BMC Bioinformatics

BioMed Central

Research

Open Access

An automated framework for understanding structural variations in the binding grooves of MHC class II molecules

Kalidas Yeturu[1], Tapani Utriainen[2], Graham JL Kemp*[2] and Nagasuma Chandra*[1]

Addresses: [1]Bioinformatics Centre, Indian Institute of Science, Bangalore, India and [2]Computer Science and Engineering, Chalmers University of Technology, SE-412 96 Göteborg, Sweden

E-mail: Kalidas Yeturu - kalidas@rishi.serc.iisc.ernet.in; Tapani Utriainen - tapani@chalmers.se; Graham JL Kemp* - kemp@chalmers.se; Nagasuma Chandra* - nchandra@serc.iisc.ernet.in
*Corresponding author

from The Eighth Asia Pacific Bioinformatics Conference (APBC 2010)
Bangalore, India 18-21 January 2010

Published: 18 January 2010

BMC Bioinformatics 2010, **11**(Suppl 1):S55 doi: 10.1186/1471-2105-11-S1-S55

This article is available from: http://www.biomedcentral.com/1471-2105/11/S1/S55

Abstract

Background: MHC/HLA class II molecules are important components of the immune system and play a critical role in processes such as phagocytosis. Understanding peptide recognition properties of the hundreds of MHC class II alleles is essential to appreciate determinants of antigenicity and ultimately to predict epitopes. While there are several methods for epitope prediction, each differing in their success rates, there are no reports so far in the literature to systematically characterize the binding sites at the structural level and infer recognition profiles from them.

Results: Here we report a new approach to compare the binding sites of MHC class II molecules using their three dimensional structures. We use a specifically tuned version of our recent algorithm, PocketMatch. We show that our methodology is useful for classification of MHC class II molecules based on similarities or differences among their binding sites. A new module has been used to define binding sites in MHC molecules. Comparison of binding sites of 103 MHC molecules, both at the whole groove and individual sub-pocket levels has been carried out, and their clustering patterns analyzed. While clusters largely agree with serotypic classification, deviations from it and several new insights are obtained from our study. We also present how differences in sub-pockets of molecules associated with a pair of autoimmune diseases, narcolepsy and rheumatoid arthritis, were captured by $PocketMatch_{13}$.

Conclusion: The systematic framework for understanding structural variations in MHC class II molecules enables large scale comparison of binding grooves and sub-pockets, which is likely to have direct implications towards predicting epitopes and understanding peptide binding preferences.

Background

Major histocompatibility complex (MHC) class II molecules are important components of the immune system and play a critical role in processes such as phagocytosis. Antigenic peptide binding by these molecules is a prerequisite for triggering immune responses. The diversity in antigen recognition is achieved through hundreds of class II alleles labelled by their serotypes, each differing from the others in terms of the residues at the binding site and their precise three dimensional arrangement.

The nature of binding site of an MHC class II molecule (Figure 1) has an important bearing on the immune system of an individual [1,2]. MHC class II molecules provide important clues in understanding autoimmune diseases (e.g. [3-5]) and susceptibility to pathogens. In the context of tuberculosis, it has been reported that different MHC alleles bind peptides from *Mycobacterium tuberculosis* with different specificities, influencing an individual's susceptibility to infection [6-8].

A thorough knowledge of the structure of the binding site is useful in designing or identifying peptide antigens for rational vaccine design. In addition, knowledge of similar or dissimilar sites aid in understanding peptide specificities. While a general appreciation of the differences between a pair of structures can be obtained through interactive molecular graphics software tools, a thorough characterization of the differences and their mapping to individual residues in the corresponding structures, and more importantly obtaining a quantitative perspective of the extent of similarities, necessarily requires a systematic method for their analysis.

We have recently reported a new algorithm *PocketMatch* [9] based on alignment of sorted distance elements binned into point-type-pair bins. An important step that precedes pocket comparison is the definition of the binding site itself. In the previous study, all residues (or any atoms in them) that were present in a 4 Å zone around any atom of the ligand were taken to constitute the site. This approach though common, is rather simplistic and more detailed methods to define the binding site need to be explored to have more accurate site definitions. Here we incorporate a new module for defining binding sites and apply it for a large scale comparison of binding sites in the MHC class II molecules.

The modified algorithm is referred to as $PocketMatch_{13}$ hereafter. Further, we show that our algorithm is useful for classification of MHC class II molecules based on binding site analysis. The algorithm captures the overall shape, detailed geometry and the chemistry at the binding sites. This analysis also aids in understanding peptide preferences by different alleles which may become the first step in the optimal design of allele specific antigens.

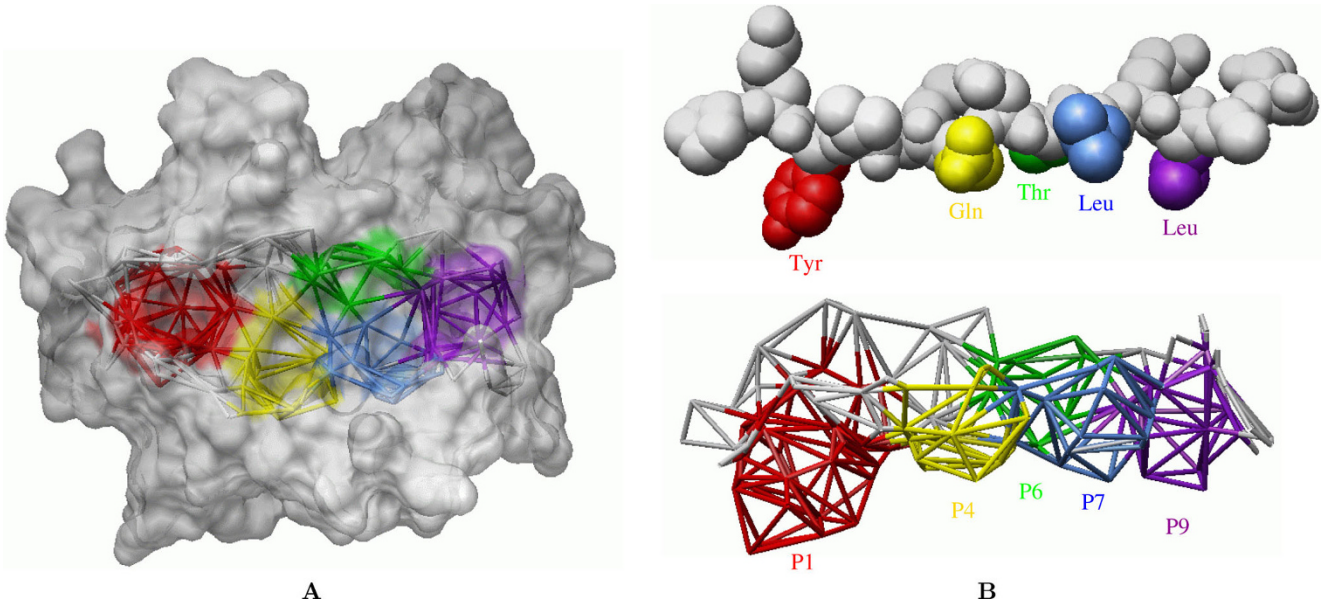

Figure 1

Structure of an MHC class II binding groove. **(A)** Binding domain of HLA-DR1 [PDB:1DLH], with the five pockets in the binding groove highlighted (P1 - red; P4 - yellow; P6 - green; P7 - blue; P9 - purple). Lines are drawn between the centres of binding site atoms that can be touched simultaneously by a probe sphere. **(B)** The influenza virus peptide from [PDB:1DLH] is shown above the binding groove, with peptide side chains shown in the same colour as the pockets into which they fit.

Results and Discussion

We report a new approach for a large scale comparison of binding sites in protein structures and apply it for comparing and classifying a set of 103 MHC class II molecules. The method, which utilizes structural features of the whole site as well as of the sub-pockets, also serves as a high resolution framework to systematically understand similarities and differences among alleles. We have used this to identify automatically intra- and inter-allelic variations in the binding grooves of molecules in the data set, and to explore the structural basis for correlations with disease.

Inter-allelic variations

To investigate similarities across MHC molecules of different types, one MHC molecule was selected from each of the 65 Protein Data Bank (PDB) entries in the dataset, and all-against-all comparisons were carried out on this set of 65 molecules (Table 1). Binding site similarity scores ($PM_{13}Scores$) were computed for all the pairs of molecules both at the level of whole groove and sub-pocket levels. Cladograms were generated to show similarities and differences in $PM_{13}Scores$ across the dataset, both at the level of the whole groove, and at the level of the five sub-pockets (Figures 2 and Figures S1-S4 in Additional file 1). In addition to considering whole binding groove, it is important to know how the similarities of the sub-pockets (P1, P4, P6, P7, P9) vary as these are the ones that determine peptide specificity.

Some MHC molecules of the same type are in different branches of the cladogram calculated for the whole groove, however clustering at the sub-pocket level was more in line with the different MHC molecule types, particularly for the P4 sub-pocket. This suggests that the P4 sub-pocket is more structurally conserved within an allele, but difference occurs across alleles. The importance of the P4 sub-pocket has been noted in many studies (e.g. [1,2,10]).

Some different MHC molecules are grouped together in the same branch in some of the sub-pocket trees. In these cases, the $PM_{13}Scores$ highlight similarities that would otherwise be difficult to spot in a large dataset. These can be followed up by looking for independent observations about these similarities that have been reported in the literature. The matching alleles, corresponding PDB codes and $PM_{13}Scores$ for pairs of sub-pockets are listed in (Table 2), where the significance of the grouping of different alleles is discussed and supporting references are presented.

Table 1: Dataset used in this study. 103 MHC class II molecules from 65 PDB files were used. #Mol – Number of molecules; #PDB – Number of PDB entries

Alleles	Supertype [21]	#PDB	PDB Identifiers	#Mol
DQA1*0102-DQB1*0602	DQ1	1	1UVQ	1
DQA1*0301/2/3-DQB1*0302	DQ8	2	1JK8, 2NNA	2
DQA1*0501/3/5/6/7/8/9-DQB1*0201/2	DQ2	1	1S9V	2
DRA*0101/2-DRB1*0101	DR1	22	1AQD, 1DLH, 1FYT, 1HXY, 1JWM, 1JWS, 1JWU, 1KG0, 1KLG, 1KLU, 1LO5, 1PYW, 1SEB, 1SJE, 1SJH, 1T5W, 1T5X, 2FSE, 2G9H, 2IAM, 2IAN, 2IPK	32
DRA*0101/2-DRB1*0301	DR3	1	1A6A	1
DRA*0101/2-DRB1*0401	DR4	2	1J8H, 2SEB	2
DRA*0101/2-DRB1*1501	DR2	3	1BX2, 1YMM, 2WBJ	5
DRA*0101/2-DRB3*0101	DR52	1	2Q6W	2
DRA*0101/2-DRB3*0301	DR52	1	3C5J	1
DRA*0101/2-DRB5*0101	DR51	4	1FV1, 1H15, 1HQR, 1ZGL	8
H2-Aa	–	1	1ES0	1
H2-Aa, H2-Ab1	–	17	1D9K, 1F3J, 1IAK, 1IAO, 1JL4, 1K2D, 1LNU, 1MUJ, 1U3H, 2IAD, 2P24, 2PXY, 2Z31, 3C5Z, 3C60, 3C6L, 3CUP	26
H2-Ea	–	2	1IEB, 1KTD	4
H2-Ea, H2-Eb1	–	7	1FNE, 1FNG, 1I3R, 1IEA, 1KT2, 1R5V, 1R5W	16

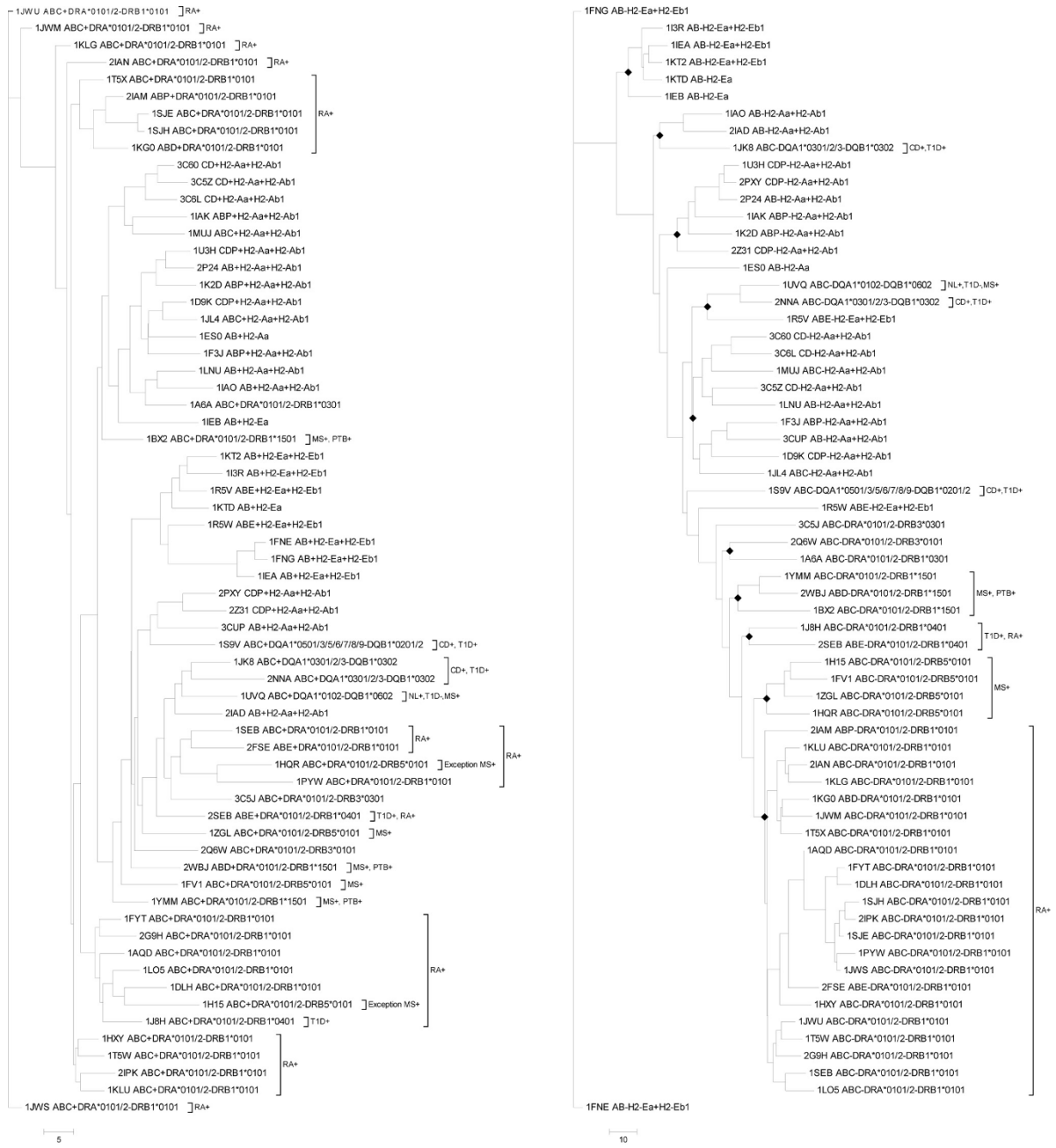

Figure 2

Cladograms based on similarities between binding sites. The cladograms for the whole groove and the P4 sub-pocket level similarities among alleles. The first four characters of the label are the PDB identifier for the structure of the MHC class II molecule. The next group of three letters are the chain identifiers used for the alpha chain, beta chain and peptide, respectively, in the original PDB file (there are only two letters in this group in cases where the peptide has been engineered to be part of one of the MHC chains). The final part of the label indicates which alpha and beta alleles are present in the MHC molecule. Branches associated with diseases are shown in *brackets* with disease label. Disease names are abbreviated as NL: Narcolepsy, T1D:Type 1 Diabetes, RA:Rheumatoid Arthritis, CD: Coeliac Disease and PTB:Pulmanory tuberculosis. The suffix '+' stands for positive association and '-' for negative association of an allele with disease. **(A)** The cladogram for the whole groove similarities. **(B)** The cladogram for the P4 similarities. Different branches are indicated by *black diamonds* to indicate net clustering.

Table 2: Summary of the analysis performed on cladograms for the whole groove and five sub-pockets. For a pair of different alleles, the pair of molecules obtaining high PM_{13} *Score* on whole groove or sub-pocket comparison is presented. Literature citation for other independent work supporting the observation is also provided wherever possible. The examples shown here refer to different alleles that appear in the same branch of the clustergarm computed either by using the whole grooves or by their individual sub-pockets as indicated in each row

Pair of alleles	(PDB1, PDB2)	Pocket	PM_{13} *Score*	Comment
DQB1*0602-DQB1*0302	(1UVQ, 1JK8)	whole	0.83	The involvement of the two alleles DQB1*0602 and DQB1*0302, negatively and positively associated with Type 1 diabetes is reported by Siebold and co-workers [22]
	(1UVQ, 2NNA)		0.75	
	(1UVQ, 2NNA)	P4	0.74	
	(1UVQ, 1JK8)	P7	0.71	
DQB1*0201, 2-DRB1*1501	(1S9V, 1BX2)	P7	0.4	
DQB1*0201, 2-DRB1*1501	(1S9V, 1BX2)	P9	0.08	
DRB1*0401-DRB1*0101	(2SEB, 2FSE)	P9	0.61	
	(1J8H, 2IPK)		0.85	The study by Rosloniec and co-workers [13] indicate the association of the two alleles are known to be associated with RA.
DRB1*1501-DRB1*0101	(2WBJ, 1LO5)	P9	0.69	These observations agree with the study by Smith and co-workers [23] that reports the similarity of the P9 sub-pocket. Study by Drouin and co-workers [24] refer to the association of the two alleles with antibiotic-refractory arthritis.
	(1YMM, 1KLU)		0.59	
	(1YMM, 2IAM)		0.66	
DRB1*1501-DRB1*0301	(1BX2, 1A6A)	whole	0.80	Both alleles came in two branches under a common root which is in accordance with a study by Zivadinov and co-workers [25] that associates the two alleles to Multiple sclerosis.
DRB3*0101-DRB1*0101	(2Q6W, 1HXY)	P1	0.66	
DRB3*0101-DRB1*0301	(2Q6W, 1A6A)	P4	0.54	The two molecules are grouped together. Though the score is only 0.54, there are no other molecules they could come similar to with matching allele types. The study by Parry and co-workers [26] indicate the expected similarity in the P4 sub-pocket and correlate the differences in other subpockets and the differences in P4 itself to difference between the two alleles in susceptibility to Type 1 diabetes.
DRB3*0301-DRB1*0401	(3C5J, 1J8H)	P6	0.43	
	(3C5J, 2SEB)		0.37	
	(3C5J, 2SEB)	P7	0.77	
DRB5*0101-DRB1*0101	(1HQR, 1PYW)	whole	0.78	
	(1HQR, 2G9H)	P1	0.9	Meinl and co-workers [27] also report similarity between the two allele types in recognition of myelin basic protein. The P1 similarity between the two alleles is reported by Jurcevic and co-workers [28].
	(1H15, 1DLH)	whole	0.8	
	(1H15, 1LO5)	P1	0.73	
	(1ZGL, 1JWM)	P1	0.8	
DRB5*0101-DRB1*0301	(1FV1, 1A6A)	P1	0.81	Though P1 score is high, the other subpocket scores are low (less than 0.34) which is in accordance with study by Texier and co-workers [29] that reports difference between the two alleles in their peptide binding properties.
		P9	0.56	
	(1FV1, 1R5W)	P9	0.6	
	(2Q6W, 3C5J)	P9	0.84	This similarity of P9 is a known feature [30] for the two, DR52a and DR52c alleles which are encoded by the DR3 gene whose alleles are all associated with autoimmune diseases.
DQB1*0602-H2-Aa, H2-Ab1	(1UVQ, 1IAK)	P6	0.64	Orthologous alleles from human and mouse [31].
	(1UVQ, 1JL4)	P9	0.71	
	(1UVQ, 2IAD)	P9	0.75	

Table 2: Summary of the analysis performed on cladograms for the whole groove and five sub-pockets. For a pair of different alleles, the pair of molecules obtaining high PM_{13} *Score* on whole groove or sub-pocket comparison is presented. Literature citation for other independent work supporting the observation is also provided wherever possible. The examples shown here refer to different alleles that appear in the same branch of the clustergarm computed either by using the whole grooves or by their individual sub-pockets as indicated in each row *(Continued)*

DQB1*0201-H2-Aa, H2-Ab1	(1S9V, 2PXY)	whole	0.73	
	(1S9V, 2Z31)		0.79	
	(1S9V, 3CUP)		0.78	
	(1S9V, 1MUJ)	P1	0.57	
DRB1*0101-H2-Aa, H2-Ab1	(1AQD, 1K2D)	P1	0.52	
	(1D9K, 1SJE)	P9	0.46	
	(1D9K, 1SJH)	P9	0.53	
DRB1*0301-H2-Aa, H2-Ab1	(1A6A, 1LNU)	whole	0.83	
	(1A6A, 1IAO)		0.83	
DRB1*1501-H2-Aa, H2-Ab1	(2WBJ, 1IAO)	P7	0.42	
	(2WBJ, 1LNU)		0.57	
DRB1*1501-H2-Ea, H2-Eb1	(1R5V, 1BX2)	P6	0.65	Orthologous alleles from human and mouse.

To analyze the net distribution of similarity scores with respect to each other for each of the five sub-pockets, a histogram is plotted for various bins of $PM_{13}Scores$ (Figure 3). Each bin corresponds to a range of $PM_{13}Scores$. For example, bin-5 corresponds to a $PM_{13}Score$ range of [0.5 to 0.6); bin-7 to the range [0.7 to 0.8) and so on. The histogram shows that P1 and P9 score highly at bin 6, corresponding to [0.6 to 0.7) of $PM_{13}Score$. The histogram gives an indication of the overall distribution of scores for each sub-pocket viewed in the context of others. This could possibly mean over-representation of data or true conservation of these two sub-pockets.

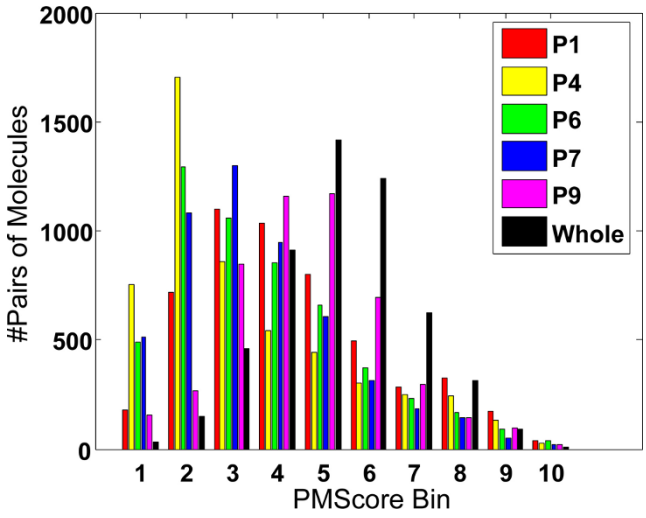

Figure 3
Frequencies of $PM_{13}Scores$ (labelled PMScore) for the five sub-pockets and whole binding groove. X-axis: 10 bins of $PM_{13}Scores$ ranging from 0 to 1.0. Red, yellow, green, blue and purple bars correspond to the P1, P4, P6, P7 and P9 sub-pockets. Black bars correspond to whole groove similarity scores.

This analysis has implications for understanding subtle differences that otherwise go undetected and aid in understanding antigen recognition preferences by different alleles and range of antigens recognized by a given allele.

Intra-allelic variations

Some MHC molecules are present more than once in the PDB entries in the dataset (Table 1). In these cases, $PocketMatch_{13}$ can be used to highlight differences in the peptide binding sites in different structures for the same allele.

The sites are first compared by considering the whole binding grooves. In many cases, as expected, $PM_{13}Scores$ are high, indicating strong similarities in the binding sites of a given allele. However, there are cases where $PM_{13}Scores$ are low for different structures of the same molecule, for example different structures of DR1 and DR5 give similarity scores as low as 0.44 (Table S1 in Additional file 1). These differences can be explored by examining the individual sub-pockets within the binding grooves (see Methods). While many pairs of corresponding sub-pockets score highly, indicating similarity in the structures of the sub-pockets, in some cases the scores are significantly lower. This can be due to differences in MHC side chain conformations giving rise to different sets of intra-site distances, or can be due to determination of which MHC atoms are accessible to a probe sphere and are thus included in sub-pocket calculations. Sub-pockets highlighted by $PocketMatch_{13}$ to be

dissimilar can then be examined in detail to identify the reason for the low PM_{13}*Scores*. Some examples of sub-pockets with low PM_{13} *Scores* are illustrated in Figure 4.

A pair of molecules belonging to DR1 exhibited low scores [PDB:1AQD, PDB:1DLH] in their P1 sub-pockets. Upon careful examination, we noticed that the P1 sub-pocket in 1DLH was wider and deeper with many more MHC atoms being included in the *PocketMatch*$_{13}$ definition of the P1 sub-pocket. Considering the set of DRA*0101-DRB1*1501 structures, the largest difference is between the P7 pockets of [PDB:1BX2] and [PDB:2WBJ] (Figure 4A). The peptide residue at the P7 position is oriented very differently in these two structures – in [PDB:1BX2], an isoleucine is oriented away from the groove, whereas in [PDB:2WBJ] a leucine is oriented "across" the top of the groove. Since the P7 peptide residue in [PDB:2WBJ] obstructs the P7 sub-pocket more than the P7 peptide residue in [PDB:1BX2], this affects the set of MHC atoms that are selected for the sub-pocket comparison calculation, and thus reduces the PM_{13}*Score* (0.06).

The two independent molecules in the crystal structure of DQ8 [PDB:1S9V] differ from each other at the P9 sub-pocket (Figure 4B); the difference between the two molecules at the P9 position is noted by [11]. This analysis indicates that *PocketMatch*$_{13}$ is sufficiently sensitive to capture subtle differences that exist among molecules belonging to the same allele.

Correlation with disease: case studies

Several MHC class II alleles are known to be either positively or negatively associated with certain diseases, and this motivates studies to identify the reasons for disease susceptibility in terms of three-dimensional molecular structure [1]. For example, Jones *et al.* [1] review the structures of alleles that are known to be positively or negatively associated with various diseases, including narcolepsy and rheumatoid arthritis (RA). We have used *PocketMatch*$_{13}$ to examine the binding grooves of alleles discussed by Jones *et al.* [1] in connection with narcolepsy and RA, using experimentally determined structures from the PDB where these are available, and model structures when they are not (see Methods). In case of Narcolepsy, the pockets of the binding groove in the experimentally determined structure of HLA-DQ6.2 (positively associated with the disease) [PDB:1UVQ],

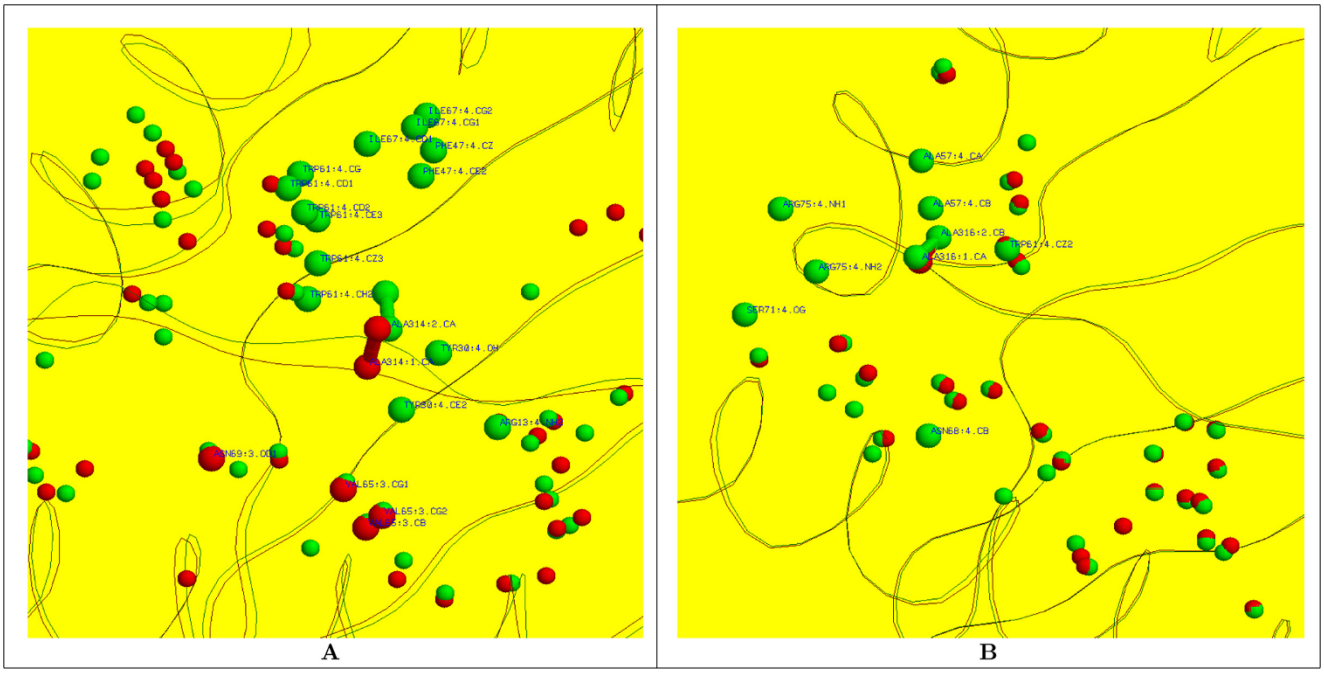

Figure 4
Examples of sub-pockets with low PM_{13} *Scores*. (A) Detail of the superposed binding grooves of [PDB:1BX2] (red) and [PDB:2WBJ] (green). MHC and peptide main chains are represented by a cartoon trace. Spheres indicate the centres of MHC atoms that are determined to be part of the binding groove (see Methods). The centres of MHC atoms that are determined to be part of the P7 sub-pocket are represented by the larger spheres; these atoms are labelled in blue. The Cα and Cβ atoms of the peptide residues at the P7 positions, remodelled as alanines, are shown with a ball-and-stick representation. **(B)** Similar to (A), but focusing on the P9 sub-pockets in [PDB:1S9V] (chains A, B, C in red; chains D, E, F in green).

were compared to those in a model structure of HLA-DQ6.1 (negatively associated with the disease). These molecules differ at only a few positions in the β chain. *PocketMatch*$_{13}$ identified the P4 sub-pocket corresponding to the Thr6 residue of the peptide to be the most dissimilar between these two structures (Table 3). The residues Ala13bβ and Tyr26β in HLA-DQ6.2 changed to Gly13β and Leu26β in HLA-DQ6.1 in the neighbourhood of peptide residue Thr6, corresponding to P4 (Figure 5A); this difference is captured by the *PocketMatch*$_{13}$ algorithm.

Table 3: Similarity scores between sub-pockets of HLA-DQ6.1 and HLA-DQ6.2. PMSMax and PMSMin are defined in Methods

Peptide residue	Pocket	PMSMin	PMSMax
Leu3	P1	1.0	1.0
Thr6	P4	0.57	0.84
Val8	P6	0.90	0.90
Ser9	P7	1.0	1.0
Ala11	P9	0.92	0.92

In case of RA, alleles HLA-DR4.1, HLA-DR4.4 and HLA-DR1 are positively associated with the disease, while HLA-DR4.2 is neutral or negative [1]. The α chains of these four MHC molecules are the same (DRA*0101), and sequence comparison of the β chains with ClustalW [12] gives sequence identities of – DR4.1:DR4.2 = 95%, DR4.1:DR4.4 = 97%, DR4.1:DR1 = 88%, DR4.2:DR4.4 = 96%, DR4.2:DR1 = 85%, DR4.4:DR1 = 88%. Given that the whole sequence similarities are not sensitive enough to capture differences at the binding site levels, we use *PocketMatch*$_{13}$ to compare the binding grooves and sub-pockets of the experimentally determined structures of HLA-DR4.1 [PDB:1J8H] and HLA-DR1 [PDB:1DLH], and model structures of HLA-DR4.2 and HLA-DR4.4.

PocketMatch$_{13}$ gives low scores for the P4 sub-pocket (Table 4A). It has been shown by Hammer and co-workers [10] that the difference in residues 70 and 71 in the β chain of the DR4.1 and DR4.2 MHCs accounts for the difference in binding specificity of the peptides.

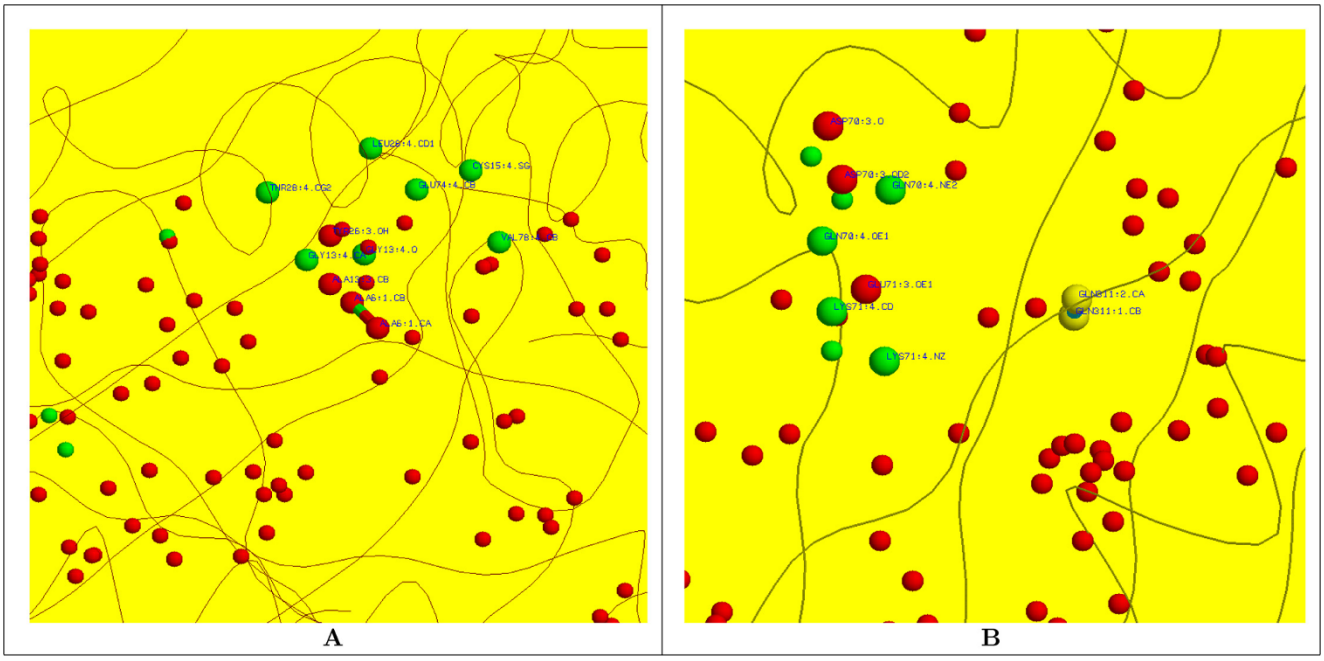

Figure 5
Detail of the binding grooves of HLA-DQ6.2 and HLA-DR4.2. **(A)** Detail of the binding groove of HLA-DQ6.2 (green). MHC and peptide main chains are represented by a cartoon trace. Spheres indicate the centres of MHC atoms that are determined to be part of the binding groove (see Methods). The centres of MHC atoms that are determined to be part of the P4 sub-pocket are represented by the larger spheres; these atoms are labelled in blue. The Cα and Cβ atoms of the peptide residue at the P7 position, remodelled as alanine, is shown with a ball-and-stick representation. Atoms that differ in binding groove of the model structure of HLA-DQ6.1 are shown in red. **(B)** Detail of the binding groove of HLA-DR4.2 (red). MHC and peptide main chains are represented by a cartoon trace. Spheres indicate the centres of MHC atoms that are determined to be part of the binding groove (see Methods). Atoms that differ in binding groove of the model structure of HLA-DR4.1 are shown in green. The centres of atoms of residues 70β and 71β are represented by the larger spheres; these atoms are labelled in blue. The Cα and Cβ atoms of the glutamine residue at the P4 position in the peptide are shown with a ball-and-stick representation (yellow).

Table 4: Sub-pocket similarities. (A) Sub-pocket similarities between a pair of alleles HLA-DR4.1 [PDB:1J8H] and HLA-DR4.2 (model) are shown. The residues of the peptide are shown on the left most column. Residue numbers 311 and 314 correspond to P4 and P7 respectively. The low $PM_{13}Scores$ are shown in boldface. (B) Variation of the P4 similarity scores among HLA-DRB1*0101 [PDB:1DLH], HLA-DR4.1 [PDB:1J8H], HLA-DR4.4 and HLA-DR4.2 are shown. (C) Variation in P7 similarities are shown for proteins mentioned under (B)

(A) - HLA-DR4.1(1J8H-abc) and HLA-DR4.2(model)

Pocket	Residue	PMSMin	PMSMax
P1	Tyr308	0.85	0.91
P4	Gln311	0.79	0.85
P6	Thr313	0.89	0.89
P7	Leu314	0.53	0.71
P9	Leu316	1.0	1.0

(B) - P4 (Gln311)-similarity

PDB/MODEL	PDB/MODEL	PMSMin	PMSMax
1DLH	1J8H	0.56	0.60
1DLH	HLA-DR4.2	0.53	0.53
1DLH	HLA-DR4.4	0.63	0.68
1J8H	HLA-DR4.2	0.79	0.85
1J8H	HLA-DR4.4	0.86	0.86
HLA-DR4.2	HLA-DR4.4	0.74	0.79

(C) - P7 (Leu314)-similarity

PDB/MODEL	PDB/MODEL	PMSMin	PMSMax
1DLH	1J8H	0.44	0.58
1DLH	HLA-DR4.2	0.50	0.50
1DLH	HLA-DR4.4	0.67	0.67
1J8H	HLA-DR4.2	0.53	0.71
1J8H	HLA-DR4.4	0.60	0.81
HLA-DR4.2	HLA-DR4.4	0.57	0.57

The low P4 scores are in line with that study. The superposition of these two alleles is shown in Figure 5B. The P4 peptide residue has Gln70β and Lys71β present in HLA-DR4.1 within 3.0 Å of the residue whereas an Asp at the position 70β and only Glu71β are present in the case of the model built for HLA-DR4.2.

All-against-all $PM_{13}Scores$ are presented in Table 4B, C. The scores indicate low $PM_{13}Score$ of [PDB:1DLH] to others in the P7 region of the binding site. Work by Rosloniec and co-workers found that mutation of the residue at the P7 position to an alanine has affected T cell stimulation more with DR4 than with DR1 [13]. The involvement of P7 sub-pocket in peptide recognition specificity is also discussed in [10]. In carrying out these case studies, model structures have been a useful supplement to the set of experimentally determined MHC class II molecules. We envisage future studies that make use of larger sets of model structures where the binding grooves have been modelled consistently using the same protocol [14].

Conclusion

A strategy for automatically comparing MHC class II binding grooves and sub-pockets based on their chemical nature and geometry is presented. Comparisons are facilitated by a pre-processing step in which MHC-peptide complexes are extracted from PDB files, and chains and structurally equivalent residue positions are relabelled consistently. Pocket similarity scores calculated by $PocketMatch_{13}$ can be used as the basis for clustering pockets based on their structural and chemical characteristics.

The framework we report can be used to carry out large scale comparison of binding grooves and sub-pockets, both to highlight differences in the binding grooves of MHC molecules of the same kind, and to identify similarities in the binding grooves of different MHC alleles. Investigations of MHC alleles associated with narcolepsy and rheumatoid arthritis demonstrate that binding grooves of alleles that are positively associated with an autoimmune disease can be compared with

those that are known to be negatively associated with the disease. The structural variations among binding pockets identified by *PocketMatch*$_{13}$ corroborate known disease associations. Future applications of this systematic framework for understanding structural variations in MHC class II molecules could have direct implications towards predicting epitopes and understanding peptide binding preferences.

Methods

Dataset preparation

103 MHC class II molecules from 65 Protein Data Bank [15] entries are used in this study (Table 1), and the sequences of the α_1 and β_1 domains from these structures were matched with allele sequences from IMGT/HLA database [16] to confirm which allele is present in the PDB entry. In this study, the focus is on MHC class II binding domains. In some cases, different alleles share identical sequences for the binding region, e.g. human alpha chains DRA*0101 and DRA*0102 have binding domains with identical sequences, so both of these alleles are listed alongside structures with this alpha chain sequence in Table 1. Similarly, many alleles have binding domains with sequences that are identical to those in [PDB:1S9V], and these are listed in Table 1.

To facilitate automatic comparison of MHC class II structures, uniform chain identifiers and residue numbers were used for all MHC-peptide complexes extracted from the PDB files. New files were written where each file contains the core parts of an α_1 domain, a β_1 and a peptide, with chains relabelled to match the chain identifiers A, B and C in [PDB:1DLH], and residues renumbered to match the numbering of residues at structurally equivalent positions in [PDB:1DLH]. Positions 5-78 of the α_1 domain and positions 5-91 of the β_1 domain were retained. A rigid body transformation was applied to superpose the the MHC binding domain complexes onto chains A and B of [PDB:1DLH<http://www.rcsb.org/pdb/cgi/explore.cgi?pdbId=1DLH>], so that all complexes are in the same frame of reference. This transformation is not necessary for the automatic comparisons that follow, but it is convenient for comparing structures using molecular graphics to review results from the automatic comparisons. Peptide residues corresponding to the 13 peptide residues in [PDB:1DLH] were identified by structural comparison, and peptide residues beyond the 13-residue peptide present in [PDB:1DLH] were removed automatically.

Comparative modelling

To enable the comparison of binding grooves of MHC class II molecules known to be positively or negatively associated with narcolepsy or RA, models of HLA-DQ6.1 consisting of alleles (HLA-DQA1*0102 and HLA-DQB1*0601), HLA-DRB4.2 (alleles HLA-DRA1*0101 and HLA-DRB1*0402) and HLA-DRB4.4 (alleles HLA-DRA1*0101 and HLA-DRB1*0404) were built interactively using the Swiss-PdbViewer [17]. [PDB:1UVQ] was used as the template structure for the model of HLA-DQ6.1 and [PDB:1J8H] was used as the template for HLA-DRB4.2 and HLA-DRB4.4.

Binding site comparison

Binding sites are represented in a frame invariant manner by distances between pairs of points, partitioned into bins, and pairs of sites are compared based on alignment of sorted sequences of distances. The sorted arrays are then aligned and scored to finally obtain comparison scores.

Molecules can be clustered based on their comparison scores.

In this study, the points used are the centres of those atoms lining the binding site. These are determined by considering accessibility to a probe sphere with radius 1.4 Å. Those MHC atoms whose accessibility is reduced by the presence of the peptide are determined to be part of the peptide binding site. Similarly, the MHC atoms that comprise individual pockets are identified as the set of atoms whose accessibility is reduced by the presence of the peptide residue at position P1, P4, P6, P7 or P9. The ProtOr radii from Table 2 of [18] are used for protein atomic groups in accessibility calculations.

The corresponding pockets between a pair of MHC binding sites are compared on large scale in an all-against-all comparison scheme. The shape signature of each pocket, capturing chemical nature and geometric distribution of atoms, is derived based on the distance lists concept used in *PocketMatch* [9].

Site comparison proceeds as follows:

- Surface atomic groups are classified into 13 types based on heavy-atom types, the number of covalently attached hydrogen atoms and the number of all covalently attached atoms, as proposed by Tsai et al. [18]: C3H0, C3H1, C4H1, C4H2, C4H3, N3H0, N3H1, N3H2, N4H3, O1H0, O2H1, S2H0, S2H1.
- Distances between all pairs of atoms are computed and binned into 13 * (13 - 1)/2 + 13 $\rightarrow$ 91 lists corresponding to each pair of atomic types (C3H0-C3H0, C3H0-C3H1, etc.)
- Each list or bin of distances is then sorted in non-decreasing order. The sorted distance elements binned into various lists according to chemical

nature of the atoms constitutes the shape descriptor of the binding pocket.

- To compare a pair of sites, each of the 91 lists is chosen in one site together with the corresponding list from the other site, and the cumulative number of similar distance elements is determined.
- A pair of distances from two lists is marked a match if the distance differ at most by a threshold of 0.5.

We call the tuned version of *PocketMatch* for the MHC class II binding site comparison, by considering solvent accessible atoms and 13 atomic group types, *PocketMatch*$_{13}$.

The numerator is simply the number of matching intra-site distances. However, the denominator can be the number of intra-site distances in either the smaller site or in the larger site – these give rise to two $PM_{13}Score$ values, referred to as PMSMax and PMSMin, respectively. Unless stated otherwise, $PM_{13}Score$ refers to the PMSMin value.

$PM_{13}Score$ values decrease as the similarity between a pair of binding grooves decreases (Figure 6). The rate at which the scores decrease is affected by the threshold chosen for site comparison, since this affects the number of matching distance elements between a pair of distance-sequences.

To illustrate the effect of perturbing the conformation of a binding groove, the coordinates of atoms in the binding groove of [PDB:1JWS] (A, B, C chains) were perturbed randomly, and an ensemble of 1000 structures was generated with root mean square deviation (RMSD) values up to 5 Å with respect to the original [PDB:1JWS] structure. We have used a similar strategy for sensitivity analysis for the original *PocketMatch* algorithm [9] and found that a threshold of 0.5 Å was adequate to distinguish between similar and dissimilar sites. Figure 6A shows the $PM_{13}Scores$ obtained by comparing the original [PDB:1JWS] structure with each of the perturbed structures in the ensemble. Rather than perturbing the atomic coordinates randomly, an alternative method for generating an ensemble of perturbed conformations would be to use conformations from a molecular dynamics trajectory. To investigate the effect of altering the chemical nature of the binding groove while retaining its original geometry, the atomic group labels of some of the atomic groups in the binding groove of [PDB:1JWS] (A, B, C chains) were re-assigned randomly, and *PocketMatch*$_{13}$ was used to compare the modified binding groove with the original one (Figure 6B). Figures 6A and 6B demonstrate that $PM_{13}Scores$ capture differences due to both the geometry and the chemical nature of the binding groove.

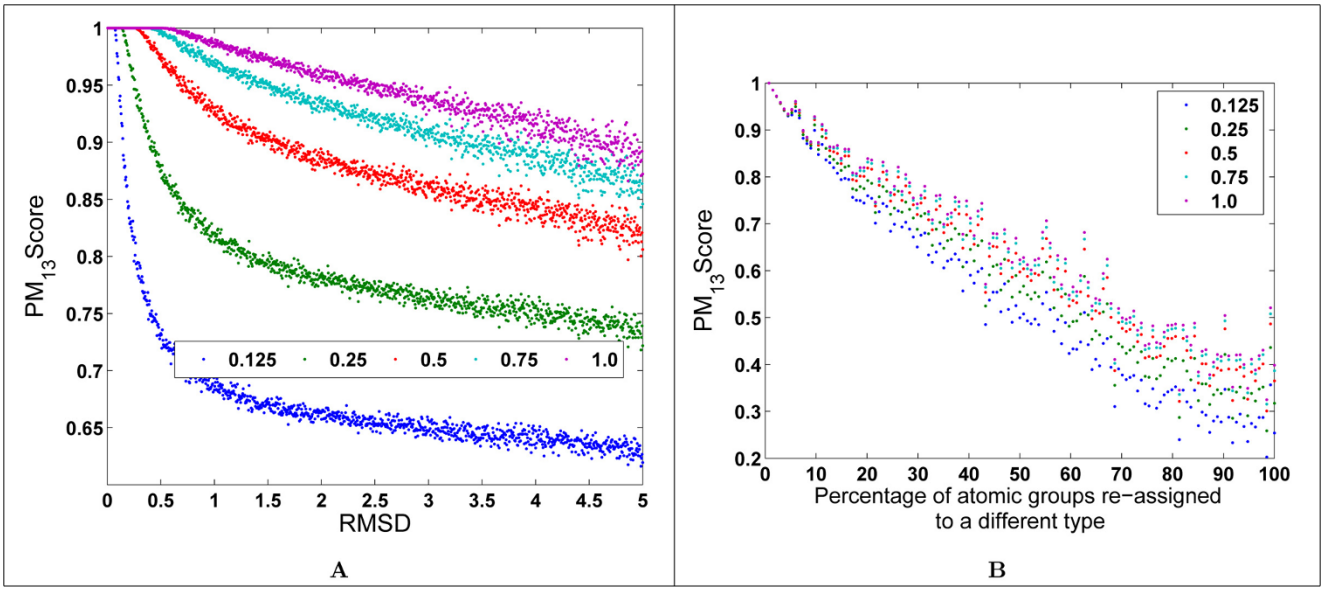

Figure 6

Effect of altering the geometry or chemistry of the binding groove on $PM_{13}Scores$. Altered binding grooves are compared with the original using *PocketMatch*$_{13}$. $PM_{13}Scores$ calculated with different distance element alignment thresholds are shown in different colours (1.0 Å in purple; 0.75 Å in cyan; 0.5 Å in red; 0.25 Å in green; 0.125 Å in blue). (A) The coordinates of atoms in the binding groove of [PDB:1JWS] (A, B, C chains) were perturbed randomly, and an ensemble of 1000 structures was generated with RMSD values up to 5 Å with respect to the original [PDB:1JWS] structure. (B) The atomic group labels of some of the atomic groups in the binding groove of [PDB:1JWS] (A, B, C chains) were re-assigned randomly.

Cladogram generation

Given a set of binding sites (whole groove or sub-pockets), one way of visualizing the relationships among these is to generate a cladogram based on distances between pairs of sites. The distance between a pair of sites is defined here to be 1-$PM_{13}Score$ between the two sites. The cladogram generation program is based on the *neighbour joining* method available in Phylip-3.67 [19] which generates trees in Newick format, which can be visualized and labelled using MEGA [20]. When generating cladograms, data were input to the program in descending order of $PM_{13}Scores$.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KY participated in implementation of the atom type version of PocketMatch, setting up of computational framework for large scale site comparisons and helped to draft the manuscript. TU participated in preparing the data set. GJLK participated in the design and coordination of the study and helped to draft the manuscript. NC participated in reviewing results, manuscript and scientific discussions.

Additional material

Additional File 1

A zip compressed archive with supplementary Figures S1-4 and Table S1.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-11-S1-S55-S1.zip]

Acknowledgements

We are grateful for support from the Kristina Stenborg Foundation. We acknowledge support from the Department of Biotechnology(DBT), Govt. of India. We also acknowledge useful comments received on preliminary results presented at ISMB/ECCB 2009 with support from a travel fellowship to KY from *BioSapiens*.

This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 1, 2010: Selected articles from the Eighth Asia-Pacific Bioinformatics Conference (APBC 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S1.

References

1. Jones EY, Fugger L, Strominger JL and Siebold C: **MHC class II proteins and disease: a structural perspective.** *Nat Rev Immunol* 2006, **6(4):**271–282.
2. Zerva L, Cizman B, Mehra NK, Alahari SK, Murali R, Zmijewski CM, Kamoun M and Monos DS: **Arginine at positions 13 or 70-71 in pocket 4 of HLA-DRB1 alleles is associated with susceptibility to tuberculoid leprosy.** *J Exp Med* 1996, **183(3):**829–836.
3. Li Y, Li H, Martin R and 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.
4. Remus N, Alcaïs A and Abel L: **Human genetics of common mycobacterial infections.** *Immunol Res* 2003, **28(2):**109–129.
5. Vergelli M, Kalbus M, Rojo SC, Hemmer B, Kalbacher H, Tranquill L, Beck H, McFarland HF, De Mars R and Long EO, *et al*: **T cell response to myelin basic protein in the context of the multiple sclerosis-associated HLA-DR15 haplotype: peptide binding, immunodominance and effector functions of T cells.** *J Neuroimmunol* 1997, **77(2):**195–203.
6. Chang ST, Linderman JJ and Kirschner DE: **Effect of multiple genetic polymorphisms on antigen presentation and susceptibility to Mycobacterium tuberculosis infection.** *Infect Immun* 2008, **76(7):**3221–3232.
7. Goldfeld AE, Delgado JC, Thim S, Bozon MV, Uglialoro AM, Turbay D, Cohen C and Yunis EJ: **Association of an HLA-DQ allele with clinical tuberculosis.** *JAMA* 1998, **279(3):**226–228.
8. Terán-Escandón D, Terán-Ortiz L, Camarena-Olvera A, González-Avila G, Vaca-Marín MA, Granados J and Selman M: **Human leukocyte antigen-associated susceptibility to pulmonary tuberculosis: molecular analysis of class II alleles by DNA amplification and oligonucleotide hybridization in Mexican patients.** *Chest* 1999, **115(2):**428–433.
9. Yeturu K and Chandra N: **PocketMatch: a new algorithm to compare binding sites in protein structures.** *BMC Bioinformatics* 2008, **9:**543–543.
10. Hammer J, Gallazzi F, Bono E, Karr RW, Guenot J, Valsasnini P, Nagy ZA and Sinigaglia F: **Peptide binding specificity of HLA-DR4 molecules: correlation with rheumatoid arthritis association.** *J Exp Med* 1995, **181(5):**1847–1855.
11. Kim CY, Quarsten H, Bergseng E, Khosla C and Sollid LM: **Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease.** *Proc Natl Acad Sci USA* 2004, **101(12):**4175–4179.
12. Thompson JD, Higgins DG and Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673–4680.
13. Rosloniec EF, Whittington KB, Zaller DM and Kang AH: **HLA-DR1 (DRB1*0101) and DR4 (DRB1*0401) use the same anchor residues for binding an immunodominant peptide derived from human type II collagen.** *J Immunol* 2002, **168:**253–259.
14. Swain MT, Brooks AJ and Kemp GJL: **An automated approach to modelling class II MHC alleles and predicting peptide binding.** *BIBE '01: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering* Washington, DC, USA: IEEE Computer Society; 2001, 81–88.
15. Berman H, Henrick K and Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nat Struct Biol* 2003, **10(12):**980–980.
16. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P and Marsh SG: **IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex.** *Nucleic Acids Res* 2003, **31:**311–314.
17. Guex N and Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18(15):**2714–2723.
18. Tsai J, Taylor R, Chothia C and Gerstein M: **The packing density in proteins: standard radii and volumes.** *J Mol Biol* 1999, **290:**253–266.
19. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164–166.
20. Tamura K, Dudley J, Nei M and Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24(8):**1596–1599.
21. Doytchinova IA and Flower DR: **In silico identification of supertypes for class II MHCs.** *J Immunol* 2005, **174(11):** 7085–7095.
22. Siebold C, Hansen BE, Wyer JR, Harlos K, Esnouf RE, Svejgaard A, Bell JI, Strominger JL, Jones EY and Fugger L: **Crystal structure of HLA-DQ0602 that protects against type I diabetes and confers strong susceptibility to narcolepsy.** *Proc Natl Acad Sci USA* 2004, **101(7):**1999–2004.
23. Smith KJ, Pyrdol J, Gauthier L, Wiley DC and Wucherpfennig KW: **Crystal structure of HLA-DR2 (DRA*0101, DRB1*1501) complexed with a peptide from human myelin basic protein.** *J Exp Med* 1998, **188(8):**1511–1520.

24. Drouin EE, Glickstein L, Kwok WW, Nepom GT and Steere AC: **Searching for borrelial T cell epitopes associated with antibiotic-refractory Lyme arthritis.** *Mol Immunol* 2008, **45(8):**2323–2332.
25. Zivadinov R, Uxa L, Bratina A, Bosco A, Srinivasaraghavan B, Minagar A, Ukmar M, Benedetto S and Zorzon M: **HLA-DRB1*1501, -DQB1*0301, -DQB1*0302, -DQB1*0602, and -DQB1*0603 alleles are associated with more severe disease outcome on MRI in patients with multiple sclerosis.** *Int Rev Neurobiol* 2007, **79:**521–535.
26. Parry CS, Gorski J and Stern LJ: **Crystallographic structure of the human leukocyte antigen DRA, DRB3*0101: models of a directional alloimmune response and autoimmunity.** *J Mol Biol* 2007, **371(2):**435–446.
27. Meinl E, Weber F, Drexler K, Morelle C, Ott M, Saruhan-Direskeneli G, Goebels N, Ertl B, Jechart G and Giegerich G: **Myelin basic protein-specific T lymphocyte repertoire in multiple sclerosis. Complexity of the response and dominance of nested epitopes due to recruitment of multiple T cell clones.** *J Clin Invest* 1993, **92(6):**2633–2643.
28. Jurcevic S, Travers PJ, Hills A, Agrewala JN, Moreno C and Ivanyi J: **Distinct conformations of a peptide bound to HLA-DR1 or DRB5*0101 suggested by molecular modelling.** *Int Immunol* 1996, **8(11):**1807–1814.
29. Texier C, Pouvelle-Moratille S, Busson M, Charron D, Ménez A and Maillère B: **Complementarity and redundancy of the binding specificity of HLA-DRB1, -DRB3, -DRB4 and -DRB5 molecules.** *Eur J Immunol* 2001, **31(6):**1837–1846.
30. Dai S, Crawford F, Marrack P and Kappler JW: **The structure of HLA-DR52c: comparison to other HLA-DRB3 alleles.** *Proc Natl Acad Sci USA* 2008, **105(33):**11893–11897.
31. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA and Mouse Genome Database Group: **The Mouse Genome Database (MGD): mouse biology and model systems.** *Nucleic Acids Res* 2008, **36 Database:** 724–728 http://www.informatics.jax.org/mgihome/other/citation.shtml, (date accessed – 2009-June-23).