An automated approach to modelling class II MHC alleles and predicting peptide binding

Martin T. Swain, Anthony J. Brooks, and Graham J.L. Kemp
Department of Computing Science
University of Aberdeen
Aberdeen, Scotland, UK, AB24 3UE

Abstract

We present an automated method for constructing 3D models of class II MHC structures that uses constraint logic programming to select side-chain conformations. The resulting models are used by a "peptide threading" program that attempts to predict peptides from a protein sequence that will bind strongly to particular MHC alleles. This method follows a comparative modelling approach in basing the model structures on experimentally determined MHC-peptide structures. However, constraints are used to ease open the peptide binding groove so that the modelled MHC structure is a less specific fit for the co-crystallised peptide in the starting structure. Preliminary results indicate that MHC models that have been constructed in this way enable the peptide threading program to make binding predictions that are comparable with those obtained when using experimentally determined MHC structures.

1 Introduction

Major histocompatibility complex (MHC) molecules, also known as HLA (human leucocyte antigen) in man, are important immune system proteins. The normal function of these molecules is to bind peptide antigens from "foreign" (or "non-self") proteins, and to present these on the surface of a cell (the antigen presenting cell). The resulting MHC-peptide complex is then recognised by a T cell receptor and this recognition event, accompanied by some other interactions, activates the T cell to carry out its programmed function. We are particularly interested in class II MHC molecules because of their importance in various autoimmune diseases caused by the anomalous presentation of "self" peptides.

There is interest in knowing which peptides are likely to associate with particular MHC alleles. In our interactions with experimentalists and clinicians investigating an autoimmune disease, a typical scenario is that a protein has been identified as the likely source of a "self" peptide, and that a statistical association has been observed between occurrence of the disease and the presence of certain MHC alleles in patients. In understanding the detailed mechanisms of the autoimmune disease it is important to know which peptide from the "self" protein is presented in the binding groove of MHC alleles associated with the particular disease. Further, if the same peptide is predicted not to bind to alleles that are not associated with that particular disease then this can be useful in guiding experimental investigations.

Release 1.10 (April 2001) of the IMGT/HLA sequence database [1] contains the amino acid sequences of 571 class II alleles; each individual has only a small subset of these. There are experimental techniques that can identify how well particular peptides bind to chosen MHC molecules. These experiments generally provide reliable results, however their main disadvantage is that they are expensive and time consuming. Further, since there are 20^{13} possible 13-residue peptides that might fit into binding groove, it is infeasible to obtain experimental binding data for all possible MHC-peptide combinations. Therefore, there is a demand for computer programs that attempt to predict which peptides are likely to bind to particular MHC alleles.

In the next section we give an overview of the structure of the class II MHC binding groove. A method for scoring MHC-peptide interactions is presented in Section 3. That method assumes that a 3D structure is available for the particular MHC allele of interest. Since only a few experimentally determined class II MHC structures are currently available, an important element of our work has been constructing 3D models of other MHC alleles that can be used with the pep-

[1] http://www.ebi.ac.uk/imgt/hla/

tide threading method described in Section 3, and in Section4 we present a novel method for building 3D models of MHC alleles. This method uses *constraint logic programming*, and we describe how this method can be adapted to produce structural models that are a less specific fit for the co-crystallised peptide in the initial structure. To assess the utility of these modelled alleles in predicting MHC-peptide interactions we have used the peptide threading program with both X-ray structures and modelled structures of the same MHC allele, and results obtained using these are presented in Section 5.

2 Class II MHC binding groove

The class II MHC peptide binding site is formed from two MHC antigen-recognition domains — one contributed from the α chain and the other from the β chain. The peptide binding site is a groove formed by two roughly parallel alpha-helices that are packed against a beta-sheet that forms the base of the groove (see Figure 1). Class II MHC molecules have the binding groove open at both ends and can theoretically bind peptides of any length, although 13-25 amino acids has been suggested as the typical length [10].

Experimentally determined X-ray structures of class II MHC-peptide complexes show some of the side chains oriented away from the peptide binding groove or located across the exterior faces of the helices flanking the binding groove, while five are directed towards the MHC allele and are accommodated in five "pockets" in the binding site. With both ends of the class II binding groove being open, peptides can lie relatively flat along the length of the binding groove, enabling hydrogen bonds to form between peptide backbone atoms and conserved residues of the binding groove alpha-helices [22].

It is the MHC β chain that contributes almost all of the binding site's variability. The differences in sequence between different MHC alleles are concentrated in the vicinity of the binding site and these differences can result in a change in the geometric and chemical character of the binding pockets, thus altering their specificity for peptide side chains and affecting the affinity with which different peptides bind to different alleles.

Regardless of the peptide that is ultimately presented on the cell surface by a particular class II MHC allele, the binding groove is occupied by a fragment the invariant chain (Ii) called CLIP (class II associated invariant chain peptide) at an intermediate stage in the maturation process [14]. In Section 4 we describe how we have used a model of the CLIP peptide in our modelling process to ensure that MHC side-chains in the vicinity of the groove are placed away from the space that a bound peptide would occupy, without biasing the model by configuring the binding groove to fit exactly a known high binding peptide.

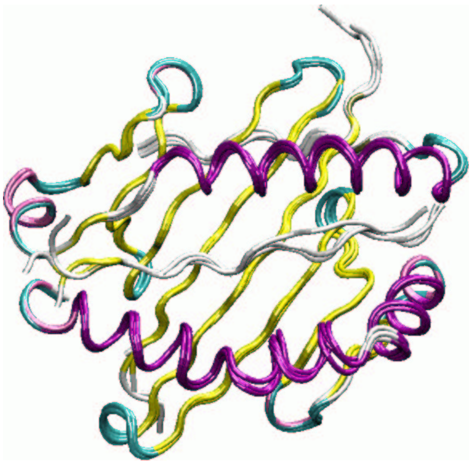


Figure 1: Backbone superposition of four class II MHC alleles HLA-DR1 [18], HLA-DR2 [21], HLA-DR3 [14], and HLA-DR4 [3]. The helix from the α chain is the long helix with a horizontal axis located towards the top of the figure. The long helix towards the bottom of the figure is from the β chain. The strands that form the base of the peptide backbone binding groove run diagonally between lower left and upper right. The backbones of the superposed peptides run horizontally between the helices, from left to right. Only the peptide binding groove of the MHC molecule is shown.

3 Peptide threading and binding prediction

The peptide binding predictions described in this paper were obtained using the peptide threading software developed by Brooks [5]. This program samples consecutive, overlapping, 13 amino acid peptides from the supplied antigen, and computes a predicted binding score for each. The peptides can then be ranked according to this score.

This program was inspired by earlier work in predicting peptide interactions with class I MHC alleles by Altuvia *et al.* [1]. In their approach, Altuvia *et al.* [1] used their previously defined list of 'assumed contacting positions' in conjunction with the HLA-DR/peptide sequences and Miyazawa and Jernigan's table of inter-residue contact potentials [17] to calculate an approximate energy value for the HLA-DR molecule/peptide complex. Scores were computed for candidate peptides by summing the table entries corresponding to the residue pairs which occurred at HLA-DR molecule and peptide positions in the list of 'assumed contacting positions'.

In contrast, the peptide threading software used here uses a scoring function which combines weighted terms representing steric overlap, the number of favourable van der Waals contacts, the hydrophobic effect, electrostatic force, and hydrogen bonds between the MHC molecule and the bound peptide.

To simplify the task of modelling the 3D structure of the peptide, we assume that all peptides binding to class II HLA-DR alleles will do so with a similar backbone conformation to that of the influenza haemagglutinin peptide from the HLA-DR1 crystal structure [22]. In that structure the peptide adopts a polyproline type-II conformation, forming hydrogen bonds between its backbone atoms and conserved residues of the HLA-DR molecule's alpha-helices [22]. It is thought that other peptides could do the same.

The backbone of the peptide is modelled with limited flexibility by creating 167 slightly different backbone conformations. To each of these backbone models, side-chains of the correct type are added, and a side-chain conformational search is then performed to find a low energy conformation for each peptide side-chain within the MHC groove. Performing such conformational searches on a "per peptide" basis would result in repeatedly calculating the same quantities. This is because when peptides are presented to the system, they may contain residues that are placed at positions in the binding groove which, for that residue type, a conformational search has already been performed. Therefore, a program has been written that pre-processes the MHC allele, performing a side-chain conformational search for each amino acid type at each peptide position. The conformations identified for each residue at each position are stored in a *side chain scanning library.* When the structural model of a peptide is subsequently required, the pre-calculated conformations corresponding to the sequence of the peptide can then be retrieved from the library and concatenated to form the peptide model. Thus, an unlimited number of peptide structural models can be built without the need to run the computationally expensive side-chain conformation search algorithm again.

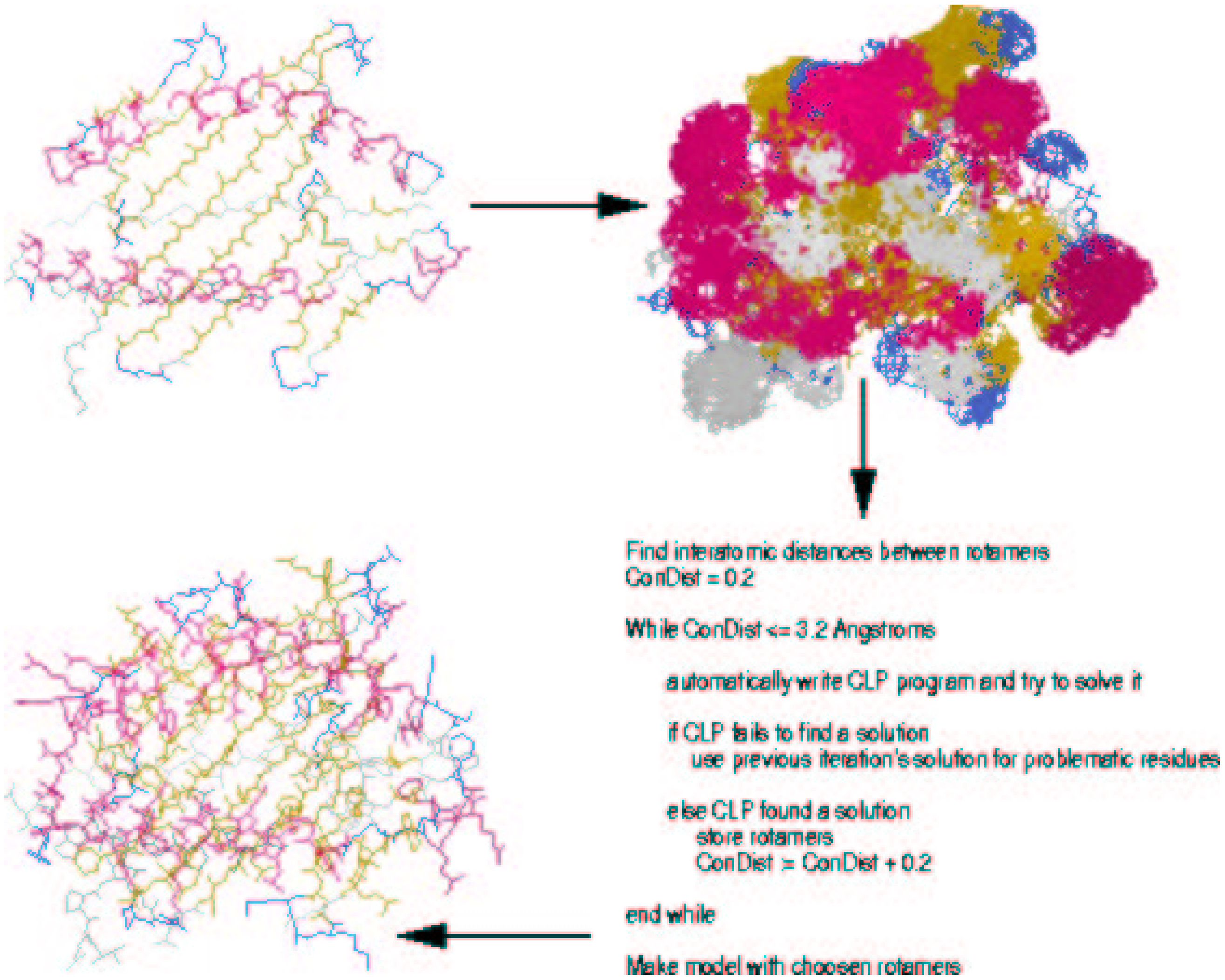
In addition to searching different peptide side chain conformations, this pre-processing program also tries different conformations for the MHC side-chains that define the peptide binding groove. However, to keep this search space a manageable size, the pre-processing program only varies the allele side chain torsions by up to +/- 30 degrees in 5 degree steps. So, while this program does have some freedom to move MHC side chains, the starting structure of the MHC allele is still important since the program only makes a small local search "centred" around the given starting conformation. In the next section we describe an automated approach to generating suitable MHC models for use in peptide threading studies.

4 Modelling class II MHC allele structures

The method for predicting MHC-peptide interactions described in Section 3 assumes that a 3D structure of the particular MHC allele of interest is available. The 3D structures of only a few class II MHC alleles are currently available. However, the utility of the prediction method can be extended by constructing model 3D structures of other alleles.

There are two features of class II MHC molecules that make their structures good candidates for homology modelling. First of all there are no insertions or deletions amongst alleles of the same type (i.e. HLA-DR or HLA-DQ, etc), and secondly the sequence analysis of class II MHC molecules shows an extremely high identity (typically $\geq 95\%$). Hence the main task in MHC homology model construction is side-chain conformational prediction.

Side-chain modelling is essentially the task of searching through a large combinatorial space of possible side-chain conformations to find a mutually consistent set — a problem for which *constraint logic programming*, or CLP, is well suited. Constraint logic programming [13, 26, 9], has been specifically developed by the artificial intelligence community to solve hard search problems, typically in areas such as planning, scheduling, packing, and resource allocation. It uses constraints with a simple built in search method: the constraints eliminate impossible alternatives and so restrict and guide the search. Here we briefly describe our CLP based method of automated side-chain modelling [24, 23], and we show how it can be used to open up the peptide binding groove of the MHC molecule.



```
Find interatomic distances between rotamers
ConDist = 0.2

While ConDist <= 3.2 Angstroms

    automatically write CLP program and try to solve it

    if CLP fails to find a solution
        use previous iteration's solution for problematic residues

    else CLP found a solution
        store rotamers
        ConDist := ConDist + 0.2

end while

Make model with choosen rotamers
```

Figure 2: The CLP method begins with a given backbone. All rotamers in the library for every backbone residue are placed onto the backbone. A series of automatically generated CLP programs are created with successively tighter side-chain packing constraints. These programs are solved by CLP to give a single rotamer for each residue.

4.1 Side-chain placement using CLP

In outline, our method starts by placing the complete set of rotamers from a rotamer library [12] onto the protein backbone. Rotamers are energetically favoured, commonly observed side-chain conformations [19, 20, 16]. They are often used by side-chain modelling algorithms because they discretise the continuous space of side-chain conformations, thus making the the combinatorial nature of the side-chain placement problem computationally tractable [11, 4]. Having placed all possible side-chain conformations on to the backbone, we then calculate steric overlaps between rotamers and the backbone, and between rotamers of different residues, using C code, and formulate them as constraints. Our method then selects compatible side-chain conformations, as illustrated in Figure 2.

The most important constraint on atomic packing in proteins that can be used to determine side-chain conformations is the avoidance of steric overlap [25].

We determine if two atoms are involved in a steric overlap by calculating the distance between the centres of two atoms: two atoms were said to be overlapping if the interatomic distance, was less than a variable called *ConDist*. We iterate over this variable ConDist, increasing its value from 0 Å to 3.2 Å. At each iteration of ConDist we create a CLP program using constraints to model steric overlaps. The CLP solver finds a set of side-chain conformations consistent with the constraints by eliminating rotamers involved in atomic clashes with the backbone, and, if a pair of rotamers are involved in a clash, by eliminating one of the clashing rotamer pair.

As ConDist increases the packing constraints become tighter and tighter. When the value of ConDist becomes greater than approximately 2.0 Å the constraints on some residues become so tight that it is not possible for any rotamer to satisfy the packing constraints. When this happens it is logically impossible to model the side-chain conformations and the CLP program will fail. To solve this problem we set the residue's side-chain conformation to that conformation determined in the previous iteration, and then treat the side-chain as though it was part of the backbone — the side-chain is fixed, and if any rotamer clashes with it then that rotamer is eliminated from the solution. When ConDist reaches the value of 3.2 Å all but a few side-chains have become fixed.

To illustrate how the accuracy of the models created depends on the ConDist variable, we have modelled the peptide binding domains of two class II MHC structures with Protein Data Bank (PDB) [2] codes 1DLH [22] and 1AQD [18]. These known structures have resolutions of 2.8 Å and 2.45 Å respectively. In remodelling the side chains of these structures, we considered residues 1-90 of both the A and B chains of the MHC molecule. The quality of the remodelled side-chains was assessed by comparing side chain torsion angles in the modelled structures with those in the known structures. The $\chi 1$ angle measures the side-chain's rotation about the bond between a residue's α- and β-carbon atoms [15], and in Figure 3 we show how the fraction of $\chi 1$ angles predicted to within 40^o for these two molecules. Both the curves show maximum accuracy at values of approximately 2.0 Å to 2.6 Å, and decrease in accuracy towards 3.2 Å. If we were simply interested in predicting the X-ray crystal structure we would set the maximum value of ConDist to 2.6 Å as we have described elsewhere [24]. However, here our intention is *not* to reproduce the side-chain conformations observed in the X-ray structure. Rather, we aim to build a model with the binding groove side-chains spread apart, leaving the groove wide open.

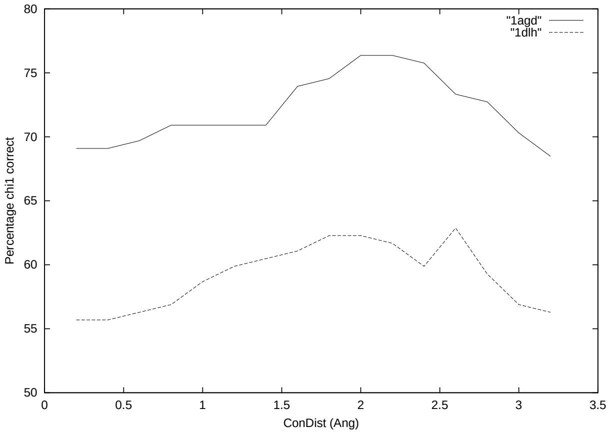


Figure 3: The average percentage of modelled side-chains with $\chi 1$ angles within 40^o of those in the two corresponding PDB files. The models were built using the CLP method.

4.2 Opening the peptide binding groove

After protein synthesis the binding groove of class II MHC molecules is filled with the CLIP peptide. Before modelling the side-chains of the MHC molecule we mutate the binding groove peptide's sequence to this CLIP peptide. Then, by letting ConDist iterate up to 3.2 Å, we impose constraints that maximise the interatomic distance between neighbouring side-chain atoms. These constraints act to repel the MHC and peptide side-chains from each other, thus leaving the binding groove wide open once the CLIP peptide has been removed.

5 Results

To assess the utility of using alleles modelled as described in Section 4 with the binding prediction method described in Section 3 we have used the peptide threading program with both X-ray structures and model structures of the same MHC allele. The results presented here are based on two X-ray structures of the HLA-DRB 0101 allele (PDB codes 1DLH and 1AQD, with resolutions of 2.8 Åand 2.45 Å respectively) and two model structures that were built by

Table 1: Here we compare the predicted peptides' rankings using two X-ray structures to models built as described in Section 4. The rankings have been converted to percentages (i.e. rank out of 100) in order to allow comparisons between sequences with different lengths.

Protein Sequence	No. Residues	Peptide	1DLH X-ray	1AQD X-ray	1DLH model	1AQD model
Influenza Haemagglutinin	566	PKYVKQNTLKLAT	9.9	4.3	3.1	4.3
Influenza Haemagglutinin	566	PDYASLRSLVASS	18.2	9.7	6.4	6.7
Human Myelin Basic	196	HFFKNIVTPRTPA	14.7	8.7	8.2	9.2
Human cartilage link	354	IKWTKLTSDYLKE	9.6	5.0	3.8	6.7
Tetanus toxin	1314	PLYKKMEAVKLRD	12.8	20.2	10.5	15.3
Tetanus toxin	1314	NAFRNVDGSGLVS	13.8	11.3	9.0	9.7
Tetanus toxin	1314	TIYQYLYAQKSPT	7.9	7.6	12.6	17.8
Tetanus toxin	1314	IYYRRLYNGLKFI	26.3	12.7	33.2	25.5
Tetanus toxin	1314	KIYSYFPSVISKV	17.6	11.4	16.6	16.4
Tetanus toxin	1314	MQYIKANSKFIGI	21.8	23.7	23.4	22.3
Glutamate decarboxylase	585	VNYAFLHATDLLP	7.9	9.1	4.4	6.7
Glutamate decarboxylase	585	LDMVGLAADWLTS	9.9	5.1	10.0	9.9
Average percentage rank			14.2	10.7	12.5	12.7

remodelling all side-chains onto the backbone conformations given in the PDB files for 1DLH and 1AQD.

Table 1 shows the ranking predictions made by our system for a set of known high binding peptides obtained from the MHCPEP database [6, 8]. This table shows that our system usually ranks high binding peptides within the top 15% of all possible peptides that can be derived from a protein sequence. The most accurate predictions are obtained using the high resolution structure (1AQD X-ray), with which the majority of the known binding peptides are ranked within the top 10%. The average predictions obtained from the models constructed using our CLP method are more accurate than those obtained from the low resolution X-ray structure (1DLH). However, when considering each peptide individually, there is a great deal of variation in the accuracy of the predictions for each of the four structures.

While more extensive testing with X-ray and model structures of other MHC alleles is needed, it is encouraging to see that the results obtained using model structures are not significantly worse than those obtained using experimentally determined MHC structures. This suggests that a combined modelling and peptide threading approach as described in this paper could be worth pursuing for alleles where no experimentally determined structures are available.

6 Discussion and future work

The peptide binding predictions presented in this paper are based on three-dimensional models of class II MHC-peptide complexes, using heuristic functions to score chemical and spatial complementarity. These heuristic scores are used to compute an overall binding score for the MHC-peptide interaction, enabling peptides to be categorised as binders or non-binders. Predictions involve the automatic construction of a model MHC structure and analysis of its binding cleft, processes which take a few hours (per allele) using a desktop workstation. Thereafter, prediction of peptide binding to that allele is accomplished at a rate of 1000 (13-mer) peptides per second.

The approach taken in this paper is different to the neural network approach described by Brusic *et al.* [7]. In their work, a neural network was trained using binding data based on 650 peptides for HLA-DR4(B1*0401). The resulting system was then validated in a number of ways, and compared to other prediction methods. Their tests have demonstrated that this approach works very well, with a highly significant association between predicted and experimental binding. In contrast, we are pursuing a method that is based on the predicted 3D conformation of the binding groove, and which does not depend upon the availability of training data, although such data would be of benefit in refining the scoring method used in evaluating MHC-peptide interactions.

Our peptide threading method relies on having a 3D structure of an MHC molecule. Sequence similarity between different alleles and observation of experimentally determined class II MHC structures suggest that different MHC alleles will have strongly conserved backbone conformations. Therefore, side-chain placement algorithms that model both MHC and peptide side-chains play a vital role in predicting MHC-peptide interactions. We believe that our modelling system removes the specificity of the MHC molecule to the bound peptide, and our preliminary results indicate that modelled MHC structures are sufficient for binding predictions. However, further investigation is needed to determine the significance of alternative side-chain conformations resulting from different peptides being bound to the MHC molecule. In particular, improvements to the modelled structures could be achieved by conserving side-chain conformations that make hydrogen bonds to the peptide backbone in experimentally derived structures.

Acknowledgements

M.T.S. is supported by a BBSRC CASE award with Biovation Ltd.

References

[1] Y. Altuvia, O. Schueler, and H. Margalit. Ranking potential binding peptides to MHC molecules by a computational threading approach. *Journal of Molecular Biology*, 249:244, 1995.

[2] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Mayer, M. D. Bruce, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.*, 112:535–542, 1977.

[3] D. R. Bolin, A. L. Swain, R. Sarabu, S. J. Berthel, P. B. Gillespie, N. J. S. Huby, R. Makofski, L. Orzechowski, A. Perrotta, K. Toth, J. P. Cooper, N. Jiang, F. Falcioni, R. Campbell, D. Cox, D. Gaizband, D. Vidovic, K. Ito, R. Crowther, U. Kammlott, R. Palermo, D. Weber, J. Guenot, Z. Nagy, and G. L. Olson. Petide and Petide mimetic inhibitors of anigen presentation by HLA-DR class II MHC molecules. Design, structure-activity relationships, and X-ray crystal structures. *J. Med. Chem.*, 43:2135, 2000.

[4] M. J. Bower, F. E. Cohen, and R. L. Dunbrack. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.*, 267:1268–1282, 1997.

[5] A. J. Brooks. *Computational Prediction of HLA-DR Binding Peptides*. PhD thesis, University of Aberdeen, 1999.

[6] V. Brusic, G. Rudy, and L. C. Harrison. MHCPEP - A Database of MHC-Binding Peptides. *Nucleic Acids Research*, 22:3663, 1994.

[7] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, and L. Harrison. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, 14:121, 1998.

[8] V. Brusic, G. Rudy, A. P. Kyne, and L. C. Harrison. MHCPEP - a database of MHC-binding peptides: update 1995. *Nucleic Acids Research*, 24:242, 1996.

[9] M. Carlsson, G. Ottosson, and B. Carlson. An open-ended finite domain constraint solver. *Proc. Programming Languages: Implementations, Logics, and Programs*, 1997.

[10] R.M. Chicz, R.G. Urban, W.S. Lane, J.C. Gorga, L.J. Stern, D.A.A. Vignali, and J.L. Strominger. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature*, 358:764, 1992.

[11] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[12] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of side-chain rotamer preferences. *Protein Science*, 6:1661–1681, 1997.

[13] T. Fruhwirth, A. Herold, V. Kuchenhoff, T. Le Provost, L. Pierre, E. Monfroy, and M. Wallace. Constraint logic programming: An informal introduction. Technical report, European Computer-Industry Research Centre, 1993.

[14] P. Ghosh, M. Amaya, E. Mellins, and D. C. Wiley. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. *Nature*, 378:457, 1995.

[15] IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. *Eur. J. Biochem.*, 17:193–201, 1970.

[16] S. C. Lovell, M. Word, J. S. Richardson, and D. C. Richardson. The Penultimate Rotamer Library. *Prot. Struct. Funct. Genet.*, 40:389–408, 2000.

[17] S. Miyazawa and R. L. Jernigan. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18:534, 1985.

[18] V. L. Murthy and L. J. Stern. The class II MHC protein HLA-DR1 in complex with an endogenous peptide implication for the structural basis of the specificity of peptide binding. *Structure*, 5:1385, 1997.

[19] J. W. Ponder and F. M. Richards. Tertiary templates for proteins. *J. Mol. Biol.*, 193:775–791, 1987.

[20] H. Schrauber. Rotamers: to be or not to be? *J. Mol. Biol.*, 230:592–612, 1993.

[21] K. J. Smith, L. Pyrdol, D. C. Gauthier, and D. C. Wiley. Crystal structure of HLA-DR2 (DRA0101, DRB11501) complexed with a peptide from human myelin basic protein. *J. Exp. Med.*, 188:1511, 1998.

[22] L.J. Stern, J.H. Brown, T.S. Jardetzky, C.G. Gorga, G.U. Robert, J.L. Strominger, and D.C. Wiley. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, 368:215, 1994.

[23] M.T. Swain and G.J.L. Kemp. A CLP approach to the protein side-chain placement problem. *Proceedings of the Seventh International Conference on Principles and Practice of Constraint Programming*, in press.

[24] M.T. Swain and G.J.L. Kemp. Modelling protein side-chain conformations using constraint logic programming. *Computers Chem.*, in press.

[25] M. Vasquez. Modeling side-chain conformations. *Curr. Opin. Struct. Biol.*, 6:217–221, 1996.

[26] M. Wallace. Constraint programming. Technical report, Imperial College, 1995.