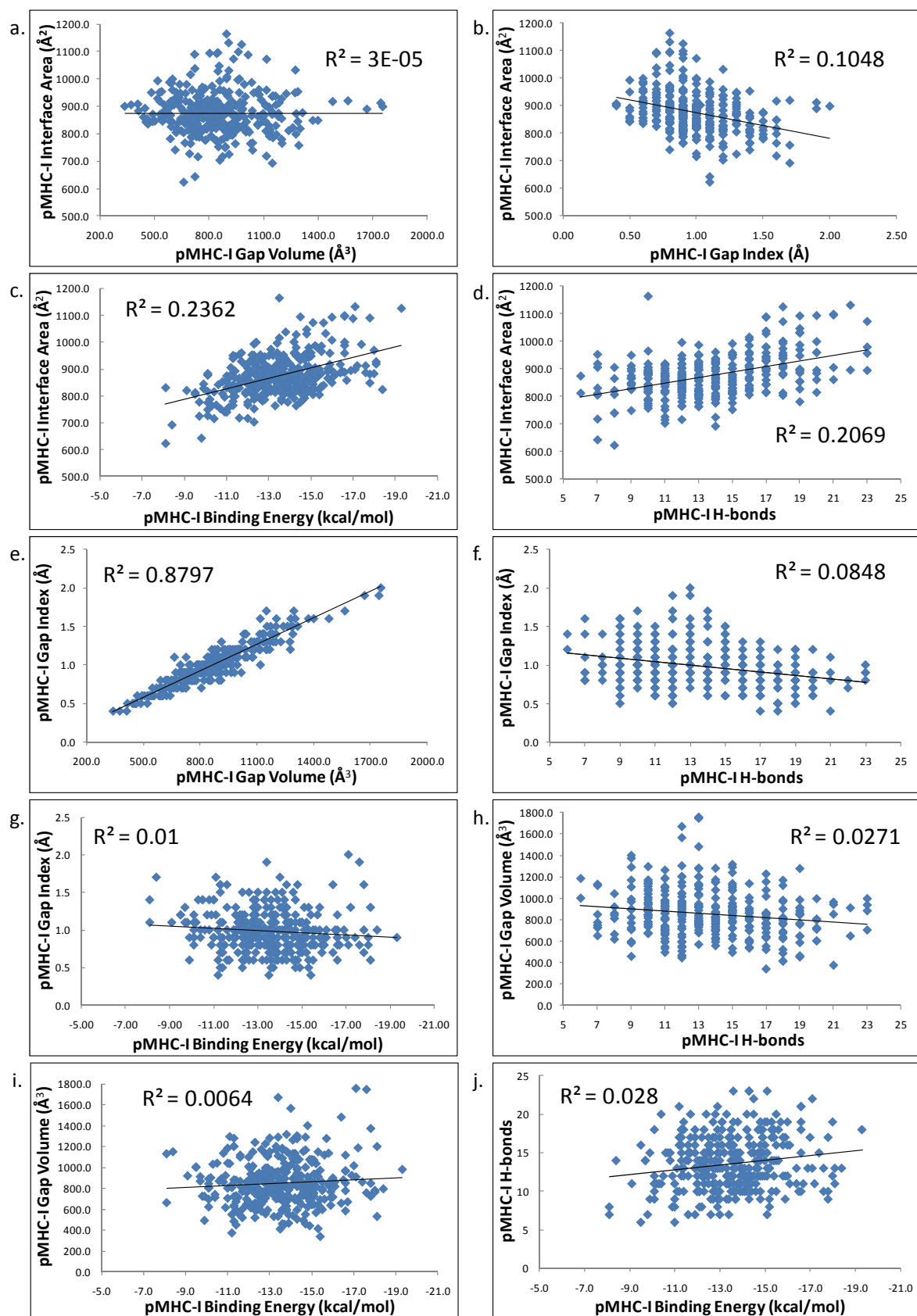
Interestingly, many sets of interaction parameters for TR/pMHC-I complexes show good correlations unlike pMHC-I structures. Among these sets are TR/pMHC-I interface area and TR/pMHC-I gap index (Figure 4.3b; $r^2=0.6239$), TR/pMHC-I interface area and TR/pMHC-I BE (Figure 4.3c; $r^2=0.4653$), TR/pMHC-I interface area and TR docking angle (Figure 4.3k; $r^2=0.628$), TR/pMHC-I gap index and TR/pMHC-I contact area (Figure 4.3p; $r^2=0.6062$), TR/pMHC-I BE and TR/pMHC-I contact area (Figure 4.3s; $r^2=0.4573$) and, TR/pMHC-I contact area and TR docking angle (Figure 4.3t; $r^2=0.6108$). Similar to pMHC-I complexes, even TR/pMHC-I structures have two sets of structural interaction parameters with excellent correlations. These sets are TR/pMHC-I gap index and TR/pMHC-I gap volume (Figure 4.3e; $r^2=0.7596$) and, TR/pMHC-I interface area and TR/pMHC-I contact area (Figure 4.3o; $r^2=0.9886$).

The above observations convey that for all TR/pMHC-I complexes analyzed here, the measure of geometric complementarity or gap index plays an indirect role (inverse relationship) in the increase or decrease of both interface area and contact area of the structures (Figures 4.3b and 4.3p). However, contrary to pMHC-I structures, the shifts in TR/pMHC-I interface areas and contact areas is not dependent on the intermolecular hydrogen bonds between the TR and the pMHC-I proteins (Figures 4.3d and 4.3r). Yet again, this highlights the noted dependency of their BE values on their interface areas and contact

areas (Figures 4.3c and 4.3s). Unlike pMHC-I complexes, the TR/pMHC-I gap volumes are in slightly inverse relationships with their contact areas and interface areas (Figure 4.3q and 4.3a), forming the basis of the observed inverse proportionalities of their gap indices with their interface areas and contact areas (Figures 4.3b and 4.3p).

A change in TR/pMHC-I gap index or gap volume is unlikely to have any effect on the formation of H-bonds between TR and pMHC-I proteins (Figures 4.3f and 4.3h), as in the case of both pMHC-I and pMHC-II complexes. The formation of TR/pMHC-I H-bonds is also unaffected by the measure of their TR docking angles (Figure 4.3n). Alarming, yet similar to pMHC-I complexes, TR/pMHC-I H-bond formation does not seem to contribute to TR/pMHC-I BE values (Figure 4.3j). Like pMHC-I complexes, TR/pMHC-I gap volumes portray no relationship with their BE values (Figure 4.3i), but contrastingly their gap indices have a little indirect contribution towards their BE values (Figure 4.3g). An increase or decrease in their TR docking angle is also inversely affected by their interface areas and contact areas (Figures 4.3k and 4.3t), shedding light on the significance of the inverse linear correlation obtained between their BE values and their TR docking angles (chapter 5). Just like their BE values, even their TR docking angles depict no dependency on their gap volumes (Figure 4.3m) but show a slight yet direct correlation with their gap indices (Figure 4.3l). Again, strong interdependencies of gap volume and gap index and, interface area and contact area are reverberant in TR/pMHC-I complexes, mirroring the behaviour of both pMHC-I and pMHC-II structures (Figures 4.3e and 4.3o).

The averages for interface area (915.8 \AA^2), contact area (913 \AA^2) and gap volume (3756.1 \AA^3) are increased for TR/pMHC-I structures compared to that of pMHC-I structures. The steep increase in the mean gap volume for TR/pMHC-I complexes meant a large mean gap index of 4.2 \AA for these structures. Their average TR docking angle of 42.10° is attributable to the increased BE (-15.6 kcal/mol) because of their linear inverse relationship (chapter 5). However, a significant decrease is notable in the average number of H-bonds (4) formed between TR and pMHC-I proteins.



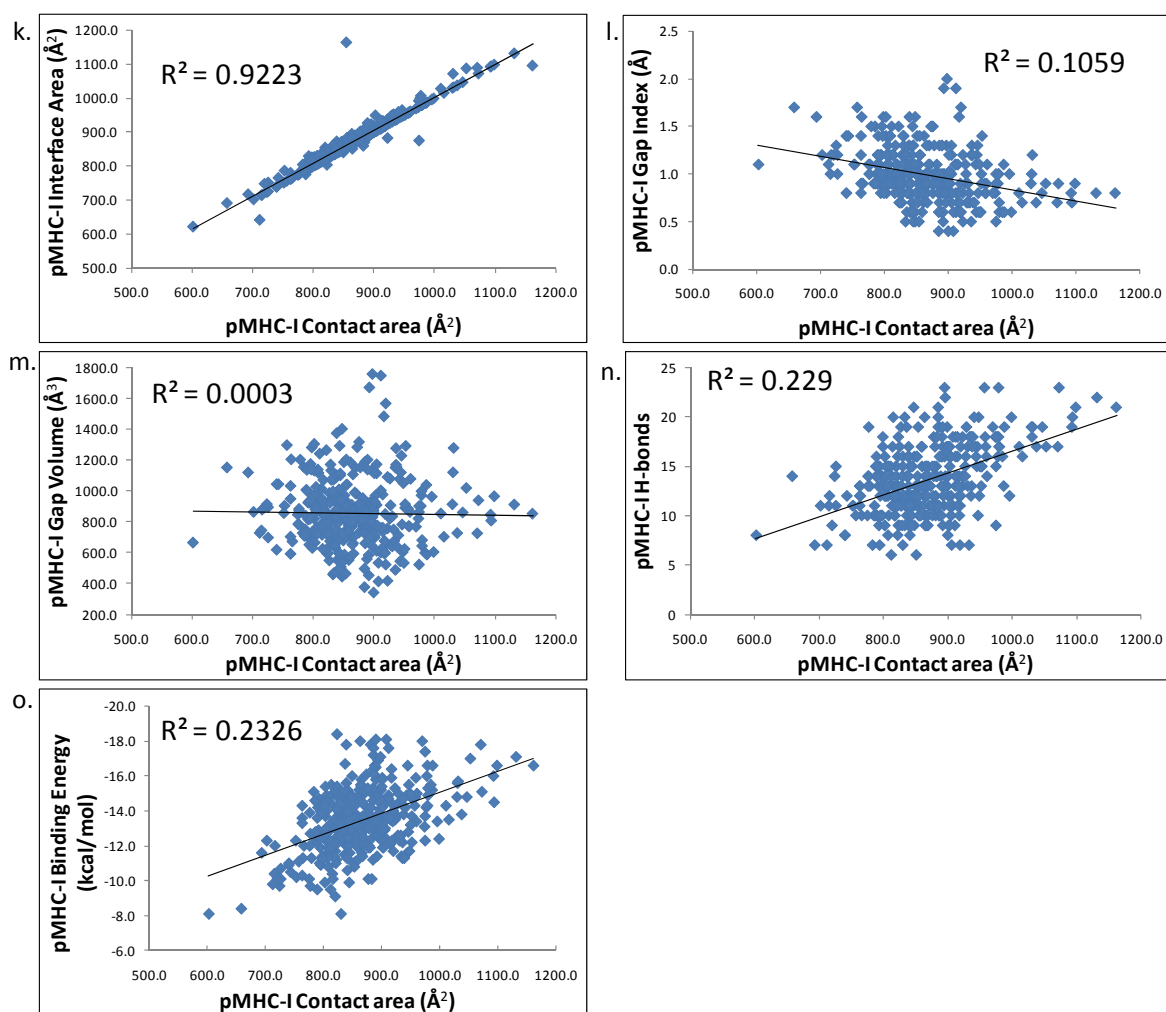
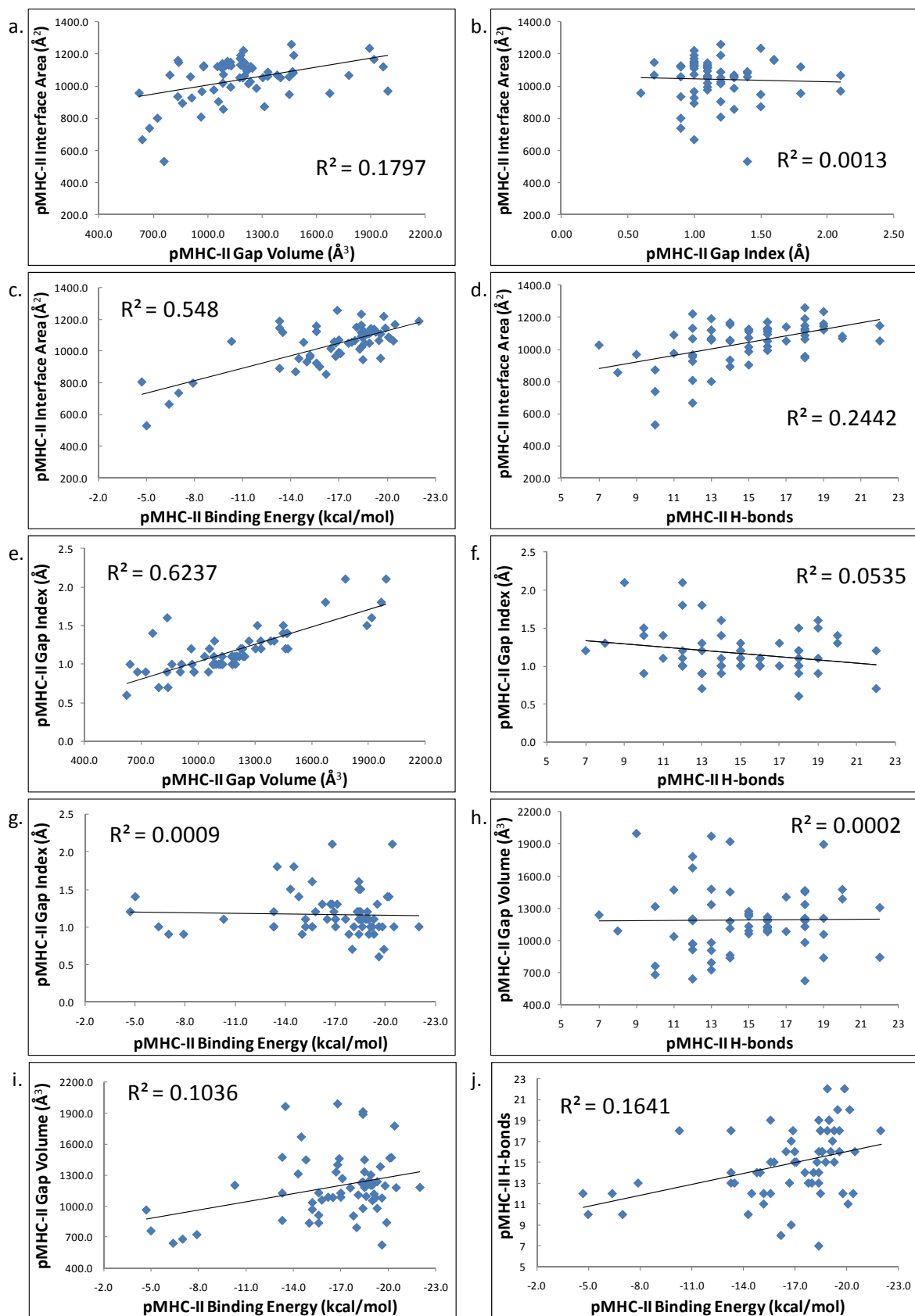


Figure 4.1: A graphical depiction of the correlation between different computed structural interaction parameters for all pMHC-I complexes in MPID-T2. a. pMHC-I interface area vs. pMHC-I gap volume. b. pMHC-I interface area vs. pMHC-I gap index. c. pMHC-I interface area vs. pMHC-I BE. d. pMHC-I interface area vs. pMHC-I H-bonds. e. pMHC-I gap index vs. pMHC-I gap volume. f. pMHC-I gap index vs. pMHC-I H-bonds. g. pMHC-I gap index vs. pMHC-I BE. h. pMHC-I gap volume vs. pMHC-I H-bonds. i. pMHC-I gap volume vs. pMHC-I BE. j. pMHC-I H-bonds vs. pMHC-I BE. k. pMHC-I interface area vs. pMHC-I contact area. l. pMHC-I gap index vs. pMHC-I contact area. m. pMHC-I gap volume vs. pMHC-I contact area. n. pMHC-I H-bonds vs. pMHC-I contact area. o. pMHC-I BE vs. pMHC-I contact area. The respective units are mentioned in the parentheses next to the names of the interaction parameters on the x and y-axes. The corresponding regression coefficients (r^2) are shown within each of the graphs.



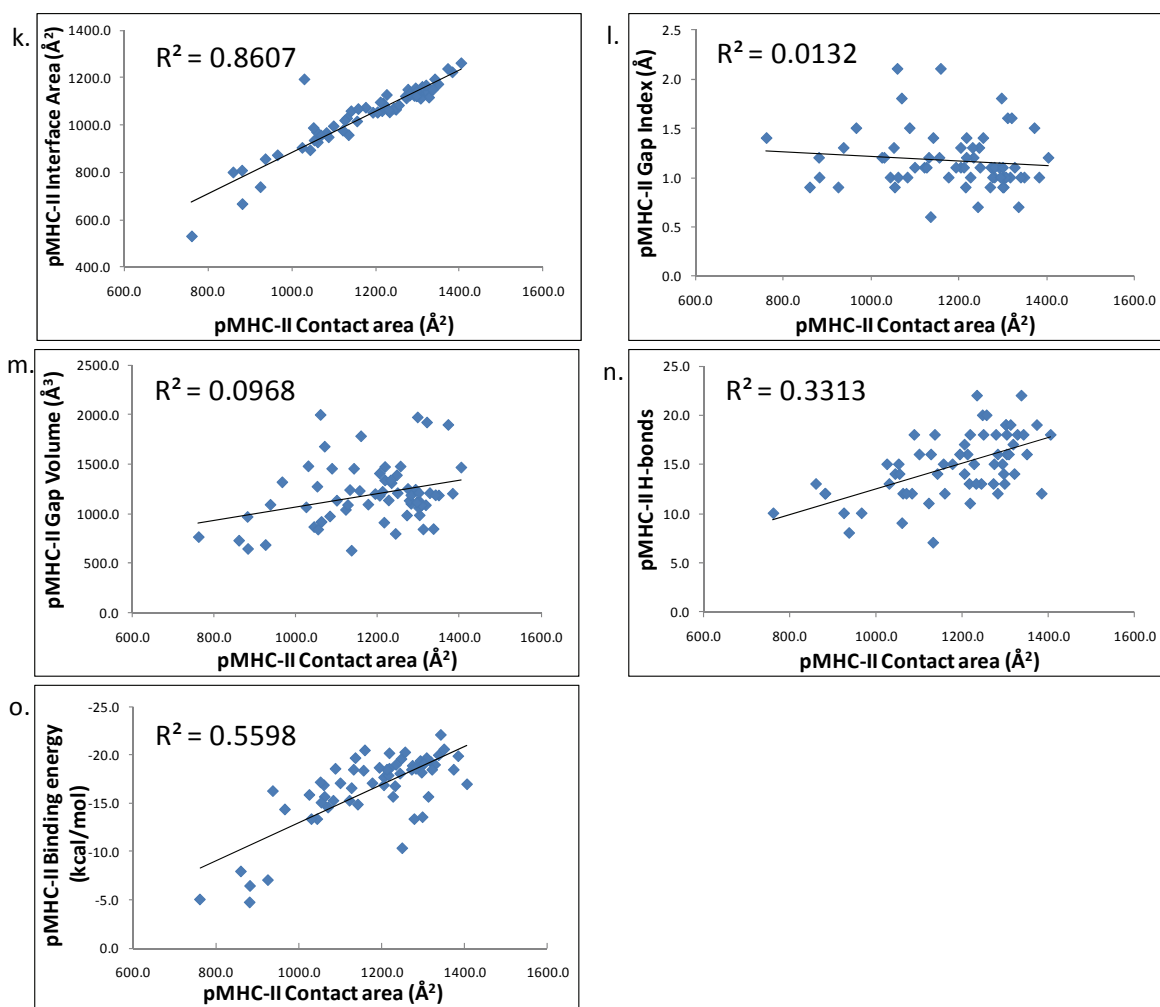
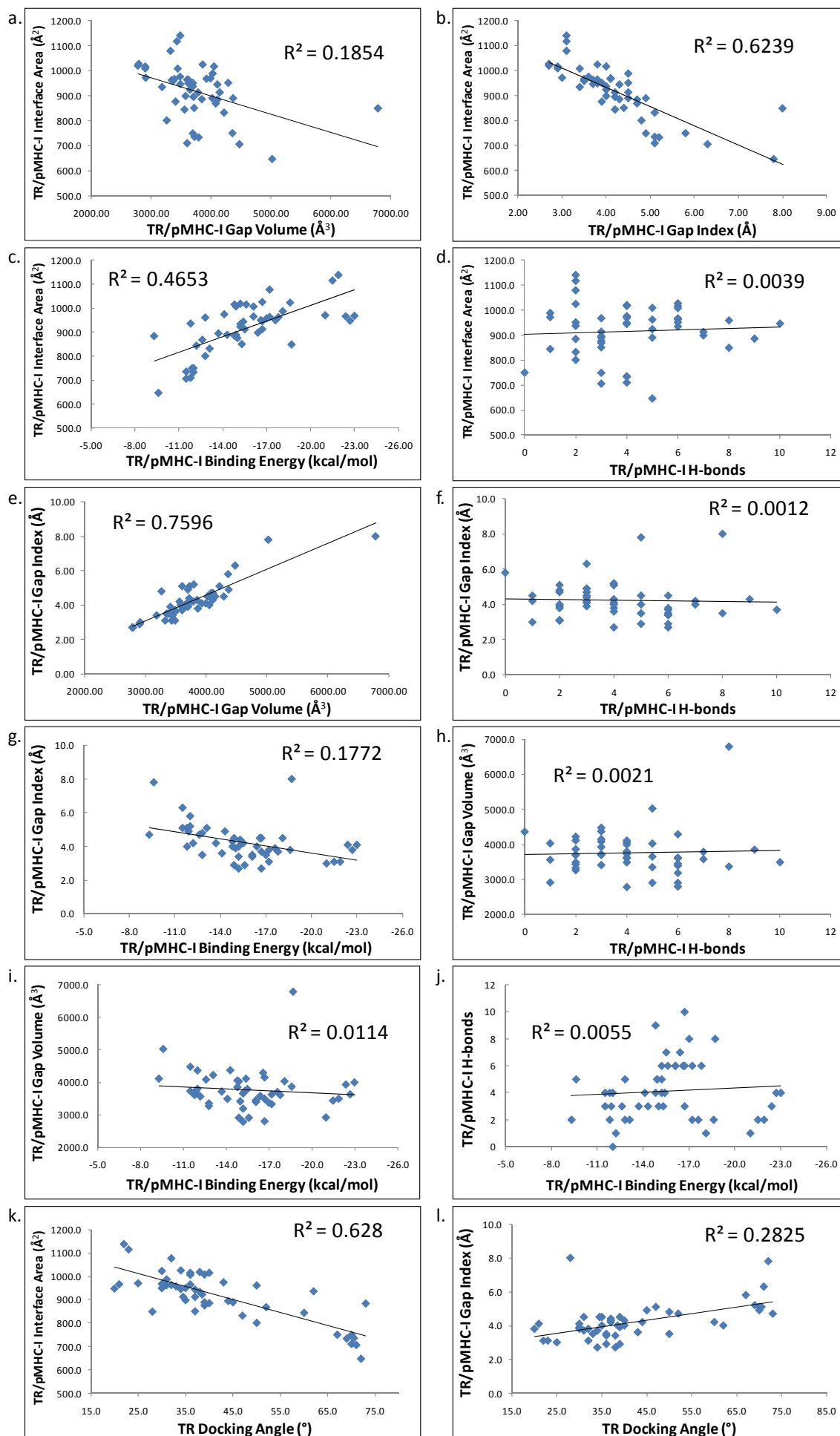


Figure 4.2: A graphical illustration of the correlation between different computed structural interaction parameters for all pMHC-II complexes in MPID-T2. a. pMHC-II interface area vs. pMHC-II gap volume. b. pMHC-II interface area vs. pMHC-II gap index. c. pMHC-II interface area vs. pMHC-II BE. d. pMHC-II interface area vs. pMHC-II H-bonds. e. pMHC-II gap index vs. pMHC-II gap volume. f. pMHC-II gap index vs. pMHC-II H-bonds. g. pMHC-II gap index vs. pMHC-II BE. h. pMHC-II gap volume vs. pMHC-II H-bonds. i. pMHC-II gap volume vs. pMHC-II BE. j. pMHC-II H-bonds vs. pMHC-II BE. k. pMHC-II interface area vs. pMHC-II contact area. l. pMHC-II gap index vs. pMHC-II contact area. m. pMHC-II gap volume vs. pMHC-II contact area. n. pMHC-II H-bonds vs. pMHC-II contact area. o. pMHC-II BE vs. pMHC-II contact area. The respective units are mentioned in the parentheses next to the names of the interaction parameters on the x and y-axes. The corresponding regression coefficients (r^2) are shown within each of the graphs.



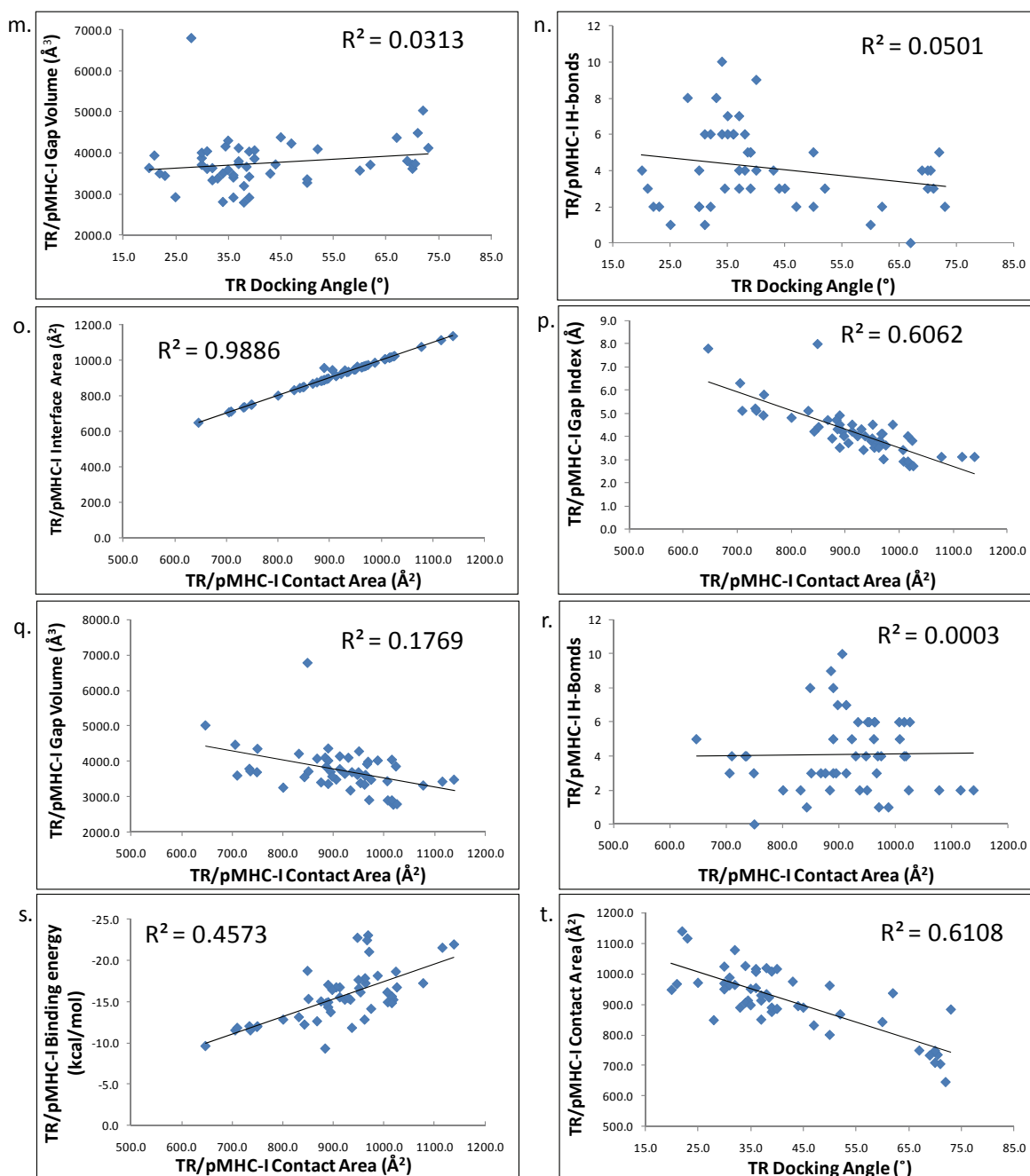
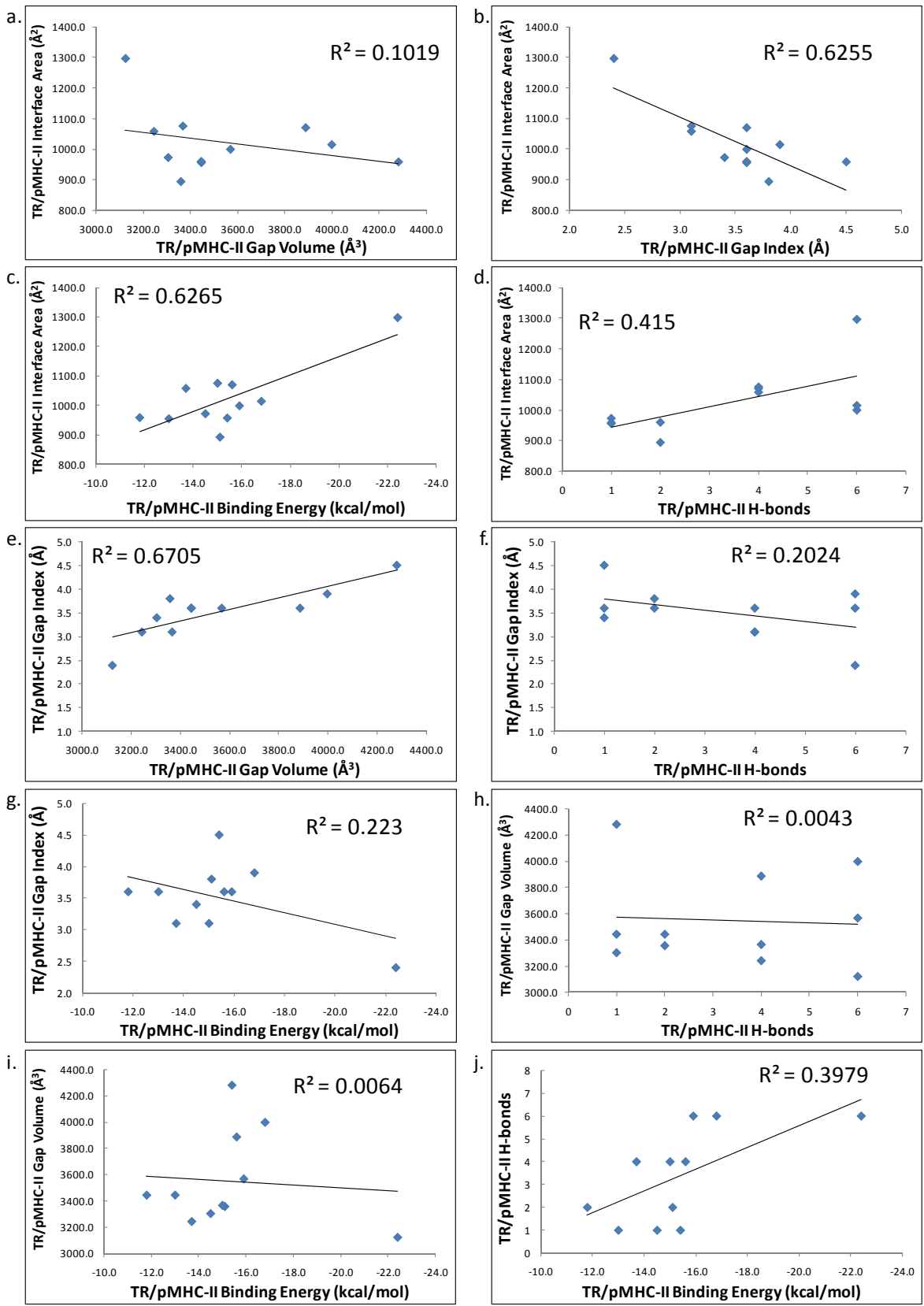


Figure 4.3: A graphical portrayal of the correlation between different computed structural interaction parameters for all TR/pMHC-I complexes in MPID-T2. a. TR/pMHC-I interface area vs. TR/pMHC-I gap volume. b. TR/pMHC-I interface area vs. TR/pMHC-I gap index. c. TR/pMHC-I interface area vs. TR/pMHC-I BE. d. TR/pMHC-I interface area vs. TR/pMHC-I H-bonds. e. TR/pMHC-I gap index vs. TR/pMHC-I gap volume. f. TR/pMHC-I gap index vs. TR/pMHC-I H-bonds. g. TR/pMHC-I gap index vs. TR/pMHC-I BE. h. TR/pMHC-I gap volume vs. TR/pMHC-I H-bonds. i. TR/pMHC-I gap volume vs. TR/pMHC-I BE. j. TR/pMHC-I H-bonds vs. TR/pMHC-I BE. k. TR/pMHC-I interface area vs. TR docking angle. l. TR/pMHC-I gap index vs. TR docking angle. m. TR/pMHC-I gap volume vs. TR docking angle. n. TR/pMHC-I H-bonds vs. TR docking angle. o. TR/pMHC-I interface area vs. TR/pMHC-I contact area. p. TR/pMHC-I gap index vs. TR/pMHC-I contact area. q. TR/pMHC-I gap volume vs. TR/pMHC-I contact area. r. TR/pMHC-I H-bonds vs.

TR/pMHC-I contact area. s. TR/pMHC-I BE vs. TR/pMHC-I contact area. t. TR/pMHC-I contact area vs. TR docking angle. The corresponding regression coefficients (r^2) are shown within each of the graphs. The respective units are mentioned in the parentheses next to the names of the interaction parameters along the x and y-axes.



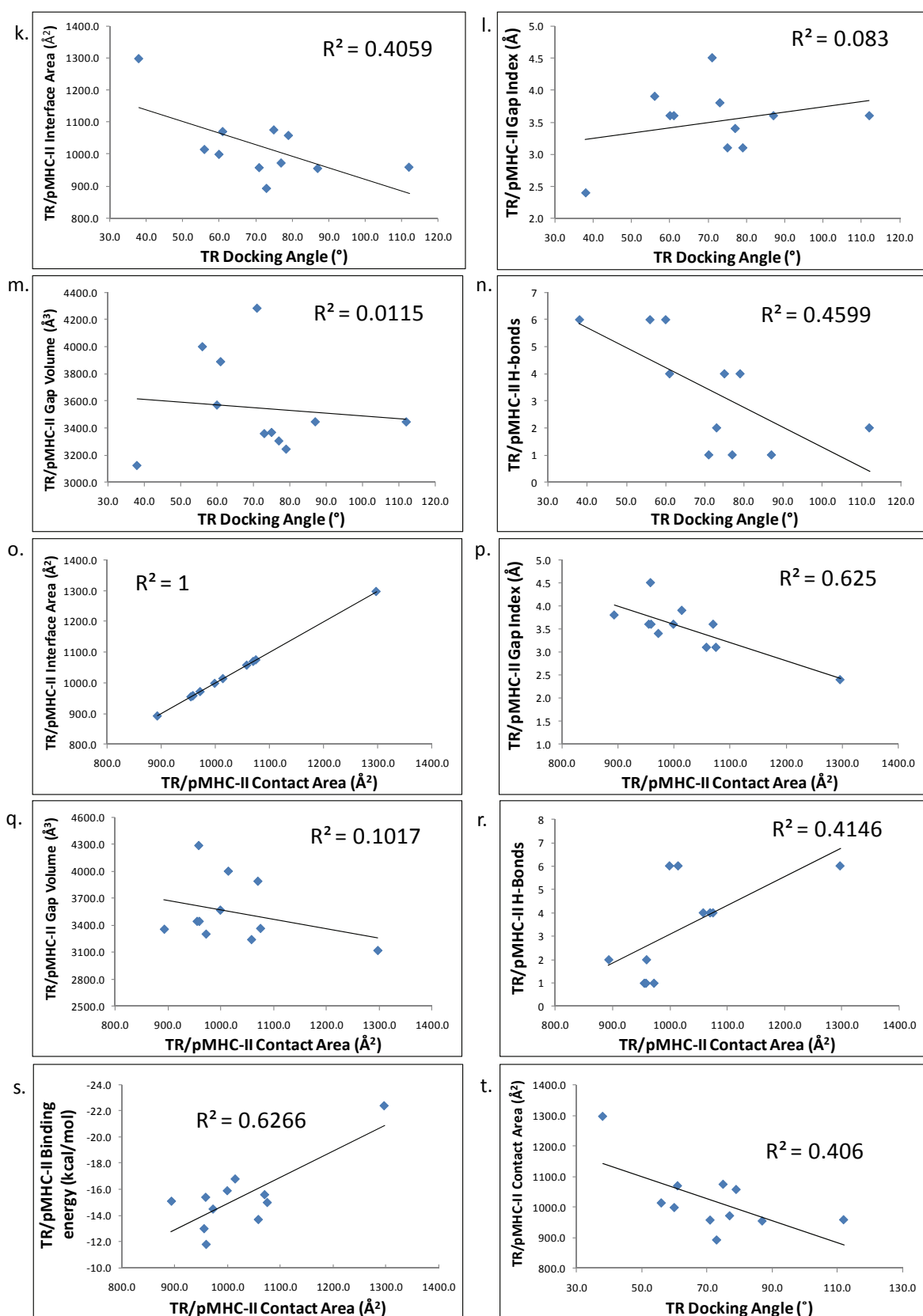


Figure 4.4: A graphical display of the correlation between different computed structural interaction parameters for all TR/pMHC-II complexes in MPID-T2. a. TR/pMHC-II interface area vs. TR/pMHC-II gap volume. b. TR/pMHC-II interface area vs. TR/pMHC-II gap index. c. TR/pMHC-II interface area vs. TR/pMHC-II BE. d. TR/pMHC-II interface area vs. TR/pMHC-II H-bonds. e. TR/pMHC-II gap index vs. TR/pMHC-II gap volume. f.

TR/pMHC-II gap index vs. TR/pMHC-II H-bonds. g. TR/pMHC-II gap index vs. TR/pMHC-II BE. h. TR/pMHC-II gap volume vs. TR/pMHC-II H-bonds. i. TR/pMHC-II gap volume vs. TR/pMHC-II BE. j. TR/pMHC-II H-bonds vs. TR/pMHC-II BE. k. TR/pMHC-II interface area vs. TR docking angle. l. TR/pMHC-II gap index vs. TR docking angle. m. TR/pMHC-II gap volume vs. TR docking angle. n. TR/pMHC-II H-bonds vs. TR docking angle. o. TR/pMHC-II interface area vs. TR/pMHC-II contact area. p. TR/pMHC-II gap index vs. TR/pMHC-II contact area. q. TR/pMHC-II gap volume vs. TR/pMHC-II contact area. r. TR/pMHC-II H-bonds vs. TR/pMHC-II contact area. s. TR/pMHC-II BE vs. TR/pMHC-II contact area. t. TR/pMHC-II contact area vs. TR docking angle. The corresponding regression coefficients (r^2) are shown within each of the graphs. The respective units are mentioned in the parentheses next to the names of the interaction parameters along the x and y-axes.

The correlations between various computed structural interaction parameters for all TR/pMHC-II structures in MPID-T2 are shown in the graphs in Figure 4.4. Extremely poor correlations between a few sets of two interaction parameters are notable. Among these are TR/pMHC-II interface area and TR/pMHC-II gap volume (Figure 4.4a; $r^2=0.1019$), TR/pMHC-II gap volume and TR/pMHC-II H-bonds (Figure 4.4h; $r^2=0.0043$), TR/pMHC-II gap volume and TR/pMHC-II BE (Figure 4.4i; $r^2=0.0064$), TR/pMHC-II gap index and TR docking angle (Figure 4.4l; $r^2=0.083$), TR/pMHC-II gap volume and TR docking angle (Figure 4.4m; $r^2=0.0115$) and, TR/pMHC-II gap volume and TR/pMHC-II contact area (Figure 4.4q; $r^2=0.1017$). Among TR/pMHC-II structures, only three sets with poor correlations are observable. These sets include TR/pMHC-II gap index and TR/pMHC-II H-bonds (Figure 4.4f; $r^2=0.2024$), TR/pMHC-II gap index and TR/pMHC-II BE (Figure 4.4g; $r^2=0.223$) and, TR/pMHC-II H-bonds and TR/pMHC-II BE (Figure 4.4j; $r^2=0.3979$).

Similar to TR/pMHC-I complexes, TR/pMHC-II structures also display several sets of interaction parameters showing good correlations with each other. These sets are TR/pMHC-II interface area and TR/pMHC-II gap index (Figure 4.4b; $r^2=0.6255$), TR/pMHC-II interface area and TR/pMHC-II BE (Figure 4.4c; $r^2=0.6265$), TR/pMHC-II interface area and TR/pMHC-II H-bonds (Figure 4.4d; $r^2=0.415$), TR/pMHC-II gap index and TR/pMHC-II gap volume (Figure 4.4e; $r^2=0.6705$), TR/pMHC-II interface area and TR docking angle (Figure 4.4k; $r^2=0.4059$), TR/pMHC-II H-bonds and TR docking angle (Figure 4.4n; $r^2=0.4599$), TR/pMHC-II gap index and TR/pMHC-II contact area (Figure 4.4p; $r^2=0.625$), TR/pMHC-II H-bonds and TR/pMHC-II contact area (Figure 4.4r; $r^2=0.4146$), TR/pMHC-II BE and TR/pMHC-II contact area (Figure 4.4s; $r^2=0.6266$) and, TR/pMHC-II contact area and TR docking angle (Figure 4.4t; $r^2=0.406$). Just like pMHC-

II structures, TR/pMHC-II structures have only one set of structural descriptors with excellent correlation and this set is TR/pMHC-II interface area and TR/pMHC-II contact area (Figure 4.4o; $r^2=1$).

For all TR/pMHC-II complexes investigated here, the above inferences imply that, just like TR/pMHC-I structures, TR/pMHC-II gap indices share a significant inverse relationship with their interface areas and contact areas (Figures 4.4b and 4.4p), thereby, highlighting the indirect role of geometric complementarity (gap index) in monitoring the BE values of TR/pMHC-II structures and/or stabilizing them (Figure 4.4g). Changes in TR/pMHC-I interface areas and contact areas are directly dependent on the number of intermolecular hydrogen bonds between TR and pMHC-II proteins (Figures 4.4d and 4.4r), which is similar to the relationships noted in pMHC-II complexes and contrary to those seen in TR/pMHC-I structures. This again points out the significant contributions of their interface areas and contact areas to their BE values (Figures 4.4c and 4.4s). Similar to TR/pMHC-I behaviour, the gap volume for TR/pMHC-II structures has little inverse relationships with their contact areas and interface areas (Figures 4.4q and 4.4a), which is different from pMHC-II structures in that their gap volumes share a direct relation with their contact areas and interface areas (Figures 4.2m and 4.2a).

As in the case of pMHC-I, pMHC-II and TR/pMHC-I complexes, a change in TR/pMHC-II gap volume is unlikely to have any effect on the formation of H-bonds between TR and pMHC-II proteins (Figure 4.4h). Surprisingly, their geometric complementarity (gap index) shares a slight inverse proportionality with TR/pMHC H-bonds (Figure 4.4f), indicating that a TR/pMHC-II structure that has a low gap index would have more H-bonds between its TR and pMHC-II proteins. However, this relationship is also attributable to the low number of currently available TR/pMHC-II crystal structures (12 with one structure, PDB code: 2icw [127] having a superantigen mediating TR/pMHC binding, rendering only 11 structures for which structural interaction parameters could be computed) as it is unlike any seen in pMHC-I, pMHC-II and TR/pMHC-I structures. Similarly, the formation of TR/pMHC-II H-bonds is also inversely affected by the measure of their TR docking angles (Figure 4.4n) contrary to what was observed for TR/pMHC-I structures (Figure 4.3n).

As expected and similar to pMHC-II complexes, TR/pMHC-II H-bond formation directly contributes to TR/pMHC-II BE values (Figure 4.4j). Unlike pMHC-II complexes and similar to pMHC-I and TR/pMHC-I structures, the gap volumes of TR/pMHC-II

complexes portray no relationship with their BE values (Figure 4.4i). Similar to the behaviour of TR/pMHC-I structures, an increase or decrease in the measure of TR docking angle for TR/pMHC-II structures is inversely affected by their interface areas and contact areas (Figures 4.4k and 4.4t), which again sheds light on the importance of the inverse linear correlation obtained between their BE values and their TR docking angles (chapter 5). TR/pMHC-II TR docking angles also do not depend on their gap volumes (Figure 4.4m) or their gap indices (Figure 4.4l). As noted for all pMHC-I, pMHC-II and TR/pMHC-I structures, the strong relationships shared by gap volume and gap index and, interface area and contact area (Figures 4.4e and 4.4o) are once again prominent among TR/pMHC-II complexes.

Compared to the averages of interface area (1040.4 \AA^2) and contact area (1181.7 \AA^2) for pMHC-II complexes, the averages of interface area (937.5 \AA^2) and contact area (937.5 \AA^2) for TR/pMHC-II complexes are lower. However, as observed for TR/pMHC-I structures, the steep rise in the mean gap volume (3252.4 \AA^3) is also noted for TR/pMHC-II structures compared to that of pMHC-II structures (1187.6 \AA^3). Just like TR/pMHC-I complexes, the sharp increase in the mean gap volume for TR/pMHC-II complexes resulted in relatively a large average gap index (3.2 \AA) for these structures. Their greater average TR docking angle (65.80°) and the lower mean of the number of H-bonds (3) formed between TR and pMHC-II proteins are attributable to their decreased BE (-14.1 kcal/mol) in comparison to that of TR/pMHC-I structures (-15.6 kcal/mol). This decrease in TR/pMHC-II mean BE value (-14.1 kcal/mol) is also relative to that of the average pMHC-II BE (-16.6 kcal/mol).

4.4 Conclusions

MPID-T2 has been developed with the aim to facilitate mining of fundamental relationships and structural descriptors hidden within TR/pMHC and pMHC interactions for in-depth characterization. The database provides a platform for the scientific fraternity to individually perform structural visualization of the MHC proteins, the bound peptides, pMHC complexes and TR/pMHC complexes, view structural alignment of both pMHC and TR/pMHC complexes (based on species, MHC allele, peptide length and TR type), access other immunology databases such as IMGT/HLA [150-156], IMGT/3Dstructure-DB [57, 58], SYFPEITHI [170, 171] and AntiJen [239] via hyperlinks for more information on each MPID-T2 record, view pre-computed schematic LIGPLOT [445] diagrams that illustrate explicit pMHC and TR/pMHC interactions, view WebLogo [446] consensus patterns among peptides of the same length, species or allele and access many

other useful resources for pMHC and TR/pMHC interactions (through the MPID-T2 help page at: <http://biolinfo.org/mpid-t2/help.html>).

Our understanding of the principles underlying TR/pMHC binding has been enhanced by the inclusion of structural descriptors like BE, molecular surface electrostatic potential (MSEP), TR docking angle and contact area (CA). These descriptors can facilitate rational development of methods to identify strong MHC-binding T cell epitopes with greater propensity to activate T cells by being used as parameters defining pMHC and TR/pMHC interactions, thereby, highlighting the significance of MPID-T2 in vaccine research. MPID-T2 also enables the user to perform TR-specific searches by based on TR types. This is the first such instance of listing computed TR/pMHC interaction characteristics and the first report on correlating different structural interaction parameters for pMHC and TR/pMHC complexes. The analysis of pMHC and TR/pMHC data from MPID-T2 has revealed various patterns for sets of two structural interaction parameters as explained above.

The present analysis suggests that the use of a large standardized set of structural interaction rules may not be applicable for all pMHC and TR/pMHC structures as interaction characteristics vary across pMHC and TR/pMHC complexes. However, a select few structural descriptors show similarities across both the datasets and can be exploited for further studies on pMHC and TR/pMHC interactions. The greater mean values for gap volume and gap index and the lower number of H-bonds formed between TR and pMHC proteins in TR/pMHC structures when compared to those for pMHC complexes, indicate a feeble TR/pMHC binding compared to pMHC binding, as alluded to earlier. Finally, the poor correlation obtained between the number of H-bonds and the BE for pMHC-I and TR/pMHC-I along with greater average TR/pMHC-I BE (-15.6 kcal/mol) compared to that of their mean pMHC-I BE (-13.5 kcal/mol), can all be attributed to the limitations of the currently available computational programs for the analysis of pMHC and TR/pMHC interactions, and present possible complexities that need to be addressed through systematic advancement in the development of computational strategies for more in-depth understanding of these vital adaptive immune system interactions.

Chapter 5: Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one

5.1 Summary

Due to its vital role in adaptive immune responses, it is extremely important to understand the basis for TR/pMHC binding. Although the first TR/pMHC structure was reported one and a half decades ago [7], TR/pMHC interaction is still an enigma. This is mainly due to the complexities of the proteins involved in this association. An in-depth investigation of this critical interaction could help us comprehend the physicochemical principles and the specificities that lie beneath TR/pMHC complex formation and hence, possibly provide clues for better awareness of TR recognition and subsequent T cell activation that triggers the adaptive immune response cascade. Hence, publication 5 explains the analysis of 61 currently available non-redundant TR/pMHC X-ray crystallographic structures collated from the MPID-T2 database (described previously in publication 4 and Chapter 4) using computed BE, TR paratope, pMHC epitope, MSEP and calculated TR docking angle (θ) to comprehend the rationale behind TR/pMHC interaction and to answer two significant questions: (i) whether there are specific energetically equivalent BE “codon” or amino acid positions associated with TR binding angles as suggested by Garcia *et al.*, [8] and; (ii) if the “germline bias” theory really holds good across a large dataset of TR/pMHC structures.

From computed MSEP of pMHC and TR interfaces, the common docking geometry of almost all TR proteins on their respective pMHC binding interfaces is rationally explained. This paper also demonstrates a novel and rational approach for θ calculation, discusses a linear correlation between BE and θ which provides an answer to our first question, highlights the possible reasons for the ability of a TR to scan many pMHC ligands yet specifically bind one, suggests a mechanism for pMHC recognition by TR leading to T cell activation and illustrates the importance of the peptide in determining TR specificity, challenging the “germline bias” theory and providing an answer to our second query. Finally, it also presents valuable new grouping (clustering) system for TR proteins based similarities on their binding site, pMHC recognition and MSEP displayed by their respective interacting pMHC interfaces, suggesting its potential use in the design of peptide based vaccines.

Understanding TR Binding to pMHC Complexes: How Does a TR Scan Many pMHC Complexes yet Preferentially Bind to One

Javed Mohammed Khan¹, Shoba Ranganathan^{1,2*}

¹ Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, Australia, ² Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Abstract

Understanding the basis of the binding of a T cell receptor (TR) to the peptide-MHC (pMHC) complex is essential due to the vital role it plays in adaptive immune response. We describe the use of computed binding (free) energy (BE), TR paratope, pMHC epitope, molecular surface electrostatic potential (MSEP) and calculated TR docking angle (θ) to analyse 61 TR/pMHC crystallographic structures to comprehend TR/pMHC interaction. In doing so, we have successfully demonstrated a novel/rational approach for θ calculation, obtained a linear correlation between BE and θ without any "codon" or amino acid preference, provided an explanation for TR ability to scan many pMHC ligands yet specifically bind one, proposed a mechanism for pMHC recognition by TR leading to T cell activation and illustrated the importance of the peptide in determining TR specificity, challenging the "germline bias" theory.

Citation: Khan JM, Ranganathan S (2011) Understanding TR Binding to pMHC Complexes: How Does a TR Scan Many pMHC Complexes yet Preferentially Bind to One. PLOS ONE 6(2): e17194. doi:10.1371/journal.pone.0017194

Editor: Vladimir Brusic, Dana-Farber Cancer Institute, United States of America

Received: October 8, 2010; **Accepted:** January 22, 2011; **Published:** February 22, 2011

Copyright: © 2011 Khan, Ranganathan. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: shoba.ranganathan@mq.edu.au

Introduction

For maximal immunological protection against a multitude of pathogens, the adaptive immune response in higher jawed vertebrates causes major histocompatibility complexes (MHC) or human leukocyte antigens (HLA) in human, to bind antigenic peptides (p) and present them as peptide-MHC (pMHC) complexes on the surface of antigen-presenting cells (APC), for recognition by T cell receptors (TR) [1]. This TR/pMHC interaction is relatively feeble compared to other important interactions between the molecules of the immune system [2], yet strong enough to trigger TR mediated activation of T cells, thereby eliciting an immediate immune response to either destroy infected cells directly (*via* CD8+ associated cytotoxic T cells) or activate (*via* CD4+ associated helper T cells) other immune system cells like B cells and macrophages to carry out the immune response. More than ten years after the first TR/pMHC structure was reported [3], the interaction between TR and pMHC complexes is still an enigma [4], due in part to the complexities of the molecules involved in this association. The two constant domains (C α and C β) of the TR are linked to variable domains (V α and V β encoded by rearranged variable (V), diversity (D) and joining (J) genes, V-J and V-D-J genes, respectively), whose CDR1, CDR2 and CDR3 loops recognize pMHC [5]. The MHC proteins are composed of two chains, α and β , with the α chain (I-ALPHA) alone forming the peptide-binding groove in MHC class I (MHC-I) proteins, while MHC class II (MHC-II) proteins have both chains α (II-APLHA) and β (II-BETA) forming the peptide binding site [6].

The mechanism responsible for the specificity of the TR/pMHC interactions remains an unsolved problem. The TR "germline bias", in which TR/pMHC binding is independent of the nature of the peptide and MHC restriction or TR specificity is based on specific conserved contacts between TR V (variable) domains and MHC proteins that co-evolve [7], has been proposed as one of the solutions. It however, is not as simple as it sounds. This is due to the mechanisms of combinatorial diversity and N-diversity of the variable domains of TR that create 1012 TR per individual [5], the very high number of MHC alleles and most of all a large number of antigenic peptides. The cross-reactivity of MHC proteins means that the TR briefly scans through several pMHC complexes before actually interacting with a specific one. While this brief scanning by the TR may provide an explanation for the feeble TR/pMHC interactions alluded to earlier, it becomes increasingly important to understand the minute aspects of this vital binding over a broad spectrum of data. Garcia and co-workers [4] have provided highly influential hypotheses using a dataset of 20 TR/pMHC structures, implying that the contacts between CDR1 and CDR2 loops of TR variable domains and MHC helices are germline-encoded leading to the conclusion that TR/pMHC binding is peptide independent. Also inferred in their study is that whatever the TR docking angle, the bound complexes have equivalent binding free energies (ΔG ; referred to here as binding energy (BE) in kcal/mol) at "codon" or amino acid positions A, B and C (as depicted inset of Figure 2b in [4]). Therefore, the main questions we address in this work are: (1) whether there are specific energetically equivalent binding energy "codon" or amino acid positions associated with TR binding angles as suggested by Garcia *et al.*, [4] and; (2) if the "germline bias"

theory really holds good across a large dataset. While addressing these questions, we have also arrived at a possible answer to another lingering question in immunology, *viz.* how can a TR scan through many pMHC complexes and yet specifically bind to one?

We have analyzed the currently available non-redundant dataset of 61 TR/pMHC X-ray crystal structures from MPID-T2 database (<http://biolinfo.org/mpid-t2>) [8], which were originally obtained from the Protein Data Bank (PDB) [9] and verified with IMGT/3Dstructure-DB (<http://www.imgt.org/3Dstructure-DB/>), the reference database for immunoglobulins, T cell receptors and MHC structures [10,11], to determine three major factors that greatly contribute to or influence TR/pMHC binding: (1) binding energy (BE) between TR and pMHC complexes [12–14]; (2) molecular surface electrostatic potential (MSEP) at TR and pMHC interfaces [15,16] and; (3) angle formed by the major axis of TR and the linear axis of the cognate peptide when TR is bound to pMHC (TR docking angle in degrees; herein referred to as ‘ θ ’ when calculated and as ‘diagonal’ when obtained from literature) [4,17]. Using *in vitro* immuno-assays, researchers have previously reported that weak BE between TR and pMHC complexes ascribe weak agonistic (T cell activating) properties to the pMHC complexes and *vice versa* [18–20]. This inference is based on the underlying idea that the strength of TR binding to pMHC plays a vital role in stabilizing the half-life of the TR/pMHC complex, consequently resulting in T cell signalling or activation. This significant finding laid the foundation for us to use BE as a useful parameter in discriminating weak-, moderate- and strong pMHC agonists. MSEP has been used in structure based drug design and in understanding protein-protein interactions by crystallographers for many years [21]. It has also been applied as a successful molecular descriptor for large assemblies of molecules such as microtubules and ribosome [22]. Not only does it include all major aspects of protein-protein interaction, it is also distinctive of molecular shapes. Therefore, we have employed MSEP as an analytical tool to dissect TR/pMHC interactions.

Using computed MSEP of pMHC and TR interacting interfaces we are able to successfully explain the common docking geometry of almost all TR proteins on their respective pMHC binding interfaces. We then discuss a linear correlation between calculated BE and θ , which provides an answer to our first question. A TR paratope (residues on TR interface that contact the pMHC) and pMHC epitope (residues on pMHC interface that contact the TR) analysis, with a focus on conserved residues among pMHC and TR interacting sequence patterns, was conducted in hope of finding certain broadly conserved structural determinants that would constitute the “smoking gun” of “MHC bias” [4]. Finally, we also discuss a new and valuable grouping (clustering) system for TR proteins based on their binding site similarities (from TR paratope analysis), pMHC recognition similarities (from pMHC epitope analysis) and similarities in MSEP displayed by their respective interacting pMHC interfaces (see Methods section for details). The results of MSEP similarity calculation at the pMHC interface along with our TR paratope and pMHC epitope analyses also suggest a weakening of “germline bias” theory over a larger dataset and highlight the significant role played by the peptide in determining TR specificity, thereby, providing an explanation to our second query. Our detailed results are as follows.

Results

BE as a determinant of weak-, moderate- and strong pMHC agonists

It has been reported earlier that lack of enough number of TR/pMHC structures makes differentiation of weak- and moderate-

agonists from strong-agonists or true-agonists from antagonists, almost impossible without immunological assays [15]. However, the availability of a relatively large dataset (61 TR/pMHC structures) together with our comprehensive BE analysis has now made it possible to discriminate strong- from weak- and moderate-agonists for both TR/pMHC-I and TR/pMHC-II structures. Figure 1 shows a plot of the calculated BE between the TR and pMHC-I structures (Figure 1a) and pMHC-II structures (Figure 1b). As seen, this graphical representation gives a clear understanding of the discriminatory power of this analysis. We have computed an overall mean of -15.5 kcal/mol and -15.4 kcal/mol and standard deviation of ± 3.3 kcal/mol and ± 2.7 kcal/mol for TR/pMHC-I and TR/pMHC-II structures, respectively. With cutoffs defined by mean and standard deviation values, we have discriminated weak-, moderate- and strong pMHC agonists. Since BE is also referred to as binding free energy, the highest negative value is considered the best. Among TR/pMHC-I complexes, weak TR agonists have a BE between 0 and -12.2 kcal/mol ($= -15.5 + 3.3$), moderate-agonists (shaded area in Figure 1a) have BE values between -12.2 and -18.8 kcal/mol ($= -15.5 - 3.3$) while strong-agonists gave BE values below -18.8 kcal/mol and are potential T cell activators. TR/pMHC-II structures with a BE between 0 and -12.7 kcal/mol ($= -15.4 + 2.7$) are classified as weak-agonists, complexes with BE between -12.7 and -18.1 kcal/mol ($= -15.4 - 2.7$) are moderate-agonists (shaded area in Figure 1b) and strong-agonists have a BE value below -18.1 kcal/mol and could be more efficient in activating the T cells.

Figure 1a shows a few TR/pMHC-I complexes (PDB codes 1lp9, 2uwe, 2j8u, 2jcc, 3kpr and 3kps in Table S1) having BE values well below -20 kcal/mol, reaching up to -23 kcal/mol. These pMHC ligands are thus very strong-agonists with greater propensity to elucidate T cell activity, concordant with the results obtained from experimental immuno-assays by Miller *et al.* [23], for the pMHC ligands in the PDB structures 2uwe and 2jcc and Macdonald *et al.* [24], for the pMHC ligands in the PDB structures 3kpr and 3kps, respectively. Overall, it was observed that there were 10 (20%) weak-, 34 (68%) moderate- and 6 (12%) strong-binding agonists amongst the TR/pMHC-I complexes. The list of 34 moderate agonists includes pMHC ligands from the PDB structures 2ak4, 2bnr and 2nx5 (Table S1) which have been previously confirmed by cytotoxicity assays [25–27]. Among the 10 weak-agonists is the pMHC from the PDB structure 2ol3, whose lower propensity to elucidate T cell activity was validated by the low level of cytotoxicity observed from cytotoxicity assays by Mazza *et al.* [28]. Similarly, Figure 1b highlights the presence of one such strong-agonist (PDB code 3mbe in Table S1) amongst TR/pMHC-II structures with a BE of -22 kcal/mol. Observations made by Yoshida *et al.* [29], from functional immuno-assays clearly indicate the strong-agonistic and T cell stimulating properties of the pMHC complex in the PDB structure 3mbe. Amidst the 11 TR/pMHC-II complexes, our analysis established 1 (~9%) weak-, 9 (~82%) moderate- and 1 (~9%) strong-binding agonist. These results suggest why a very small percentage (9–12% from our results) of peptide antigens that are predicted to be T cell epitopes by computational methodologies can actually elicit T cell response *in vitro* [30].

pMHC interfaces display a ring of charged amino acids for recognition by complementarily charged TR V α and V β domain interfaces

Most TR proteins that recognize pMHC complexes bind on the central regions of G-ALPHA1 and G-ALPHA2 helices (Figure 2a) for pMHC-I and G-ALPHA and G-BETA helices (Figure 2e) for

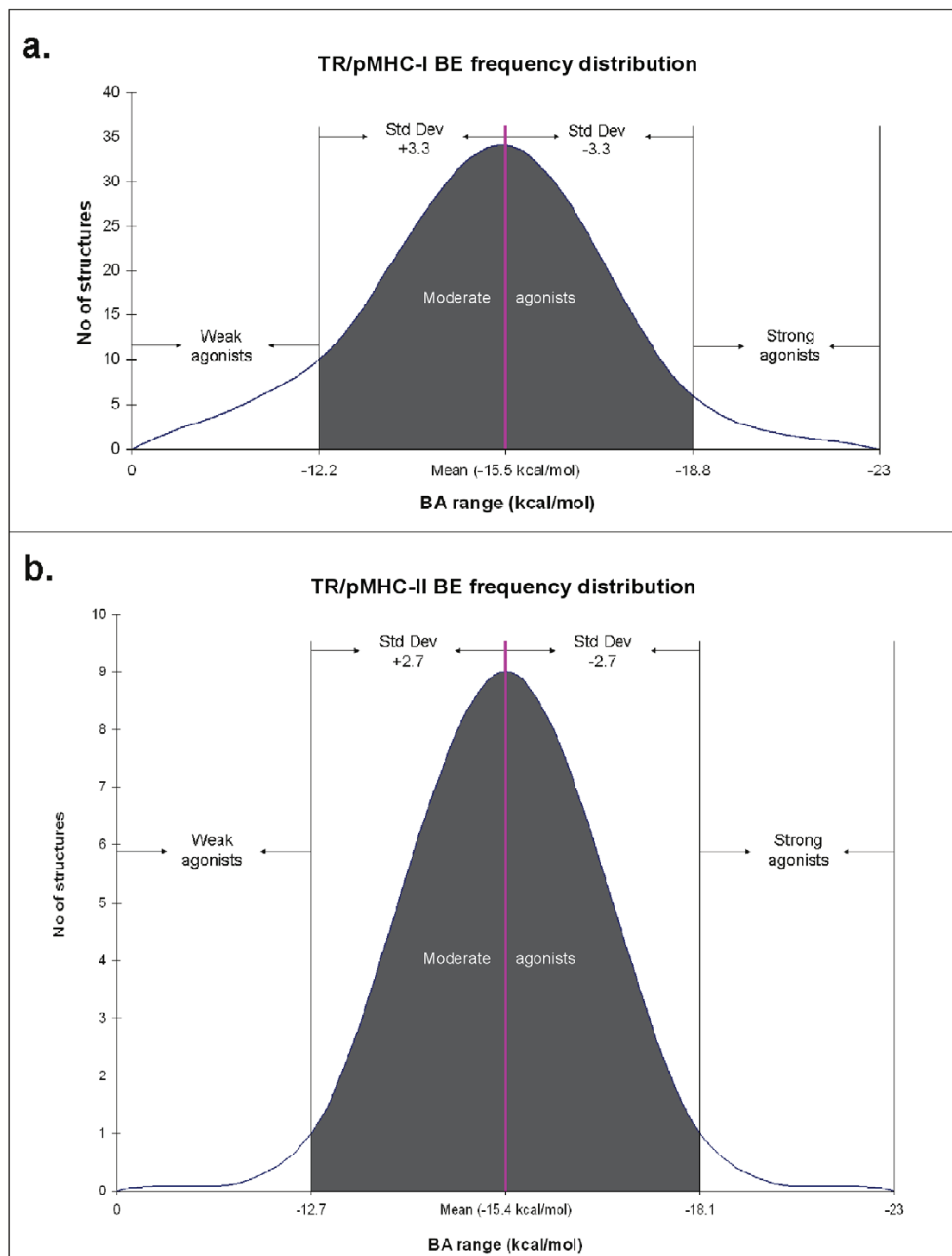


Figure 1. Standard curves for the frequency of computed BE between the TR and pMHC complexes for a. TR/pMHC-I complexes and b. TR/pMHC-II complexes. On the X-axis is the range of BE and on the Y-axis is the number of structures having their BE within these ranges. The pink lines signify the mean BE values. Standard deviation on either side of mean values is represented by shaded area (moderate agonists) in the graphs.
doi:10.1371/journal.pone.0017194.g001

pMHC-II proteins [6]. MSEP displayed by the helices of a pMHC-I (PDB code 2e7l; Figure 2a) and pMHC-II (PDB code 1u3h; Figure 2e) clearly depict a sequential clockwise ring of positively and negatively charged residues on G-ALPHA1 and G-ALPHA2 helices (MHC-I), G-ALPHA and G-BETA helices (MHC-II) which interact with complementarily charged residues

on CDR1 and CDR2 loops of TR α and β variable domains (Figure 2b, f). This was the case in almost all pMHC and TR interacting regions that were analyzed. Interestingly, previous characterization studies on TR/pMHC complexes have revealed molecular interactions along similar regions on the TR and pMHC interfaces [31,32], thereby, supporting our MSEP driven

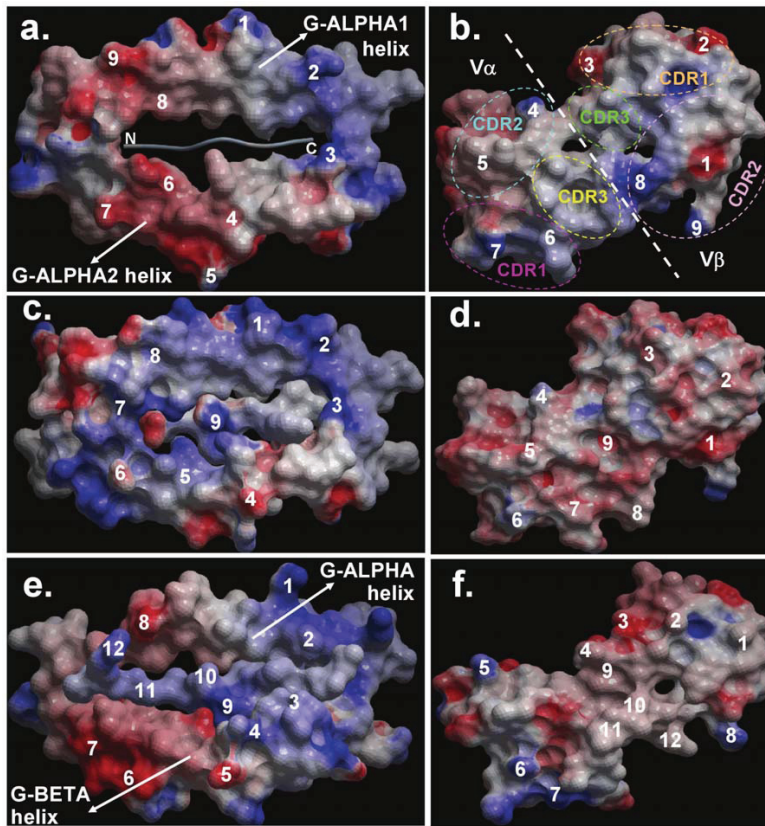


Figure 2. An aerial view of the MSEP displayed by the pMHC interfaces of TR/pMHC-I complexes a. 2e7l (PDB code), c. 1mwa (PDB code) and that of TR/pMHC-II complex e. 1u3h (PDB code) along with b, d, f. their respective contacting TR V α and V β domain interfaces rotated 180° along their interacting axis to visualize their binding interface. The charged residues on the pMHC interfaces are numbered, which interact with the corresponding complementary charges (numbered accordingly) on their respective TR V α and V β domain interfaces. These V α and V β domain interfaces are collectively formed by the CDR1, CDR2 and CDR3 (shown as coloured dotted ovals in b.) loops that interact with the pMHC. The locations of CDR1, 2 and 3 loops in b. are the same for the TR interacting regions in d. and f.
doi:10.1371/journal.pone.0017194.g002

interactions theory. However, in very few pMHC-I cases, such as 1mwa (PDB code), the MHC helices exhibit a ring of mostly positive residues with one/two negative residues on either helix contributing towards TR docking (Figure 2c). In such complexes, the corresponding binding TR interface is almost completely negatively charged, with one/two positive residues on either variable domain (Figure 2d). Across the entire dataset, the positive and negative arrangement seems to be by far more preferred than a ring with predominantly a single charge. It was also observed that negative charges on the two helices of both MHC-I and MHC-II structures occur around the N-termini of bound peptides whereas positive charges are located around their C-termini (Figure 2a, c and e).

A *vice versa* arrangement of charges is seen on TR interacting regions (Figure 2b, d and f). A noteworthy observation is that, MSEP presented by almost all pMHC interfaces are overall similar, suggesting that the ability of a TR to scan through many pMHC interfaces is attributable to the common electrostatic rings displayed on pMHC interfaces. Interestingly, a few, possibly key positions on pMHC interfaces vary in the charges displayed across the entire dataset. This is significant in the context of TR/pMHC interaction because mutating specific charged interacting residues on pMHC interfaces is known to cause increase or decrease in

experimentally determined TR/pMHC binding affinity due to increased or decreased electrostatic interactions between the TR and pMHC leading to an enhanced or reduced T cell response, respectively [29]. As concluded in many earlier studies [16, 20, 28 and 33], our results confirm the importance of peptide in TR/pMHC binding, opposing the notion that TR/pMHC interaction is independent of peptide [4,34]. A proof of this is the fact that various peptides display different combinations of positive and negative residues (Figure 2c and e) which interact with corresponding complementarily charged residues on highly variable CDR3 loops of TR V α and V β domains (Figure 2d and f). Thus, the most variable regions of TR (CDR3) are positioned in the center of binding interface where they contact the peptide, whereas the more conserved regions of TR (CDR1 and CDR2) and the tops of MHC helices engage in contacts that surround the central CDR3-peptide region like a “gasket” [4]. Therefore, MHC helices along with bound peptides, present a set of electrostatic charges that are recognised by specific TR domains.

However, these surfaces should also not be too highly charged or they would bind other counter-ions that may need to be removed and hence might compete with TR for interaction. To support our theory, some short-(salt bridges) to long range (>4 Å

distance) electrostatic interactions have been found in TR/pMHC crystal structures. For example, between the D10 TR V α residue Lys68 (IMGT unique numbering {referred to as IMGT} 82; [35]) and murine MHC-II (I-Ak) G-BETA residue Asp76 (IMGT 72) in the PDB structure 1d9k [36] or between the A6 (PDB code 1ao7; [3]), B7 (PDB code 1bd2; [37]) and 2C (PDB code 2ckb; [38]) TR V α residue Lys68 (IMGT 82) and the murine/human MHC-I (H2-Kb/HLA-A2) G-ALPHA2 residue Glu166 (IMGT 76) [6,39]. Amongst other examples, are the electrostatic interactions between Glu52 (IMGT 63) residue of V β CDR2 loop and Arg79 (IMGT 79) residue of HLA-B8 in TR/pMHC-I complex LC13/EBV/HLA-B8 (PDB code 1mi5; [40]) and the interactions between the human MHC-II (HLA-DR1 and HLA-DR4; PDB codes 1fyt and 1j8h, respectively) G-ALPHA residue Lys39 (IMGT 43) (in a loop projecting up and away from the floor of β -sheet that forms the base of MHC binding groove) and the V β residue Glu56 (IMGT 67) of HA1.7 TR [16,41]. A recent molecular modeling study proved that a single point mutation (G95R; IMGT 107) in V β CDR3 loop of 2C TR increased its affinity to QL9/Ld pMHC by a factor of 1000. This, they suggest, is most likely due to direct electrostatic interaction of Arg95 side chain with an Asp8 (IMGT 8) residue in the QL9 peptide nonamer [42]. Thus, electrostatic effects can work at a distance [43], especially for orienting purposes, so their role in orienting TR relative to pMHC at an early stage during antigen recognition is vital.

It has been reported earlier that diagonal angle of TR docking on pMHC varies between 22°–71° spanning a range of about 50° [17]. Charges displayed on MHC helices, when considered together, seem to present themselves at an angle. Utilizing the location of these charges, we have computed the corresponding TR docking angle (θ) on each pMHC interface (see Methods section for details). Our TR docking angle calculation results show that apart from the PDB structure 1ymm (θ of 112°; Table S1), whose diagonal TR docking angle (110°) has been reported to be of an unusually high value [44], θ varies between 20°–87° over the entire dataset (Figure 3), clearly overlapping the previously reported range of 22°–71° [17] and extending it in both directions. These results provide further evidence for docking of TR onto pMHC interface at an angle such that the TR appears almost “diagonally” [17] attached to the pMHC surface. θ for TR/pMHC-II structures was generally around 72° while for TR/pMHC-I complexes it was 42° on average. We note that when a TR docks onto pMHC interface with a low θ , the area covered by TR paratope on pMHC interface is greater due to the increased number of possible contacts between TR and pMHC interfaces (Figure 4a), therefore, implying that smaller the θ , stronger the binding interaction between TR and pMHC and *vice versa* (Figure 4b). This could possibly be one of the underlying reasons as to why a recent TR-like antibody designing study has yielded a Fab 3M4E5-based “Fab T1” antibody which gives a 20-fold affinity improvement compared to Fab 3M4E5 (PDB code 3hae; [45]) itself and exceeds the affinity of the original TR (1G4; PDB code 2bnr; [26]) by 1,000-fold, thereby, resulting in increased T cell cytotoxic activity [45]. The Fab 3M4E5 antibody (which itself has a 100-fold improvement in affinity compared to the original 1G4 TR [45]) binds the peptide/HLA-A*0201 complex (PDB code 3hae) at an angle of 40° [45] when compared to the diagonal TR docking angle of 69° (θ by our calculations is 39°) for the original 1G4 TR (PDB code 2bnr) [26,45] and it makes more contacts with the pMHC compared to the 1G4 TR causing increased T cell cytotoxicity [45]. These additional interactions are between the A*0201 G-ALPHA2 residue A158 (IMGT 69) and the Fab 3M4E5 VH domain residues G56 & T58 (IMGT 63 and 65), A*0201 G-ALPHA2 residue Y159 (IMGT 70) and Fab

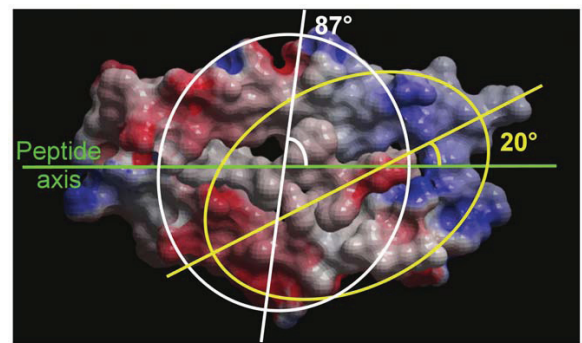


Figure 3. TR docking angle (θ) range computed using charge distribution on pMHC interfaces with reference to the axis of cognate peptide. Charges displayed on pMHC interface are located at an angle (θ) with respect to the axes of linear peptides (green), ranging from 20° (yellow ellipse) to 87° (white ellipse) (spanning 68°) over the entire dataset, which is similar to and overlaps the range of diagonal angles (50°; 22°–71°) for TR docking reported earlier [17]. doi:10.1371/journal.pone.0017194.g003

3M4E5 VH domain residue S57 (IMGT 64), A*0201 G-ALPHA2 domain residue T163 (IMGT 73) and Fab 3M4E5 VH domain residues G55 & S57 (IMGT 62 and 64), A*0201 G-ALPHA2 domain residues E166 & W177 (IMGT 76 and 77) and Fab 3M4E5 VH domain residue S54 (IMGT 59), which cause a change in the angle with which the antibody binds the pMHC complex [45], thereby supporting our hypothesis.

BE is inversely proportional to θ

Utilizing TR BE values computed for pMHC-I and pMHC-II weak-, moderate- and strong agonists and θ calculated using MSEP on their pMHC binding interfaces, we have established a significant correlation between BE and θ , as shown in Figure 5. Evidently, weak-agonists have a higher θ when compared to moderate-agonists and strong-agonists. Strong-agonists have the least θ amongst both TR/pMHC-I and TR/pMHC-II structures. This observation clearly highlights the significance of the derived correlation suggesting that for a given pMHC complex, TR BE is inversely proportional to θ and implying that, lower the θ stronger the binding between pMHC ligand and the respective TR and *vice versa*. Graphs in Figure 5 are explanatory of the above said correlation. Pearson correlation coefficient (r) between BE and θ for TR/pMHC-I complexes is 0.92 with a regression coefficient (r^2) of 0.841. Similarly, for TR/pMHC-II complexes, Pearson correlation coefficient (r) is 0.91 and regression coefficient $r^2 = 0.821$. Interestingly, one TR/pMHC-I structure (1lp9; cyan in Figure 5a) seems to be an outlier from our correlation despite being classified as a strong-agonist. This was primarily owing to the collaborative contribution of the V α CDR1, 2 and 3 loops which bind strongly to the MHC G-ALPHA2 residues 154–167 (IMGT 65–77) and MHC G-ALPHA1 residues 65–69 (IMGT 65–69) [46]. Comparatively, the binding exhibited by V β CDR1 which only binds to the peptide residue F6 (IMGT 6) and V β CDR2 loops that bind to MHC G-ALPHA1 residues 65–72 (IMGT 65–72), respectively, is weak with only V β CDR3 loops binding strongly to MHC G-ALPHA2 residues 146–155 (IMGT 58–66), resulting in an overall greater diagonal TR docking angle [46]. Therefore, the strong binding of V α CDR1, 2, 3 and V β CDR3 loops with MHC G-ALPHA1 and G-ALPHA2 residues coupled with the tilt in the TR paratope due lack of interactions between V β CDR1 and MHC residues and weak interactions

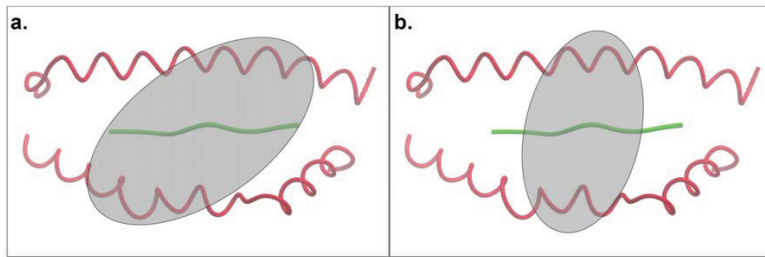


Figure 4. Relationship between θ and area covered by TR paratope on pMHC interface. **a.** Small θ value leading to a large interaction area compared to **b.** Large θ , resulting in a smaller paratope area. pMHC binding interface is shown as C α trace with MHC helices in red and cognate peptide in green. Ellipses represent TR paratopes on pMHC, which are at distinct small and large θ with respect to the axis of bound peptides (angle calculation is shown previously in Figure 3). Shaded regions within the ellipses denote corresponding areas covered by TR paratopes. These areas clearly suggest large and small number of contacts that TR could make with pMHC in **a.** and **b.**, respectively.
doi:10.1371/journal.pone.0017194.g004

between V β CDR2 loops with MHC G-ALPHA1 resulted in our observations of the 1lp9 structure having an overall high TR/pMHC BE and a relatively higher θ value compared to other

strong-agonists. Hence, this outlier was removed from our depicted correlation for TR/pMHC-I structures in Figure 5a. Upon inclusion of the outlier, the Pearson correlation coefficient (r)

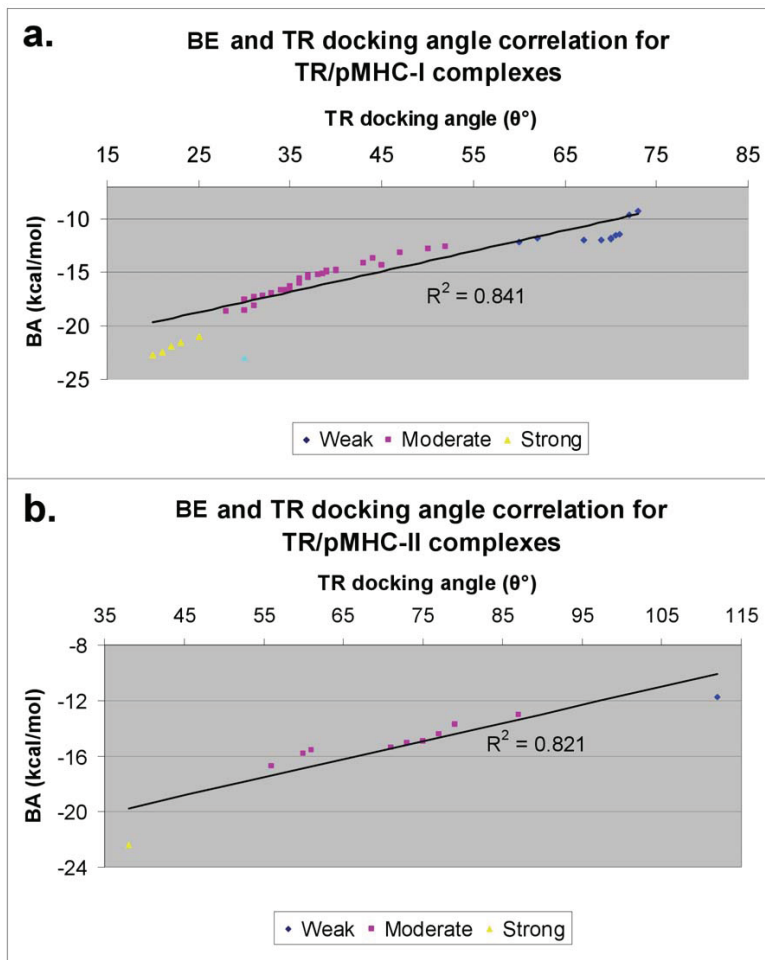


Figure 5. Correlation between BE and θ for a. pMHC-I agonists and b. pMHC-II complexes. The regression coefficients $r^2 = 0.841$ for pMHC-I agonists and $r^2 = 0.821$ for pMHC-II complexes are shown. The single outlier (PDB code 1lp9) in **a.** is highlighted in cyan.
doi:10.1371/journal.pone.0017194.g005

between BE and θ for TR/pMHC-I complexes decreases to 0.90 with a reduced regression coefficient (r^2) of 0.808.

TR paratope and pMHC epitope analyses reveal conserved positions

Residues on TR variable domains that contact the residues on pMHC interface are collectively referred to as “TR paratope”. Similarly, residues on pMHC interface that contact the residues on TR variable domains are collectively termed as “pMHC epitope”. Analyzing TR paratope and pMHC epitope across a wide dataset such as this is an important aspect in our quest to uncover the physicochemical basis of TR specificity and pMHC selectivity. Our results reaffirm the results of Garcia *et al.*, [4] and Rudolf *et al.*, [15] that there were no major conserved contacts observed between TR variable domains and pMHC interfaces over the entire dataset. However, we note that there are sets of pMHC ligands which have strikingly similar, even identical, patterns of interacting residues. Same is the case with TR variable domains which seem to fall into sets which show highly conserved patterns of interacting residues. These sets, along with MSEP based cluster dendrograms (Fig. S1) and heat maps (Fig. S2) for pMHC interfaces obtained from our MSEP analysis, were used to cluster TR proteins (see Methods section for details). This characteristic was prominent in both TR/pMHC-I and TR/pMHC-II sequences.

One, very significant and highly conserved contact was observed on all 11 pMHC-II interfaces. This residue was Gln (Q) 57 (IMGT 65), while Gly (G) 58 (IMGT 66) was mostly conserved on MHC G-ALPHA helix (labeled in Figure 6c). These residues are of utmost importance, as it could be this pair along with a few peptide residues that the TR variable domains could be looking for TR/pMHC complex formation in TR/pMHC-II structures. Amongst TR/pMHC-II complexes, these residues, perhaps serve as an alarm for TR signaling. Besides these conserved residues, we identified several conserved positions on the peptides, G-ALPHA1 and G-ALPHA2 MHC-I helices (Figure 6a), G-ALPHA and G-BETA MHC-II helices (Figure 6c), CDR1, CDR2 and CDR3 loops of respective pMHC-I and pMHC-II binding TR V α and V β domains (Figure 6 b and d). These conserved residues and positions identified are listed in Table 1.

At this stage there are no absolutely conserved residues found in the interacting regions of TR/pMHC-I structures on the whole, but, as said above, there seems to be grouping and a definite pattern of conserved positions on interacting regions of both pMHC and TR, which present different combination of residues according to complementary MSEP displayed on corresponding interacting regions. Therefore, specificity of TR for one pMHC could possibly come from the specific pattern of interacting residues exhibited by that particular pMHC ligand at the above described conserved positions for both pMHC-I and pMHC-II. Based on our observations, we suggest that conserved residues along with residue variations at conserved positions form the basis of TR selectivity and specificity. Hence, these results, together with the common electrostatic rings seen on pMHC interfaces, explain the ability of a TR to survey many pMHC complexes before actually binding to one specific pMHC. Interestingly, number of conserved positions for TR/pMHC-I structures, are less compared to that of TR/pMHC-II structures. One fact that could be attributed to such a result is the small proportion of TR/pMHC-II structures (11) when compared to TR/pMHC-I (50) structures in the current data. Nevertheless, one could easily comprehend that with the increase in number of TR/pMHC-II structures, the number of conserved positions would eventually decrease.

Combining the results from our TR paratope, pMHC epitope and TR docking angle analyses, it is obvious that when a TR docks onto a pMHC binding interface with an overall small θ , the number of contacts between pMHC and TR are greater, thereby, increasing the area covered on pMHC interface by TR V α and V β domains (TR paratope; Figure 4a), compared to the area covered when the TR docks with an overall large θ (Figure 4b), hence proving our earlier inference. This increase or decrease in number of contacts between pMHC and TR according to the decrease and increase in θ , respectively, has a direct consequence on BE between pMHC and TR as shown in the above correlation.

TR grouping is allele and species dependent but TR specificity is peptide dependent

Calculation of MSEP similarities for all pMHC interfaces using webPIPSA server [47] and CLUSTALX [48] multiple sequence alignment of all TR paratopes and pMHC epitopes, have together provided us substantial evidence to define grouping (clustering) among TR proteins (see Methods section for details). These analyses formed the basis of our understanding of TR/pMHC binding and pMHC recognition similarities shown by TR proteins. webPIPSA uses the software R [49] for statistical computing and analytical grouping to produce a dendrogram (Fig. S1) and generate a heat map (Fig. S2). Table S1 portrays a clear clustering amongst TR proteins obtained by summarizing the results of webPIPSA analysis and multiple sequence alignment for TR paratopes and pMHC epitopes. By initial mapping of respective MHC alleles onto cluster dendrograms in Figure S1, it was evident that similarities in MSEP displayed by pMHC interfaces were allele based.

Further investigation by mapping corresponding TR types (names for all TR proteins obtained from the literature) onto cluster dendrograms alongside MHC alleles revealed that many TR proteins bind to same MHC allele which in turn is bound to different peptides (Table S1). This implies that TR specificity is perhaps primarily peptide dependent rather than completely allele dependent, shedding light on the impact of peptide properties in this significant immunological synapse, thus, further enforcing our earlier conclusion and weakening the “TR-MHC germline bias” theory. As seen, there were three clusters identified among pMHC-I binding TR proteins. Cluster I.1 comprises of six different types of TR proteins all of which are known to bind pMHC with murine MHC alleles. Cluster I.2 is made up of eight TR types which behave in a more diverse fashion by binding to pMHC with human alleles other than A*0201. Eight types of TR proteins which recognize pMHC-I with A*0201 allele fall under Cluster I.3. pMHC-II binding TR proteins were segregated into two distinct clusters, where, Cluster II.1 has five types of TR proteins which are associated with murine I-Au, I-Ag7 and I-Ak alleles and Cluster II.2 includes four TR types associated with human DR-alleles. These results are also noted to be species specific since all murine pMHC structures are clustered together implying that all TR types associated with murine MHC alleles are clustered together. This adds another dimension to this significant TR grouping system. It is worth noting that at the TR level the MHC supertype definitions do not apply.

Interestingly, there are multiple PDB structures for a single TR/pMHC complex, showing different TR binding angles, where we have tested the validity of our inverse relationship between calculated BE and θ . 2f54 and 2bnr (PDB code; bold in Table S1) form one such pair. Here, θ for 2f54 was computed to be 36° which is 3° smaller than that of 2bnr (39°). The calculated BE values for the two structures are −15.6 kcal/mol (2f54) and −14.9 kcal/mol (2bnr), respectively, which are inversely related to the θ values. These subtle changes in θ and BE are due to the underlying fact that the side chain

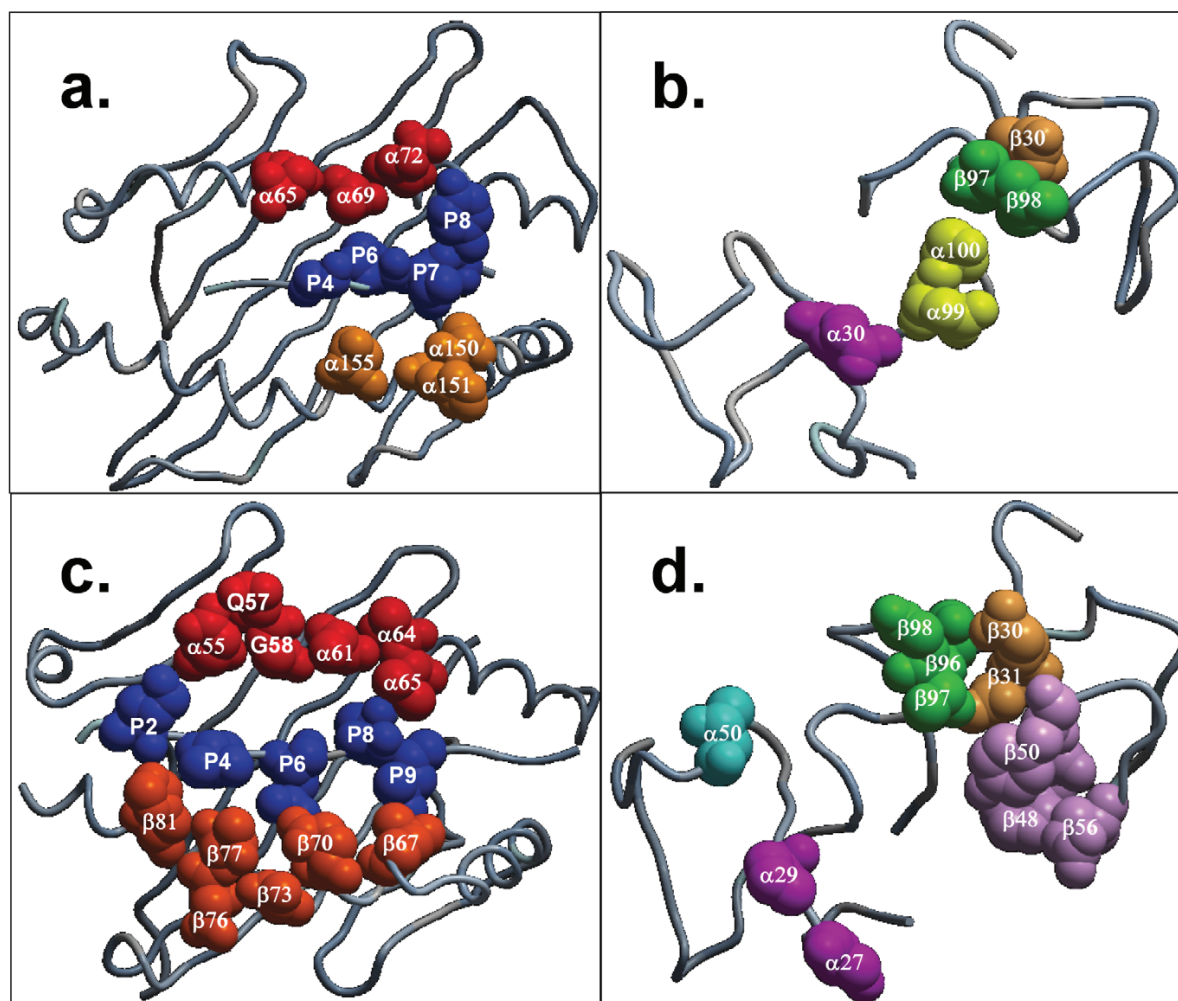


Figure 6. Residue conservation at pMHC and TR interfaces for a. pMHC-I ligands. b. pMHC-I binding TR. c. pMHC-II complexes and d. pMHC-II-binding TR. Conserved residue Q57 (IMGT 65) and mostly conserved residue G58 (IMGT 66) on G-ALPHA helix of pMHC-II interface in c are labelled. Conserved positions are labelled according to their chain locations on pMHC and TR interfaces. Highlighted in red are conserved positions, a conserved residue and a mostly conserved residue on G-ALPHA1 helix of pMHC-I and G-ALPHA helix of pMHC-II interfaces in a. and c., respectively. Conserved positions on G-ALPHA2 helix of pMHC-I in a. are in gold. Residue positions on peptides are in blue and on G-BETA helix of pMHC-II in c. are in orange. Conserved residues and positions in b. and d. are coloured according to their CDR loops as follows: Vα CDR1: pink, CDR2: cyan, CDR3: yellow, Vβ CDR1: pale orange, CDR2: pale pink and CDR3: green. The colouring scheme used for CDR loops is the same used in Figure 2b. Protein backbones are represented as Cα trace in grey.
doi:10.1371/journal.pone.0017194.g006

of Q155 (IMGT 66) residue from MHC G-ALPHA2 domain forms a hydrogen bond with the side chain of TR Vα residue S51 (IMGT 58) in 2f54 [50] resulting in a well ordered Q155 (IMGT 66) side chain, when compared to its relatively disordered side chain orientation due to hydrogen bond formation with the side chain of TR Vα residue T95 (IMGT 109) in 2bnr [26]. Similarly, 2vlj, 2vll and 1oga (bold and italics in Table S1) represent the same TR/pMHC complex, with different TR docking orientations. Compared to that of 1oga (69°; [17]), the diagonal TR docking angles for 2vlj and 2vll are reported to be roughly up to 5° larger [51], whereas our computed θ values are 1° and 1.5° larger than both the diagonal TR docking angle and the computed θ value for 1oga (69°), respectively. Their respective calculated BE values are −11.7 kcal/mol (2vlj), −11.4 kcal/mol (2vll) and −11.9 kcal/mol (1oga), which are in

accord with our computed θ values and the diagonal TR docking angles reported. Yet again, the core residues involved in TR/pMHC interaction are conserved in all three of these structures and slight variations in θ and BE are a direct consequence of the subtle positional changes accommodated by the peripheral residues at the binding interface through regulations in their side chain conformations [51]. These are mainly MHC G-ALPHA1 residue Q72 (IMGT 72), MHC G-ALPHA2 domain residue Q155 (IMGT 66) and the TR Vβ residue I53 (IMGT 58) [51].

Discussion

We have analyzed available TR/pMHC structures using a number of physicochemical characteristics to understand any basic

Table 1. List of conserved residues and positions.

MHC Class	Structural Location	Loop	Conserved Residues	Conserved Positions
I	MHC G-ALPHA1 helix	-	-	$\alpha 65$, $\alpha 69$ and $\alpha 72$
	MHC G-ALPHA2 helix	-	-	$\alpha 150$, $\alpha 151$ and $\alpha 155$
	Peptide	-	-	P4, P6, P7 and P8
	TR V α	CDR1	-	$\alpha 30$
		CDR2	-	-
		CDR3	-	$\alpha 99$ and $\alpha 100$
	TR V β	CDR1	-	$\beta 30$
		CDR2	-	-
		CDR3	-	$\beta 97$ and $\beta 98$
II	MHC G-ALPHA helix	-	Q57 and G58 (mostly conserved)	$\alpha 61$, $\alpha 64$ and $\alpha 65$
	MHC G-BETA helix	-	-	$\beta 67$, $\beta 70$, $\beta 73$, $\beta 76$, $\beta 77$ and $\beta 81$
	Peptide	-	-	P2, P4, P6, P8 and P9
	TR V α	CDR1	-	$\alpha 27$, $\alpha 29$
		CDR2	-	$\alpha 50$
		CDR3	-	-
	TR V β	CDR1	-	$\beta 30$ and $\beta 31$
		CDR2	-	$\beta 48$, $\beta 50$ and $\beta 56$
		CDR3	-	$\beta 96$, $\beta 97$ and $\beta 98$

doi:10.1371/journal.pone.0017194.t001

differences between pMHC-I and pMHC-II interactions with TR. The avidity of TR/pMHC interaction has been classified as weak-, moderate-, and strong-, based on the BE values that were computed for pMHC and TR binding interfaces. Using BE as a discriminator between weak-, moderate- and strong-agonists will add value to prediction methods enabling them to successfully predict true T cell epitopes or strong-agonists that are highly likely to initiate T cell response. Also, it would be interesting to decompose BE into electrostatic and van der Waals components to get an insight into the energetic contributions and correlate these with the differing amino acids at the TR and pMHC interfaces. We have also proposed a novel and rational approach to computing θ value by mapping charged rings formed from MSEP on the pMHC interface. Here, we note from literature that, although for some TR/pMHC crystal structures the entire TR paratope is used to calculate the diagonal TR docking angle [17], using the central mass of TR V α and V β domains as a reference to draw an axis [46,52] that cuts the cognate peptide axis at an angle (generally much greater than the angle obtained by using the entire paratope) appears to be the common practice of diagonal TR docking angle calculation for most crystal structures. Hence, the fact that we employ TR paratope, pMHC epitope and MSEP at pMHC interfaces to procure the θ values, could be the fundamental reason for our θ values being extremely close or fairly distant to the diagonal TR docking angles reported for some structures (Table S1). Results from our MSEP analysis explain the common TR docking geometry on pMHC interface, seen in all TR/pMHC structures. None of the structures available to us for analysis has a glycan molecule at or near the TR/pMHC interface. However, some of these molecules have a glycan shield around them which may also contribute towards docking by excluding certain modes of binding and helping in orientation of TR [53]. This is a possible complexity that needs to be factored in as more data becomes available. Using MSEP in epitope prediction methods could further accelerate the progress of

structure-based prediction techniques besides minimizing false positives and true negatives from actual agonistic peptides in a given set of peptide antigens. We have reported a strong correlation between BE values and θ across the entire dataset which solves the first query addressed in this manuscript (described earlier in Introduction section). Analysis of TR paratopes and pMHC epitopes revealed that although there are no absolutely conserved residues found in interacting regions of both TR and pMHC ligands, there are vital conserved positions on both interfaces across TR/pMHC-I and TR/pMHC-II structures that could have fundamental implication for peptide vaccine design. Identification of conserved residues/positions on pMHC and TR interacting regions provides clues to the positional specificity of TR proteins. Furthermore, we have clustered TR proteins based on their binding site similarities, pMHC recognition similarities and similarities in MSEP on their respective interacting pMHC interfaces, to dissect TR/pMHC binding requirements. MSEP similarity calculation at the pMHC interface together with TR paratope and pMHC epitope analyses have thus given us enough evidence to suggest a weakening of “germline bias” theory over a larger dataset and highlight the significant role played by the peptide in determining TR specificity, thereby, answering our second question (see Introduction section for details).

Based on our findings, we wish to propose a mechanism for TR/pMHC binding and TR activation which explains the phenomenon of pMHC recognition by TR and TR specificity simultaneously. We suggest that, after peptide binding to MHC, many similar pMHC complexes are presented on the cell surface which exhibit similar charged rings of MSEP (explained earlier in the results of our TR and pMHC interface MSEP analysis) thereby signalling or attracting the TR towards them through long-range electrostatic steering. Due to their electrostatic similarity, the TR actually surveys many pMHC complexes. This is possible by temporary interactions between the rings of charged residues displayed on MHC helices and on CDR1 and CDR2

loops of TR V α and V β domains. This phenomenon is followed by the recognition of specific arrangements of pMHC residues (at conserved positions) by CDR3 loops. Once this recognition occurs, the TR localizes itself on the pMHC such that the half-life of TR/pMHC complex is sufficiently stabilized for T cell activation. Therefore, the entire process of pMHC recognition and TR signalling is possibly governed by two factors, the electrostatic ring displayed by pMHC interface and a specific arrangement of residues presented by pMHC.

From our extensive studies on TR/pMHC interactions we have defined structural features that can be analyzed as parameters governing TR/pMHC complex formation relevant for immune system activation. These parameters are MSEP of TR and pMHC interfaces and TR docking angle (θ), which, when coupled with the knowledge of specific arrangement of residues at conserved positions on TR and pMHC interfaces, could be used as discriminants for *in silico* identification of strong-agonistic pMHC complexes. Results of these analyses could be used to develop and/or enhance methods to successfully predict T cell epitopes in accordance with their MHC and TR binding specificities. This could greatly improve the efficacy of T cell epitope prediction models in separating true T cell epitopes from a large number of predicted MHC-binding peptides. This kind of structure-based screening helps overcome the barriers of insufficient training data and lack of peptide binding motifs, especially for MHC-II alleles, thereby cutting down the lead time involved in experimental vaccine development methods, resulting in production of effective and highly specific peptide vaccines with a wide population coverage. Our results will facilitate the rational development of peptide vaccines, capable of eliciting T cell response, for immunotherapies to protect against or combat infectious, autoimmune, allergic and graft *vs.* host diseases.

Methods

Data

The data used in this study comprises of 61 non-redundant TR/pMHC structures from the MPID-T2 database (<http://biolinfo.org/mpid-t2>) [8], which were originally obtained from the Protein Data Bank (PDB) [9] and verified with the IMGT/3Dstructure-DB (<http://www.imgt.org/3Dstructure-DB/>) database [10,11]. The PDB structure 2icw was not included in this study as it has a superantigen between the TR and the pMHC which prevents actual TR/pMHC interaction by mediating the TR/pMHC binding [54]. Out of the 61 structures, 50 were MHC-I complexes spanning 9 alleles from human (7) and mouse (2) and 11 MHC-II complexes spanning 7 alleles, again from human (4) and mouse (3). When there is more than one structure with the same peptide sequence, MHC allele and TR type, mutations in the MHC α (I-ALPHA) chain (MHC-I), TR V α and V β CDR2 & 3 loops and the degree of tilt or relative change (compared to the first structure with similar TR type, MHC allele and peptide sequence in Table S1) in θ were taken into account as primary criteria to consider the structures non-redundant. Coordinates for truncated versions of the X-ray structures, encompassing single structural complexes of the pMHC binding interfaces and the variable domains of the TR were extracted for TR paratope, pMHC epitope analyses and MSEP calculations.

BE calculation

The interaction of most ligands with their binding sites can be characterized in terms of binding free energy or binding energy (BE). In general, high energy TR/pMHC binding results from greater intermolecular force between the pMHC and its TR while low energy ligand binding involves less intermolecular force

between the pMHC and its TR. High energy binding involves a longer residence time for the TR on its respective pMHC than in the case of low energy binding. High energy binding of pMHC to a TR is often physiologically important as some of the BE can be used to cause a conformational change in the TR, resulting in a physiological response or T cell response [55,56]. Since BE is also referred to as binding free energy, the most negative value is considered the best. In literature, BE (ΔG) is usually derived from the binding constants of the interaction such as K_d and K_a .

The general thermodynamic formulae used are as follows:

$$\Delta G = RT \ln K_d \quad (1)$$

$$K_d = 1/K_a \quad (2)$$

where K_d is the dissociation constant, R is the universal gas constant, T is the absolute temperature and K_a is the association constant. BE values between the pMHC and TR for all TR/pMHC structures were calculated using the program DCOMPLEX [57], which uses DFIRE-based potentials [58]. The program first calculates the total atom-atom potential of mean force, G , for each structure, which is given by:

$$G = \frac{1}{2} \sum_{ij} \bar{u}(i_j, r_{ij}) \quad (3)$$

where \bar{u} is the atom-atom potential of mean force between two atoms, i and j that are a distance r apart, the summation is over atomic pairs that are not in the same residue and a factor of $1/2$ is used to avoid double-counting of residue-residue and atom-atom interactions [57].

The binding free energy between two interacting proteins A and B can also be obtained by using:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_A + G_B) \quad (4)$$

where A and B are considered as two rigid bodies whose interface residues contribute most to ΔG_{bind} [57]. Therefore, the final equation used by DCOMPLEX [57] to calculate BE is as follows:

$$\Delta G_{\text{bind}} = \frac{1}{2} \sum_{ij}^{\text{interface}} \bar{u}(i_j, r_{ij}) \quad (5)$$

DCOMPLEX provides an overall BE, without details of specific components for electrostatic, van der Waals, hydrophobic and entropic terms.

MSEP similarity calculation

MSEP in proteins is a result of charged side chains of the amino acid residues and bound ions. These potentials play a vital role in protein folding, stability, enzyme catalysis and specific protein-protein recognitions. MSEP similarity between any two protein molecules is a measure of the similarity in their composition of charged residues. Interactions between the TR and pMHC in all the structures depend vastly on the charges that the binding site on the pMHC displays. Thus, the web server webPIPSA [47] was used to calculate the MSEP and compare the electrostatic interaction properties of only the pMHC binding interfaces in all the structures. The algorithm begins with calculation of the protein MSEP and then calculates similarity indices for all pairs of proteins based on the electrostatic similarity.

The similarity indices are then converted to electrostatic distances which are then displayed as a colour coded matrix called as the heat map (Fig. S2) and as a tree or a cluster dendrogram (Fig. S1). These cluster dendrograms and heat maps were consequently used for TR clustering (described below). Structural models of only the pMHC interfaces were used for this analysis. ICM [59,60] was then used to visually analyse the electrostatic images of all the structures.

Calculation of TR docking angles (θ)

Similarly, we generated and visualized electrostatic images of the TR binding interfaces ($V\alpha$ and $V\beta$ domains). The respective pMHC and TR interfaces were then matched for complementarities of charges and the corresponding charges were numbered accordingly on both the interfaces (Figure 2). These charged residues were cross verified with the list of pMHC and TR interacting residues collated for TR paratope and pMHC epitope residue conservation analyses. The charged residues missing from these lists were omitted and the charges were renumbered for consistency in results. A line was drawn which connects the numbers on each of the pMHC interfaces using ICM [59,60]. Once connected, the numbers on a given pMHC interface formed an ellipsoidal shape, which determines the TR paratope on the pMHC (Figure 3). These ellipses were noticed to be at a certain angle with respect to the $C\alpha$ backbone axes of the respective cognate peptides across the entire dataset. Finally, straight lines were drawn diagonally across the ellipses which cut the axes of the bound peptides at a given angle (Figure 3). These angles were measured using ICM [59,60] and are called TR docking angle (θ) on the pMHC interfaces (Figure 3).

TR paratope and pMHC epitope residue conservation analyses

These analyses required us to manually extrapolate and list the interacting residues of the pMHC and TR for each structure either from the literature or by using ICM [59,60] computer program. CLUSTALX [48] was later used to perform multiple sequence alignment in the hope of identifying any conserved patterns in the interacting residues of pMHC and TR interfaces.

TR grouping

Initially, the sets of pMHC and TR interfaces, obtained from our TR paratope and pMHC epitope residue conservation analyses, showing similar pattern of interacting residues (mentioned earlier in the Results section), were matched against the cluster dendrograms (Fig. S1) and heat maps (Fig. S2), to verify if the structures that display the sets observed in residue conservation analyses, are present within distinct clusters of pMHC complexes (Fig. S1 and S2). After this confirmation, the respective MHC alleles and corresponding TR types were mapped onto the cluster dendrograms which clearly indicated the grouping (clustering) amongst the TR molecules based

on similarities in their binding site, pMHC recognition properties and MSEP displayed on their respective interacting pMHC interfaces.

Supporting Information

Table S1 Grouping of TR proteins. Mutations in MHC α (I-ALPHA) chain and TR $V\beta$ domain (MHC-I; TR Cluster I.2 and I.3), TR mutant names and the degree of tilt or relative change (compared to the first structure with similar TR type, MHC allele and peptide sequence) in θ are mentioned in parentheses (see Methods section for details). (PDF)

Figure S1 Cluster dendrograms for all pMHC interfaces based on their MSEP similarities. a. pMHC-I complexes clustered into three distinct clusters. b. pMHC-II ligands clustered into two distinct clusters. Each pMHC interface is denoted by its corresponding PDB code. Every pMHC is mapped onto its respective MHC allele and the interacting TR type (TR name). This clearly indicates the clustering amongst the TR proteins. The three distinct clusters of pMHC-I binding TR proteins are coloured yellow: cluster I.1, green: cluster I.2 and orange: cluster I.3. The two clusters amongst pMHC-II binding TR proteins are highlighted in light blue: cluster II.1 and lavender: cluster II.2. TR grouping (clustering) is in accordance with Table S1. (PDF)

Figure S2 Heat maps for all pMHC interfaces based on the calculated MSEP values depicted as a colour coded matrix showing clustering amongst pMHC complexes in a reverse order as compared to the cluster dendrograms in Figure S1. a. pMHC-I complexes clustered into three. b. pMHC-II structures in two distinct clusters. Each pMHC interface is again denoted by its corresponding PDB code. Inset, are the legends showing the color key used to create heat matrices and the MSEP value ranges for pMHC interfaces. Also shown is the formula used to calculate electrostatic distances for clustering. (PDF)

Acknowledgments

JMK gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship. We also thank Mr. H.R. Cheruku, Macquarie University, Sydney for assistance. Open access publication charges were borne by Macquarie University.

Author Contributions

Conceived and designed the experiments: SR. Performed the experiments: JMK. Analyzed the data: JMK SR. Contributed reagents/materials/analysis tools: JMK SR. Wrote the paper: JMK SR.

References

- Mueller DL (2010) Mechanisms maintaining peripheral tolerance. *Nat Immunol* 11: 21–27.
- Lo WL, Felix NJ, Walters JJ, Rohrs H, Gross ML, et al. (2009) An endogenous peptide positively selects and augments the activation and survival of peripheral CD4⁺ T cells. *Nat Immunol* 10: 1155–1161.
- Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, et al. (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* 384: 134–141.
- Garcia KC, Adams JJ, Feng D, Ely LK (2009) The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol* 10: 143–147.
- Lefranc MP, Lefranc G (2001) The T cell receptor FactsBook. San Diego: Academic Press.
- Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriot A, et al. (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol* 29: 917–938.
- Jerne NK (2004) The somatic generation of immune recognition. 1971. *Eur J Immunol* 34: 1234–1242.
- Khan JM, Ranganathan S (2010) pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res* 6(Suppl 1): S2.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242.
- Kaas Q, Ruiz M, Lefranc MP (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* 32: D208–210.
- Ehrenmann F, Kaas Q, Lefranc MP (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* 38: D301–307.
- Armstrong KM, Insaioo FK, Baker BM (2008) Thermodynamics of T-cell receptor-peptide/MHC interactions: progress and opportunities. *J Mol Recognit* 21: 275–287.

13. Hulsmeijer M, Chames P, Hillig RC, Stanfield RL, Held G, et al. (2005) A major histocompatibility complex-peptide-restricted antibody and T cell receptor molecules recognize their target by distinct binding modes: crystal structure of human leukocyte antigen (HLA)-A1-MAGE-A1 in complex with FAB-HYB3. *J Biol Chem* 280: 2972–2980.
14. Reiser JB, Darnault C, Guimezanes A, Gregoire C, Mosser T, et al. (2000) Crystal structure of a T cell receptor bound to an allogeneic MHC molecule. *Nat Immunol* 1: 291–297.
15. Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* 24: 419–466.
16. Hennecke J, Carfi A, Wiley DC (2000) Structure of a covalently stabilized complex of a human alpha/beta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1. *EMBO J* 19: 5611–5624.
17. Stewart-Jones GB, McMichael AJ, Bell JI, Stuart DI, Jones EY (2003) A structural basis for immunodominant human T cell receptor recognition. *Nat Immunol* 4: 657–663.
18. Ding YH, Baker BM, Garboczi DN, Biddison WE, Wiley DC (1999) Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. *Immunity* 11: 45–56.
19. Li Y, Huang Y, Lue J, Quandt JA, Martin R, et al. (2005) Structure of a human autoimmune TCR bound to a myelin basic protein self-peptide and a multiple sclerosis-associated MHC class II molecule. *EMBO J* 24: 2968–2979.
20. Deng L, Langley RJ, Brown PH, Xu G, Teng L, et al. (2007) Structural basis for the recognition of mutant self by a tumor-specific, MHC class II-restricted T cell receptor. *Nat Immunol* 8: 398–408.
21. Weiner PK, Langridge R, Blaney JM, Schaefer R, Kollman PA (1982) Electrostatic potential molecular surfaces. *Proc Natl Acad Sci U S A* 79: 3754–3758.
22. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98: 10037–10041.
23. Miller PJ, Pazy Y, Conti B, Riddle D, Appella E, et al. (2007) Single MHC mutation eliminates enthalpy associated with T cell receptor binding. *J Mol Biol* 373: 315–327.
24. Macdonald WA, Chen Z, Gras S, Archbold JK, Tynan FE, et al. (2009) T cell allorecognition via molecular mimicry. *Immunity* 31: 897–908.
25. Tynan FE, Burrows SR, Buckle AM, Clements CS, Borg NA, et al. (2005) T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide. *Nat Immunol* 6: 1114–1122.
26. Chen JL, Stewart-Jones G, Bossi G, Lissin NM, Wooldridge L, et al. (2005) Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *J Exp Med* 201: 1243–1255.
27. Tynan FE, Reid HH, Kjer-Nielsen L, Miles JJ, Wilce MC, et al. (2007) A T cell receptor flattens a bulged antigenic peptide presented by a major histocompatibility complex class I molecule. *Nat Immunol* 8: 268–276.
28. Mazza C, Auphan-Anezin N, Gregoire C, Guimezanes A, Kellenberger C, et al. (2007) How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides? *EMBO J* 26: 1972–1983.
29. Yoshida K, Corper AL, Herro R, Jabri B, Wilson IA, et al. (2010) The diabetogenic mouse MHC class II molecule I-Ag7 is endowed with a switch that modulates TCR affinity. *J Clin Invest* 120: 1578–1590.
30. Yewdell JW, Bennink JR (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol* 17: 51–88.
31. Kaas Q, Lefranc MP (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol* 5: 505–528.
32. Kaas Q, Duprat E, Tourneur G, Lefranc MP (2008) IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Schoenbach C, Ranganathan S, Brusci V, eds. *Immunoinformatics*, Immunomics Reviews Series. New York: Springer, pp 19–49.
33. Reiser JB, Gregoire C, Darnault C, Mosser T, Guimezanes A, et al. (2002) A T cell receptor CDR3beta loop undergoes conformational changes of unprecedented magnitude upon binding to a peptide/MHC class I complex. *Immunity* 16: 345–354.
34. Zerrahn J, Held W, Raulet DH (1997) The MHC reactivity of the T cell repertoire prior to positive and negative selection. *Cell* 88: 627–636.
35. Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, et al. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 27: 55–77.
36. Reinherz EL, Tan K, Tang L, Kern P, Liu J, et al. (1999) The crystal structure of a T cell receptor in complex with peptide and MHC class II. *Science* 286: 1913–1921.
37. Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, et al. (1998) Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. *Immunity* 8: 403–411.
38. Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, et al. (1998) Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science* 279: 1166–1172.
39. Wilson IA (1999) Perspectives: protein structure. Class-conscious TCR? *Science* 286: 1867–1868.
40. Kjer-Nielsen L, Clements CS, Purcell AW, Brooks AG, Whistock JC, et al. (2003) A structural basis for the selection of dominant alpha/beta T cell receptors in antiviral immunity. *Immunity* 18: 53–64.
41. Hennecke J, Wiley DC (2002) Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alleloreactivity. *J Exp Med* 195: 571–581.
42. Chlewicki LK, Holler PD, Monti BC, Clutter MR, Kranz DM (2005) High-affinity, peptide-specific T cell receptors can be generated by mutations in CDR1, CDR2 or CDR3. *J Mol Biol* 346: 223–239.
43. McCoy AJ, Chandana Epa V, Colman PM (1997) Electrostatic complementarity at protein/protein interfaces. *J Mol Biol* 268: 570–584.
44. Hahn M, Nicholson MJ, Pyrdol J, Wucherpfennig KW (2005) Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor. *Nat Immunol* 6: 490–496.
45. Stewart-Jones G, Wadle A, Hombach A, Shenderov E, Held G, et al. (2009) Rational development of high-affinity T-cell receptor-like antibodies. *Proc Natl Acad Sci U S A* 106: 5784–5788.
46. Buslepp J, Wang H, Biddison WE, Appella E, Collins EJ (2003) A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection. *Immunity* 19: 595–606.
47. Richter S, Wenzel A, Stein M, Gabsdoulne RR, Wade RC (2008) webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res* 36: W276–280.
48. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876–4882.
49. Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J Comput and Graph Stat* 5: 299–314.
50. Dunn SM, Rizkallah PJ, Baston E, Mahon T, Cameron B, et al. (2006) Directed evolution of human T cell receptor CDR2 residues by phage display dramatically enhances affinity for cognate peptide-MHC without increasing apparent cross-reactivity. *Protein Sci* 15: 710–721.
51. Ishizuka J, Stewart-Jones GB, van der Merwe A, Bell JI, McMichael AJ, et al. (2008) The structural dynamics and energetics of an immunodominant T cell receptor are programmed by its Vbeta domain. *Immunity* 28: 171–182.
52. Archbold JK, Macdonald WA, Gras S, Ely LK, Miles JJ, et al. (2009) Natural micropolymerism in human leukocyte antigens provides a basis for genetic control of antigen recognition. *J Exp Med* 206: 209–219.
53. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA (2001) Glycosylation and the immune system. *Science* 291: 2370–2376.
54. Wang L, Zhao Y, Li Z, Guo Y, Jones LL, et al. (2007) Crystal structure of a complete ternary complex of TCR, superantigen and peptide-MHC. *Nat Struct Mol Biol* 14: 169–171.
55. Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, et al. (2004) T cell cross-reactivity and conformational changes during TCR engagement. *J Exp Med* 200: 1455–1466.
56. Armstrong KM, Piepenbrink KH, Baker BM (2008) Conformational changes and flexibility in T-cell receptor recognition of peptide-MHC complexes. *Biochem J* 415: 183–196.
57. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* 56: 93–101.
58. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 11: 2714–2726.
59. Abagyan RA, Totrov MM, Kuznetsov DA (1994) ICM: A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comp Chem* 15: 488–506.
60. Abagyan RA, Totrov MM (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235: 983–1002.

Supplementary Table S1

Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one

Javed M. Khan and Shoba Ranganathan

Table S1. Grouping of TR proteins.

Mutations in MHC α (I-ALPHA) chain and TR V β domain (MHC-I; TR Cluster I.2 and I.3), TR mutant names and the degree of tilt or relative change (compared to the first structure with similar TR type, MHC allele and peptide sequence) in θ are mentioned in parentheses (see Methods section for details).

S.No.	PDB code	Species	Peptide Sequence	MHC Allele	BE (kcal/mol)	Experimental binding affinity (K_d in μ M)	θ (°)	Diagonal TR docking angle (°) from literature	TR Type (TR name)	TR Cluster
TR/pMHC-I										
1	2oi9	Murine	QLSPFPFDL	H2-Ld	-13.13	-	47	-	2C	I.1
2	1g6r	Murine	SIYRYGGL	H2-Kb	-12.57	-	52	44 [46]	2C	
3	1mwa	Murine	EQYKFYSV	H2-Kbm3	-14.12	-	43	-	2C	
4	2ekb	Murine	EQYKFYSV	H2-Kb	-11.78	10	62	22 [17]	2C	
5	1fo0	Murine	INFDFTI	H2-Kb	-9.61	2.6	72	58 [46]	BM3.3	
6	1nam	Murine	RGYVYQGL	H2-Kb	-11.99	114	67	-	BM3.3	
7	2ol3	Murine	SQYYNSL	H2-Kb	-11.47	112	71	-	BM3.3	
8	2e7l	Murine	QLSPFPFDL	H2-Ld	-13.73	2 [23]	44	44	M6	
9	3e2h	Murine	QLSPFPFDL	H2-Ld	-12.16	4.6	60	-	M67	
10	3e3q	Murine	QLSPFPFDL	H2-Ld	-12.76	11.6 x 10 ⁻²	50	-	M13	
11	1kj2	Murine	KVITFIDL	H2-Kb	-15.30	-	37	-	KB5-C20	

S.No.	PDB code	Species	Peptide Sequence	MHC Allele	BE (kcal/mol)	Experimental binding affinity (K_d in μ M)	θ (°)	Diagonal TR docking angle (°) from literature	TR Type (TR name)	TR Cluster
TR/pMHC-I										
12	1w72	Human	EADPTGHSY	A*0101	-9.26	1.4×10^{-2}	73	-	FAB – HYB3	I.2
13	1mi5	Human	FLRGRAYGL	B*0801	-17.19	-	32	-	LC13	
14	3kpr	Human	EEYLKAWTF	B*4405	-21.88	1.54	22	-	LC13	
15	3kps	Human	EEYLQAFTY	B*4405	-21.52	49	23	-	LC13	
16	2esv	Human	VMAPRTLIL	E*0101	-12.79	30.2	50	50	KK50.4	
17	2nx5	Human	EPLPQQQLTAY	B*3501	-14.76	-	40	-	ELS4	
18	2ak4	Human	LPEPLPQQQLTAY	B*3508	-14.83	9.9	40	-	SB27	
19	3kxf	Human	LPEPLPQQQLTAY	B*3508 (Q65A, T69A, Q155A)	-18.70	-	28	-	SB27	
20	3ffc	Human	FLRGRAYGL	B*0801	-18.60	8.9	30	58	CF34	
21	3dxa	Human	EENLLDFVRF	B*4405	-18.10	0.3	31	80	DM1	
22	3mv7	Human	HPVGEADYFEY	B*3501	-16.96	-	33	-	TK3	I.3
23	3mv8	Human	HPVGEADYFEY	B*3501	-16.68	-	34	-	TK3 (V β Q55H)	
24	3mv9	Human	HPVGEADYFEY	B*3501	-16.36	-	35	-	TK3 (V β Q55A)	
25	1bd2	Human	LLFGYPVYV	A*0201	-14.33	-	45	52 [17]	B7	
26	1lp9	Human	ALWGFFPVL	A*0201	-22.97	10	30	89	AHIII 12.2	
27	2uwe	Human	ALWGFFPVL	A*0201 (T163A)	-22.44	4.7	21	-	AHIII 12.2	

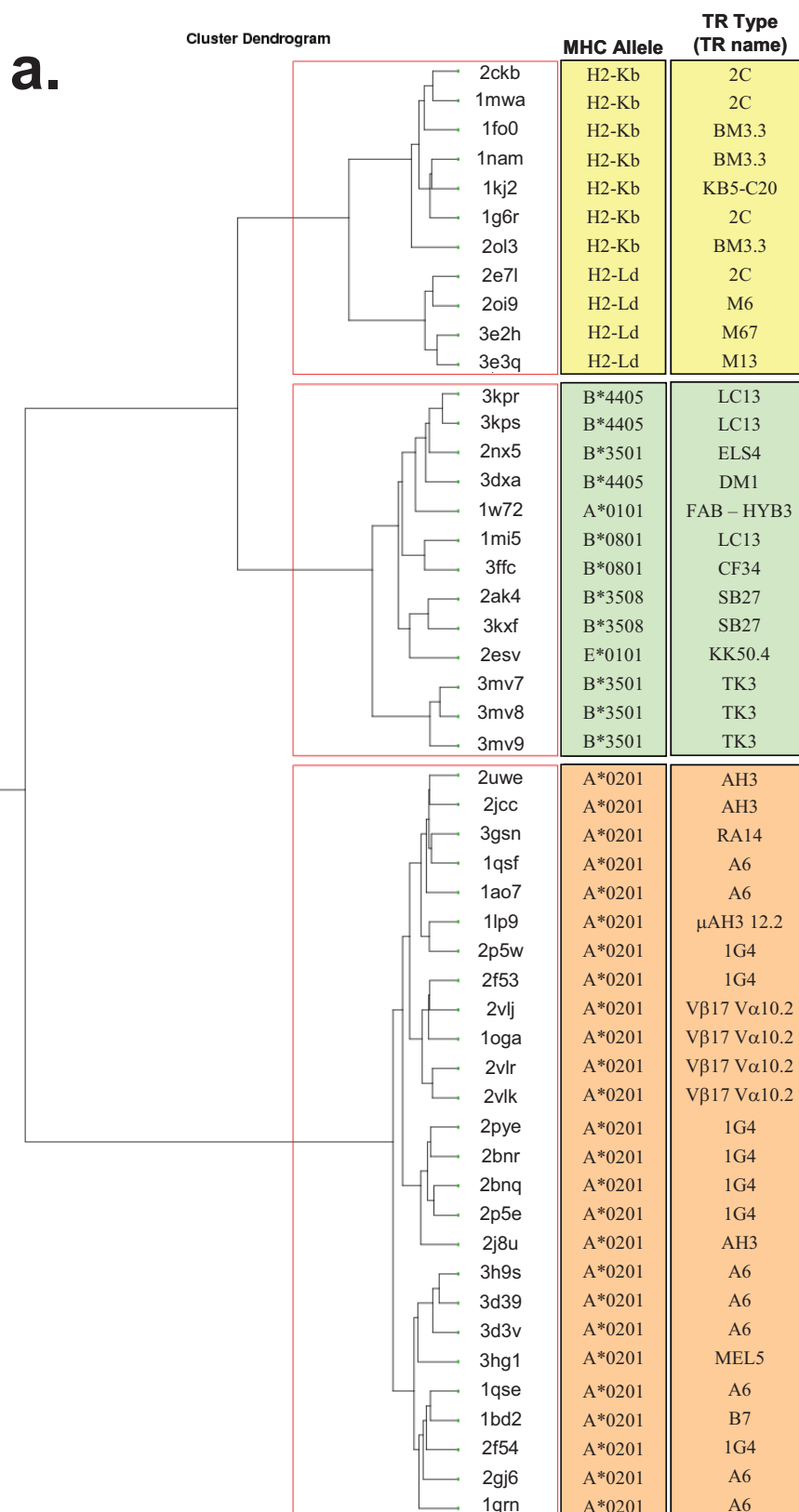
S.No.	PDB code	Species	Peptide Sequence	MHC Allele	BE (kcal/mol)	Experimental binding affinity (K_d in μ M)	θ (°)	Diagonal TR docking angle (°) from literature	TR Type (TR name)	TR Cluster
TR/pMHC-I										
28	2j8u	Human	ALWGFFPVL	A*0201 (K66A)	-21.01	31.8	25	-	AHIII 12.2	I.3 (contd.)
29	2jcc	Human	ALWGFFPVL	A*0201 (W167A)	-22.74	15.4	20	-	AHIII 12.2	
30	1ao7	Human	LLFGYPVYV	A*0201	-17.20	-	32	37 [17]	A6	
31	1qrm	Human	LLFGYAVYV	A*0201	-16.59	100	35	-	A6	
32	1qse	Human	LLFGYPRYV	A*0201	-17.28	8	31	-	A6	
33	1qsf	Human	LLFGYPVAV	A*0201	-14.99	-	39	-	A6	
34	2gj6	Human	LLFGKPVYV	A*0201	-15.49	160	37	-	A6	
35	3d39	Human	LLFGFPVYV (Y5{4-fluoroPhenylalanine})	A*0201	-15.16	0.64	38.5	-	A6	
36	3d3v	Human	LLFGFPVYV (Y5{3,4-difluoroPhenylalanine})	A*0201	-15.40	0.46	37	-	A6	
37	3h9s	Human	MLWGYLQYV	A*0201	-14.91	-	39	-	A6	
38	2bnr	Human	SLLMWITQC	A*0201	-14.93	13.3	39	69	1G4	
39	2f54	Human	SLLMWITQC	A*0201	-15.59	15	36	-	1G4 (-3°)	
40	2f53	Human	SLLMWITQC	A*0201	-16.06	1×10^{-3}	36	-	1G4 (c49c50)	
41	2p5e	Human	SLLMWITQC	A*0201	-16.08	48×10^{-6}	36	-	1G4 (c58c61)	
42	2p5w	Human	SLLMWITQC	A*0201	-16.69	-	34	-	1G4 (c58c62)	
43	2pye	Human	SLLMWITQC	A*0201	-15.21	8.16×10^{-2}	38	-	1G4 (c5c1)	
44	2bnq	Human	SLLMWITQV	A*0201	-15.23	5	38	68.1	1G4	

S.No.	PDB code	Species	Peptide Sequence	MHC Allele	BE (kcal/mol)	Experimental binding affinity (K_d in μ M)	θ (°)	Diagonal TR docking angle (°) from literature	TR Type (TR name)	TR Cluster
TR/pMHC-I										
45	1oga	Human	GILGFVFTL	A*0201	-11.99	-	69	69	V β 17 V α 10.2	I.3 (contd.)
46	2vij	Human	GILGFVFTL	A*0201	-11.77	5.2	70	-	V β 17 V α 10.2 (+5°)	
47	2vik	Human	GILGFVFTL	A*0201	-11.48	5.2	70.5	-	V β 17 V α 10.2 (+10°)	
48	2v1r	Human	GILGFVFTL	A*0201	-11.88	4.9	70	-	V β 17 V α 10.2 (V β S99A)	
49	3gsn	Human	NLVPMVATV	A*0201	-16.66	27.7	34.5	35	RA14	
50	3hg1	Human	ELAGIGILTV	A*0201	-17.58	18	30	35	MEL5	
TR/pMHC-II										
1	1u3h	Murine	SRGGASQYRPSQ	I-Au	-15.06	-	73	-	172.10 V β 8.2	II.1
2	2z31	Murine	RGGASQYRPSQ	I-Au	-13.01	25.3	87	-	cl 19 V β 8.2	
3	2pxy	Murine	RGGASQYRPSQ	I-Au	-14.47	25.4	77	-	1934.4 V β 8.2	
4	3mbe	Murine	GAMKRHGLDNYRGYSLGN	I-Ag7	-22.42	16.2 x 10 ⁻²	38	48	21.30	
5	1d9k	Murine	GNSHRGAIEWEGIESG	I-Ak	-15.42	-	71	80	D10	
6	1j8h	Human	PKYVKQNTLKLAT	DRB1*0401	-16.76	-	56	-	HA1.7	II.2
7	1fyt	Human	PKYVKQNTLKLAT	DRB1*0101	-15.87	-	60	70	HA1.7	
8	1zgl	Human	VHFFKNIVTPRTPG	DRA*0101	-15.59	-	61	47	3A6	
9	2iam	Human	GELIGILNAAKVPAD	DRB1*0101	-13.72	-	79	71	E8	
10	2ian	Human	GELIGTLNAAKVPAD	DRB1*0101	-14.96	-	75	71	E8	
11	1ymm	Human	ENPVVHFFKNIVTP	DRB1*1501	-11.76	-	112	110	Ob.1A12	

Supplementary Figure S1

Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one

Javed M. Khan and Shoba Ranganathan



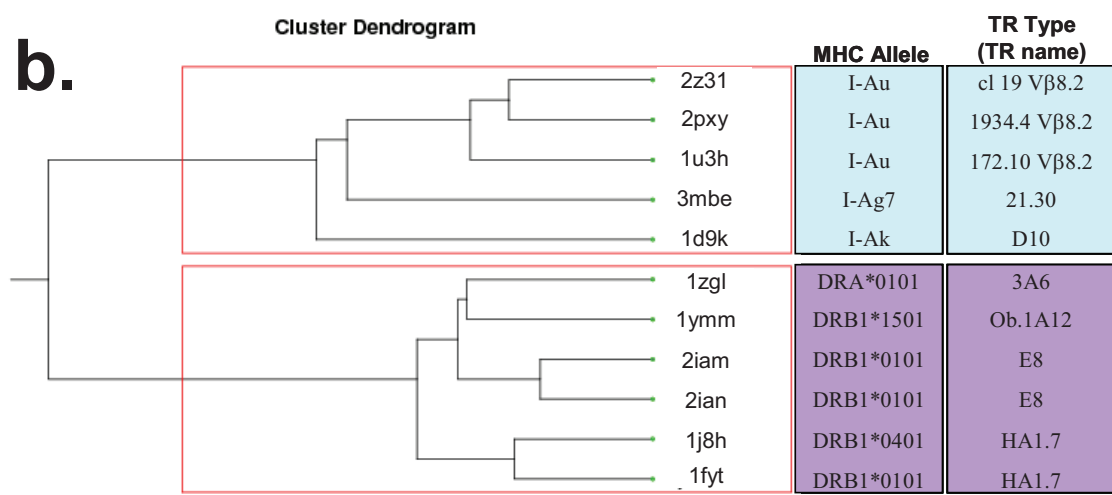
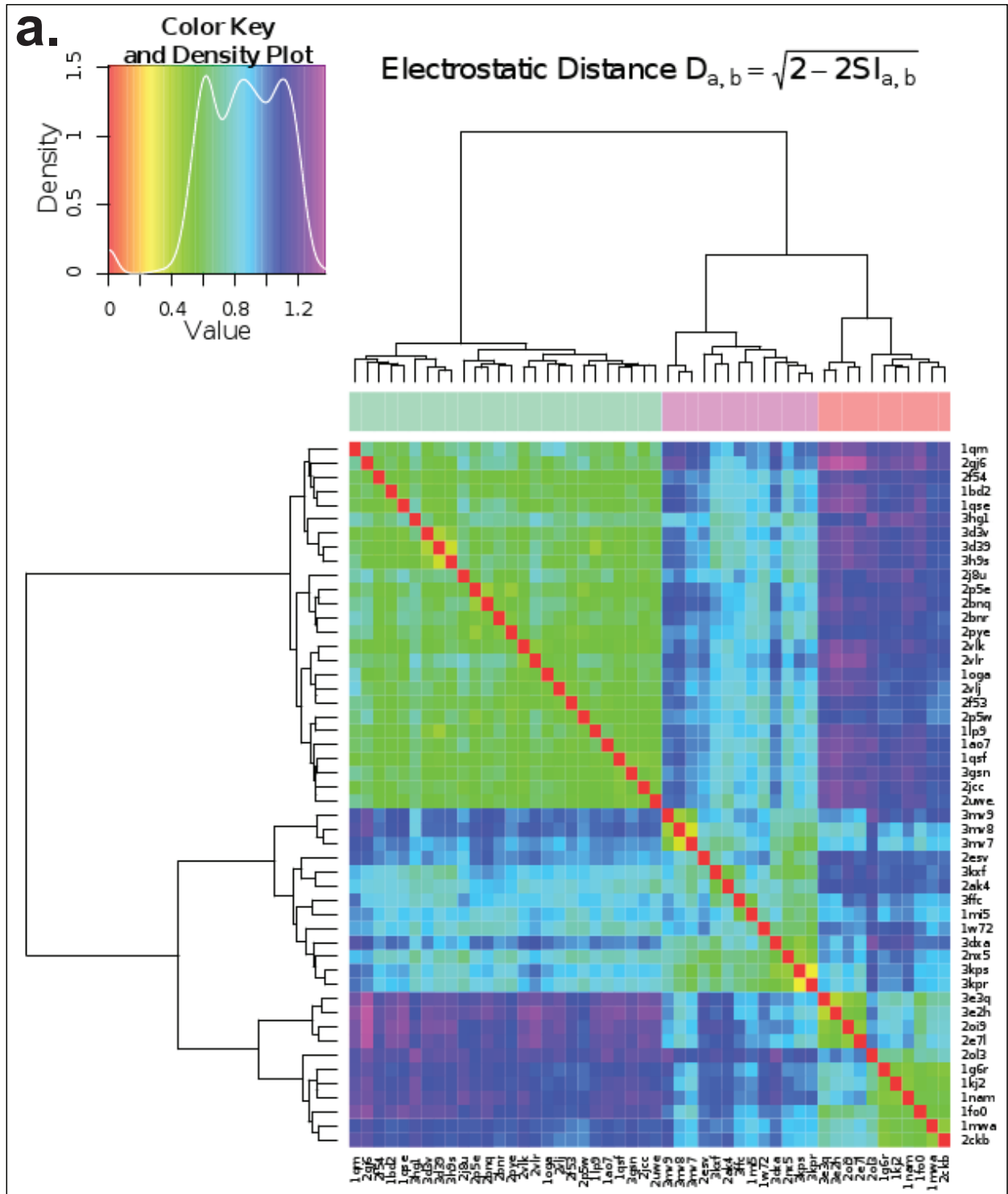


Figure S1. Cluster dendrograms for all pMHC interfaces based on their MSEP similarities.
a. pMHC-I complexes clustered into three distinct clusters. b. pMHC-II ligands clustered into two distinct clusters. Each pMHC interface is denoted by its corresponding PDB code. Every pMHC is mapped onto its respective MHC allele and the interacting TR type (TR name). This clearly indicates the clustering amongst the TR proteins. The three distinct clusters of pMHC-I binding TR proteins are coloured yellow – cluster I.1, green – cluster I.2 and orange – cluster I.3. The two clusters amongst pMHC-II binding TR proteins are highlighted in light blue – cluster II.1 and lavender – cluster II.2. TR grouping (clustering) is in accordance with Table 1.

Supplementary Figure S2

Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one

Javed M. Khan and Shoba Ranganathan



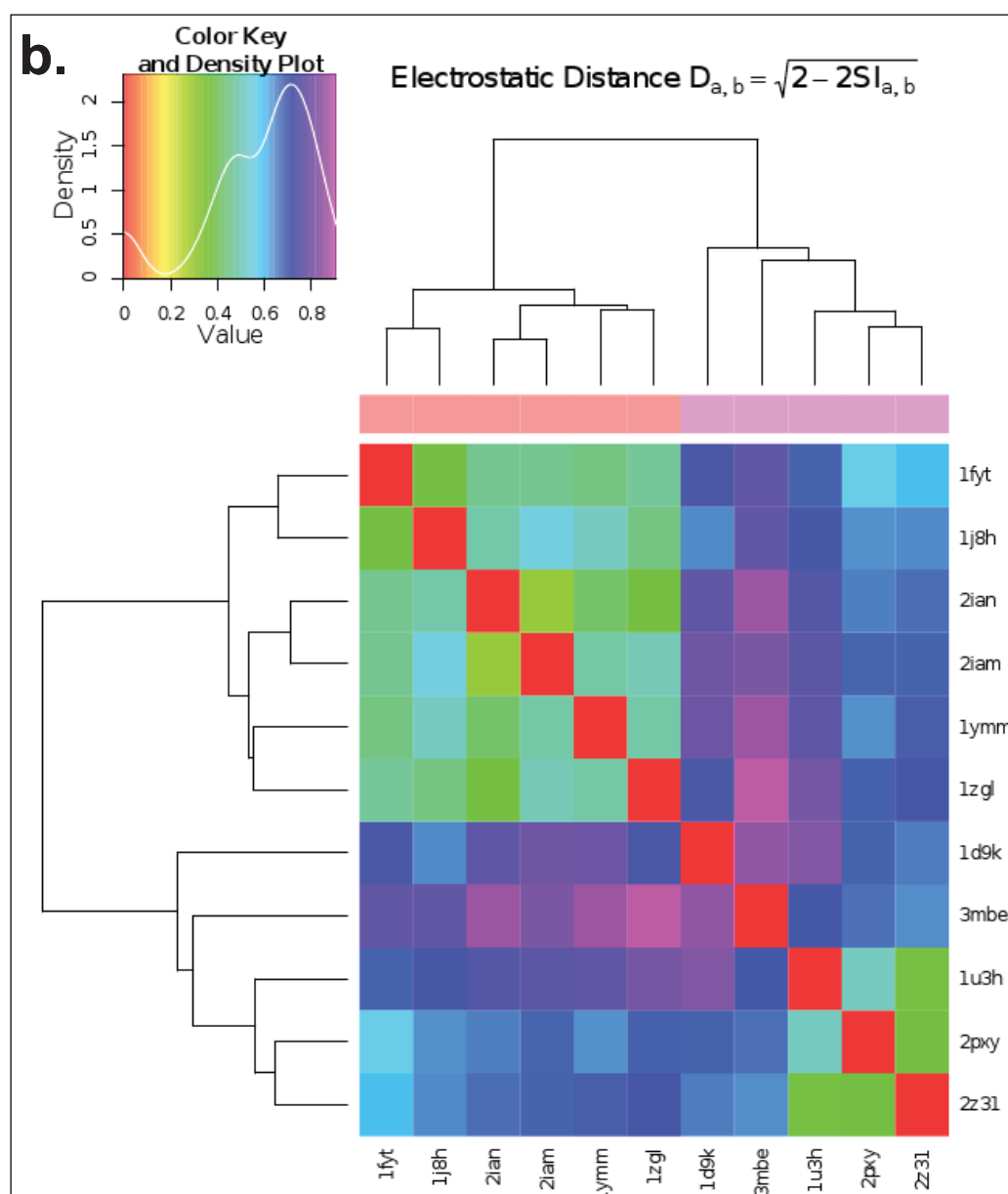


Figure S2. Heat maps for all pMHC interfaces based on the calculated MSEP values depicted as a colour coded matrix showing clustering amongst pMHC complexes in a reverse order as compared to the cluster dendograms in Supplementary Figure 1. a. pMHC-I complexes clustered into three. b. pMHC-II structures in two distinct clusters. Each pMHC interface is again denoted by its corresponding PDB code. Inset, are the legends showing the color key used to create heat matrices and the MSEP value ranges for pMHC interfaces. Also shown is the formula used to calculate electrostatic distances for clustering.

5.2 Conclusions

A number of physicochemical characteristics have been utilized to analyze all available TR/pMHC structures such that any basic differences between pMHC-I and pMHC-II interactions with TR proteins are understood. Based on the computed TR/pMHC BE values, the avidity of TR/pMHC interaction has been classified as weak-, moderate-, and strong-. By mapping charged rings formed from MSEP on the pMHC interface, a novel and rational approach to computing θ value has been described. No absolute conserved residues were found in interacting regions of both TR and pMHC from the analysis of TR paratopes and pMHC epitopes, yet vital conserved positions were observed on both interfaces across TR/pMHC-I and TR/pMHC-II structures. These conserved positions could have fundamental implication for peptide vaccine design and could potentially provide clues to the positional specificity of TR proteins. Furthermore, TR/pMHC binding requirements have been dissected by clustering the TR proteins.

The findings suggest that the entire process of pMHC recognition and TR signalling is possibly governed by two factors, the electrostatic ring displayed by pMHC interface and a specific arrangement of residues presented by pMHC, thereby, explaining the phenomenon of pMHC recognition by TR and TR specificity simultaneously. The extensive studies on TR/pMHC interactions have helped define structural features, especially MSEP, that can be analyzed as parameters governing TR/pMHC complex formation relevant for immune system activation. These parameters could be used to develop, enhance and/or accelerate the progress of structure-based prediction techniques to successfully predict T cell epitopes in accordance with their MHC and TR binding specificities besides minimizing false positives (FP) and true negatives (TN) from actual agonistic peptides in a given set of peptide antigens.

Chapter 6: *In silico* prediction of immunogenic T cell epitopes for HLA-DQ8

6.1 Summary

MHC-II proteins play a critical role in adaptive immune responses. They bind antigenic peptide fragments and present them on the APC surface for recognition by the CD4⁺ helper T cells and subsequent immune response. While MHC-I alleles have been extensively studied [2, 314, 331], investigations pertaining to MHC-II alleles have been hindered, especially in the context of MHC-II restricted T cell epitope prediction, primarily due to the lack of MHC-II related biochemical, functional and crystallographic data [11, 15, 18]. Nevertheless, development of T cell epitope prediction methods applicable to MHC-II proteins [11, 13, 14] was made possible by recent growth in both experimental and structural data for MHC-II alleles. Many MHC-II alleles such as HLA-DQ are known to be associated with pathogenesis of autoimmune disorders [447] and hypersensitivity reactions [448, 449]. Due to its association with various human autoimmune [450, 451] and hypersensitivity disorders [448, 449], HLA-DQ8 is an allele of particular interest among all HLA-DQ alleles. Sequence-based computational techniques for predicting HLA-DQ8-restricted T cell epitopes [452-454], have encountered limited success, with Wang *et al.* [455] recently reporting the average area under the receiver operating characteristic (ROC) curve, A_{ROC} , of 0.88 (for HLA-DR, DP and DQ alleles), whereas the accuracy and efficiency of a recently developed structure-based model [11] need to be enhanced. Hence, publication 6 describes a combined structure-based prediction model for DQ8-restricted T cell epitope prediction using pDOCK [49], and MSEP-based clustering (as described in Chapter 5) of peptide docked pMHC binding interfaces to predict immunogenic T cell epitopes. It also highlights the use of both pMHC and TR/pMHC interaction knowledge and parameters to identify T cell activating peptide epitopes. The prediction model was rigorously trained, tested and validated using experimentally binding and non-binding data for DQ8. High prediction accuracy (average $A_{ROC}>0.94$) for DQ8-binders is verified against experimental data. 77 % (24 out of 31) accuracy is recorded for the prediction of known T cell activators and all peptide binding registers were accurately predicted using this novel prediction model. The binding patterns of DQ8-binding peptides were also studied and our results reconfirm that peptide epitopes that do not conform to binding motifs exist and are precisely identified by the developed T cell prediction model.

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed Mohammed Khan¹, Gaurav Kumar¹ and Shoba Ranganathan^{1,2,*}¹Department of Chemistry and Biomolecular Sciences & ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia.²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Motivation: HLA-DQ alleles are involved in the pathogenesis of hypersensitivity reactions, with HLA-DQ8 associated with several human autoimmune disorders. Limited success has been achieved using sequence-based computational techniques for predicting HLA-DQ8-restricted T cell epitopes while accuracy and efficiency of recently developed structure-based models need to be improved.

Methods: We describe a combined structure-based prediction approach for DQ8-restricted T cell epitope prediction using a recently developed fast and accurate docking protocol, pDOCK, and molecular surface electrostatic potential (MSEP)-based clustering of pMHC binding interfaces. The prediction model was rigorously trained, tested and validated using experimentally verified DQ8 binding and non-binding peptides.

Results: High prediction accuracy (average area under the ROC curve, average $A_{ROC} > 0.94$) is validated against experimental data. Our model also predicts all binding registers correctly and known T cell activators with 77% accuracy. We also studied the patterns of DQ8-binding peptides and reassure the existence of epitopes not conforming to binding motifs.

1 INTRODUCTION

Among many important proteins that take part in adaptive immune responses, major histocompatibility complex (MHC) proteins arguably play the most crucial role. They bind and present short antigenic peptides on the cell surface, as peptide-MHC (pMHC) complexes, for recognition by T cell receptor (TR) proteins to form T cell receptor-peptide-MHC (TR/pMHC) complexes which subsequently activate the T cells to carry out the immune response (Rammensee, *et al.*, 1993; Lefranc and Lefranc, 2001). Both of these steps trigger a series of immunological events essential for initiation and regulation of immune responses (Khan and Ranganathan, 2010; Khan, *et al.*, 2010; Khan and Ranganathan, 2011).

Broadly classified into two types, MHC class I (MHC-I) proteins bind and present endogenous (processed within the cell) peptides for recognition by the CD8⁺ cytotoxic T cells whereas MHC class II (MHC-II) proteins prefer exogenous (processed outside the cell)

peptides for recognition by the CD4⁺ helper T cells (Khan and Ranganathan, 2010). While MHC-I alleles have been extensively studied (Rammensee, *et al.*, 1993; Reche, *et al.*, 2002; Hoof, *et al.*, 2009), investigations pertaining to MHC-II alleles have been hindered, especially in the context of MHC-II restricted T cell epitope prediction, primarily due to the lack of MHC-II related biochemical, functional and crystallographic data (Khan, *et al.*, 2010). However, recent growth in both experimental and structural data for MHC-II alleles has facilitated their analysis for the development of T cell epitope prediction methods applicable to MHC-II proteins (Tong, *et al.*, 2006a; 2006b; 2007).

Currently available computational protocols for the identification of MHC-II restricted T cell epitopes can be categorized into sequence and structure-based methods. Sequence-based methods are relatively advanced in predicting T cell epitopes for MHC-II alleles, such as HLA-DR (Brusic, *et al.*, 2004; Nielsen, *et al.*, 2008; Dimitrov, *et al.*, 2010), with abundant biochemical peptide binding data. Nonetheless, for MHC-II alleles with limited peptide data, such as HLA-DQ, these approaches have been used with varying degree of success (Godkin, *et al.*, 1997; 1998; Harfouch-Hammoud, *et al.*, 1999; Rammensee, *et al.*, 1999), with Wang *et al.*, (2010) recently reporting an average area under the ROC curve (A_{ROC}) of 0.88 (for HLA-DR, DP and DQ alleles), owing to their dependence on experimental data for training purposes. On the other hand, structure-based procedures such as docking (Tong, *et al.*, 2004; Tong, *et al.*, 2006b) have been successfully applied to predict T cell epitopes even for MHC-II alleles with very limited peptide data while addressing the dual issues of docking and scoring for MHC-II binding peptides (Tong, *et al.*, 2006a; 2006b; 2007). However, similar to all other methods, even this combined docking and scoring-based approach, utilizes only pMHC interaction data for T cell epitope prediction which affects its accuracy as only 50% of strong MHC-binding peptides are known to activate T cells (Yewdell and Bennink, 1999). Therefore, given the significance of TR/pMHC binding in T cell mediated immunity, it becomes extremely important to factor in TR/pMHC interaction knowledge in conjunction with pMHC binding data for improved prediction of immunogenic T cell epitopes. Also, the speed and efficiency of the docking protocol need to be improved for high-throughput screening of MHC-binding peptides to identify high-binders.

*To whom correspondence should be addressed.

Many HLA-DQ alleles are known to be involved in the pathogenesis of hypersensitivity reactions (Neeno, *et al.*, 1996; Krco, *et al.*, 2000) and autoimmune disorders (Klein, *et al.*, 2000). Among these, an allele of particular interest is HLA-DQ8 (made up of the haplotypes DQA1*0301 and DQB1*0302, and also known as HLA-DQ3.2 β) due to its association with various human autoimmune disorders such as insulin-dependent diabetes mellitus (IDDM) (Erlich, *et al.*, 1993; Nepom and Kwok, 1998), autoimmune encephalomyelitis (Mangalam, *et al.*, 2009), autoimmune polyendocrine syndrome type II (APS-II) (Robles, *et al.*, 2002), IDDM-associated periodontal disease (Faustman, *et al.*, 1991) and celiac disease (Sollid and Thorsby, 1993) and hypersensitivity disorders including house dust mite allergy (Neeno, *et al.*, 1996; Krco, *et al.*, 2000). DQ8 is found in approximately 20-30% of the human population (Gonzalez-Galarza, *et al.*, 2011) and is prevalent in about 86% of IDDM patients (Graham, *et al.*, 2002). Hence, in order to elucidate the role of DQ8 both in autoimmunity and allergenicity, enhanced understanding of DQ8-restricted pMHC and TR/pMHC binding is essential.

Recently, we have developed pDOCK (Khan and Ranganathan, 2010) which is a robust new protocol for rapid and accurate fully-flexible docking of peptides to MHC-I and MHC-II alleles. Benchmarking pDOCK with the previous docking technique (Tong, *et al.*, 2004; 2006b) revealed a 2.5 fold and ~60% increase in its accuracy and speed, respectively. Upon validation against previously published studies, a seven-fold increase was recorded in pDOCK accuracy. pDOCK also accurately determined the binding registers of all MHC-I and MHC-II binding peptides used in that study. Following which, we have also very recently analyzed 61 (50 TR/pMHC-I and 11 TR/pMHC-II) available TR/pMHC crystal structures (Khan and Ranganathan, 2011) collated from the MPID-T2 (Khan, *et al.*, 2011) database and identified certain structural interaction characteristics such as molecular surface electrostatic potential (MSEP) that can be used as parameters governing TR/pMHC complex formation for T cell epitope prediction. We have now combined the power of pDOCK to successfully identify strong MHC-binding peptides using a previously developed complementary scoring function (Tong, *et al.*, 2006b) and the efficient MSEP-based clustering of pMHC binding interfaces (Khan and Ranganathan, 2011) to predict DQ8-restricted immunogenic T cell epitopes with high accuracy and correct binding registers. We also investigated the binding patterns of DQ8-restricted peptides and confirm the existence of peptide epitopes that do not conform to binding motifs, as reported earlier (Tong, *et al.*, 2006b).

2 METHODS

2.1 Data

2.1.1 Structural data The crystal structure of Insulin B9-23-DQ8 pMHC complex (Lee, *et al.*, 2001), with the Protein Data Bank or PDB (Berman, *et al.*, 2002) code 1jk8, was used to extrapolate the structural coordinates of the DQ8 allele. Internal Coordinate Mechanics (ICM) package version 3.6-1 (Abagyan, *et al.*, 1994) was then used to relax the extracted structure by conjugate gradient minimization.

2.1.2 Biochemical and functional peptide binding data The dataset of 1719 peptides with known binding affinity values for DQ8, used by Wang *et al.*, (2010), is both not publicly available and/or listed in the pub-

lished article. Therefore, we have used the available set of peptides known to bind DQ8 and/or elicit T cell proliferation, for this study. The experimental data used for this investigation was primarily divided into two datasets: (i) peptides from biochemical studies with experimental IC₅₀ values and (ii) peptides from functional T cell assays that are known to cause T cell activation.

127 peptides with experimentally determined IC₅₀ values, obtained from biochemical studies (Godkin, *et al.*, 1998; Sidney, *et al.*, 2002; Suri, *et al.*, 2005; Chang and Unanue, 2009), comprised dataset I. These peptides with known IC₅₀ values were further classified as high-affinity MHC-binders: IC₅₀ ≤ 500 nM, medium-affinity binders: 500 nM < IC₅₀ ≤ 1500 nM, low-affinity binders: 1500 < IC₅₀ ≤ 5000 nM and non-binders: 5000 < IC₅₀. Therefore, dataset I was made up of 70 high-affinity, 14 medium-affinity, 29 low-affinity binders and 14 non-binders. Although 14 peptides were considered non-binding based on their IC₅₀ values, some of them have reported binding registers. Similarly, some of the peptides that are regarded as binders (high-, medium- and low-affinity binders; 113 peptides) did not have any known binding registers. Therefore, 87 (84 binding and three non-binding) peptides in this dataset had experimentally determined binding registers and 40 (29 binding and 11 non-binding) peptides had no known binding registers. This dataset was further divided into the training set (Supplementary Table S1; 57 peptides with 43 high-, five medium-, five low- and four non-binders) and test set 1 (Supplementary Table S2; remaining 70 peptides with 27 high-, nine medium-, 24 low- and 10 non-binders).

36 DQ8-specific peptides, out of which, 31 were known to cause DQ8-restricted T cell proliferation and five were known to not activate T cells, formed dataset II. These peptides were derived from functional studies (Neeno, *et al.*, 1996; Krco, *et al.*, 2000; Paisansinsup, *et al.*, 2002; Chang and Unanue, 2009) that conducted *in vitro* immuno-assays to detect T cell activity and were subsequently used as test set 2 (Supplementary Table S3) in this study.

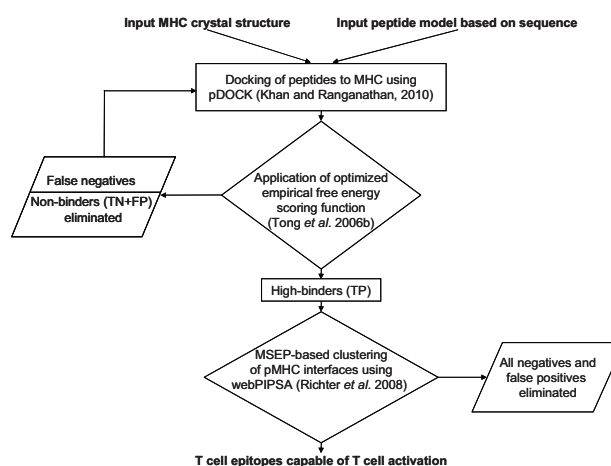


Fig. 1. Flowchart of the prediction model used in this work.

2.2 Prediction Model

DQ8-binding and non-binding peptide sequences were docked into the MHC peptide binding groove of the X-ray crystallographic structure for DQ8 using pDOCK (Khan and Ranganathan, 2010). Following this, a previously reported customized free energy scoring function (Tong, *et al.*, 2006b) was utilized to improve the predictive performance of the model. Finally, MSEP-based clustering (Khan and Ranganathan, 2011) of the peptide docked pMHC binding interfaces from test set 2 was performed to enhance the accuracy of the model and effectively predict DQ8-restricted immunogenic T cell epitopes. Figure 1 illustrates the prediction model developed using the combined approach in this study.

2.2.1 Docking of peptides to DQ8 Docking of all peptides to the extracted template crystal structure of DQ8 was performed using pDOCK (Khan and Ranganathan, 2010). pDOCK utilizes the ICM (Abagyan, *et al.*, 1994) optimal-bias Monte Carlo minimization procedure (Abagyan and Totrov, 1999) which in turn uses the Merck Molecular Force Field or MMFF (Halgren, 1995) and Empirical Conformational Energy Program for Peptides 3 (ECEPP/3) force field parameters (Nemethy, *et al.*, 1992) to perform each docking. In brief, the pDOCK protocol involves: (i) preparatory step 1: MHC receptor modeling and/or positioning using the ICM global optimization algorithm (Abagyan, *et al.*, 1994); (ii) preparatory step 2: determining the docking grid by defining the grid dimensions (length x breadth x height) based on standardized values (Khan and Ranganathan, 2010) for MHC supertypes (MHC-I and MHC-II; MHC-II in this case) for peptide placement and grid map generation within the vicinity of the MHC peptide binding site and; (iii) a single consolidated final docking and refinement step: peptide positioning within the grid, fully flexible docking of the peptides into the peptide binding groove followed by iterative *ab initio* refinements of all peptide residues along with the backbone and peptide interacting side-chain dihedral angles of the MHC binding site residues to eliminate or minimize atomic clash regions at the pMHC interface, using the ICM global optimization docking algorithm (Abagyan, *et al.*, 1994) and a biased Monte Carlo procedure (Abagyan and Totrov, 1999). The preparatory steps were together used to generate the MHC receptor maps and the final single docking and refinement step was used to carry out peptide docking, generate the final least energy docked peptide conformation and further refine the product. pDOCK was run on a 2 CPU 3.20 GHz 3 GB RAM workstation.

2.2.2 Empirical free energy scoring function A previously reported scoring function (Tong, *et al.*, 2006b) is employed in this investigation. Originally based on the free energy potential (Abagyan and Totrov, 1999) in the ICM 3.6-1 software package (Abagyan, *et al.*, 1994), this adopted scoring function has its binding free energy calculated as the difference between the energy of the solvated pMHC complex and the sum of the energy of the solvated MHC receptor and the peptide. The fully relaxed conformation of the free peptide in water (Schapira, *et al.*, 1999) is chosen as the reference state for a given peptide. The MHC and the peptide are separated after docking and their relaxed energies are computed, following energy minimization in water for all binding energy calculations. Therefore, the binding free energy (ΔG_{bind}) function used here is expressed as follows:

$$\Delta G_{\text{bind}} = \alpha \Delta G_{\text{EL}} + \beta \Delta G_{\text{H}} + \gamma \Delta G_{\text{EN}} + C \quad (1)$$

where, ΔG_{EL} is the electrostatic contributions from the desolvation of partial charges transferred from an aqueous medium to a protein core environment and the pMHC coulombic interactions. Using an implementation of the boundary element algorithm (Bharadwaj, *et al.*, 1995; Schapira, *et al.*, 1999), the numeric solution of the Poisson equation determines ΔG_{EL} . The hydrophobic energy (ΔG_{H}) is composed of the product of solvent accessible surface area (determined by rolling a sphere of radius 1.4 Å along the surface of the molecule) by the surface tension. The entropic term of the protein side-chains is denoted by ΔG_{EN} and is computed from the maximal burial entropies for each type of amino acid and their relative accessibilities. Entropy change in the system due to the decrease of free molecular concentration and the loss of rotational/translational degrees of freedom upon binding (Schapira, *et al.*, 1999), is accounted for by the constant term C or K (Rognan, *et al.*, 1999). Generally, physical parameters that are independent of the dataset used represent C . It has been noted (Janin, 1995) that, among various research groups, there are great variations in the value used for C . To obtain the best separation of binders and non-binders, the coefficients (α , β , γ) assigned to each energy term in this scoring function were optimized. Many previous studies (Krystek, *et al.*, 1993; Novotny, *et al.*, 1997; Schapira, *et al.*, 1999; Tong, *et al.*, 2006b) have successfully used this separation schema consisting of the most significant potentials contributing to protein-protein, protein-ligand and protein-peptide interactions.

2.2.3 Optimizing the scoring function A similar approach to that employed by Tong *et al.*, (2006b) was again utilized for optimization of the above described scoring function. Initially, the concentration of ligand required to saturate half of the available binding sites of the protein (Bock and Gough, 2002), in other words, the reported IC_{50} values (for dataset I), were considered to be similar to the equilibrium dissociation constants (K_d) since the concentration of the ligand in the unbound state is much lower than K_d of the ligand in the binding assay, such that $\Delta G_{\text{bind}} \approx -RT \ln(\text{IC}_{50})$ (Rognan, *et al.*, 1999). This was followed by recalibration of the coefficients for different energy terms by standard least-square multivariate regression analysis, as previously described by Wang *et al.*, (2002), of the training set to improve the discriminative power of the scoring function. Subsequently, quality of the scoring function was assessed using 10-fold cross-validation (Figure 2) (Bock and Gough, 2002). The technique utilized here is called k -fold cross-validation, where a scoring function is trained on $(k-1)$ partitions by constructing k random, (approximately) equal-sized, disjoint partitions of the sample data, and tested on the excluded partition. After k such experiments, the results are averaged and an estimate of the error rate expected upon generalization to new data is given by the observed error rate. Finally, the cross-validation coefficient q^2 and the standard error of prediction s_{press} were used to evaluate the predictive power of the scoring function. Further evaluation using evolutionary regression analysis (Wang, *et al.*, 2002) with different subsets representing 5-fold, 4-fold, 3-fold and 2-fold cross-validation (Figure 2), was also conducted to assess the robustness of the scoring function.

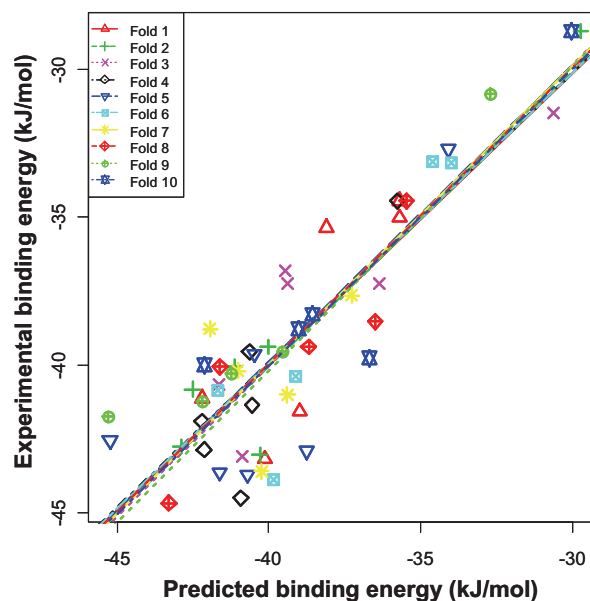


Fig. 2. Relationship between experimental and predicted binding energies from 1-fold to 10-fold cross-validations. Although the figure depicts folds 1-10, cross-validation results only for folds 2, 3, 4, 5 and 10 are discussed.

2.2.4 Clustering of pMHC interfaces This step was carried out as recently outlined by us (Khan and Ranganathan, 2011). However, in this case the peptide docked pMHC complexes from test set 2 were used to extract the coordinates for truncated versions of the pMHC complexes, encompassing the pMHC binding interfaces. These pMHC interfaces were subjected to MSEP-based clustering along with known human T cell activating pMHC-II binding interfaces (Khan and Ranganathan, 2011) from all six currently available human TR/pMHC-II crystal structures to identify MSEP similarities between the peptide docked pMHC interfaces and the human pMHC-II interfaces known to activate T cells depending on the electro-

static distances between them. The webPIPSA (Richter, *et al.*, 2008) server was used to calculate MSEP and compare electrostatic interaction properties of the pMHC interfaces. The web-server begins with calculation of pMHC interface MSEP using the University of Houston Brownian Dynamics (UHBD) program (Madura, *et al.*, 1995) and then compares their electrostatic properties by calculating similarity indices for all pairs of pMHC interfaces based on their electrostatic similarity, using the PIPSA algorithm (Blomberg, *et al.*, 1999). These similarity indices are then converted to electrostatic distances which are clustered and displayed as a colour coded matrix called heat map (Supplementary Figure S1) using the R (Ihaka and Gentleman, 1996) software package. This clustering output (Supplementary Figure S1) was divided into five groups with the six known human T cell activating pMHC-II interfaces in group A, the test set 2 pMHC interfaces nearest to group A forming group B (regarded as strong-agonists), group C comprising the moderate-agonists which are test set 2 pMHC interfaces next to group B, test set 2 pMHC interfaces next to group C making up group D (considered as weak-agonists) and group E being composed of test set 2 pMHC interfaces furthest from group A making them non-agonists.

2.3 Training, testing and validation

The bound conformations of binding peptides with experimentally determined registers and the best conformations of non-binding peptides without any preferred registers were sampled to initially train the DQ8 prediction model. Among the 57 peptides in the training set, 55 (53 binding and two non-binding) peptides had known binding registers and 2 non-binding peptides had no known binding conformations. After optimization of the empirical free energy scoring function by statistical analyses performed on the training set, the optimized scoring function was tested on test set 1 to further assess its predictive ability. Test set 1 had 32 (31 binding and one non-binding) peptides with known binding registers and 38 (29 binding and nine non-binding) peptides with no known binding registers. Following this, the optimized scoring function and the MSEP-based clustering approach were together applied on test set 2 to improve the overall accuracy of the prediction model, thereby, validating it against experimental T cell activation data. Test set 2 had 22 DQ8-binding peptides with known binding registers and 14 DQ8-binding peptides with no known binding registers.

Similar to the method reported by Tong *et al.*, (2006b), we performed sensitivity (SE), specificity (SP) and receiver operating characteristic (ROC) analyses, described previously by Brusic *et al.*, (2002), on test set 1 to evaluate the efficiency of the optimized scoring function. The percentages of correctly predicted binders and non-binders are given by $SE = TP / (TP + FN)$ and $SP = TN / (TN + FP)$, respectively. Experimental binders with at least one predicted binding register and experimental non-binders with no predicted binding register are represented by true positives (TP) and true negatives (TN), respectively. Whereas, experimental binders predicted as non-binders and experimental non-binders predicted as binders, are denoted by false negatives (FN) and false positives (FP), respectively. ROC analysis, where the ROC curve is generated by plotting SE as a function of (1-SP) for various classification thresholds, was used to verify the accuracy of our predictions. A measure of the prediction accuracy is provided by the area under the ROC curve (A_{ROC}), where $A_{ROC} < 70\%$ denotes poor, $A_{ROC} > 80\%$ is for good and $A_{ROC} > 90\%$ represents excellent predictions (Brusic, *et al.*, 2002). The values of $SP \geq 80\%$ are considered useful in practice (Tong, *et al.*, 2006b). Thus, SE values for three values of SP (80%, 90% and 95%) in test set 1, were assessed.

3 RESULTS AND DISCUSSION

Evaluation of the accuracy of the DQ8 prediction model was carried out in two steps: (i) assessment of efficiency of the optimized scoring function using test set 1; and (ii) verification of the overall prediction accuracy of the model using test set 2. The accuracy of

our model partially relies on the scoring function used. Reasonable correlation ($r^2=0.79$, $s=2.05$ kJ/mol) between the predicted binding energy values (from docking) and the experimental binding free energy values (computed using IC_{50} values), was obtained for the training set by using default ICM coefficients ($\alpha=\beta=\gamma=1$; $C=0$) in the scoring function. Better correlation ($r^2=0.82$, $s=1.95$ kJ/mol) was achieved after recalibration of the scoring function by fitting to the training data using multiple linear regression thereby significantly improving the discriminative power of the scoring function. Following 10-fold cross-validation ($N=51$, $q^2=0.80$, $s_{press}=2.20$ kJ/mol), the optimal scoring function is:

$$\Delta G_{bind} = 0.015\Delta G_{EL} - 0.859\Delta G_H + 0.827\Delta G_{EN} - 1.91 \quad (2)$$

The entropic and the electrostatic terms are positive, while the overall computed binding energy and the hydrophobic term are negative.

Rognan *et al.*, (1999) performed a leave-one-out cross-validation on training datasets of five and 37 pMHC complexes. However, the current training set of 57 complexes is comparatively larger for such analyses. Contrastingly, Wang *et al.*, (2002) and Bock and Gough (2002) used training sets of 200 and 2617 complexes, respectively, for extensive cross-validation analyses. Yet again, our training set is too small for extensive cross-validation analyses. It is also worth noting that the standard error in the training set after recalibration of the scoring function ($s=1.95$ kJ/mol= 0.46 kcal/mol) is less than the standard error after 10-fold cross-validation ($s_{press}=2.20$ kJ/mol= 0.52 kcal/mol) as expected and unlike the higher standard error after recalibration ($s=4.77$ kJ/mol= 1.13 kcal/mol) than the standard error after 10-fold cross-validation ($s_{press}=2.20$ kJ/mol= 0.52 kcal/mol) reported by Tong *et al.*, (2006b), highlighting the ability of pDOCK (Khan and Ranganathan, 2010) to handle noise in data and showcasing its robustness. Also, our standard error values both before and after recalibration ($s=2.05$ kJ/mol and $s=1.95$ kJ/mol, respectively) are significantly lower to the ones ($s=2.91$ kJ/mol and $s=4.77$ kJ/mol, respectively) documented by Tong *et al.*, (2006b). We have also carried out evolutionary regression analysis, similar to the one carried out previously (Tong, *et al.*, 2006b), to estimate the robustness of the scoring function for 5-fold ($N=46$, $q^2=0.79$, $s_{press}=2.09$ kJ/mol), 4-fold ($N=43$, $q^2=0.77$, $s_{press}=2.07$ kJ/mol), 3-fold ($N=38$, $q^2=0.78$, $s_{press}=2.05$ kJ/mol) and 2-fold ($N=29$, $q^2=0.74$, $s_{press}=2.03$ kJ/mol) cross-validations and once again the standard error values for our training set were comparatively lower. Importantly, the cross-validation coefficient q^2 and the standard error of prediction s_{press} are stable all through, with mean values of $q^2=0.78$ and $s_{press}=2.09$ kJ/mol, and the respective standard deviation values of 0.02 and 0.07 kJ/mol. These results do not indicate any unusual increase in the standard error values for any of the folds contrasting to the reports of an uncharacteristic increase in the error value for 2-fold cross-validation by Tong *et al.*, (2006b). The internal consistency of the optimized scoring function used in this prediction model is therefore validated by this iterative regression procedure, rendering it suitable for the identification of MHC-binders within the test datasets and hence for use in our prediction model. In order to evaluate the efficiency of the optimized scoring function, three decision threshold binding energy values (Table 1), which define levels of specificities suitable for practical applications (Brusic, *et al.*, 2002), were used to determine the correspond-

ing sensitivity values on different subsets – H (high-affinity binders only; $A_{ROC}=0.89$); MH (medium- and high-affinity binders; $A_{ROC}=0.96$); and LMH (low-, medium-, high-affinity binders; $A_{ROC}=0.98$) from test set 1. The suitable use of structural data for discriminating MHC-II binding peptides from the background with almost excellent accuracy ($A_{ROC}\geq 0.89$) is advocated by these outcomes. In general, very few false positives and a large number of true positives are observed at $SP=0.95$ contrasting to the previous report of fewer true positives at $SP=0.95$ (Tong, *et al.*, 2006b), shedding light on pDOCK's efficiency even at higher levels of specificity. High-sensitivity predictions are commonly expected at $SP=0.80$ (Tong, *et al.*, 2006b). Our MHC-binding prediction results for test set 1 (Table 1) fit almost perfectly with the expected binding patterns of DQ8-binding peptides, providing a sensitivity of 99% (at $SP=0.80$ for MH and H). With higher levels of specificity, a gradual decrease in sensitivity values (at $SP=0.90$, $SE=0.97$ for LMH and MH and $SE=0.96$ for H; at $SP=0.95$, $SE=0.97$, 0.94 and 0.89 for LMH, MH and H, respectively) is observed. On an average however, the sensitivity values are above 96%, with $S=0.89$ (89% of high-affinity binders are correctly identified) being the worst case scenario. The efficacy of pDOCK (Khan and Ranganathan, 2010) in accurately detecting binding registers was also evaluated with experimentally determined registers. Our findings reconfirm our earlier observation (Khan and Ranganathan, 2010) that pDOCK accurately determines binding registers for all the peptides docked from the training set and test set 1 (Supplementary Tables S1 and S2, respectively). All 22 experimentally determined registers from test set 2 (Supplementary Table S3) were also correctly predicted by pDOCK.

Table 1. Identification of MHC-binders to DQ8: sensitivity values and binding energy thresholds for specificity levels of 0.80, 0.90 and 0.95

Specificity (SP) Level	Group	Sensitivity (SE)	Binding Energy Threshold (kJ/mol)
SP = 0.80	LMH	0.98	-29.55
	MH	0.99	-34.00
	H	0.99	-35.20
SP = 0.90	LMH	0.97	-29.63
	MH	0.97	-34.90
	H	0.96	-36.50
SP = 0.95	LMH	0.97	-29.70
	MH	0.94	-35.25
	H	0.89	-37.91

Finally, using the binding energy decision threshold (-37.91 kJ/mol) defined above for high-affinity binders at the specificity of 95% ($SE=0.89$), the predictive performance and the accuracy of the scoring function was tested on a functional dataset of 36 peptides (test set 2) known to bind DQ8, out of which, 31 were T cell activators and 5 were non-activators. However, the top 31 predictions by applying the scoring function included the 5 non-activators. Through structural analysis of all available TR/pMHC crystal structures, we have recently shown that MSEP at both pMHC and TR binding interfaces play a major role in TR/pMHC complex formation and thus in T cell activation (Khan and Ranganathan, 2011). Therefore, we applied this knowledge and

performed MSEP-based clustering of all peptide docked pMHC interfaces from test set 2 along with pMHC interfaces from all six available TR/pMHC-II structures using the webPIPSA (Richter, *et al.*, 2008) server, to identify electrostatic similarities between them. The electrostatic distances for all pMHC interfaces clustered, varied from 0.063 to 1.187 (colour key and density plot-inset in Supplementary Figure S1). Clustering identified nine pMHC interfaces (group E in Supplementary Figure S1) as non-agonists which included two known non-activators (pMHC 4 and 8 in group E from Supplementary Figure S1; peptides 34 and 35 in Supplementary Table S3). Hence, combining the results from the application of the optimized scoring function and from clustering, those peptides that were predicted to be T cell activators in both instances were selected, resulting in 24 true positives (Supplementary Table S4). The predicted binding energy values for these 24 peptides range from -50.15 to -36.96 kJ/mol and all of these are known T cell activators. This is in accordance with existing reports that high-affinity binders have a greater chance of stimulating T cells (Deng, *et al.*, 1997; Keogh, *et al.*, 2001; Jensen, 2007) and that they are critical for peptide vaccine design. Nonetheless, TR agonistic properties of pMHC interfaces obtained from MSEP-based clustering are used as primary criterion and predicted pMHC binding energy is used as secondary parameter in ranking the peptides in Supplementary Table S4. This is primarily because although strong pMHC binding is a prerequisite for TR recognition (Jensen, 2007), it does not necessarily mean T cell activation as there is only a 50% chance of immunogenicity (Yewdell and Bennink, 1999) even among strong MHC-binding peptides.

Our prediction model successfully identified the Dermatophagoides pteronyssinus (*Der p 2*) allergenic peptide 31-50 (pMHC 4 in group E from Supplementary Figure S1; peptide 34 in Supplementary Table S3) as a non-activator as opposed to its identification as a T cell activator by Tong *et al.*, (2006b). Similarly, the *Der p 2* peptide 41-60 (Supplementary Table S4) is ranked 4 in our prediction which is in complete agreement with the experimental study (Krco, *et al.*, 2000) and is again contrasting to its last (#12) ranking prediction made by Tong *et al.*, (2006b). The above two examples shed light on the efficiency of our prediction model. This combined approach correctly predicts 24 of the 31 T cell activators, resulting in 77% accuracy for successful prediction of immunogenic T cell epitopes. The specificity and sensitivity results are also consistent with the results obtained from ROC analysis. Therefore, the current prediction model is suitable to screen for high-affinity binders at $SP=0.95$ and then, to identify immunogenic T cell epitopes among the high-binders.

3.1 Epitopes not conforming to binding motifs exist

Identification of potential immunodominant epitopes within autoantigenic proteins has been done for many MHC-II alleles by developing allele-specific consensus peptide-binding motifs. Nevertheless, there have been reports that the existence of these motifs in a given peptide does not necessarily render allele-specificity to it (Harfouch-Hammoud, *et al.*, 1999). This study has revealed that considering all relevant residue positions (P1, P4, P6, P7, P9) for peptides, all 70 peptide sequences (Supplementary Table S5) have one or more amino acid residues and 56 peptides (Supplementary Table S5) have two or more residues that do not conform to available DQ8 binding motifs (Godkin, *et al.*, 1997; Rammensee, *et al.*,

1999), in test set 1 alone. Despite using existing DQ8 binding motifs, the peptides A-gliadin 49-63 (#1), VP16 (#23) and MHC Ia 46-63 (#24) are generally considered negatives, however, from supplementary table S5 it is evident that these T cell epitopes are easily identified just by using our scoring function, thereby, reaffirming earlier observations by Tong *et al.*, (2006b). This yet again proves that many other factors such as peptide and MHC binding groove physicochemical composition have to be considered in T cell epitope prediction systems and binding motifs by themselves are inadequate for identifying T cell epitopes. Despite rapid advances in peptide vaccine development, identifying allele-specific T cell epitopes, especially MHC-II restricted epitopes, suitable for designing vaccines and immunotherapies remains a challenging prospect. Various excellent approaches (Brusic, *et al.*, 1998; Mallios, 2001; Doytchinova and Flower, 2003) have been adopted by researchers to address this issue. However, training of predictive models using peptide nonamers preselected based on existing binding motifs renders them unable to predict epitopes that do not conform to binding motifs.

Our overall significant outcomes along with increasing evidence for inadequacy of binding motifs in defining T cell epitopes, suggest that we have developed a model that can be successfully applied as a generic protocol for easy *in silico* identification of potential immunogenic T cell epitopes. The current model is therefore applicable for screening vaccine candidates irrespective of sequence motifs. Also, pDOCK (Khan and Ranganathan, 2010) accurately predicts all binding registers, eliminating the use of the nine-residue sliding window approach used by Tong *et al.*, (2006b) resulting in multiple registers within candidate DQ8-binding peptides. We have also illustrated efficient discrimination of different categories of peptide binders from non-binders, using the scoring function, as well as different categories of pMHC agonists from non-agonists, using MSEP, while accurately predicting the binding register of DQ8-restricted peptides. This combined approach provides a set of sensitive and specific computational tools to facilitate high-throughput screening of peptides for immunotherapeutic applications such as controlling allergic and autoimmune responses.

ACKNOWLEDGEMENTS

JMK gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship. We also thank Mr. H.R. Cheruku, Macquarie University, Sydney for assistance with supplementary tables and Dr. Vladimir Brusic, Dana-Farber Cancer Institute, Boston for help with binding energy threshold determination. Open access publication charges were borne by Macquarie University.

REFERENCES

- Abagyan, R.A. and Totrov, M.M. (1999) *Ab Initio* Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure, *J Comput Phys*, **151**, 402-421.
- Abagyan, R.A., *et al.* (1994) ICM – a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation., *J Comput Chem*, **15**, 488-506.
- Berman, H.M., *et al.* (2002) The Protein Data Bank, *Acta Crystallogr D Biol Crystallogr*, **58**, 899-907.
- Bharadwaj, R., *et al.* (1995) The fast multipole boundary element method for molecular electrostatics: An optimal approach for large systems, *Journal of Computational Chemistry*, **16**, 898-913.
- Blomberg, N., *et al.* (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity, *Proteins*, **37**, 379-387.
- Bock, J.R. and Gough, D.A. (2002) A new method to estimate ligand-receptor energetics, *Mol Cell Proteomics*, **1**, 904-910.
- Brusic, V., *et al.* (2004) Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications, *Methods*, **34**, 436-443.
- Brusic, V., *et al.* (2002) Prediction of promiscuous peptides that bind HLA class I molecules, *Immunol Cell Biol*, **80**, 280-285.
- Brusic, V., *et al.* (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, *Bioinformatics*, **14**, 121-130.
- Chang, K.Y. and Unanue, E.R. (2009) Prediction of HLA-DQ8beta cell peptidome using a computational program and its relationship to autoreactive T cells, *Int Immunol*, **21**, 705-713.
- Deng, Y., *et al.* (1997) MHC affinity, peptide liberation, T cell repertoire, and immunodominance all contribute to the paucity of MHC class I-restricted peptides recognized by antiviral CTL, *J Immunol*, **158**, 1507-1515.
- Dimitrov, I., *et al.* (2010) Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis, *Eur J Med Chem*, **45**, 236-243.
- Doytchinova, I.A. and Flower, D.R. (2003) Towards the *in silico* identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction, *Bioinformatics*, **19**, 2263-2270.
- Erlich, H.A., *et al.* (1993) HLA class II alleles and susceptibility and resistance to insulin dependent diabetes mellitus in Mexican-American families, *Nat Genet*, **3**, 358-364.
- Faustman, D., *et al.* (1991) Linkage of faulty major histocompatibility complex class I to autoimmune diabetes, *Science*, **254**, 1756-1761.
- Godkin, A., *et al.* (1997) Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2), *Int Immunol*, **9**, 905-911.
- Godkin, A.J., *et al.* (1998) Use of complete eluted peptide sequence data from HLA-DR and -DQ molecules to predict T cell epitopes, and the influence of the nonbinding terminal regions of ligands in epitope selection, *J Immunol*, **161**, 850-858.
- Gonzalez-Galarza, F.F., *et al.* (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations, *Nucleic Acids Res*, **39**, D913-919.
- Graham, J., *et al.* (2002) Genetic effects on age-dependent onset and islet cell autoantibody markers in type 1 diabetes, *Diabetes*, **51**, 1346-1355.
- Halgren, T.A. (1995) Merck Molecular Force Field. I-V., *J Comp Chem*, **17**, 490-641.
- Harfouch-Hammoud, E., *et al.* (1999) Identification of peptides from autoantigens GAD65 and IA-2 that bind to HLA class II molecules predisposing to or protecting from type 1 diabetes, *Diabetes*, **48**, 1937-1947.
- Hoof, I., *et al.* (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans, *Immunogenetics*, **61**, 1-13.

- Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*, **5**, 299-314.
- Janin, J. (1995) Protein-protein recognition, *Prog Biophys Mol Biol*, **64**, 145-166.
- Jensen, P.E. (2007) Recent advances in antigen processing and presentation, *Nat Immunol*, **8**, 1041-1048.
- Keogh, E., *et al.* (2001) Identification of new epitopes from four different tumor-associated antigens: recognition of naturally processed epitopes correlates with HLA-A*0201-binding affinity, *J Immunol*, **167**, 787-796.
- Khan, J.M., *et al.* (2011) MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures, *Bioinformatics*, **27**, 1192-1193.
- Khan, J.M. and Ranganathan, S. (2010) pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes, *Immunome Res*, **6** Suppl 1, S2.
- Khan, J.M. and Ranganathan, S. (2011) Understanding TR binding to pMHC complexes: how does the TR scan many pMHC molecules yet preferentially bind to one., *PLoS One*, **6**, e17194.
- Khan, J.M., *et al.* (2010) Structural Immunoinformatics: Understanding MHC-peptide-TR binding. In Davies, M.N., Ranganathan, S. and Flower, D.R. (eds), *Bioinformatics for Immunomics*. Springer, New York, 77-94.
- Klein, L., *et al.* (2000) Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells, *Nat Med*, **6**, 56-61.
- Krco, C.J., *et al.* (2000) Immune response of HLA-DQ transgenic mice to house dust mite allergen p2: identification of HLA-DQ restricted minimal epitopes and critical residues, *Clin Immunol*, **97**, 154-161.
- Krystek, S., *et al.* (1993) Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures, *J Mol Biol*, **234**, 661-679.
- Lee, K.H., *et al.* (2001) Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes, *Nat Immunol*, **2**, 501-507.
- Lefranc, M.P. and Lefranc, G. (2001) *The T cell receptor Facts-Book*. Academic Press, San Diego.
- Madura, J.D., *et al.* (1995) Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program, *Computer Physics Communications*, **91**, 57-95.
- Mallios, R.R. (2001) Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm, *Bioinformatics*, **17**, 942-948.
- Mangalam, A., *et al.* (2009) HLA-DQ8 (DQB1*0302)-restricted Th17 cells exacerbate experimental autoimmune encephalomyelitis in HLA-DR3-transgenic mice, *J Immunol*, **182**, 5131-5139.
- Neeno, T., *et al.* (1996) HLA-DQ8 transgenic mice lacking endogenous class II molecules respond to house dust allergens: identification of antigenic epitopes, *J Immunol*, **156**, 3191-3195.
- Nemethy, G., *et al.* (1992) Energy parameters in polypeptides, 10: Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides., *J Phys Chem*, **96**, 6472-6484.
- Nepom, G.T. and Kwok, W.W. (1998) Molecular basis for HLA-DQ associations with IDDM, *Diabetes*, **47**, 1177-1184.
- Nielsen, M., *et al.* (2008) Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan, *PLoS Comput Biol*, **4**, e1000107.
- Novotny, J., *et al.* (1997) Empirical free energy calculations: a blind test and further improvements to the method, *J Mol Biol*, **268**, 401-411.
- Paisansinsup, T., *et al.* (2002) HLA class II influences the immune response and antibody diversification to Ro60/Sjogren's syndrome-A: heightened antibody responses and epitope spreading in mice expressing HLA-DR molecules, *J Immunol*, **168**, 5876-5884.
- Rammensee, H., *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs, *Immunogenetics*, **50**, 213-219.
- Rammensee, H.G., *et al.* (1993) Peptides naturally presented by MHC class I molecules, *Annu Rev Immunol*, **11**, 213-244.
- Reche, P.A., *et al.* (2002) Prediction of MHC class I binding peptides using profile motifs, *Hum Immunol*, **63**, 701-709.
- Richter, S., *et al.* (2008) webPIPSA: a web server for the comparison of protein interaction properties, *Nucleic Acids Research*, **36**, W276-W280.
- Robles, D.T., *et al.* (2002) The genetics of autoimmune polyendocrine syndrome type II, *Endocrinol Metab Clin North Am*, **31**, 353-368, vi-vii.
- Rognan, D., *et al.* (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins, *J Med Chem*, **42**, 4650-4658.
- Schapira, M., *et al.* (1999) Prediction of the binding energy for small molecules, peptides and proteins, *J Mol Recognit*, **12**, 177-190.
- Sidney, J., *et al.* (2002) The HLA molecules DQA1*0501/B1*0201 and DQA1*0301/B1*0302 share an extensive overlap in peptide binding specificity, *J Immunol*, **169**, 5098-5108.
- Sollid, L.M. and Thorsby, E. (1993) HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis, *Gastroenterology*, **105**, 910-922.
- Suri, A., *et al.* (2005) Natural peptides selected by diabetogenic DQ8 and murine I-A(g7) molecules show common sequence specificity, *J Clin Invest*, **115**, 2268-2276.
- Tong, J.C., *et al.* (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex, *Protein Sci*, **13**, 2523-2532.
- Tong, J.C., *et al.* (2006a) Prediction of desmoglein-3 peptides reveals multiple shared T-cell epitopes in HLA DR4- and DR6-associated pemphigus vulgaris, *BMC Bioinformatics*, **7** Suppl 5, S7.
- Tong, J.C., *et al.* (2006b) Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides, *Bioinformatics*, **22**, 1232-1238.
- Tong, J.C., *et al.* (2007) In silico characterization of immunogenic epitopes presented by HLA-Cw*0401, *Immunome Res*, **3**, 7.
- Wang, P., *et al.* (2010) Peptide binding predictions for HLA DR, DP and DQ molecules, *BMC Bioinformatics*, **11**, 568.
- Wang, R., *et al.* (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction, *J Comput Aided Mol Des*, **16**, 11-26.
- Yewdell, J.W. and Bennink, J.R. (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses, *Annu Rev Immunol*, **17**, 51-88.

Supplementary Table S1

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S1. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used in the training set for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#55).

S.No	Description	Peptide sequence	IC ₅₀ (nM)	Reference
1	Thyroid per 632-645Y	IDV <u>WLGGLAEN</u> FLPY	39.00	Sidney et al. 2002
2	Thyroid per 632-645Y analog	IDV <u>WLGGLAENV</u> LPY	22.97	Sidney et al. 2002
3	Thyroid per 632-645Y analog	IDV <u>WLGGLAESF</u> LPY	17.94	Sidney et al. 2002
4	Thyroid per 632-645Y analog	IDV <u>WLGGLAEDF</u> LPY	33.85	Sidney et al. 2002
5	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPD	25.74	Sidney et al. 2002
6	Thyroid per 632-645Y analog	IDV <u>LLGGLAENF</u> LPY	119.29	Sidney et al. 2002
7	Thyroid per 632-645Y analog	IDV <u>WLGGLAEYF</u> LPY	24.58	Sidney et al. 2002
8	Thyroid per 632-645Y analog	IDV <u>WLGGLAENY</u> LPY	30.49	Sidney et al. 2002
9	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPL	32.18	Sidney et al. 2002
10	Thyroid per 632-645Y analog	IDV <u>WLGGLAEKF</u> LPY	31.47	Sidney et al. 2002
11	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> YPY	57.61	Sidney et al. 2002
12	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> DPY	25.35	Sidney et al. 2002
13	Thyroid per 632-645Y analog	IDV <u>WLGGLAEND</u> LPY	16.77	Sidney et al. 2002
14	Thyroid per 632-645Y analog	IDV <u>WLGYLEAENF</u> LPY	325.00	Sidney et al. 2002
15	Thyroid per 632-645Y analog	IDV <u>SLGGLAENF</u> LPY	72.43	Sidney et al. 2002
16	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LSY	195.00	Sidney et al. 2002
17	Thyroid per 632-645Y analog	IDV <u>WLGGLAEQF</u> LPY	34.50	Sidney et al. 2002
18	Thyroid per 632-645Y analog	IDV <u>WLGGVAENF</u> LPY	92.86	Sidney et al. 2002
19	Thyroid per 632-645Y analog	IDV <u>WYGGLAENF</u> LPY	62.95	Sidney et al. 2002
20	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LLY	139.29	Sidney et al. 2002
21	Thyroid per 632-645Y analog	IDV <u>WLLGLAENF</u> LPY	130.00	Sidney et al. 2002
22	Thyroid per 632-645Y analog	IDV <u>WSGGLAENF</u> LPY	35.88	Sidney et al. 2002
23	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LKY	278.57	Sidney et al. 2002
24	Thyroid per 632-645Y analog	IDV <u>WLGGSAENF</u> LPY	216.67	Sidney et al. 2002
25	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> VPY	76.82	Sidney et al. 2002
26	Thyroid per 632-645Y analog	IDV <u>WLYGLAENF</u> LPY	125.81	Sidney et al. 2002
27	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPK	66.00	Sidney et al. 2002
28	Thyroid per 632-645Y analog	IDV <u>WLGGALN</u> FLPY	130.00	Sidney et al. 2002
29	Thyroid per 632-645Y analog	IDV <u>WLGGKAENF</u> LPY	177.27	Sidney et al. 2002
30	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> LPF	50.47	Sidney et al. 2002
31	Thyroid per 632-645Y analog	IDV <u>WLGGYAENF</u> LPY	139.29	Sidney et al. 2002
32	Thyroid per 632-645Y analog	IDV <u>WLKGLAENF</u> LPY	325.00	Sidney et al. 2002
33	Thyroid per 632-645Y analog	IDV <u>WLDGLAENF</u> LPY	100.00	Sidney et al. 2002

S.No	Description	Peptide	IC ₅₀ (nM)	Reference
34	Thyroid per 632-645Y analog	IDV <u>WLGGLAEN</u> FLPS	69.57	Sidney et al. 2002
35	Thyroid per 632-645Y analog	IDV <u>WLGGLAENK</u> LPY	1677.00	Sidney et al. 2002
36	Thyroid per 632-645Y analog	IDV <u>KLGGGLAEN</u> FLPY	2028.00	Sidney et al. 2002
37	Thyroid per 632-645Y analog	IDV <u>WDGGLAEN</u> FLPY	83.22	Sidney et al. 2002
38	Thyroid per 632-645Y analog	IDV <u>WKGGLAEN</u> FLPY	96.97	Sidney et al. 2002
39	Thyroid per 632-645Y analog	IDV <u>WLSGLAEN</u> FLPY	105.41	Sidney et al. 2002
40	Thyroid per 632-645Y analog	IDV <u>WLGSGLAEN</u> FLPY	78.00	Sidney et al. 2002
41	Thyroid per 632-645Y analog	IDV <u>WLGDGLAEN</u> FLPY	105.41	Sidney et al. 2002
42	Thyroid per 632-645Y analog	IDV <u>WLGGLSEN</u> FLPY	390.00	Sidney et al. 2002
43	Thyroid per 632-645Y analog	IDV <u>WLGGLDEN</u> FLPY	177.27	Sidney et al. 2002
44	Thyroid per 632-645Y analog	IDV <u>WLGGLAEN</u> FLYY	108.33	Sidney et al. 2002
45	Thyroid per 632-645Y analog	IDV <u>WLGGLAEN</u> FSPY	53.94	Sidney et al. 2002
46	MHC II E8 51-65	FDG <u>DEIFHVDIE</u> KSE	1000.00	Suri et al. 2005
47	MHC II E8 51-65 analog	FDG <u>DEIAHVDIE</u> KSE	3300.00	Suri et al. 2005
48	TRAIL receptor 2 364-380 analog	GRFTY <u>QNAAAQPA</u> TGPG	1000.00	Suri et al. 2005
49	Nicastrin 65-78	ISG <u>DTGVIHVVE</u> KE	1000.00	Suri et al. 2005
50	Nicastrin 65-78 analog	ISG <u>DTGVIHVVA</u> KE	4300.00	Suri et al. 2005
51	E25B protein 112-126	YQTI <u>EENIKIFEED</u> A	800.00	Suri et al. 2005
52	E25B protein 112-126 analog	YQTI <u>EENIKIFE</u> EKA	1700.00	Suri et al. 2005
53	ZnT8 diabetic autoantigen	LYP <u>DYQIQAGIM</u> IT	700.00	Chang and Unanue, 2009
54	ZnT8 diabetic autoantigen	AVD <u>GVISVHSLHI</u> W	18000.00	Chang and Unanue, 2009
55	ZnT8 diabetic autoantigen	SKRL <u>TFGWYRAE</u> IL	20200.00	Chang and Unanue, 2009
56	B2m 91–104	TPTEKDEYCARVNH	10000.00	Sidney et al. 2002
57	Artificial sequence	YARFQSQTTLKQKT	10000.00	Sidney et al. 2002

Supplementary Table S2

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S2. HLA-DQ8 specific peptides with experimentally determined IC₅₀ values used as test set 1 for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#32).

S.No.	Description	Peptide sequence	IC ₅₀ (nM)	Reference
1	Thyroid per 632-645Y analog	IDV <u>WLGGLAELF</u> LPY	22.71	Sidney et al. 2002
2	Thyroid per 632-645Y analog	IDV <u>WLGGLAENS</u> LPY	34.94	Sidney et al. 2002
3	Thyroid per 632-645Y analog	IDV <u>DLGGLAENF</u> LPY	20.28	Sidney et al. 2002
4	Thyroid per 632-645Y analog	IDV <u>WLGGLYENF</u> LPY	195.00	Sidney et al. 2002
5	Thyroid per 632-645Y analog	IDV <u>WLGGLAYNF</u> LPY	100.00	Sidney et al. 2002
6	Thyroid per 632-645Y analog	IDV <u>VLGGLAENF</u> LPY	52.00	Sidney et al. 2002
7	Thyroid per 632-645Y analog	IDV <u>WVGGLAENF</u> LPY	44.30	Sidney et al. 2002
8	Thyroid per 632-645Y analog	IDV <u>WLGGLAENFL</u> DY	41.94	Sidney et al. 2002
9	Thyroid per 632-645Y analog	IDV <u>WLGGLASNF</u> LPY	354.55	Sidney et al. 2002
10	Thyroid per 632-645Y analog	IDV <u>WLGLLAENF</u> LPY	50.65	Sidney et al. 2002
11	Thyroid per 632-645Y analog	IDV <u>WLGGLAENF</u> KPY	74.56	Sidney et al. 2002
12	E25B protein 112-126 analog	YQTI <u>EENIKIFKE</u> DA	1000.00	Suri et al. 2005
13	E25B protein 112-126 analog	YQTI <u>EENIKIFE</u> DA	1800.00	Suri et al. 2005
14	E25B protein 112-126 analog	YQTI <u>EENIKIFE</u> AAA	2500.00	Suri et al. 2005
15	E25B protein 112-126 analog	YQTI <u>EENIKIFE</u> AAAA	1700.00	Suri et al. 2005
16	E25B protein 112-126 analog	YQTI <u>KENIKIFE</u> EDA	3800.00	Suri et al. 2005
17	TRAIL receptor 2 364-380 analog	GRFT <u>KQNAAAQPE</u> TGPG	3700.00	Suri et al. 2005
18	TRAIL receptor 2 364-380 analog	GRFT <u>AQNAAAQPA</u> TGPG	3100.00	Suri et al. 2005
19	TRAIL receptor 2 364-380 analog	GRFT <u>AQNAAAQPE</u> TGPG	1700.00	Suri et al. 2005
20	TRAIL receptor 2 364-380 analog	GRFT <u>YQNAAAQPE</u> TGPG	1700.00	Suri et al. 2005
21	Nicastrin 65-78 analog	ISG <u>DTGVHVV</u> AKE	4300.00	Suri et al. 2005
22	Nicastrin 65-78 analog	ISG <u>ATGVHVV</u> EKE	2300.00	Suri et al. 2005

S.No.	Description	Peptide sequence	IC ₅₀ (nM)	Reference
23	Superoxide dimutase 1 90-103	AGK <u>DGVANVSIEDR</u>	2000.00	Suri et al. 2005
24	Superoxide dimutase 1 90-103 analog	AGK <u>AGVANVSIEDR</u>	1800.00	Suri et al. 2005
25	Superoxide dimutase 1 90-103 analog	AGK <u>DGVANASIEDR</u>	2800.00	Suri et al. 2005
26	MHC II E8 51-65 analog	FDG <u>KEIFHVDIEKSE</u>	2800.00	Suri et al. 2005
27	MHC II E8 51-65 analog	FDG <u>AEIFHVDIEKSE</u>	2000.00	Suri et al. 2005
28	MHC II E8 51-65 analog	FDG <u>DEIFHADIEKSE</u>	3100.00	Suri et al. 2005
29	ZnT8 diabetic autoantigen	LT <u>QIESAADQDPS</u>	2500.00	Chang and Unanue, 2009
30	ZnT8 diabetic autoantigen	RTG <u>IAQALSSFDLH</u>	2500.00	Chang and Unanue, 2009
31	ZnT8 diabetic autoantigen	ILS <u>VHVATAASQDS</u>	4900.00	Chang and Unanue, 2009
32	ZnT8 diabetic autoantigen	AIL <u>TDAAHLLIDLT</u>	7200.00	Chang and Unanue, 2009
33	A-gliadin 30-44	FPGQQQQFPPQQPYP	600.00	Godkin et al. 1998
34	A-gliadin 34-48	QQQFPPQQPYPQPQP	10000.00	Godkin et al. 1998
35	A-gliadin 41-55	QPYPQPQPFPSQQPY	1120.00	Godkin et al. 1998
36	A-gliadin 49-63	FPSQQPYLQLQFPFQ	20.00	Godkin et al. 1998
37	A-gliadin 56-70	LQLQFPFQPQFPPL	20.00	Godkin et al. 1998
38	A-gliadin 77-91	SFPPQQPYPQPQPQY	370.00	Godkin et al. 1998
39	A-gliadin 196-210	PSSQFQQPLQQYPLG	10000.00	Godkin et al. 1998
40	A-gliadin 201-215	QQPLQQYPLQGGSFR	2180.00	Godkin et al. 1998
41	A-gliadin 207-221	YPLGQGSFRPSQQNP	100.00	Godkin et al. 1998
42	A-gliadin 227-241	VQPQQQLPQFEIRNL	73.00	Godkin et al. 1998
43	34P3A	IARAKMFPAVAEK	541.00	Sidney et al. 2002
44	Artificial sequence	AAAAVAEAY	48.00	Sidney et al. 2002
45	Artificial sequence	YARFQRQTTLKAAA	10000.00	Sidney et al. 2002
46	Artificial sequence (ROIV)	YAHAAHAAHAAHAA	2942.00	Sidney et al. 2002
47	CD20 249–262 analog	EEDIEIPIQEEY	21.00	Sidney et al. 2002
48	CLIP 95-102	KPVSKMRMATPLLMQALP	650.00	Sidney et al. 2002

S.No.	Description	Peptide sequence	IC ₅₀ (nM)	Reference
49	CLIP 96-114	KLPKPPKPVSKMRMATPLL	10000.00	Sidney et al. 2002
50	DQa1 0501 16-30	YQSYGPSGQYTHEFD	10000.00	Sidney et al. 2002
51	FceR 104–122	SQDLELSWNLNGLQADLSS	123.00	Sidney et al. 2002
52	FceR 104–122 analog	SQDLELSWNLNGLQAY	118.00	Sidney et al. 2002
53	GAD 101–115	CDGERPTLAFLQDVM	69.00	Sidney et al. 2002
54	MHC Ia 46-63	EPRAPWIEQEGPEYW	519.00	Sidney et al. 2002
55	GAD65 253–265	IARFKMFPEVKEK	3712.00	Sidney et al. 2002
56	HA 255–271Y	FESTGNLIAPEYGFKISY	62.00	Sidney et al. 2002
57	HSV	DMTPADALDDFDL	173.00	Sidney et al. 2002
58	IA-2 499-509	GVAGLLVALAV	95.00	Sidney et al. 2002
59	IA-2 499–509	MSSGSFINISV	2470.00	Sidney et al. 2002
60	Insulin B 5–15	FVNQHLCGSHLVEAL	10000.00	Sidney et al. 2002
61	Lamba repressor 12–24	LEDARRLKAIYEK	717.00	Sidney et al. 2002
62	Lol p1 101–120	APYHFDLSGHAFGSMAKKGE	3602.00	Sidney et al. 2002
63	MHC Ia 51–63 analog	YPFIEQEGPEFFDQE	1156.00	Sidney et al. 2002
64	ML LSR2 5–17	GVTYEIDLTNKN	10000.00	Sidney et al. 2002
65	OVA 267-276 Y	LTEWTSSNVMEERY	62.00	Sidney et al. 2002
66	p21 51–66; C out	LLDILDTAGLEEYSAMRD	202.00	Sidney et al. 2002
67	Pf ABRA 487–506	DSNIMNSINNVMEIDFFEK	171.00	Sidney et al. 2002
68	Pf cp 379–396 truncated analog	IEKKIAKMEKASY	10000.00	Sidney et al. 2002
69	Pf MSP-1 250-271	FGYRKPLDNIKDNVGKMEDYIKK	10000.00	Sidney et al. 2002
70	VP16	PPLYATGRLSQAQLMPSPPM	538.00	Sidney et al. 2002

Supplementary Table S3

***In silico* prediction of immunogenic T cell epitopes for HLA-DQ8**

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S3. HLA-DQ8 specific peptides used as test set 2 for this study.

The nonamer in the binding groove is underlined in bold font for peptides with experimentally determined binding registers (#1-#22). Peptides that do not elicit T cell response after binding to the MHC are in italics.

S.No.	Description	Amino acid residue numbers	Peptide sequence	Reference
1	Dermatophagoides ptermyssinus (Der p 2) allergenic peptide	1-20	DQVDVKDCANHEIKKVLVPG	Neeno et al. 1996 and Kroo et al. 2000
2	Human GAD65	126-140	FDRSTKVIDFH YYPNE	Chang and Unanue, 2009
3	Human GAD65	206-220	TYE IA PV FV LL LE YVT	Chang and Unanue, 2009
4	Human GAD65	231-250	PGSGDGIFSPG GAIS NM YA	Chang and Unanue, 2009
5	Human GAD65	431-445	KHYDLSYDTGDKALQ	Chang and Unanue, 2009
6	Human GAD65	461-475	AKGTTGFE AHVD KCL	Chang and Unanue, 2009
7	Human GAD65	471-490	VDKCLELAELY LYNIIKN REG	Chang and Unanue, 2009
8	Human GAD65	536-550	RMM EY GTTMVSYQPL	Chang and Unanue, 2009
9	Islet Antigen-2	601-618	RQHARQQDK ERLAALGPE	Chang and Unanue, 2009
10	Islet Antigen-2	616-633	GPE GA HGDTTFEYQDL CR	Chang and Unanue, 2009
11	Islet Antigen-2	646-663	EGPPEP SRVSSVSSQ FSD	Chang and Unanue, 2009
12	Islet Antigen-2	661-678	FSD AAQAQ SPSSHSSSTPSW	Chang and Unanue, 2009
13	Islet Antigen-2	721-738	AEPNTC ATAQ GE GN IKN	Chang and Unanue, 2009
14	Islet Antigen-2	826-843	DE GAS LYHVYEVN LV SEH	Chang and Unanue, 2009
15	Islet Antigen-2	931-948	KGV KEIDIA AT LE HV RD Q	Chang and Unanue, 2009
16	Islet Antigen-2	961-979	FALT AVAE EEV NAILKALP	Chang and Unanue, 2009
17	Preproinsulin	1-24	MALWMRLPL LLALLWGP DPAAA	Chang and Unanue, 2009

S.No.	Description	Amino acid residue numbers	Peptide sequence	Reference
18	Preproinsulin	14-33	LALWGPDPA AAAFVNQ HLCGSHLVE	Chang and Unanue, 2009
19	Preproinsulin	34-53	HLV EALYLV CGGERGFFYTPKTRRE	Chang and Unanue, 2009
20	Preproinsulin	44-63	GERGFFYTPKTR REAED LQVGQVE	Chang and Unanue, 2009
21	Preproinsulin	74-93	AGSLQPLALE GLSLQKRGIVEQCCT	Chang and Unanue, 2009
22	Preproinsulin	94-110	QCCT SICS LYQLENYCN	Chang and Unanue, 2009
23	Der p 2 allergenic peptide	41-60	FEAVQNTKTAKIEIKASIDG	Neeno et al. 1996 and Krcó et al. 2000
24	Der p 2 allergenic peptide	51-70	KIEIKASIDGLEVDVPGIDP	Neeno et al. 1996 and Krcó et al. 2000
25	Der p 2 allergenic peptide	61-80	LEVDVPGIDPNACHYMKCPL	Neeno et al. 1996 and Krcó et al. 2000
26	Der p 2 allergenic peptide	91-110	TWNVPKIAPKSENVVTVKV	Neeno et al. 1996 and Krcó et al. 2000
27	Der p 2 allergenic peptide	101-120	SENVVTVKVMGDDGVLACA	Neeno et al. 1996 and Krcó et al. 2000
28	Der p 2 allergenic peptide	111-129	MGDDGVLACAIATHAKIRD	Neeno et al. 1996 and Krcó et al. 2000
29	Human Ro60	401-425	MVVTRTEKDSYVVASFDEMVP CPVT	Paisansinsup et al. 2002
30	Human Ro60	466-485	VFTDNETFAGGVHPAIALRE	Paisansinsup et al. 2002
31	Human Ro60	501-525	MTSNGFTIADPDDRGMLDMCGFD TG	Paisansinsup et al. 2002
32	Der p 2 allergenic peptide	11-30	HEIKKVL VPGCHGSEPC/IIN	Neeno et al. 1996 and Krcó et al. 2000
33	Der p 2 allergenic peptide	21-40	CHGSEPC/IHRGKPFQLEAV	Neeno et al. 1996 and Krcó et al. 2000
34	Der p 2 allergenic peptide	31-50	RGKPFQLEAVFEAVQNTKTA	Neeno et al. 1996 and Krcó et al. 2000
35	Der p 2 allergenic peptide	71-90	HACHYMKCPLVKGQYDIDKY	Neeno et al. 1996 and Krcó et al. 2000
36	Der p 2 allergenic peptide	81-100	VKGQQYDIKYTWNVPKIAPK	Neeno et al. 1996 and Krcó et al. 2000

Supplementary Figure S1

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

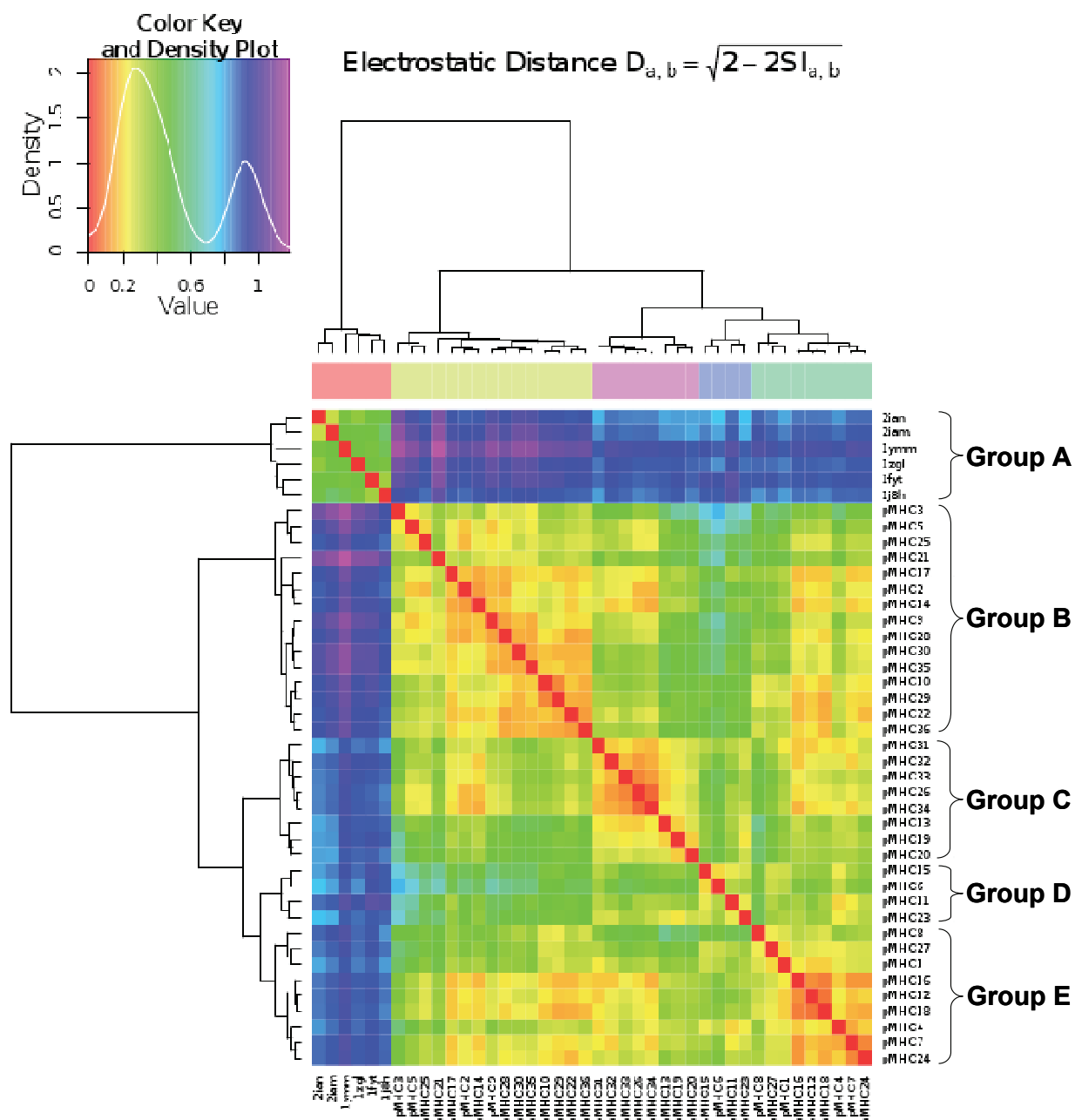


Figure S1. Heat map showing the clustering output for all pMHC interfaces from test set 2 along with all six available human pMHC-II interfaces from TR/pMHC-II crystal structures based on the calculated MSEP values depicted as a colour coded matrix. Groups A, B, C, D and E represent the six available human pMHC-II interfaces from TR/pMHC-II crystal structures, strong-agonists (SA), moderate-agonists (MA), weak-agonists (WA) and non-agonists.

Supplementary Table S4

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S4. Top 24 predictions from test set 2 using our prediction model.

All predictions are known T cell activators, indicated by the ‘+’ sign. Experimental binding registers are in underlined bold font. IA-2, Der p 2, PPI, Ro60, and GAD65 refer to insulinoma-associated (islet) antigen-2, dermatophagoides ptermyssinus allergenic peptides, pre-proinsulin, human ribonucleoprotein-60 and human glutamic acid decarboxylase-65, respectively. Similarly, SA, MA and WA denote strong-agonist, moderate-agonist and weak-agonist, respectively.

Description	Peptide sequence	Reference	T cell response	Predicted binding energy (kJ/mol)	pMHC clustering result	Rank
IA-2 601-618	<u>RQHARQQDKER</u> LALGPE	Chang and Unanue, 2009	+	-50.15	SA	1
GAD65 471-490	VDKCLELAEY <u>LYNIIKNREG</u>	Chang and Unanue, 2009	+	-46.13	SA	2
PPI 44-63	GERGFFY <u>TPKTRREAE</u> DLQVGQVE	Chang and Unanue, 2009	+	-43.21	SA	3
Der p 2 41-60	FEAVQNTKTAKIEIKASIDG	Neeno <i>et al.</i> 1996 and Krco <i>et al.</i> 2000	+	-42.88	SA	4
GAD65 126-140	<u>FDRSTKVIDFHYPNE</u>	Chang and Unanue, 2009	+	-40.70	SA	5
Der p 2 91-110*	TWNVPKIAPKSENVVVTVKV	Neeno <i>et al.</i> 1996 and Krco <i>et al.</i> 2000	+	-40.01	SA	6
PPI 14-33	LALWGPD <u>PAAAFVNQHLCGSHLVE</u>	Chang and Unanue, 2009	+	-39.59	SA	7
IA-2 646-663	EGPPEP <u>SRVSSVSQFSD</u>	Chang and Unanue, 2009	+	-38.60	SA	8
IA-2 661-678	FSD <u>AAQASPSH</u> SSSTPSW	Chang and Unanue, 2009	+	-38.05	SA	9

Description	Peptide sequence	Reference	T cell response	Predicted binding energy (kJ/mol)	pMHC clustering result	Rank
GAD65 231-250	PGGSGDGI <u>FSPGG</u> AI <u>SNM</u> YA	Chang and Unanue, 2009	+	-37.30	SA	10
IA-2 616-633	GPEGAHGD <u>TTFEY</u> QDLCR	Chang and Unanue, 2009	+	-37.22	SA	11
Ro60 466-485	VFTDNETFAGGVHPAIALRE	Paisansinsup <i>et al.</i> 2002	+	-37.02	SA	12
GAD65 461-475	AKG <u>TTGF</u> EAHV <u>DK</u> CL	Chang and Unanue, 2009	+	-41.59	MA	13
PPI 1-24	MALWMRLPL <u>LLALLALW</u> GPDPAAA	Chang and Unanue, 2009	+	-41.04	MA	14
Ro60 401-425	MVVTRTEKDSYVVASFDEMVP <u>CP</u> VT	Paisansinsup <i>et al.</i> 2002	+	-40.76	MA	15
PPI 94-110	QCCT <u>SICSLYQ</u> LENYCN	Chang and Unanue, 2009	+	-40.37	MA	16
GAD65 431-445	KHYDLSYDTGDKALQ	Chang and Unanue, 2009	+	-40.01	MA	17
PPI 34-53	HLV <u>EALYLV</u> CGERGFYTPKTRRE	Chang and Unanue, 2009	+	-39.82	MA	18
IA-2 826-843	DEGASLYHVYE <u>VNL</u> VSEH	Chang and Unanue, 2009	+	-38.54	MA	19
PPI 74-93	AGSLQPLALEGSLQKRGIVEQCCT	Chang and Unanue, 2009	+	-36.96	MA	20
Der p 2 101-120	SENVVVTVKVMGDDGVLACA	Neeno <i>et al.</i> 1996 and Krco <i>et al.</i> 2000	+	-41.83	WA	21
IA-2 931-948	KGVKEIDIAATLEHV <u>RD</u> Q	Chang and Unanue, 2009	+	-41.53	WA	22
Ro60 501-525	MTSNGFTIADPPDDRGMLDMCGFDTG	Paisansinsup <i>et al.</i> 2002	+	-38.88	WA	23
Der p 2 51-70*	KIEIKASIDGLEVDVPGIDP	Neeno <i>et al.</i> 1996 and Krco <i>et al.</i> 2000	+	-38.52	WA	24

In silico prediction of immunogenic T cell epitopes for HLA-DQ8

Javed M. Khan, Gaurav Kumar and Shoba Ranganathan

Table S5. Analysis of DQ8 binding motifs for peptides from test set 1.

The residues concurrent with available binding motifs for DQ8 (Godkin *et al.* 1998; Rammensee *et al.* 1999) are underlined in bold font. All peptides have one or more residues that do not conform to binding motifs and 56 peptides have 2 or more such residues. Y and N refer to yes and no, respectively.

S.No.	Position					2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1	4	6	7	9				
Binding Motif	T	V	RR	R	R				
	S	D	VV	E					
	W	M	DD	N					
	K	A	AA	G					
	E	I	II	D					
	D		YY	P					
	Q			Q					
	F								
	L								
	M								
Epitopes	1	FPSQQ	PYLQLQ	PPFPQ		Y	A-gliadin 49-63	-49.00	Godkin <i>et al.</i> 1998
2	IDV	<u>W</u> LGGL	<u>A</u> ENFKPY			Y	Thyroid per 632-645Y analog	-49.12	Sidney <i>et al.</i> 2002
3	IDV	<u>W</u> LGGL	<u>A</u> ENFLPY			Y	Thyroid per 632-645Y analog	-48.74	Sidney <i>et al.</i> 2002
4	IDV	<u>W</u> LGGL	<u>A</u> ENFLDY			Y	Thyroid per 632-645Y analog	-48.03	Sidney <i>et al.</i> 2002
5	IDV	<u>W</u> VGGGL	<u>A</u> ENFLPY			Y	Thyroid per 632-645Y analog	-46.79	Sidney <i>et al.</i> 2002

S.No.	Position					2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1	4	6	7	9				
6	I D V Y L G G L <u>A</u> E N F L P Y					Y	Thyroid per 632-645Y analog	-45.00	Sidney <i>et al.</i> 2002
7	L Q L Q P <u>F</u> P Q P Q P F P <u>P</u> L					Y	A-gliadin 56-70	-44.55	Godkin <i>et al.</i> 1998
8	I D V <u>W</u> L G G L <u>A</u> Y N F L P Y					Y	Thyroid per 632-645Y analog	-44.35	Sidney <i>et al.</i> 2002
9	L T E <u>W</u> T S S N <u>V</u> M E <u>E</u> R Y					Y	OVA 267-276 Y	-43.73	Sidney <i>et al.</i> 2002
10	V <u>Q</u> P Q Q L P Q F E I R N L					Y	A-gliadin 227-241	-43.44	Godkin <i>et al.</i> 1998
11	Y P <u>L</u> G Q G S F <u>R</u> P S Q Q N P					Y	A-gliadin 207-221	-43.12	Godkin <i>et al.</i> 1998
12	E E <u>D</u> I E <u>I</u> P I <u>Q</u> <u>E</u> E E Y					N	CD20 249-262 analog	-42.25	Sidney <i>et al.</i> 2002
13	I D V <u>W</u> L G G L <u>A</u> S N F L P Y					Y	Thyroid per 632-645Y analog	-42.25	Sidney <i>et al.</i> 2002
14	I D V <u>W</u> L G G L <u>Y</u> E N F L P Y					Y	Thyroid per 632-645Y analog	-41.81	Sidney <i>et al.</i> 2002
15	A A A A <u>A</u> V <u>A</u> A E A Y					Y	Artificial sequence	-41.75	Sidney <i>et al.</i> 2002
16	G V A G L L <u>V</u> A L A V					Y	IA-2 499-509	-41.34	Sidney <i>et al.</i> 2002
17	C D G E R P T L <u>A</u> F L Q D V M					Y	GAD 101-115	-41.25	Sidney <i>et al.</i> 2002
18	F E S <u>T</u> G N L I <u>A</u> P E Y G F K I S Y					Y	HA 255-271Y	-41.20	Sidney <i>et al.</i> 2002
19	S Q D L E L S <u>W</u> N L N G L Q A <u>D</u> L S S					Y	FceR 104-122	-41.19	Sidney <i>et al.</i> 2002
20	D S N I <u>M</u> N S <u>I</u> N N <u>V</u> M <u>D</u> E I D F F E K					N	Pf ABRA 487-506	-40.53	Sidney <i>et al.</i> 2002
21	S Q D L E L <u>S</u> W N L N G L Q A Y					Y	FceR 104-122 analog	-40.12	Sidney <i>et al.</i> 2002
22	I D V <u>D</u> L G G L <u>A</u> E N F L P Y					Y	Thyroid per 632-645Y analog	-40.11	Sidney <i>et al.</i> 2002
23	P P <u>L</u> Y A T G <u>R</u> L S <u>Q</u> A Q L M P S P P M					Y	VP16	-38.87	Sidney <i>et al.</i> 2002
24	<u>E</u> P R <u>A</u> P W <u>I</u> <u>Q</u> E G P E Y W					N	MHC Ia 46-63	-38.74	Sidney <i>et al.</i> 2002
25	S <u>F</u> P P Q Q P <u>Y</u> P <u>Q</u> P Q P Q Y					Y	A-gliadin 77-91	-38.26	Godkin <i>et al.</i> 1998
26	I D V <u>W</u> L G G L <u>A</u> E L F L P Y					Y	Thyroid per 632-645Y analog	-37.93	Sidney <i>et al.</i> 2002
27	K P V <u>S</u> K M R M <u>A</u> T P L L M Q A L P					Y	CLIP 95-102	-37.56	Sidney <i>et al.</i> 2002
28	I D V <u>W</u> L G G L <u>A</u> E N S L P Y					Y	Thyroid per 632-645Y analog	-36.95	Sidney <i>et al.</i> 2002

S.No.	Position					2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference													
	1	4	6	7	9																	
29	Y	Q	T	I	<u>E</u>	<u>E</u>	N	<u>I</u>	<u>K</u>	<u>I</u>	<u>F</u>	<u>K</u>	<u>E</u>	D	A	E25B protein 112-126 analog	-36.76	Suri <i>et al.</i> 2005				
30	L	L	I	L	<u>D</u>	<u>I</u>	<u>L</u>	<u>T</u>	<u>A</u>	<u>G</u>	<u>L</u>	<u>E</u>	<u>E</u>	Y	S	A	M	R	D	p21 51-66; C out	-36.50	Sidney <i>et al.</i> 2002
31	Y	P	F	I	<u>E</u>	<u>Q</u>	<u>E</u>	<u>G</u>	<u>P</u>	<u>E</u>	<u>F</u>	<u>F</u>	<u>D</u>	<u>Q</u>	<u>E</u>					MHC Ia 51-63 analog	-36.41	Sidney <i>et al.</i> 2002
32	L	E	D	<u>A</u>	<u>R</u>	<u>R</u>	<u>L</u>	<u>K</u>	<u>A</u>	<u>I</u>	<u>Y</u>	<u>E</u>	<u>K</u>							Lamba repressor 12-24	-36.10	Sidney <i>et al.</i> 2002
33	I	A	R	A	K	M	F	P	<u>A</u>	<u>V</u>	<u>A</u>	<u>E</u>	<u>K</u>							34P3A	-35.68	Sidney <i>et al.</i> 2002
34	D	<u>M</u>	<u>T</u>	<u>P</u>	<u>A</u>	<u>D</u>	<u>A</u>	<u>L</u>	<u>D</u>	<u>D</u>	<u>F</u>	<u>D</u>	<u>L</u>							HSV	-35.67	Sidney <i>et al.</i> 2002
35	L	T	I	<u>Q</u>	<u>I</u>	<u>E</u>	<u>S</u>	<u>A</u>	<u>A</u>	<u>D</u>	<u>Q</u>	<u>D</u>	<u>P</u>	<u>S</u>						ZnT8 diabetic autoantigen	-35.46	Chang and Unanue, 2009
36	R	T	G	I	A	<u>Q</u>	<u>A</u>	<u>L</u>	<u>S</u>	<u>S</u>	<u>F</u>	<u>D</u>	<u>L</u>	<u>H</u>						ZnT8 diabetic autoantigen	-35.01	Chang and Unanue, 2009
37	F	P	G	<u>Q</u>	<u>Q</u>	<u>Q</u>	<u>F</u>	<u>P</u>	<u>P</u>	<u>Q</u>	<u>Q</u>	<u>P</u>	<u>P</u>	<u>Y</u>	<u>P</u>					A-gliadin 30-44	-34.98	Godkin <i>et al.</i> 1998
38	Y	Q	T	I	<u>E</u>	<u>E</u>	N	<u>I</u>	<u>K</u>	<u>I</u>	<u>F</u>	<u>A</u>	<u>A</u>	<u>A</u>						E25B protein 112-126 analog	-34.92	Suri <i>et al.</i> 2005
39	G	R	F	T	A	Q	N	<u>A</u>	<u>A</u>	<u>Q</u>	<u>P</u>	<u>E</u>	<u>T</u>	<u>G</u>	<u>P</u>	<u>G</u>				TRAIL receptor 2 364-380 analog	-34.87	Suri <i>et al.</i> 2005
40	Q	P	Y	<u>P</u>	<u>Q</u>	<u>P</u>	<u>P</u>	<u>F</u>	<u>P</u>	<u>S</u>	<u>Q</u>	<u>Q</u>	<u>P</u>	<u>Y</u>						A-gliadin 41-55	-34.78	Godkin <i>et al.</i> 1998
41	I	L	S	V	H	V	<u>A</u>	<u>T</u>	<u>A</u>	<u>A</u>	<u>S</u>	<u>Q</u>	<u>D</u>	<u>S</u>						ZnT8 diabetic autoantigen	-34.21	Chang and Unanue, 2009
42	M	<u>S</u>	<u>S</u>	<u>G</u>	<u>S</u>	<u>F</u>	<u>I</u>	<u>N</u>	<u>I</u>	<u>S</u>	<u>V</u>									IA-2 499-509	-34.09	Sidney <i>et al.</i> 2002
43	G	R	F	T	A	Q	N	<u>A</u>	<u>A</u>	<u>Q</u>	<u>P</u>	<u>A</u>	<u>T</u>	<u>G</u>	<u>P</u>	<u>G</u>				TRAIL receptor 2 364-380 analog	-33.94	Suri <i>et al.</i> 2005
44	G	R	F	T	Y	Q	N	<u>A</u>	<u>A</u>	<u>Q</u>	<u>P</u>	<u>E</u>	<u>T</u>	<u>G</u>	<u>P</u>	<u>G</u>				TRAIL receptor 2 364-380 analog	-33.84	Suri <i>et al.</i> 2005
45	F	D	G	A	E	I	F	H	<u>V</u>	<u>D</u>	<u>I</u>	<u>E</u>	<u>K</u>	<u>S</u>	<u>E</u>					MHC II E8 51-65 analog	-33.63	Suri <i>et al.</i> 2005
46	A	P	Y	H	<u>F</u>	<u>D</u>	<u>L</u>	<u>S</u>	<u>G</u>	<u>H</u>	<u>A</u>	<u>F</u>	<u>G</u>	<u>S</u>	<u>M</u>	<u>A</u>	<u>K</u>	<u>K</u>	<u>G</u>	Lol p1 101-120	-33.62	Sidney <i>et al.</i> 2002
47	Y	Q	T	I	<u>E</u>	<u>E</u>	N	<u>I</u>	<u>K</u>	<u>I</u>	<u>F</u>	<u>E</u>	<u>A</u>	<u>D</u>	<u>A</u>					E25B protein 112-126 analog	-33.48	Suri <i>et al.</i> 2005

S.No.	Position					2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1	4	6	7	9				
48	YQTII	E	N	I	K	I	FEAAA	-33.23	Suri <i>et al.</i> 2005
49	AGKAGV	A	N	V	S	I	E DR	-33.19	Suri <i>et al.</i> 2005
50	IAR	F	K	M	F	P	V K E K	-32.91	Sidney <i>et al.</i> 2002
51	YQTI	K	E	N	I	K	I FE E DA	-32.81	Suri <i>et al.</i> 2005
52	FDG	K	E	I	F	H	V D I E KSE	-32.76	Suri <i>et al.</i> 2005
53	ISGAT	V	I	H	V	V	E KE	-32.43	Suri <i>et al.</i> 2005
54	FDG	D	E	I	F	H	A D I E KSE	-31.71	Suri <i>et al.</i> 2005
55	AGK	D	G	V	A	N	A S I E DR	-31.67	Suri <i>et al.</i> 2005
56	AGK	D	G	V	A	N	V S I E DR	-30.94	Suri <i>et al.</i> 2005
57	YAHAAHAA	A	A	A	A	A	HAAHAA	-30.92	Sidney <i>et al.</i> 2002
58	GRFT	K	Q	N	A	A	Q P E TGPG	-29.86	Suri <i>et al.</i> 2005
59	YARF	Q	R	Q	T	T	LKAAA	-29.67	Sidney <i>et al.</i> 2002
60	QQP	L	Q	Q	Y	P	LGG G SFR	-29.57	Godkin <i>et al.</i> 1998
61	FGYRKP	L	D	N	I	K	D N V G KMEDYIKK	-29.55	Sidney <i>et al.</i> 2002
62	FVNQH	L	C	G	S	H	L V EAL	-29.26	Sidney <i>et al.</i> 2002
63	ISG	D	T	G	V	I	H V VAKE	-29.11	Suri <i>et al.</i> 2005
64	PS	S	Q	F	Q	P	L Q Q YPLG	-28.82	Godkin <i>et al.</i> 1998
65	GV	T	Y	E	I	D	L T NKN	-28.46	Sidney <i>et al.</i> 2002
66	A I L	T	D	A	H	L	L I D L T	-26.68	Chang and Unanue, 2009

S.No.	Position						2 or more residues do not conform to binding motifs	Source	Predicted Binding Energy (kJ/mol)	Reference
	1	4	6	7	9					
67	Q	Q	Q	<u>F</u>	P	P	Y	A-gliadin 34-48	-26.49	Godkin <i>et al.</i> 1998
68	Y	Q	S	Y	G	P	Y	DQa1 0501 16-30	-24.86	Sidney <i>et al.</i> 2002
69	K	L	P	K	P	P	Y	CLIP 96-114	-24.64	Sidney <i>et al.</i> 2002
70	I	E	K	K	I	A	Y	Pf cp 379-396 truncated analog	-23.89	Sidney <i>et al.</i> 2002

6.2 Conclusions

The current prediction model is efficient at screening for high-affinity binders at SP=0.95 and identifying immunogenic T cell epitopes among the high-binders. The discriminatory power of this model is also highlighted by efficient discrimination of both different categories of binders from non-binders and different categories of pMHC agonists from non-agonists while accurately predicting the binding registers of all DQ8-restricted peptides. The increasing evidence for inadequacy of binding motifs in defining T cell epitopes and our significant results indicate that we have developed a model that can be successfully applied as a generic protocol for easy *in silico* identification of potential immunogenic T cell epitopes for other MHC alleles. The current model is therefore applicable for screening of vaccine candidates irrespective of sequence motifs. This combined approach provides a set of sensitive and specific computational tools to facilitate high-throughput screening of peptides for immunotherapeutic applications such as controlling allergic and autoimmune responses. Due to precise predictions of all binding registers by pDOCK [49], we have eliminated using the previously described [11] approach that utilizes a sliding window of size nine residues to identify multiple registers within candidate DQ8-binding peptides.

Chapter 7: Conclusions and future directions

7.1 Summary

This thesis is divided into seven chapters. Chapter 1 provides a literature survey on MHC and TR biology and diversity, the complexities involved in identifying T cell epitopes, existing bioinformatics resources and applications that are available for the study of MHC proteins, pMHC and TR/pMHC complexes and prediction of T cell epitopes. The second chapter lists the publications included in this thesis and the respective chapters they are included in as a table for cross reference purposes. A new rapid, accurate, robust and generic protocol (pDOCK) for docking peptides to MHC-I and MHC-II proteins is described in Chapter 3. The accuracy of the docking protocol was assessed against a large dataset of non-redundant pMHC complexes for which 3D structures are available. The method was also benchmarked with the earlier multi-step docking technique [10, 11] and validated against previously published studies [419, 424, 433, 435-437, 442, 443]. This procedure forms the methodological basis for subsequent T cell epitope prediction for specific MHC alleles and hence vaccine design.

This is followed by a description of the MPID-T2 database that stores, disseminates and depicts pMHC and TR/pMHC binding and sequence-structure-function information, in Chapter 4. Data analysis based on the correlation of all predefined and newly characterized structural interaction parameters is also presented in this chapter. Chapter 5 describes the use of structural descriptors such as computed BE, TR paratope, pMHC epitope, MSEP and calculated TR docking angle to analyse 61 TR/pMHC crystallographic structures to comprehend TR/pMHC interaction. It also demonstrates a novel/rational approach for θ calculation, a linear correlation between BE and θ , an explanation for TR ability to scan many pMHC ligands yet specifically bind to one, a proposed mechanism for pMHC recognition by TR leading to T cell activation and illustrates the importance of peptide in TR/pMHC interaction.

Chapter 6 details the use pDOCK, a complementary scoring function, and a MSEP-based clustering of pMHC interfaces to develop a prediction model or a predictive approach for functional prediction of HLA-DQ8 restricted T cell epitopes. High prediction accuracy of MHC-II binding immunogenic peptides was validated by experimental biochemical and functional data obtained from the literature. This approach successfully identified known

antigenic peptide epitopes including the ones that lacked any conserved binding motifs. Chapter 7 highlights the innovations, significance and contributions of this thesis and draws conclusions from the bioinformatic-based approach to TR/pMHC structural analysis and T cell epitope prediction. This chapter also discusses future directions. The work presented in this thesis has been published in a series of book chapters and journal articles including the development of an interaction database for pMHC and TR/pMHC crystallographic structures (Chapter 4).

7.2 Conclusions

This thesis reports a series of pioneering work in the field of structural immunoinformatics through the use of 3D X-ray crystallographic structures and structural models of pMHC and TR/pMHC complexes. In conclusion, the following inferences can be drawn:

1. Through systematic improvements in speed, accuracy and hence the efficiency compared to our previous docking technique and existing methodologies, I have developed a new robust pMHC docking protocol (pDOCK; Chapter 3), that can be applied as a generic methodology for high-throughput screening of peptides for easy *in silico* identification of promiscuous strong-MHC binding peptide epitopes which can then be subjected to further filtering through the use of newly developed TR/pMHC interaction parameters (Chapter 5) to identify true immunogenic peptide epitopes with greater propensity to bind to MHC proteins and consequently activate T cells making them key targets for the design of vaccines and immunotherapies.
2. The extremely high polymorphism of MHC alleles [157] and diversity of TR proteins [3, 46-48, 108] has been a confounding factor in the study of TR/pMHC binding specificities. For a TR protein to recognize a specific pMHC complex, geometric and electrostatic complementarity between the receptor (TR) and its corresponding ligand (pMHC) is essential for the formation of chemical bonds between their functional groups, which in turn determines the net stability of the TR/pMHC complex. To this end, I have successfully introduced the use of structural interaction information to analyse high-level relationships hidden within TR/pMHC crystallographic structures and demonstrated the existence of different interaction characteristics among different pMHC and TR types (Chapter 5). The result of this analysis paves the way for more accurate inferences about TR/pMHC binding specificities.

3. An innovative systematic stepwise application of pDOCK and the MSEP-based pMHC interface clustering (Chapter 6) for the analysis of the human MHC-II allele HLA-DQ8 binding and non-binding peptides for subsequent T cell epitope prediction, has demonstrated the utility of the methods and protocols developed in this thesis besides addressing the issue of degeneracy in peptide binding to MHC-II proteins by reassuring the existence peptide epitopes which lack conserved binding motifs within a candidate MHC-II binding peptide. This provides new insights to the binding specificities of MHC-II alleles, suggesting that pMHC binding and/or T cell activation is not necessarily peptide motif or MHC germline dependent. Rather, the accurate prediction of T cell epitopes in test set-II, that are known immunogenic antigens or T cell activators, by using the above mentioned two-step procedure (Chapter 6) strengthens the idea that T cell activation is primarily dependent on the electrostatic ring displayed by pMHC interface and a specific arrangement of residues presented by pMHC interface.
4. These results support the applicability of the above mentioned two-step prediction model to other disease-implicated alleles for successful identification of true or immunogenic T cell epitopes. The outcomes of this study will therefore facilitate the rational development of effective and highly specific peptide vaccines capable of eliciting T cell response with wide population coverage, for immunotherapies to protect against or combat infectious, autoimmune, allergic and graft vs. host diseases, thereby, cutting down the lead time involved in experimental vaccine development methods.

7.3 Innovations

This thesis highlights original findings from application of bioinformatic tools to the study of TR/pMHC interactions and its significance in celiac disease and insulin-dependent diabetes mellitus (IDDM) associated HLA-DQ8 allele. Several novel aspects are presented in this thesis. The new docking protocol, pDOCK, developed as a part of this work, is a fast, accurate and robust method for high-throughput screening of pathogenic sequences, based on fully flexible docking of peptides to MHC-I and MHC-II proteins. Besides this, many other innovations such as use of MSEP of the pMHC binding interfaces to calculate the TR docking angle, analysis of pMHC epitopes and TR paratopes across all available TR/pMHC structures to identify conserved positions that could contribute significantly to

TR/pMHC binding, use of computed or calculated TR/pMHC BE as a discriminator for weak-, medium- and strong- pMHC agonists and an excellent correlations between the BE and TR docking angle, are presented in this thesis.

This is, to the best of my knowledge, the first study of its kind, where structural interaction parameters have been used for the analysis of TR/pMHC crystal structures. Structural interaction characteristics among different pMHC and TR types have been discovered, using which, a novel and rational grouping system for TR proteins has been developed. Finally, the innovative use of pDOCK and MSEP of pMHC interfaces to predict immunogenic epitopes for the disease-implicated human MHC-II allele HLA-DQ8 is also presented.

7.4 Significance and contributions

This work reverberates with inherent importance. Among many other significances and contributions of this thesis, a few critical ones are listed below:

1. The thesis presents an improvement in the speed and accuracy of pMHC docking methodology through a new docking technique called pDOCK (Chapter 3).
2. It offers compelling insights into physicochemical basis associated with TR/pMHC interaction, TR specificity and T cell activation (Chapter 5).
3. It lists new structural descriptors that could be used to improve T cell epitope prediction efficacy (Chapter 4 and Chapter 5).
4. It outlines the rationale behind the ability of a TR protein to survey many pMHC interfaces and yet specifically bind to one (Chapter 5).
5. It provides evidence for the vital role of peptide in TR/pMHC recognition and binding (Chapter 5).
6. It describes an improved epitope prediction accuracy based on improvement in the speed and efficiency of the docking protocol and application of the new TR/pMHC analytical parameters characterized (Chapter 6). The overall approach in Chapter 6 includes free energy estimates used as a first step to successfully identify

immunogenic epitopes and brings together a range of methods from fast coarse-grained docking to detailed computation. The success of this methodology for peptides is encouraging, with possible applicability as a generalized protocol for larger biomolecular systems.

7. It helps in understanding the molecular basis of immune function in defense and in disease, in the light of allelic variability in human population.
8. It helps further accelerate the development of structure based prediction techniques.
9. It contributes towards rational development of vaccines for prevention and immunotherapy to combat infectious, autoimmune, allergic and graft vs. host diseases
10. It assists in clinical vaccine development and hence cuts down the lead time and effort involved in classical vaccine design.

7.5 Future directions

The studies presented in this thesis could lead to advancements in many directions for better understanding of TR/pMHC interactions. The methodology (pDOCK) described in Chapter 3 could be automated for high-throughput identification of strong-MHC binding peptides and combined with an automated MSEP clustering approach to develop a fully automated structure-based T cell prediction model. This fully automated prediction model can then be implemented as a research tool or a web-server that provides service to the scientific community, especially to immunologists and computational biologists. Preliminary work for the implementation of the research tool has already begun. The database (MPID-T2) covered in Chapter 4 paves way for further developments that will facilitate the extraction of high-level relationships hidden within TR/pMHC interactions by both extending the currently presented work to larger datasets as more TR/pMHC structural data becomes available and extrapolating new structural descriptors that can be factored in as parameters for TR/pMHC binding. Future developments will include listing post translational modifications (PTM) for peptides to help understand the effect of PTM on TR/pMHC binding and interaction.

The analysis done in Chapter 5 has revealed a series of interesting features that could potentially be applied for more in-depth analysis of TR/pMHC complexes. Although the current methodology focuses on the use of existing crystallographic data for analysis, this work may be extended to theoretical models for alleles without experimental structures. Such analysis will prove useful as the majority of MHC alleles have not been crystallized and much remains unknown with regards to the binding mechanisms underlying both pMHC and TR/pMHC interactions. The classification of TR types into clusters may be further formulated by taking into account more TR/pMHC structural descriptors and any new interaction information or characteristics. This will allow finer selection of representative proteins that can effectively cover TR/pMHC specificity space.

The analysis and T cell epitope prediction exhibited in Chapter 6 serve as an essential preliminary step towards better understanding the pathology of disease-related peptide antigens by focusing on one such disease-implicated human MHC-II allele HLA-DQ8. A similar approach may be applied for the analysis of other disease-associated alleles and their related antigenic peptides. This will also provide valuable insights into disease pathology and facilitates the fine profiling of T cell epitope repertoire among peptides binding to disease-implicated alleles. Furthermore, an automated methodology for TR docking angle calculation based on pMHC interface MSEP as described in Chapter 5 can also be developed and included into the above mentioned prediction model for further fine-tuning and accuracy enhancement in T cell epitope prediction. There is also scope to explore and/or venture into other modes of docking using the technique that underlies pDOCK, for example, docking of peptides to TR interfaces and perhaps then modeling the MHC proteins around the docked peptides as an alternative to MD simulations of the entire immune complex [416] for subsequent identification of immunogenic T cell epitopes.

Moreover, it would be interesting to investigate whether the small TR docking angle (θ) and high binding energy reported in Chapter 5 and publication 5 are a result of the alignment between macroscopic dipoles. Finally, it would be interesting to analyse other aspects of both pMHC and TR/pMHC complex formation, such as the burial of polar regions while taking into account the incurred desolvation penalties.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: **The adaptive immune system**. In: *Molecular biology of the cell*. 4th edn. New York: Garland Science; 2002: 1363-1421.
2. Rammensee HG, Falk K, Rotzschke O: **Peptides naturally presented by MHC class I molecules**. *Annu Rev Immunol* 1993, **11**:213-244.
3. Lefranc MP, Lefranc G: **The T cell receptor FactsBook**. San Diego: Academic Press; 2001.
4. Mueller DL: **Mechanisms maintaining peripheral tolerance**. *Nat Immunol* 2010, **11**(1):21-27.
5. Lo WL, Felix NJ, Walters JJ, Rohrs H, Gross ML, Allen PM: **An endogenous peptide positively selects and augments the activation and survival of peripheral CD4+ T cells**. *Nat Immunol* 2009, **10**(11):1155-1161.
6. van der Merwe PA, Davis SJ: **Molecular interactions mediating T cell antigen recognition**. *Annu Rev Immunol* 2003, **21**:659-684.
7. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC: **Structure of the complex between human T-cell receptor, viral peptide and HLA-A2**. *Nature* 1996, **384**(6605):134-141.
8. Garcia KC, Adams JJ, Feng D, Ely LK: **The molecular basis of TCR germline bias for MHC is surprisingly simple**. *Nat Immunol* 2009, **10**(2):143-147.
9. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, Peters B: **Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries**. *Immunome Res* 2008, **4**:2.
10. Tong JC, Tan TW, Ranganathan S: **Modeling the structure of bound peptide ligands to major histocompatibility complex**. *Protein Sci* 2004, **13**(9):2523-2532.
11. Tong JC, Zhang GL, Tan TW, August JT, Brusica V, Ranganathan S: **Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides**. *Bioinformatics* 2006, **22**(10):1232-1238.
12. Tong JC, Bramson J, Kanduc D, Chow S, Sinha AA, Ranganathan S: **Modeling the bound conformation of Pemphigus vulgaris-associated peptides to MHC Class II DR and DQ alleles**. *Immunome Res* 2006, **2**:1.

13. Tong JC, Tan TW, Sinha AA, Ranganathan S: **Prediction of desmoglein-3 peptides reveals multiple shared T-cell epitopes in HLA DR4- and DR6-associated pemphigus vulgaris.** *BMC Bioinformatics* 2006, **7 Suppl 5**:S7.
14. Tong JC, Zhang ZH, August JT, Brusica V, Tan TW, Ranganathan S: **In silico characterization of immunogenic epitopes presented by HLA-Cw*0401.** *Immunome Res* 2007, **3**:7.
15. Khan JM, Tong JC, Ranganathan S: **Structural Immunoinformatics: Understanding MHC-peptide-TR binding.** In: *Bioinformatics for Immunomics*. Edited by Davies MN, Ranganathan S, Flower DR, vol. 3. New York: Springer; 2010: 77-94.
16. Yewdell JW, Bennink JR: **Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses.** *Annu Rev Immunol* 1999, **17**:51-88.
17. Yu K, Petrovsky N, Schonbach C, Koh JY, Brusica V: **Methods for prediction of peptide binding to MHC molecules: a comparative study.** *Mol Med* 2002, **8**(3):137-148.
18. Tong JC, Tan TW, Ranganathan S: **Methods and protocols for prediction of immunogenic epitopes.** *Brief Bioinform* 2007, **8**(2):96-108.
19. Gorer PA: **The detection of a hereditary antigenic difference in the blood of mice by means of human group A serum.** *J Genet* 1936, **32**:17-31.
20. Gorer PA: **The detection of antigenic differences in mouse erythrocytes by the employment of immune sera.** *Br J Exp Pathol* 1936, **17**:42-50.
21. Gorer PA: **The genetic and antigenic basis of tumour transplantation.** *J Pathol Bacteriol* 1937, **44**:691-697.
22. Gorer PA: **Further studies on antigenic differences in mouse erythrocytes.** *Br J Exp Pathol* 1937, **18**:31-36.
23. Gorer PA: **The antigenic basis of tumour transplantation.** *J Pathol Bacteriol* 1938, **47**:231-252.
24. Gorer PA, Mikulska ZB: **The antibody response to tumor inoculation; improved methods of antibody detection.** *Cancer Res* 1954, **14**(9):651-655.
25. Gorer PA, O'Gorman P: **The cytotoxic activity of isoantibodies in mice.** *Transplant Bull* 1956, **3**:142-143.
26. Snell GD: **Methods for the study of histocompatibility genes.** *J Genet* 1948, **49**(2):87-108.

27. Snell GD, Higgins GF: **Alleles at the histocompatibility-2 locus in the mouse as determined by tumor transplantation.** *Genetics* 1951, **36**(3):306-310.
28. Snell GD: **A fifth allele at the histocompatibility-2 locus of the mouse as determined by tumor transplantation.** *J Natl Cancer Inst* 1951, **11**(6):1299-1305.
29. Elgert KD: **Immunology: understanding the immune system**, 2nd edn. New Jersey: Wiley-BlackWell; 2009.
30. McDevitt HO, Benacerraf B: **Genetic control of specific immune responses.** *Adv Immunol* 1969, **11**:31-74.
31. Zinkernagel RM, Doherty PC: **Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system.** *Nature* 1974, **248**(450):701-702.
32. Weiss A: **Discovering the TCR beta-chain by subtraction.** *J Immunol* 2005, **175**(5):2769-2770.
33. Allison JP, McIntyre BW, Bloch D: **Tumor-specific antigen of murine T-lymphoma defined with monoclonal antibody.** *J Immunol* 1982, **129**(5):2293-2300.
34. Meuer SC, Fitzgerald KA, Hussey RE, Hodgdon JC, Schlossman SF, Reinherz EL: **Clonotypic structures involved in antigen-specific human T cell function. Relationship to the T3 molecular complex.** *J Exp Med* 1983, **157**(2):705-719.
35. Haskins K, Kubo R, White J, Pigeon M, Kappler J, Marrack P: **The major histocompatibility complex-restricted antigen receptor on T cells. I. Isolation with a monoclonal antibody.** *J Exp Med* 1983, **157**(4):1149-1169.
36. Allison JP, McIntyre BW, Bloch D: **Tumor-specific antigen of murine T-lymphoma defined with monoclonal antibody. 1982.** *J Immunol* 2005, **174**(3):1144-1151.
37. Hedrick SM, Cohen DI, Nielsen EA, Davis MM: **Isolation of cDNA clones encoding T cell-specific membrane-associated proteins.** *Nature* 1984, **308**(5955):149-153.
38. Hedrick SM, Nielsen EA, Kavaler J, Cohen DI, Davis MM: **Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins.** *Nature* 1984, **308**(5955):153-158.
39. Hedrick SM, Cohen DI, Nielsen EA, Davis MM: **Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. 1984.** *J Immunol* 2005, **175**(5):2771-2775.

40. Yanagi Y, Yoshikai Y, Leggett K, Clark SP, Aleksander I, Mak TW: **A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains.** *Nature* 1984, **308**(5955):145-149.
41. Chien YH, Gascoigne NR, Kavaler J, Lee NE, Davis MM: **Somatic recombination in a murine T-cell receptor gene.** *Nature* 1984, **309**(5966):322-326.
42. Wright P, Nimgaonkar VL, Donaldson PT, Murray RM: **Schizophrenia and HLA: a review.** *Schizophr Res* 2001, **47**(1):1-12.
43. Tong JC: **Structural immunoinformatics.** Singapore: National University of Singapore; 2006.
44. Accolla RS, Auffray C, Singer DS, Guardiola J: **The molecular biology of MHC genes.** *Immunol Today* 1991, **12**(4):97-99.
45. Accolla RS, Adorini L, Sartoris S, Sinigaglia F, Guardiola J: **MHC: orchestrating the immune response.** *Immunol Today* 1995, **16**(1):8-11.
46. Kronenberg M, Siu G, Hood LE, Shastri N: **The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition.** *Annu Rev Immunol* 1986, **4**:529-591.
47. Davis MM: **Molecular genetics of the T cell-receptor beta chain.** *Annu Rev Immunol* 1985, **3**:537-560.
48. Bonilla FA, Oettgen HC: **Adaptive immunity.** *J Allergy Clin Immunol* 2010, **125**(2 Suppl 2):S33-40.
49. Khan JM, Ranganathan S: **pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes.** *Immunome Res* 2010, **6** Suppl 1:S2.
50. Kaas Q, Lefranc MP: **T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB.** *In Silico Biol* 2005, **5**(5-6):505-528.
51. Kaas Q, Duprat E, Tourneur G, Lefranc MP: **IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes.** In: *Immunoinformatics.* Edited by Schoenbach C, Ranganathan S, Brusica V. New York: Springer; 2008: 19-49.
52. Armstrong KM, Insaiddoo FK, Baker BM: **Thermodynamics of T-cell receptor-peptide/MHC interactions: progress and opportunities.** *J Mol Recognit* 2008, **21**(4):275-287.

53. Stewart-Jones GB, McMichael AJ, Bell JI, Stuart DI, Jones EY: **A structural basis for immunodominant human T cell receptor recognition.** *Nat Immunol* 2003, **4**(7):657-663.
54. Rudolph MG, Stanfield RL, Wilson IA: **How TCRs bind MHCs, peptides, and coreceptors.** *Annu Rev Immunol* 2006, **24**:419-466.
55. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: **Structure of the human class I histocompatibility antigen, HLA-A2.** *Nature* 1987, **329**(6139):506-512.
56. Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC: **Structure of the human class I histocompatibility antigen, HLA-A2.** *J Immunol* 2005, **174**(1):6-19.
57. Kaas Q, Ruiz M, Lefranc MP: **IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data.** *Nucleic Acids Res* 2004, **32**(Database issue):D208-210.
58. Ehrenmann F, Kaas Q, Lefranc MP: **IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF.** *Nucleic Acids Res* 2010, **38**(Database issue):D301-307.
59. Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriout A, Lefranc G: **IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN.** *Dev Comp Immunol* 2005, **29**(11):917-938.
60. Gao GF, Tormo J, Gerth UC, Wyer JR, McMichael AJ, Stuart DI, Bell JI, Jones EY, Jakobsen BK: **Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2.** *Nature* 1997, **387**(6633):630-634.
61. Liu Y, Xiong Y, Naidenko OV, Liu JH, Zhang R, Joachimiak A, Kronenberg M, Cheroutre H, Reinherz EL, Wang JH: **The crystal structure of a TL/CD8alphaalpa complex at 2.1 A resolution: implications for modulation of T cell activation and memory.** *Immunity* 2003, **18**(2):205-215.
62. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**(1):235-242.
63. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S *et al*: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **58**(Pt 6 No 1):899-907.

64. Groh V, Bahram S, Bauer S, Herman A, Beauchamp M, Spies T: **Cell stress-regulated human major histocompatibility complex class I gene expressed in gastrointestinal epithelium.** *Proc Natl Acad Sci U S A* 1996, **93**(22):12445-12450.
65. Kim CY, Masli S, Streilein JW: **Qa-1, a nonclassical MHC molecule with immunomodulatory functions, is ubiquitously expressed in the immune-privileged anterior chamber of the eye.** *Ocul Immunol Inflamm* 2005, **13**(4):271-277.
66. Kim S, Poursine-Laurent J, Truscott SM, Lybarger L, Song YJ, Yang L, French AR, Sunwoo JB, Lemieux S, Hansen TH *et al*: **Licensing of natural killer cells by host major histocompatibility complex class I molecules.** *Nature* 2005, **436**(7051):709-713.
67. Yewdell JW: **Immunology. Hide and seek in the peptidome.** *Science* 2003, **301**(5638):1334-1335.
68. Hanada K, Yewdell JW, Yang JC: **Immune recognition of a human renal cancer antigen through post-translational protein splicing.** *Nature* 2004, **427**(6971):252-256.
69. Yewdell JW, Reits E, Neefjes J: **Making sense of mass destruction: quantitating MHC class I antigen presentation.** *Nat Rev Immunol* 2003, **3**(12):952-961.
70. Suh WK, Cohen-Doyle MF, Fruh K, Wang K, Peterson PA, Williams DB: **Interaction of MHC class I molecules with the transporter associated with antigen processing.** *Science* 1994, **264**(5163):1322-1326.
71. Ortmann B, Androlewicz MJ, Cresswell P: **MHC class I/beta 2-microglobulin complexes associate with TAP transporters before peptide binding.** *Nature* 1994, **368**(6474):864-867.
72. Yewdell JW, Tschärke DC: **Inside the professionals.** *Nature* 2002, **418**(6901):923-924.
73. Tschärke DC, Yewdell JW: **T cells bite the hand that feeds them.** *Nat Med* 2003, **9**(6):647-648.
74. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC: **Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1.** *Nature* 1993, **364**(6432):33-39.
75. Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, Wiley DC: **Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide.** *Nature* 1994, **368**(6468):215-221.

76. Stern LJ, Wiley DC: **Antigenic peptide binding by class I and class II histocompatibility proteins.** *Structure* 1994, **2**(4):245-251.
77. Cresswell P: **Assembly, transport, and function of MHC class II molecules.** *Annu Rev Immunol* 1994, **12**:259-293.
78. Hahn M, Nicholson MJ, Pyrdol J, Wucherpfennig KW: **Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor.** *Nat Immunol* 2005, **6**(5):490-496.
79. Peters PJ, Raposo G, Neefjes JJ, Oorschot V, Leijendekker RL, Geuze HJ, Ploegh HL: **Major histocompatibility complex class II compartments in human B lymphoblastoid cells are distinct from early endosomes.** *J Exp Med* 1995, **182**(2):325-334.
80. Rudensky AY, Maric M, Eastman S, Shoemaker L, DeRoos PC, Blum JS: **Intracellular assembly and transport of endogenous peptide-MHC class II complexes.** *Immunity* 1994, **1**(7):585-594.
81. Sette A, Livingston B, McKinney D, Appella E, Fikes J, Sidney J, Newman M, Chesnut R: **The development of multi-epitope vaccines: epitope identification, vaccine design and clinical evaluation.** *Biologicals* 2001, **29**(3-4):271-276.
82. Sette A, Newman M, Livingston B, McKinney D, Sidney J, Ishioka G, Tangri S, Alexander J, Fikes J, Chesnut R: **Optimizing vaccine design for cellular processing, MHC binding and TCR recognition.** *Tissue Antigens* 2002, **59**(6):443-451.
83. Govindarajan KR, Kanguane P, Tan TW, Ranganathan S: **MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules.** *Bioinformatics* 2003, **19**(2):309-310.
84. Tong JC, Kong L, Tan TW, Ranganathan S: **MPID-T: database for sequence-structure-function information on T-cell receptor/peptide/MHC interactions.** *Appl Bioinformatics* 2006, **5**(2):111-114.
85. Kanguane P, Sakharkar MK, Kolatkar PR, Ren EC: **Towards the MHC-peptide combinatorics.** *Hum Immunol* 2001, **62**(5):539-556.
86. Garrett TP, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC: **Specificity pockets for the side chains of peptide antigens in HLA-Aw68.** *Nature* 1989, **342**(6250):692-696.
87. Chlewicki LK, Holler PD, Monti BC, Clutter MR, Kranz DM: **High-affinity, peptide-specific T cell receptors can be generated by mutations in CDR1, CDR2 or CDR3.** *J Mol Biol* 2005, **346**(1):223-239.

88. Stewart-Jones G, Wadle A, Hombach A, Shenderov E, Held G, Fischer E, Kleber S, Nuber N, Stenner-Liewen F, Bauer S *et al*: **Rational development of high-affinity T-cell receptor-like antibodies.** *Proc Natl Acad Sci U S A* 2009, **106**(14):5784-5788.
89. Madden DR, Gorga JC, Strominger JL, Wiley DC: **The three-dimensional structure of HLA-B27 at 2.1 Å resolution suggests a general mechanism for tight peptide binding to MHC.** *Cell* 1992, **70**(6):1035-1048.
90. Madden DR, Garboczi DN, Wiley DC: **The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2.** *Cell* 1993, **75**(4):693-708.
91. Guo HC, Jardetzky TS, Garrett TP, Lane WS, Strominger JL, Wiley DC: **Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle.** *Nature* 1992, **360**(6402):364-366.
92. Collins EJ, Garboczi DN, Karpusas MN, Wiley DC: **The three-dimensional structure of a class I major histocompatibility complex molecule missing the alpha 3 domain of the heavy chain.** *Proc Natl Acad Sci U S A* 1995, **92**(4):1218-1221.
93. Tynan FE, Burrows SR, Buckle AM, Clements CS, Borg NA, Miles JJ, Beddoe T, Whisstock JC, Wilce MC, Silins SL *et al*: **T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide.** *Nat Immunol* 2005, **6**(11):1114-1122.
94. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC: **The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site.** *J Immunol* 2000, **164**(12):6398-6405.
95. Bolin DR, Swain AL, Sarabu R, Berthel SJ, Gillespie P, Huby NJ, Makofske R, Orzechowski L, Perrotta A, Toth K *et al*: **Peptide and peptide mimetic inhibitors of antigen presentation by HLA-DR class II MHC molecules. Design, structure-activity relationships, and X-ray crystal structures.** *J Med Chem* 2000, **43**(11):2135-2148.
96. Murthy VL, Stern LJ: **The class II MHC protein HLA-DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding.** *Structure* 1997, **5**(10):1385-1396.
97. Li Y, Li H, Martin R, Mariuzza RA: **Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 proteins.** *J Mol Biol* 2000, **304**(2):177-188.

98. Mak TW: **The T cell antigen receptor: "The Hunting of the Snark"**. *Eur J Immunol* 2007, **37 Suppl 1**:S83-93.
99. Li Y, Huang Y, Lue J, Quandt JA, Martin R, Mariuzza RA: **Structure of a human autoimmune TCR bound to a myelin basic protein self-peptide and a multiple sclerosis-associated MHC class II molecule**. *EMBO J* 2005, **24**(17):2968-2979.
100. Gras S, Burrows SR, Kjer-Nielsen L, Clements CS, Liu YC, Sullivan LC, Bell MJ, Brooks AG, Purcell AW, McCluskey J *et al*: **The shaping of T cell receptor recognition by self-tolerance**. *Immunity* 2009, **30**(2):193-203.
101. Hulsmeyer M, Chames P, Hillig RC, Stanfield RL, Held G, Coulie PG, Alings C, Wille G, Saenger W, Uchanska-Ziegler B *et al*: **A major histocompatibility complex-peptide-restricted antibody and t cell receptor molecules recognize their target by distinct binding modes: crystal structure of human leukocyte antigen (HLA)-A1-MAGE-A1 in complex with FAB-HYB3**. *J Biol Chem* 2005, **280**(4):2972-2980.
102. Yi Q, Osterborg A: **Idiotypic-specific T cells in multiple myeloma: targets for an immunotherapeutic intervention?** *Med Oncol* 1996, **13**(1):1-7.
103. Raitakari M, Brown RD, Sze D, Yuen E, Barrow L, Nelson M, Pope B, Esdale W, Gibson J, Joshua DE: **T-cell expansions in patients with multiple myeloma have a phenotype of cytotoxic T cells**. *Br J Haematol* 2000, **110**(1):203-209.
104. Sze DM, Giesajtis G, Brown RD, Raitakari M, Gibson J, Ho J, Baxter AG, Fazekas de St Groth B, Basten A, Joshua DE: **Clonal cytotoxic T cells are expanded in myeloma and reside in the CD8(+)CD57(+)CD28(-) compartment**. *Blood* 2001, **98**(9):2817-2827.
105. Sze DM, Brown RD, Yuen E, Gibson J, Ho J, Raitakari M, Basten A, Joshua DE, Fazekas de St Groth B: **Clonal cytotoxic T cells in myeloma**. *Leuk Lymphoma* 2003, **44**(10):1667-1674.
106. Okamoto M, Inaba T, Yamada N, Uchida R, Fuchida SI, Okano A, Shimazaki C, Taniwaki M: **Expression and role of MHC class I-related chain in myeloma cells**. *Cytotherapy* 2006, **8**(5):509-516.
107. Qian J, Xie J, Hong S, Yang J, Zhang L, Han X, Wang M, Zhan F, Shaughnessy JD, Jr., Epstein J *et al*: **Dickkopf-1 (DKK1) is a widely expressed and potent tumor-associated antigen in multiple myeloma**. *Blood* 2007, **110**(5):1587-1594.
108. Kjer-Nielsen L, Clements CS, Purcell AW, Brooks AG, Whisstock JC, Burrows SR, McCluskey J, Rossjohn J: **A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity**. *Immunity* 2003, **18**(1):53-64.

109. Bentley GA, Mariuzza RA: **The structure of the T cell antigen receptor.** *Annu Rev Immunol* 1996, **14**:563-590.
110. Buck CA: **Immunoglobulin superfamily: structure, function and relationship to other receptor molecules.** *Semin Cell Biol* 1992, **3**(3):179-188.
111. Bjorkman PJ: **MHC restriction in three dimensions: a view of T cell receptor/ligand interactions.** *Cell* 1997, **89**(2):167-170.
112. DeFranco AL: **Structure and function of the B cell antigen receptor.** *Annu Rev Cell Biol* 1993, **9**:377-410.
113. Reth M: **B cell antigen receptors.** *Curr Opin Immunol* 1994, **6**(1):3-8.
114. Yoshida K, Corper AL, Herro R, Jabri B, Wilson IA, Teyton L: **The diabetogenic mouse MHC class II molecule I-Ag7 is endowed with a switch that modulates TCR affinity.** *J Clin Invest* 2010, **120**(5):1578-1590.
115. Jones LL, Colf LA, Stone JD, Garcia KC, Kranz DM: **Distinct CDR3 conformations in TCRs determine the level of cross-reactivity for diverse antigens, but not the docking orientation.** *J Immunol* 2008, **181**(9):6255-6264.
116. Weintraub BC, Goodnow CC: **Immune Responses: costimulatory receptors have their say.** *Curr Biol* 1998, **8**(16):R575-577.
117. Garcia KC, Scott CA, Brunmark A, Carbone FR, Peterson PA, Wilson IA, Teyton L: **CD8 enhances formation of stable T-cell receptor/MHC class I molecule complexes.** *Nature* 1996, **384**(6609):577-581.
118. Wulfig C, Davis MM: **A receptor/cytoskeletal movement triggered by costimulation during T cell activation.** *Science* 1998, **282**(5397):2266-2269.
119. Garcia KC, Teyton L, Wilson IA: **Structural basis of T cell recognition.** *Annu Rev Immunol* 1999, **17**:369-397.
120. Garcia KC, Degano M, Speir JA, Wilson IA: **Emerging principles for T cell receptor recognition of antigen in cellular immunity.** *Rev Immunogenet* 1999, **1**(1):75-90.
121. Dustin ML, Cooper JA: **The immunological synapse and the actin cytoskeleton: molecular hardware for T cell signaling.** *Nat Immunol* 2000, **1**(1):23-29.
122. Hennecke J, Wiley DC: **T cell receptor-MHC interactions up close.** *Cell* 2001, **104**(1):1-4.
123. Garcia KC, Degano M, Stanfield RL, Brunmark A, Jackson MR, Peterson PA, Teyton L, Wilson IA: **An alphabeta T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex.** *Science* 1996, **274**(5285):209-219.

124. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC: **Structure of the complex between human T-cell receptor, viral peptide and HLA-A2.** *Nature*. 1996. **384**: 134-141. *J Immunol* 2010, **185**(11):6394-6401.
125. Garcia KC, Degano M, Pease LR, Huang M, Peterson PA, Teyton L, Wilson IA: **Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen.** *Science* 1998, **279**(5354):1166-1172.
126. Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC: **Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids.** *Immunity* 1998, **8**(4):403-411.
127. Wang L, Zhao Y, Li Z, Guo Y, Jones LL, Kranz DM, Mourad W, Li H: **Crystal structure of a complete ternary complex of TCR, superantigen and peptide-MHC.** *Nat Struct Mol Biol* 2007, **14**(2):169-171.
128. Jerne NK: **The somatic generation of immune recognition.** *Eur J Immunol* 1971, **1**(1):1-9.
129. Jerne NK: **The somatic generation of immune recognition. 1971.** *Eur J Immunol* 2004, **34**(5):1234-1242.
130. Reinherz EL, Tan K, Tang L, Kern P, Liu J, Xiong Y, Hussey RE, Smolyar A, Hare B, Zhang R *et al*: **The crystal structure of a T cell receptor in complex with peptide and MHC class II.** *Science* 1999, **286**(5446):1913-1921.
131. Wilson IA: **Perspectives: protein structure. Class-conscious TCR?** *Science* 1999, **286**(5446):1867-1868.
132. Colf LA, Bankovich AJ, Hanick NA, Bowerman NA, Jones LL, Kranz DM, Garcia KC: **How a single T cell receptor recognizes both self and foreign MHC.** *Cell* 2007, **129**(1):135-146.
133. Hennecke J, Wiley DC: **Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity.** *J Exp Med* 2002, **195**(5):571-581.
134. Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, Dorrell L, Douek DC, van der Merwe PA, Jones EY, McMichael AJ: **T cell cross-reactivity and conformational changes during TCR engagement.** *J Exp Med* 2004, **200**(11):1455-1466.
135. Dunn SM, Rizkallah PJ, Baston E, Mahon T, Cameron B, Moysey R, Gao F, Sami M, Boulter J, Li Y *et al*: **Directed evolution of human T cell receptor CDR2**

- residues by phage display dramatically enhances affinity for cognate peptide-MHC without increasing apparent cross-reactivity. *Protein Sci* 2006, **15**(4):710-721.
136. Hennecke J, Carfi A, Wiley DC: **Structure of a covalently stabilized complex of a human alphabeta T-cell receptor, influenza HA peptide and MHC class II molecule, HLA-DR1.** *EMBO J* 2000, **19**(21):5611-5624.
 137. Deng L, Langley RJ, Brown PH, Xu G, Teng L, Wang Q, Gonzales MI, Callender GG, Nishimura MI, Topalian SL *et al*: **Structural basis for the recognition of mutant self by a tumor-specific, MHC class II-restricted T cell receptor.** *Nat Immunol* 2007, **8**(4):398-408.
 138. Mazza C, Auphan-Anezin N, Gregoire C, Guimezanes A, Kellenberger C, Roussel A, Kearney A, van der Merwe PA, Schmitt-Verhulst AM, Malissen B: **How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides?** *EMBO J* 2007, **26**(7):1972-1983.
 139. Reiser JB, Gregoire C, Darnault C, Mosser T, Guimezanes A, Schmitt-Verhulst AM, Fontecilla-Camps JC, Mazza G, Malissen B, Housset D: **A T cell receptor CDR3beta loop undergoes conformational changes of unprecedented magnitude upon binding to a peptide/MHC class I complex.** *Immunity* 2002, **16**(3):345-354.
 140. Buslepp J, Wang H, Biddison WE, Appella E, Collins EJ: **A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection.** *Immunity* 2003, **19**(4):595-606.
 141. Rudolph MG, Wilson IA: **The specificity of TCR/pMHC interaction.** *Curr Opin Immunol* 2002, **14**(1):52-65.
 142. Ishizuka J, Stewart-Jones GB, van der Merwe A, Bell JI, McMichael AJ, Jones EY: **The structural dynamics and energetics of an immunodominant T cell receptor are programmed by its Vbeta domain.** *Immunity* 2008, **28**(2):171-182.
 143. Turner SJ, Doherty PC, McCluskey J, Rossjohn J: **Structural determinants of T-cell receptor bias in immunity.** *Nat Rev Immunol* 2006, **6**(12):883-894.
 144. Maynard J, Petersson K, Wilson DH, Adams EJ, Blondelle SE, Boulanger MJ, Wilson DB, Garcia KC: **Structure of an autoimmune T cell receptor complexed with class II peptide-MHC: insights into MHC bias and antigen specificity.** *Immunity* 2005, **22**(1):81-92.

145. Feng D, Bond CJ, Ely LK, Maynard J, Garcia KC: **Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'**. *Nat Immunol* 2007, **8**(9):975-983.
146. Reiser JB, Darnault C, Gregoire C, Mosser T, Mazza G, Kearney A, van der Merwe PA, Fontecilla-Camps JC, Housset D, Malissen B: **CDR3 loop flexibility contributes to the degeneracy of TCR recognition**. *Nat Immunol* 2003, **4**(3):241-247.
147. Burrows SR, Chen Z, Archbold JK, Tynan FE, Beddoe T, Kjer-Nielsen L, Miles JJ, Khanna R, Moss DJ, Liu YC *et al*: **Hard wiring of T cell receptor specificity for the major histocompatibility complex is underpinned by TCR adaptability**. *Proc Natl Acad Sci U S A* 2010, **107**(23):10608-10613.
148. Gras S, Saulquin X, Reiser JB, Debeaupuis E, Echasserieau K, Kissenpfennig A, Legoux F, Chouquet A, Le Gorrec M, Machillot P *et al*: **Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope**. *J Immunol* 2009, **183**(1):430-437.
149. Borbulevych OY, Piepenbrink KH, Gloor BE, Scott DR, Sommese RF, Cole DK, Sewell AK, Baker BM: **T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility**. *Immunity* 2009, **31**(6):885-896.
150. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SG: **IMGT/HLA database--a sequence database for the human major histocompatibility complex**. *Tissue Antigens* 2000, **55**(3):280-287.
151. Robinson J, Marsh SG: **The IMGT/HLA sequence database**. *Rev Immunogenet* 2000, **2**(4):518-531.
152. Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SG: **IMGT/HLA Database--a sequence database for the human major histocompatibility complex**. *Nucleic Acids Res* 2001, **29**(1):210-213.
153. Marsh SG: **HLA nomenclature and the IMGT/HLA sequence database**. *Novartis Found Symp* 2003, **254**:165-173; discussion 173-166, 216-122, 250-162.
154. Robinson J, Marsh SG: **The IMGT/HLA database**. *Methods Mol Biol* 2007, **409**:43-60.
155. Robinson J, Waller MJ, Fail SC, McWilliam H, Lopez R, Parham P, Marsh SG: **The IMGT/HLA database**. *Nucleic Acids Res* 2009, **37**(Database issue):D1013-1017.

156. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SG: **The IMGT/HLA database.** *Nucleic Acids Res* 2011, **39**(Database issue):D1171-1176.
157. Williams TM: **Human leukocyte antigen gene polymorphism and the histocompatibility laboratory.** *J Mol Diagn* 2001, **3**(3):98-104.
158. Sette A, Vitiello A, Rehman B, Fowler P, Nayarsina R, Kast WM, Melief CJ, Oseroff C, Yuan L, Ruppert J *et al*: **The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes.** *J Immunol* 1994, **153**(12):5586-5592.
159. Feltkamp MC, Vierboom MP, Kast WM, Melief CJ: **Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity.** *Mol Immunol* 1994, **31**(18):1391-1401.
160. Doytchinova IA, Flower DR: **Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201.** *J Med Chem* 2001, **44**(22):3572-3581.
161. Doytchinova IA, Blythe MJ, Flower DR: **Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A*0201.** *J Proteome Res* 2002, **1**(3):263-272.
162. Doytchinova IA, Flower DR: **Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction.** *Bioinformatics* 2003, **19**(17):2263-2270.
163. Doytchinova IA, Guan P, Flower DR: **Identifying human MHC supertypes using bioinformatic methods.** *J Immunol* 2004, **172**(7):4314-4323.
164. Doytchinova IA, Walshe VA, Jones NA, Gloster SE, Borrow P, Flower DR: **Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes.** *J Immunol* 2004, **172**(12):7495-7502.
165. Doytchinova IA, Flower DR: **In silico identification of supertypes for class II MHCs.** *J Immunol* 2005, **174**(11):7085-7095.
166. Sette A: **The immune epitope database and analysis resource: from vision to blueprint.** *Genome Inform* 2004, **15**(2):299.
167. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O *et al*: **The immune epitope database and analysis resource: from vision to blueprint.** *PLoS Biol* 2005, **3**(3):e91.

168. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J *et al*: **Immune epitope database analysis resource (IEDB-AR)**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W513-518.
169. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B: **The immune epitope database 2.0**. *Nucleic Acids Res* 2010, **38**(Database issue):D854-862.
170. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs**. *Immunogenetics* 1999, **50**(3-4):213-219.
171. Schuler MM, Nastke MD, Stevanovic S: **SYFPEITHI: database for searching and T-cell epitope prediction**. *Methods Mol Biol* 2007, **409**:75-93.
172. Schirle M, Weinschenk T, Stevanovic S: **Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens**. *J Immunol Methods* 2001, **257**(1-2):1-16.
173. Zhao Y, Pinilla C, Valmori D, Martin R, Simon R: **Application of support vector machines for T-cell epitopes prediction**. *Bioinformatics* 2003, **19**(15):1978-1984.
174. Brusic V, Schonbach C, Takiguchi M, Ciesielski V, Harrison LC: **Application of genetic search in derivation of matrix models of peptide binding to MHC molecules**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:75-83.
175. Brusic V, Rudy G, Harrison LC: **Prediction of MHC binding peptides using artificial neural networks**. In: *Complex Systems: Mechanism of Adaptation*. Edited by Stonier RJ, Yu XS. Amsterdam: IOS Press; 1994: 253-260.
176. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L: **Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network**. *Bioinformatics* 1998, **14**(2):121-130.
177. Brusic V, Petrovsky N, Zhang G, Bajic VB: **Prediction of promiscuous peptides that bind HLA class I molecules**. *Immunol Cell Biol* 2002, **80**(3):280-285.
178. Hammerstrom D: **Neural networks at work**. *IEEE Spectrum* 1993, **30**:26-32.
179. Karplus K: **Evaluating regularizers for estimating distributions of amino acids**. *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:188-196.
180. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology**. *Comput Appl Biosci* 1996, **12**(4):327-345.

181. Eskin E, Noble WS, Singer Y: **Using substitution matrices to estimate probability distributions for biological sequences.** *J Comput Biol* 2002, **9**(6):775-791.
182. Yu YK, Altschul SF: **The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions.** *Bioinformatics* 2005, **21**(7):902-911.
183. Knutson KL, Schiffman K, Disis ML: **Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients.** *J Clin Invest* 2001, **107**(4):477-484.
184. Lopez JA, Weilenman C, Audran R, Roggero MA, Bonelo A, Tiercy JM, Spertini F, Corradin G: **A synthetic malaria vaccine elicits a potent CD8(+) and CD4(+) T lymphocyte immune response in humans. Implications for vaccination strategies.** *Eur J Immunol* 2001, **31**(7):1989-1998.
185. Roberts JD, Niedzwiecki D, Carson WE, Chapman PB, Gajewski TF, Ernstoff MS, Hodi FS, Shea C, Leong SP, Johnson J *et al*: **Phase 2 study of the g209-2M melanoma peptide vaccine and low-dose interleukin-2 in advanced melanoma: Cancer and Leukemia Group B 509901.** *J Immunother* 2006, **29**(1):95-101.
186. Bourdette DN, Edmonds E, Smith C, Bowen JD, Guttmann CR, Nagy ZP, Simon J, Whitham R, Lovera J, Yadav V *et al*: **A highly immunogenic trivalent T cell receptor peptide vaccine for multiple sclerosis.** *Mult Scler* 2005, **11**(5):552-561.
187. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R: **UniProt archive.** *Bioinformatics* 2004, **20**(17):3236-3237.
188. Apweiler R, Bairoch A, Wu CH: **Protein sequence databases.** *Curr Opin Chem Biol* 2004, **8**(1):76-80.
189. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: **Infrastructure for the life sciences: design and implementation of the UniProt website.** *BMC Bioinformatics* 2009, **10**:136.
190. The UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Res* 2010, **38**(Database issue):D142-148.
191. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33**(Database issue):D154-159.
192. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R *et al*: **The Universal Protein Resource (UniProt):**

- an expanding universe of protein information.** *Nucleic Acids Res* 2006, **34**(Database issue):D187-191.
193. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot.** *Methods Mol Biol* 2007, **406**:89-112.
 194. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between evolution and stability.** *Brief Bioinform* 2004, **5**(1):39-55.
 195. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998.** *Nucleic Acids Res* 1998, **26**(1):38-42.
 196. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.** *Nucleic Acids Res* 1999, **27**(1):49-54.
 197. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A: **Protein variety and functional diversity: Swiss-Prot annotation in its biological context.** *C R Biol* 2005, **328**(10-11):882-899.
 198. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**(10):1282-1288.
 199. Cote RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H: **The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases.** *BMC Bioinformatics* 2007, **8**:401.
 200. Stoesser G, Sterk P, Tuli MA, Stoeck P, Cameron GN: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 1997, **25**(1):7-14.
 201. Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 1998, **26**(1):8-15.
 202. Stoesser G, Tuli MA, Lopez R, Sterk P: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 1999, **27**(1):18-24.
 203. Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 2000, **28**(1):19-23.
 204. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Lombard V, Lopez R, Parkinson H *et al*: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 2001, **29**(1):17-21.
 205. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V *et al*: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res* 2002, **30**(1):21-26.

206. Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R *et al*: **The EMBL Nucleotide Sequence Database: major new developments**. *Nucleic Acids Res* 2003, **31**(1):17-22.
207. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R *et al*: **The EMBL Nucleotide Sequence Database**. *Nucleic Acids Res* 2004, **32**(Database issue):D27-30.
208. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G *et al*: **The EMBL Nucleotide Sequence Database**. *Nucleic Acids Res* 2005, **33**(Database issue):D29-33.
209. Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A *et al*: **EMBL Nucleotide Sequence Database: developments in 2005**. *Nucleic Acids Res* 2006, **34**(Database issue):D10-15.
210. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P *et al*: **EMBL Nucleotide Sequence Database in 2006**. *Nucleic Acids Res* 2007, **35**(Database issue):D16-20.
211. Putnam NC: **Searching MEDLINE free on the Internet using the National Library of Medicine's PubMed**. *Clin Excell Nurse Pract* 1998, **2**(5):314-316.
212. Vastag B: **NIH launches PubMed Central**. *J Natl Cancer Inst* 2000, **92**(5):374.
213. Roberts RJ: **PubMed Central: The GenBank of the published literature**. *Proc Natl Acad Sci U S A* 2001, **98**(2):381-382.
214. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature**. *Database (Oxford)* 2011, **2011**:baq036.
215. Giudicelli V, Chaume D, Bodmer J, Muller W, Busin C, Marsh S, Bontrop R, Marc L, Malik A, Lefranc MP: **IMGT, the international ImMunoGeneTics database**. *Nucleic Acids Res* 1997, **25**(1):206-211.
216. Giudicelli V, Chaume D, Mennessier G, Althaus HH, Muller W, Bodmer J, Malik A, Lefranc MP: **IMGT, the international ImMunoGeneTics database: a new design for immunogenetics data access**. *Stud Health Technol Inform* 1998, **52 Pt 1**:351-355.
217. Lefranc MP, Giudicelli V, Busin C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Chaume D: **IMGT, the International ImMunoGeneTics database**. *Nucleic Acids Res* 1998, **26**(1):297-303.

218. Lefranc MP, Giudicelli V, Ginestoux C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Barbie V, Chaume D: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 1999, **27**(1):209-212.
219. Ruiz M, Giudicelli V, Ginestoux C, Stoeckl P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G *et al*: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 2000, **28**(1):219-221.
220. Lefranc MP: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 2001, **29**(1):207-209.
221. Lefranc MP: **IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology.** *Dev Comp Immunol* 2002, **26**(8):697-705.
222. Warr GW, Clem LW, Soderhall K: **The international ImMunoGeneTics Database IMGT.** *Dev Comp Immunol* 2003, **27**(1):1.
223. Lefranc MP: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 2003, **31**(1):307-310.
224. Lefranc MP, Giudicelli V, Ginestoux C, Chaume D: **IMGT, the international ImMunoGeneTics information system, <http://imgt.cines.fr>: the reference in immunoinformatics.** *Stud Health Technol Inform* 2003, **95**:74-79.
225. Lefranc MP: **IMGT, the international ImMunoGeneTics information system, <http://imgt.cines.fr>.** *Novartis Found Symp* 2003, **254**:126-136; discussion 136-142, 216-122, 250-122.
226. Lefranc MP: **IMGT, The International ImMunoGeneTics Information System, <http://imgt.cines.fr>.** *Methods Mol Biol* 2004, **248**:27-49.
227. Lefranc MP, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, Ginestoux C, Clement O, Chaume D, Lefranc G: **IMGT, the international ImMunoGeneTics information system.** *Nucleic Acids Res* 2005, **33**(Database issue):D593-597.
228. Lefranc MP: **IMGT, the international ImMunoGeneTics information system: a standardized approach for immunogenetics and immunoinformatics.** *Immunome Res* 2005, **1**:3.
229. Lefranc MP: **IMGT, the international ImMunoGeneTics information system for Immunoinformatics. Methods for querying IMGT databases, tools, and Web resources in the context of immunoinformatics.** *Methods Mol Biol* 2007, **409**:19-42.

230. Lefranc MP: **IMGT, the International ImMunoGeneTics Information System for Immunoinformatics : methods for querying IMGT databases, tools, and web resources in the context of immunoinformatics.** *Mol Biotechnol* 2008, **40**(1):101-111.
231. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J *et al*: **IMGT, the international ImMunoGeneTics information system.** *Nucleic Acids Res* 2009, **37**(Database issue):D1006-1012.
232. Helmborg W, Dunivin R, Feolo M: **The reagent database at dbMHC.** *Tissue Antigens* 2004, **63**(2):142-148.
233. Bhasin M, Singh H, Raghava GP: **MHCBN: a comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics* 2003, **19**(5):665-666.
234. Lata S, Bhasin M, Raghava GP: **MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes.** *BMC Res Notes* 2009, **2**:61.
235. Brusica V, Rudy G, Harrison LC: **MHCPEP: a database of MHC-binding peptides.** *Nucleic Acids Res* 1994, **22**(17):3663-3665.
236. Brusica V, Rudy G, Kyne AP, Harrison LC: **MHCPEP--a database of MHC-binding peptides: update 1995.** *Nucleic Acids Res* 1996, **24**(1):242-244.
237. Brusica V, Rudy G, Kyne AP, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1996.** *Nucleic Acids Res* 1997, **25**(1):269-271.
238. Brusica V, Rudy G, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1997.** *Nucleic Acids Res* 1998, **26**(1):368-371.
239. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwigama CK, Flower DR: **AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data.** *Immunome Res* 2005, **1**(1):4.
240. Giudicelli V, Chaume D, Lefranc MP: **IMGT/LIGM-DB: a systematized approach for ImMunoGeneTics database coherence and data distribution improvement.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:59-68.
241. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, Lefranc MP: **IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences.** *Nucleic Acids Res* 2006, **34**(Database issue):D781-784.

242. Giudicelli V, Chaume D, Lefranc MP: **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.** *Nucleic Acids Res* 2005, **33**(Database issue):D256-261.
243. Robinson J, Waller MJ, Stoeckl P, Marsh SG: **IPD--the Immuno Polymorphism Database.** *Nucleic Acids Res* 2005, **33**(Database issue):D523-526.
244. Robinson J, Waller MJ, Fail SC, Marsh SG: **The IMGT/HLA and IPD databases.** *Hum Mutat* 2006, **27**(12):1192-1199.
245. Robinson J, Marsh SG: **IPD: the Immuno Polymorphism Database.** *Methods Mol Biol* 2007, **409**:61-74.
246. Robinson J, Mistry K, McWilliam H, Lopez R, Marsh SG: **IPD--the Immuno Polymorphism Database.** *Nucleic Acids Res* 2010, **38**(Database issue):D863-869.
247. Reche PA, Zhang H, Glutting JP, Reinherz EL: **EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology.** *Bioinformatics* 2005, **21**(9):2140-2141.
248. Jongeneel V: **Towards a cancer immunome database.** *Cancer Immun* 2001, **1**:3.
249. Schlessinger A, Ofra Y, Yachdav G, Rost B: **Epitome: database of structure-inferred antigenic epitopes.** *Nucleic Acids Res* 2006, **34**(Database issue):D777-780.
250. **HIV Molecular Immunology.** In: *Theoretical Biology and Biophysics*. Edited by Korber BTM, Brander C, Haynes BF, Koup R, Moore JP, Walker BD, Watkins DI. Los Alamos, New Mexico: Los Alamos National Laboratory; 2006/2007.
251. Giudicelli V, Lefranc MP: **Ontology for immunogenetics: the IMGT-ONTOLOGY.** *Bioinformatics* 1999, **15**(12):1047-1054.
252. Lefranc MP, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V *et al*: **IMGT-ONTOLOGY for immunogenetics and immunoinformatics.** *In Silico Biol* 2004, **4**(1):17-29.
253. Lefranc MP: **IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics.** *Mol Immunol* 2004, **40**(10):647-660.
254. Lefranc MP, Clement O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D *et al*: **IMGT-Choreography for immunogenetics and immunoinformatics.** *In Silico Biol* 2005, **5**(1):45-60.
255. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc MP, Giudicelli V: **IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm.** *Biochimie* 2008, **90**(4):570-583.

256. Pappalardo F, Lefranc MP, Lollini PL, Motta S: **A novel paradigm for cell and molecule interaction ontology: from the CMM model to IMGT-ONTOLOGY.** *Immunome Res* 2010, **6**(1):1.
257. Lefranc MP, Pommie C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G: **IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.** *Dev Comp Immunol* 2003, **27**(1):55-77.
258. Lefranc MP: **Nomenclature of the human immunoglobulin genes.** *Curr Protoc Immunol* 2001, **Appendix 1**:Appendix 1P.
259. Lefranc MP: **Nomenclature of the human T cell receptor genes.** *Curr Protoc Immunol* 2001, **Appendix 1**:Appendix 1O.
260. Giudicelli V, Chaume D, Lefranc MP: **IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W435-440.
261. Brochet X, Lefranc MP, Giudicelli V: **IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W503-508.
262. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP: **IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs.** *Bioinformatics* 2004, **20 Suppl 1**:i379-385.
263. Alamyar E, Giudicelli V, Duroux P, Lefranc MP: **IMGT/HighV-QUEST: A High-Throughput System and Web Portal for the Analysis of Rearranged Nucleotide Sequences of Antigen Receptors - High-Throughput Version of IMGT/V-QUEST** In: *Journées Ouvertes en Biologie, Informatique et Mathématiques: 2010; Montpellier, France; 2010.*
264. Elemento O, Lefranc MP: **IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes.** *Dev Comp Immunol* 2003, **27**(9):763-779.
265. Baum TP, Pasqual N, Thuderoz F, Hierle V, Chaume D, Lefranc MP, Jouvin-Marche E, Marche PN, Demongeot J: **IMGT/GeneInfo: enhancing V(D)J recombination database accessibility.** *Nucleic Acids Res* 2004, **32**(Database issue):D51-54.

266. Baum TP, Hierle V, Pasqual N, Bellahcene F, Chaume D, Lefranc MP, Jouvin-Marche E, Marche PN, Demongeot J: **IMGT/GeneInfo: T cell receptor gamma TRG and delta TRD genes in database give access to all TR potential V(D)J recombinations.** *BMC Bioinformatics* 2006, **7**:224.
267. Ruiz M, Lefranc MP: **IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures.** *Immunogenetics* 2002, **53**(10-11):857-883.
268. Kaas Q, Lefranc MP: **IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains.** *Current Bioinformatics* 2007, **2**:21-30.
269. Kaas Q, Ehrenmann F, Lefranc MP: **IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles?** *Brief Funct Genomic Proteomic* 2007, **6**(4):253-264.
270. Poirion C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc MP: **IMGT/mAb-DB: the IMGT® database for therapeutic monoclonal antibodies.** . In: *Journées Ouvertes en Biologie, Informatique et Mathématiques: 2010; Montpellier, France*; 2010.
271. Jerome JB, Alessandri MC: **Nonproprietary Names; National and International.** *JAMA* 1965, **192**:405-408.
272. Jerome JB, Luback PM: **International Nonproprietary Names (INN). A World Health Organization activity.** *J Am Med Womens Assoc* 1972, **27**(10):536-538.
273. Wehrli A: **The selection and protection of international nonproprietary names for pharmaceutical substances.** *WHO Chron* 1981, **35**(5):172-175.
274. Kopp-Kubel S: **International Nonproprietary Names (INN) for pharmaceutical substances.** *Bull World Health Organ* 1995, **73**(3):275-279.
275. Johnson G, Wu TT: **Kabat database and its applications: 30 years after the first variability plot.** *Nucleic Acids Res* 2000, **28**(1):214-218.
276. Johnson G, Wu TT: **Kabat Database and its applications: future directions.** *Nucleic Acids Res* 2001, **29**(1):205-206.
277. Johnson G, Wu TT: **The Kabat database and a bioinformatics example.** *Methods Mol Biol* 2004, **248**:11-25.
278. Helmberg W, Dunivin R, Feolo M: **The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W173-175.

279. Gourraud PA, Mano S, Barnetche T, Carrington M, Inoko H, Cambon-Thomsen A: **Integration of microsatellite characteristics in the MHC region: a literature and sequence based analysis.** *Tissue Antigens* 2004, **64**(5):543-555.
280. Gourraud PA, Feolo M, Hoffman D, Helmberg W, Cambon-Thomsen A: **The dbMHC microsatellite portal: a public resource for the storage and display of MHC microsatellite information.** *Tissue Antigens* 2006, **67**(5):395-401.
281. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
282. Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung CS, Bilofsky HS: **The GenBank nucleic acid sequence database.** *Comput Appl Biosci* 1985, **1**(4):225-233.
283. Burks C: **The GenBank database and the flow of sequence data for the human genome.** *Basic Life Sci* 1988, **46**:51-56.
284. Ouellette BF: **The GenBank sequence database.** *Methods Biochem Anal* 1998, **39**:16-45.
285. Karsch-Mizrachi I, Ouellette BF: **The GenBank sequence database.** *Methods Biochem Anal* 2001, **43**:45-63.
286. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2005, **33**(Database issue):D34-38.
287. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2007, **35**(Database issue):D21-25.
288. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2008, **36**(Database issue):D25-30.
289. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2009, **37**(Database issue):D26-31.
290. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2010, **38**(Database issue):D46-51.
291. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2011, **39**(Database issue):D32-37.
292. McKusick VA: **Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders**, 12th edn. Baltimore: Johns Hopkins University Press; 1998.
293. Boyadjev SA, Jabs EW: **Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders.** *Clin Genet* 2000, **57**(4):253-266.

294. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM)**. *Hum Mutat* 2000, **15**(1):57-61.
295. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2002, **30**(1):52-55.
296. Amladi S: **Online Mendelian Inheritance in Man 'OMIM'**. *Indian J Dermatol Venereol Leprol* 2003, **69**(6):423-424.
297. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33**(Database issue):D514-517.
298. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *Am J Hum Genet* 2007, **80**(4):588-604.
299. Amberger J, Bocchini CA, Scott AF, Hamosh A: **McKusick's Online Mendelian Inheritance in Man (OMIM)**. *Nucleic Acids Res* 2009, **37**(Database issue):D793-796.
300. Flower DR, Macdonald IK, Ramakrishnan K, Davies MN, Doytchinova IA: **Computer aided selection of candidate vaccine antigens**. *Immunome Res* 2010, **6 Suppl 2**:S1.
301. Akdis CA, Akdis M, Blesken T, Wymann D, Alkan SS, Muller U, Blaser K: **Epitope-specific T cell tolerance to phospholipase A2 in bee venom immunotherapy and recovery by IL-2 and IL-15 in vitro**. *J Clin Invest* 1996, **98**(7):1676-1683.
302. Stienekemeier M, Falk K, Rotzschke O, Weishaupt A, Schneider C, Toyka KV, Gold R, Strominger JL: **Vaccination, prevention, and treatment of experimental autoimmune neuritis (EAN) by an oligomerized T cell epitope**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13872-13877.
303. Lazoura E, Apostolopoulos V: **Insights into peptide-based vaccine design for cancer immunotherapy**. *Curr Med Chem* 2005, **12**(13):1481-1494.
304. Mocellin S, Pilati P, Nitti D: **Peptide-based anticancer vaccines: recent advances and future perspectives**. *Curr Med Chem* 2009, **16**(36):4779-4796.
305. Reche PA, Keskin DB, Hussey RE, Ancuta P, Gabuzda D, Reinherz EL: **Elicitation from virus-naïve individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes**. *Med Immunol* 2006, **5**:1.
306. Tchernev G, Orfanos CE: **Antigen mimicry, epitope spreading and the pathogenesis of pemphigus**. *Tissue Antigens* 2006, **68**(4):280-286.

307. Draenert R, Altfeld M, Brander C, Basgoz N, Corcoran C, Wurcel AG, Stone DR, Kalams SA, Trocha A, Addo MM *et al*: **Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T cell responses.** *J Immunol Methods* 2003, **275**(1-2):19-29.
308. Jensen PE: **Recent advances in antigen processing and presentation.** *Nat Immunol* 2007, **8**(10):1041-1048.
309. Lafuente EM, Reche PA: **Prediction of MHC-peptide binding: a systematic and comprehensive overview.** *Curr Pharm Des* 2009, **15**(28):3209-3220.
310. Doytchinova IA, Flower DR: **Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study.** *Proteins* 2002, **48**(3):505-518.
311. Jojic N, Reyes-Gomez M, Heckerman D, Kadie C, Schueler-Furman O: **Learning MHC I-peptide binding.** *Bioinformatics* 2006, **22**(14):e227-235.
312. Reche PA, Reinherz EL: **PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W138-142.
313. Reche PA, Reinherz EL: **Definition of MHC supertypes through clustering of MHC peptide-binding repertoires.** *Methods Mol Biol* 2007, **409**:163-173.
314. Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs.** *Hum Immunol* 2002, **63**(9):701-709.
315. Reche PA, Glutting JP, Zhang H, Reinherz EL: **Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.** *Immunogenetics* 2004, **56**(6):405-419.
316. Reche PA, Reinherz EL: **Prediction of peptide-MHC binding using profiles.** *Methods Mol Biol* 2007, **409**:185-200.
317. Parker KC, Bednarek MA, Coligan JE: **Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains.** *J Immunol* 1994, **152**(1):163-175.
318. Doytchinova IA, Guan P, Flower DR: **EpiJen: a server for multistep T cell epitope prediction.** *BMC Bioinformatics* 2006, **7**:131.
319. Schafer JR, Jesdale BM, George JA, Kouttab NM, De Groot AS: **Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix.** *Vaccine* 1998, **16**(19):1880-1884.
320. Singh H, Raghava GP: **ProPred1: prediction of promiscuous MHC Class-I binding sites.** *Bioinformatics* 2003, **19**(8):1009-1014.

321. Singh H, Raghava GP: **ProPred: prediction of HLA-DR binding sites.** *Bioinformatics* 2001, **17**(12):1236-1237.
322. Hakenberg J, Nussbaum AK, Schild H, Rammensee HG, Kuttler C, Holzhutter HG, Kloetzel PM, Kaufmann SH, Mollenkopf HJ: **MAPPP: MHC class I antigenic peptide processing prediction.** *Appl Bioinformatics* 2003, **2**(3):155-158.
323. Peters B, Tong W, Sidney J, Sette A, Weng Z: **Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules.** *Bioinformatics* 2003, **19**(14):1765-1772.
324. Dimitrov I, Garnev P, Flower DR, Doytchinova I: **EpiTOP--a proteochemometric tool for MHC class II binding prediction.** *Bioinformatics* 2010, **26**(16):2066-2068.
325. Dimitrov I, Garnev P, Flower DR, Doytchinova I: **Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis.** *Eur J Med Chem* 2010, **45**(1):236-243.
326. Guan P, Doytchinova IA, Zygouri C, Flower DR: **MHCPred: bringing a quantitative dimension to the online prediction of MHC binding.** *Appl Bioinformatics* 2003, **2**(1):63-66.
327. Guan P, Doytchinova IA, Zygouri C, Flower DR: **MHCPred: A server for quantitative prediction of peptide-MHC binding.** *Nucleic Acids Res* 2003, **31**(13):3621-3624.
328. Guan P, Hattotuagama CK, Doytchinova IA, Flower DR: **MHCPred 2.0: an updated quantitative T-cell epitope prediction server.** *Appl Bioinformatics* 2006, **5**(1):55-61.
329. Bhasin M, Raghava GP: **A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes.** *J Biosci* 2007, **32**(1):31-42.
330. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O *et al*: **NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence.** *PLoS One* 2007, **2**(8):e796.
331. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, Buus S, Nielsen M: **NetMHCpan, a method for MHC class I binding prediction beyond humans.** *Immunogenetics* 2009, **61**(1):1-13.
332. Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O: **Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan.** *PLoS Comput Biol* 2008, **4**(7):e1000107.

333. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S: **NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure.** *Immunome Res* 2010, **6**:9.
334. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V: **MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W172-179.
335. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S: **Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach.** *Tissue Antigens* 2003, **62**(5):378-384.
336. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O: **Reliable prediction of T-cell epitopes using neural networks with novel sequence representations.** *Protein Sci* 2003, **12**(5):1007-1017.
337. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: **Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach.** *Bioinformatics* 2004, **20**(9):1388-1397.
338. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M: **NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W509-512.
339. Lundegaard C, Lund O, Nielsen M: **Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers.** *Bioinformatics* 2008, **24**(11):1397-1398.
340. Nielsen M, Lundegaard C, Lund O: **Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method.** *BMC Bioinformatics* 2007, **8**:238.
341. Nielsen M, Lund O: **NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction.** *BMC Bioinformatics* 2009, **10**:296.
342. Jacob L, Vert JP: **Efficient peptide-MHC-I binding prediction for alleles with few known binders.** *Bioinformatics* 2008, **24**(3):358-366.
343. Bhasin M, Raghava GP: **SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence.** *Bioinformatics* 2004, **20**(3):421-423.

344. Tung CW, Ho SY: **POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties.** *Bioinformatics* 2007, **23**(8):942-949.
345. Donnes P, Elofsson A: **Prediction of MHC class I binding peptides, using SVMHC.** *BMC Bioinformatics* 2002, **3**:25.
346. Donnes P, Kohlbacher O: **SVMHC: a server for prediction of MHC-binding peptides.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W194-197.
347. Liu W, Meng X, Xu Q, Flower DR, Li T: **Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models.** *BMC Bioinformatics* 2006, **7**:182.
348. Wan J, Liu W, Xu Q, Ren Y, Flower DR, Li T: **SVRMHC prediction server for MHC-binding peptides.** *BMC Bioinformatics* 2006, **7**:463.
349. Liu W, Wan J, Meng X, Flower DR, Li T: **In silico prediction of peptide-MHC binding affinity using SVRMHC.** *Methods Mol Biol* 2007, **409**:283-291.
350. Swain MT, Brooks AJ, Kemp GJL: **Proceedings 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering** In: *2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2001): 2001; Bethesda, Maryland: IEEE Computer Society; 2001: 81-88.*
351. Schueler-Furman O, Altuvia Y, Sette A, Margalit H: **Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles.** *Protein Sci* 2000, **9**(9):1838-1846.
352. Falk K, Rotzschke O, Deres K, Metzger J, Jung G, Rammensee HG: **Identification of naturally processed viral nonapeptides allows their quantification in infected cells and suggests an allele-specific T cell epitope forecast.** *J Exp Med* 1991, **174**(2):425-434.
353. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG: **Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.** *Nature* 1991, **351**(6324):290-296.
354. Rotzschke O, Falk K, Stevanovic S, Jung G, Walden P, Rammensee HG: **Exact prediction of a natural T cell epitope.** *Eur J Immunol* 1991, **21**(11):2891-2894.
355. Pamer EG, Harty JT, Bevan MJ: **Precise prediction of a dominant class I MHC-restricted epitope of *Listeria monocytogenes*.** *Nature* 1991, **353**(6347):852-855.
356. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG: **Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.** **1991.** *J Immunol* 2006, **177**(5):2741-2747.

357. Jardetzky TS, Lane WS, Robinson RA, Madden DR, Wiley DC: **Identification of self peptides bound to purified HLA-B27.** *Nature* 1991, **353**(6342):326-329.
358. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH: **Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.** *Science* 1992, **255**(5049):1261-1263.
359. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH: **Pillars article: Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.** *Science* 1992, **255**: 1261-1263. *J Immunol* 2007, **179**(5):2669-2671.
360. Zhang QJ, Gavioli R, Klein G, Masucci MG: **An HLA-A11-specific motif in nonamer peptides derived from viral and cellular proteins.** *Proc Natl Acad Sci U S A* 1993, **90**(6):2217-2221.
361. Lipford GB, Hoffman M, Wagner H, Heeg K: **Primary in vivo responses to ovalbumin. Probing the predictive value of the Kb binding motif.** *J Immunol* 1993, **150**(4):1212-1222.
362. Sette A, Sidney J, Oseroff C, del Guercio MF, Southwood S, Arrhenius T, Powell MF, Colon SM, Gaeta FC, Grey HM: **HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions.** *J Immunol* 1993, **151**(6):3163-3170.
363. Sidney J, Oseroff C, del Guercio MF, Southwood S, Krieger JI, Ishioka GY, Sakaguchi K, Appella E, Sette A: **Definition of a DQ3.1-specific binding motif.** *J Immunol* 1994, **152**(9):4516-4525.
364. Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F: **Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning.** *J Exp Med* 1994, **180**(6):2353-2358.
365. Rammensee HG, Friede T, Stevanović S: **MHC ligands and peptide motifs: first listing.** *Immunogenetics* 1995, **41**(4):178-228.
366. Meister GE, Roberts CG, Berzofsky JA, De Groot AS: **Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences.** *Vaccine* 1995, **13**(6):581-591.
367. D'Amato J, Houbiers JG, Drijfhout JW, Brandt RM, Schipper R, Bavinck JN, Melief CJ, Kast WM: **A computer program for predicting possible cytotoxic T**

- lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum Immunol* 1995, **43**(1):13-18.
368. Jameson SC, Bevan MJ: **Dissection of major histocompatibility complex (MHC) and T cell receptor contact residues in a Kb-restricted ovalbumin peptide and an assessment of the predictive power of MHC-binding motifs.** *Eur J Immunol* 1992, **22**(10):2663-2667.
 369. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A: **Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules.** *Cell* 1993, **74**(5):929-937.
 370. Chen W, Khilko S, Fecondo J, Margulies DH, McCluskey J: **Determinant selection of major histocompatibility complex class I-restricted antigenic peptides is explained by class I-peptide affinity and is strongly influenced by nondominant anchor residues.** *J Exp Med* 1994, **180**(4):1471-1483.
 371. Rajapakse M, Schmidt B, Brusica V: **Multi-Objective Evolutionary Algorithm for Discovering Peptide Binding Motifs.** In: *Applications of Evolutionary Computing*. Edited by Rothlauf F, Branke J, Cagnoni S, Costa E, Cotta C, Drechsler R, Lutton E, Machado P, Moore J, Romero J *et al*, vol. 3907: Springer Berlin / Heidelberg; 2006: 149-158.
 372. Bouvier M, Wiley DC: **Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules.** *Science* 1994, **265**(5170):398-402.
 373. Scott CA, Peterson PA, Teyton L, Wilson IA: **Crystal structures of two I-Ad-peptide complexes reveal that high affinity can be achieved without large anchor residues.** *Immunity* 1998, **8**(3):319-329.
 374. Martin W, Sbai H, De Groot AS: **Bioinformatics tools for identifying class I-restricted epitopes.** *Methods* 2003, **29**(3):289-298.
 375. Davenport MP, Ho Shon IA, Hill AV: **An empirical method for the prediction of T-cell epitopes.** *Immunogenetics* 1995, **42**(5):392-397.
 376. Gulukota K, Sidney J, Sette A, DeLisi C: **Two complementary methods for predicting peptides binding major histocompatibility complex molecules.** *J Mol Biol* 1997, **267**(5):1258-1267.
 377. De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G: **An interactive Web site providing major histocompatibility ligand predictions: application to HIV research.** *AIDS Res Hum Retroviruses* 1997, **13**(7):529-531.
 378. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.

379. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:21-29.
380. Rajapakse M, Wyse L, Schmidt B, Brusica V: **Deriving Matrix of Peptide-MHC Interactions in Diabetic Mouse by Genetic Algorithm.** In: *Intelligent Data Engineering and Automated Learning - IDEAL 2005*. Edited by Gallagher M, Hogan J, Maire F, vol. 3578: Springer Berlin / Heidelberg; 2005: 440-447.
381. Mallios RR: **Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 1999, **15**(6):432-439.
382. Stryhn A, Pedersen LO, Romme T, Holm CB, Holm A, Buus S: **Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding.** *Eur J Immunol* 1996, **26**(8):1911-1918.
383. Udaka K, Wiesmuller KH, Kienle S, Jung G, Tamamura H, Yamagishi H, Okumura K, Walden P, Suto T, Kawasaki T: **An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries.** *Immunogenetics* 2000, **51**(10):816-828.
384. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothe BR, Chisari FV, Watkins DI, Sette A: **Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications.** *Immunogenetics* 2005, **57**(5):304-314.
385. Peters B, Sette A: **Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method.** *BMC Bioinformatics* 2005, **6**:132.
386. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F *et al*: **Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices.** *Nat Biotechnol* 1999, **17**(6):555-561.
387. Bian H, Hammer J: **Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE.** *Methods* 2004, **34**(4):468-475.
388. Duda RO, Hart PE, Stork DG: **Pattern Classification.** *Journal of Classification* 2001, **24**(2):305-307.
389. Kingsford C, Salzberg SL: **What are decision trees?** *Nat Biotechnol* 2008, **26**(9):1011-1013.

390. Savoie CJ, Kamikawaji N, Sasazuki T, Kuhara S: **Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs.** *Pac Symp Biocomput* 1999;182-189.
391. Segal MR, Cummings MP, Hubbard AE: **Relating amino acid sequence to phenotype: analysis of peptide-binding data.** *Biometrics* 2001, **57**(2):632-642.
392. Zhu S, Udaka K, Sidney J, Sette A, Aoki-Kinoshita KF, Mamitsuka H: **Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules.** *Bioinformatics* 2006, **22**(13):1648-1655.
393. Presnell SR, Cohen FE: **Artificial neural networks for pattern recognition in biochemical sequences.** *Annu Rev Biophys Biomol Struct* 1993, **22**:283-298.
394. Adams HP, Koziol JA: **Prediction of binding to MHC class I molecules.** *J Immunol Methods* 1995, **185**(2):181-190.
395. Milik M, Sauer D, Brunmark AP, Yuan L, Vitiello A, Jackson MR, Peterson PA, Skolnick J, Glass CA: **Application of an artificial neural network to predict specific class I MHC binding peptide sequences.** *Nat Biotechnol* 1998, **16**(8):753-756.
396. Honeyman MC, Brusica V, Stone NL, Harrison LC: **Neural network-based prediction of candidate T-cell epitopes.** *Nat Biotechnol* 1998, **16**(10):966-969.
397. Soam SS, Khan F, Bhasker B, Mishra BN: **Prediction of MHC class I binding peptides using probability distribution functions.** *Bioinformation* 2009, **3**(9):403-408.
398. Schuster-Bockler B, Bateman A: **An introduction to hidden Markov models.** *Curr Protoc Bioinformatics* 2007, **Appendix 3**:Appendix 3A.
399. Rose RC, Juang BH: **Hidden Markov models for speech and signal recognition.** *Electroencephalogr Clin Neurophysiol Suppl* 1996, **45**:137-152.
400. Mamitsuka H: **Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models.** *Proteins* 1998, **33**(4):460-474.
401. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, Brusica V, Kobayashi T: **Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules.** *J Biosci Bioeng* 2002, **94**(3):264-270.
402. Zhang C, Bickis MG, Wu FX, Kusalik AJ: **Optimally-connected hidden markov models for predicting MHC-binding peptides.** *J Bioinform Comput Biol* 2006, **4**(5):959-980.

403. Wistrand M, Sonnhammer EL: **Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER.** *BMC Bioinformatics* 2005, **6**:99.
404. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids.**: Cambridge University Press; 1998.
405. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ: **Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach.** *Nucleic Acids Res* 2004, **32**(21):6437-6444.
406. Byvatov E, Schneider G: **Support vector machine applications in bioinformatics.** *Appl Bioinformatics* 2003, **2**(2):67-77.
407. Bhasin M, Zhang H, Reinherz EL, Reche PA: **Prediction of methylated CpGs in DNA sequences using a support vector machine.** *FEBS Lett* 2005, **579**(20):4302-4308.
408. Bhasin M, Reinherz EL, Reche PA: **Recognition and classification of histones using support vector machine.** *J Comput Biol* 2006, **13**(1):102-112.
409. Bozic I, Zhang GL, Brusica V: **Predictive Vaccinology: Optimisation of Predictions Using Support Vector Machine Classifiers.** *Intelligent Data Engineering and Automated Learning - IDEAL 2005* 2005, **3578**:127-132.
410. Bhasin M, Raghava GP: **Prediction of CTL epitopes using QM, SVM and ANN techniques.** *Vaccine* 2004, **22**(23-24):3195-3204.
411. Salomon J, Flower DR: **Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores.** *BMC Bioinformatics* 2006, **7**:501.
412. Akutsu T, Sim KL: **Protein Threading Based on Multiple Protein Structure Alignment.** *Genome Inform Ser Workshop Genome Inform* 1999, **10**:23-29.
413. Swindells MB: **Structure prediction and modelling.** *Curr Opin Biotechnol* 1992, **3**(4):338-347.
414. Altuvia Y, Schueler O, Margalit H: **Ranking potential binding peptides to MHC molecules by a computational threading approach.** *J Mol Biol* 1995, **249**(2):244-250.
415. Hammer J, Gallazzi F, Bono E, Karr RW, Guenot J, Valsasnini P, Nagy ZA, Sinigaglia F: **Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association.** *J Exp Med* 1995, **181**(5):1847-1855.

416. Flower DR, Phadwal K, Macdonald IK, Coveney PV, Davies MN, Wan S: **T-cell epitope prediction and immune complex simulation using molecular dynamics: state of the art and persisting challenges.** *Immunome Res* 2010, **6 Suppl 2**:S4.
417. Swindells MB, Thornton JM: **Structure prediction and modelling.** *Curr Opin Biotechnol* 1991, **2**(4):512-519.
418. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234**(3):779-815.
419. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V: **Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins.** *J Med Chem* 1999, **42**(22):4650-4658.
420. Zhang C, Anderson A, DeLisi C: **Structural principles that govern the peptide-binding motifs of class I MHC molecules.** *J Mol Biol* 1998, **281**(5):929-947.
421. Michielin O, Luescher I, Karplus M: **Modeling of the TCR-MHC-peptide complex.** *J Mol Biol* 2000, **300**(5):1205-1235.
422. Logean A, Sette A, Rognan D: **Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions.** *Bioorg Med Chem Lett* 2001, **11**(5):675-679.
423. Michielin O, Karplus M: **Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a simulation analysis.** *J Mol Biol* 2002, **324**(3):547-569.
424. Sezerman U, Vajda S, DeLisi C: **Free energy mapping of class I MHC molecules and structural determination of bound peptides.** *Protein Sci* 1996, **5**(7):1272-1281.
425. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J Mol Biol* 1996, **256**(3):623-644.
426. Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H: **A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets.** *Hum Immunol* 1997, **58**(1):1-11.
427. Schueler-Furman O, Elber R, Margalit H: **Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes.** *Fold Des* 1998, **3**(6):549-564.

428. Kanguane P, Sakharkar MK, Lim KS, Hao H, Lin K, Chee RE, Kolatkar PR: **Knowledge-based grouping of modeled HLA peptide complexes.** *Hum Immunol* 2000, **61**(5):460-466.
429. Betancourt MR, Thirumalai D: **Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes.** *Protein Sci* 1999, **8**(2):361-369.
430. Zhao B, Mathura VS, Rajaseger G, Mochhala S, Sakharkar MK, Kanguane P: **A novel MHCp binding prediction model.** *Hum Immunol* 2003, **64**(12):1123-1143.
431. Singh SP, Mishra BN: **Ranking of binding and nonbinding peptides to MHC class I molecules using inverse folding approach: implications for vaccine design.** *Bioinformation* 2008, **3**(2):72-82.
432. Mohanapriya A, Lulu S, Kayathri R, Kanguane P: **Class II HLA-peptide binding prediction using structural principles.** *Hum Immunol* 2009, **70**(3):159-169.
433. Bordner AJ, Abagyan R: **Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes.** *Proteins* 2006, **63**(3):512-526.
434. Caflisch A, Niederer P, Anliker M: **Monte Carlo docking of oligopeptides to proteins.** *Proteins* 1992, **13**(3):223-230.
435. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: **Computing the structure of bound peptides. Application to antigen recognition by class I major histocompatibility complex receptors.** *J Mol Biol* 1993, **234**(3):515-521.
436. Rosenfeld R, Zheng Q, Vajda S, DeLisi C: **Flexible docking of peptides to class I major-histocompatibility-complex receptors.** *Genet Anal* 1995, **12**(1):1-21.
437. Desmet J, De Maeyer M, Spriet J, Lasters I: **Flexible docking of peptide ligands to proteins.** *Methods Mol Biol* 2000, **143**:359-376.
438. Fagerberg T, Cerottini JC, Michielin O: **Structural prediction of peptides bound to MHC class I.** *J Mol Biol* 2006, **356**(2):521-546.
439. Rognan D, Scapozza L, Folkers G, Daser A: **Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes.** *Biochemistry* 1994, **33**(38):11476-11485.
440. Lim JS, Kim S, Lee HG, Lee KY, Kwon TJ, Kim K: **Selection of peptides that bind to the HLA-A2.1 molecule by molecular modelling.** *Mol Immunol* 1996, **33**(2):221-230.
441. Sieker F, Springer S, Zacharias M: **Comparative molecular dynamics analysis of tapasin-dependent and -independent MHC class I alleles.** *Protein Science* 2007, **16**(2):299-308.

442. Wei HY, Tsai KC, Lin TH: **Modeling ligand-receptor interaction for some MHC class II HLA-DR4 peptide mimetic inhibitors using several molecular docking and 3D QSAR techniques.** *J Chem Inf Model* 2005, **45**(5):1343-1351.
443. Liu Z, Dominy BN, Shakhnovich EI: **Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential.** *J Am Chem Soc* 2004, **126**(27):8515-8528.
444. Khan JM, Ranganathan S: **A multi-species comparative structural bioinformatics analysis of inherited mutations in alpha-D-mannosidase reveals strong genotype-phenotype correlation.** *BMC Genomics* 2009, **10** Suppl 3:S33.
445. Wallace AC, Laskowski RA, Thornton JM: **LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions.** *Protein Eng* 1995, **8**(2):127-134.
446. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14**(6):1188-1190.
447. Klein L, Klugmann M, Nave KA, Tuohy VK, Kyewski B: **Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells.** *Nat Med* 2000, **6**(1):56-61.
448. Neeno T, Krco CJ, Harders J, Baisch J, Cheng S, David CS: **HLA-DQ8 transgenic mice lacking endogenous class II molecules respond to house dust allergens: identification of antigenic epitopes.** *J Immunol* 1996, **156**(9):3191-3195.
449. Krco CJ, Harders J, Chapoval S, David CS: **Immune response of HLA-DQ transgenic mice to house dust mite allergen p2: identification of HLA-DQ restricted minimal epitopes and critical residues.** *Clin Immunol* 2000, **97**(2):154-161.
450. Robles DT, Fain PR, Gottlieb PA, Eisenbarth GS: **The genetics of autoimmune polyendocrine syndrome type II.** *Endocrinol Metab Clin North Am* 2002, **31**(2):353-368, vi-vii.
451. Mangalam A, Luckey D, Basal E, Jackson M, Smart M, Rodriguez M, David C: **HLA-DQ8 (DQB1*0302)-restricted Th17 cells exacerbate experimental autoimmune encephalomyelitis in HLA-DR3-transgenic mice.** *J Immunol* 2009, **182**(8):5131-5139.
452. Godkin A, Friede T, Davenport M, Stevanovic S, Willis A, Jewell D, Hill A, Rammensee HG: **Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2).** *Int Immunol* 1997, **9**(6):905-911.

453. Godkin AJ, Davenport MP, Willis A, Jewell DP, Hill AV: **Use of complete eluted peptide sequence data from HLA-DR and -DQ molecules to predict T cell epitopes, and the influence of the nonbinding terminal regions of ligands in epitope selection.** *J Immunol* 1998, **161**(2):850-858.
454. Harfouch-Hammoud E, Walk T, Otto H, Jung G, Bach JF, van Endert PM, Caillat-Zucman S: **Identification of peptides from autoantigens GAD65 and IA-2 that bind to HLA class II molecules predisposing to or protecting from type 1 diabetes.** *Diabetes* 1999, **48**(10):1937-1947.
455. Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B: **Peptide binding predictions for HLA DR, DP and DQ molecules.** *BMC Bioinformatics* 2010, **11**:568.