J Comput Aided Mol Des (2009) 23:301–307
DOI 10.1007/s10822-009-9259-2

A critical cross-validation of high throughput structural binding prediction methods for pMHC

Bernhard Knapp · Ulrich Omasits · Sophie Frantal · Wolfgang Schreiner

Received: 6 October 2008 / Accepted: 13 January 2009 / Published online: 5 February 2009

Abstract T-cells recognize antigens via their T-cell receptors. The major histocompatibility complex (MHC) binds antigens in a specific way, transports them to the surface and presents the peptides to the TCR. Many in silico approaches have been developed to predict the binding characteristics of potential T-cell epitopes (peptides), with most of them being based solely on the amino acid sequence. We present a structural approach which provides insights into the spatial binding geometry. We combine different tools for side chain substitution (threading), energy minimization, as well as scoring methods for protein/peptide interfaces. The focus of this study is on high data throughput in combination with accurate results. These methods are not meant to predict the accurate binding free energy but to give a certain direction for the classification of peptides into peptides that are potential binders and peptides that definitely do not bind to a given MHC structure. In total we performed approximately 83,000 binding affinity prediction runs to evaluate interactions between peptides and MHCs, using different combinations of tools. Depending on the tools used, the prediction quality ranged from almost random to around 75% of accuracy for correctly predicting a peptide to be either a binder or a non-binder. The prediction quality strongly depends on all three evaluation steps, namely, the threading of the peptide, energy minimization and scoring.

Keywords T cell epitope prediction · Scoring · Energy minimization · Threading · Substitution

B. Knapp (✉) · U. Omasits · W. Schreiner
Unit for Medical Statistics and Informatics—Section for Biomedical Computersimulation and Bioinformatics, Medical University of Vienna—General Hospital, Spitalgasse 23, Room: BT88—88.03.712, 1090 Wien, Austria
e-mail: bernhard.knapp@meduniwien.ac.at

S. Frantal
Unit for Medical Statistics and Informatics—Section for Medical Statistics, Medical University of Vienna—General Hospital, Spitalgasse 23, 1090 Wien, Austria

Introduction

A major prerequisite for the activation of $CD8^+$ T-cells is the binding between a peptide and the major histocompatibility complex (MHC) class I molecule. The proteasome cleaves proteins into small peptide fragments. These are transported through the membrane of the endoplasmic reticulum (ER) by the "transporter associated with antigen processing" (TAP) system where they are loaded into the MHC binding groove. Some peptides cannot bind to the MHC molecule and some peptides do bind but are unstable. Only the binding peptides are transported via the Golgi apparatus to the cell surface, where the peptide/MHC (pMHC) can be recognized by $CD8^+$ T-cells.

For all these steps prediction tools exist. For details regarding these tools the reader is referred to the reviews mentioned in [1–4]. The majority of prediction methods for the interaction between a peptide and the MHC molecule are sequence-based. These methods obtain good results as long as enough experimental calibration data is available. Some of these methods are even over-fitted to a single allele, to such an extent that the area under the receiver operating characteristic (ROC) curve is almost 1. Especially for MHC class II molecules a low cross-validation performance over different alleles and types of antigens has recently been shown to obtain (cf. [5]). On the other hand, structure-based approaches have to accommodate the challenges that lie in the structural details of the interaction

between peptide and MHC molecule, where even a single amino acid (AA) mutation in the MHC [6] or the epitope [7, 8] can decide between binding and non-binding, or between the activation of the T-cell or lack thereof. Therefore, with respect to well-known alleles, structure-based approaches are usually less effective than sequence-based methods. However, they are superior to the sequence-based analyses if the allele changes. Moreover, they provide an answer to the question of why a specific peptide is binding or non-binding.

In this study we present a critical cross-validation of structural high throughput binding prediction methods for pMHC interactions. We compare threading, energy minimization and structural scoring methods in their different combinations to provide some insights into the performance of methods that are not specifically calibrated for pMHC binding prediction.

Methods

We designed a benchmark starting from a given peptide sequence and the spatial structure of the MHC allele. The binding affinity between peptide and MHC was evaluated in 4 steps. First, the peptide is threaded into the MHC binding groove (1), then the modeled structure is energetically minimized (2). After that, the binding affinity of the modeled complex is predicted (3) and the predicted score is compared to the experimental binding data (4). The workflow of this evaluation process is illustrated in Fig. 1.

The structure of the MHC is taken from the PDB [9], while the peptide sequences are extracted from the community set [10], which provides an experimental IC_{50} value for each peptide listed in the community set. In our case there are 3,089 nonameric peptides listed for HLA-A*0201, out of which 2,122 have an exactly determined experimental IC_{50} value and make up the "core set". The other peptides are merely characterized as "$IC_{50} >$ 5,000 nM" or "$IC_{50} >$ 20,000 nM". This raises the problem that a value of ">5,000 nM" may be larger than a second value classified as ">20,000 nM", which makes it difficult to compare such values. Such values are calculated as being "equal to" rather than "greater than" the indicated value, and are defined as "extended set". In total there are 1,181 peptides known to bind to HLA-A*0201 ($IC_{50} <$ 500 nM) and 1,908 which are known to not bind to HLA*0201 ($IC_{50} \geq$ 500 nM). Peptides with a length of ten AAs are excluded from the set since the vast majority of binding peptides in the community set are nonameric. On the basis of these data, the four major computational steps mentioned above have been implemented as described in detail in the following sections.

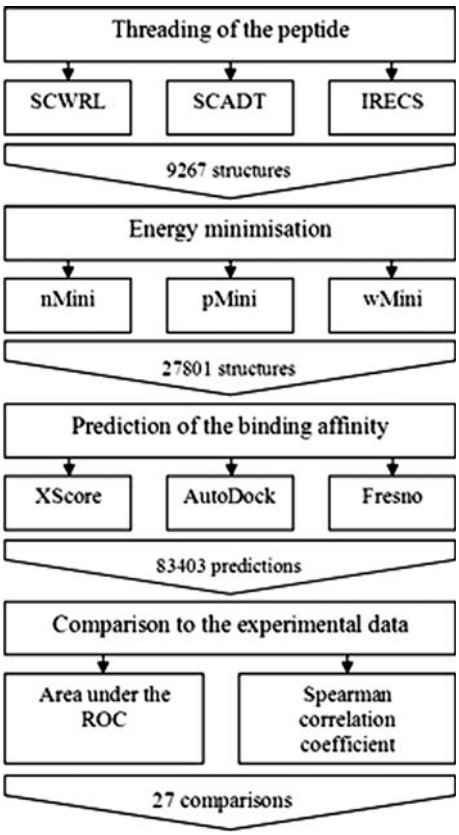

Fig 1 Workflow of the algorithm

Threading of the peptide

The consequences of removing chains or even parts of domains that are not directly involved in the regions under investigation are discussed controversially in the literature. Most authors hold that this process does not affect the results [11–15]. However, this view is not shared by everyone (cf. [16]). In order to improve runtime, we decided to remove the TCR, the β_2 microglobulin as well as the α_3 domain of the MHC. In our previous study [17], we were able to show that for the substitution of a peptide, it is sufficient to model the side chains of the new peptide. Thus the backbone of the peptide remains untouched (although minor adjustments are made during the energy minimization phase). We use the three tools which, as shown in our previous study, proved most appropriate for the modeling of the complexes: SCWRL [18], SCATD [19] and IRECS [20]. The tools are used to substitute the peptide side chains of the complex with the PDB identification *1qsf*. The result of this step is a set of 9,267 (3,089 × 3) modeled complexes in the PDB format. Thus, all the complexes from which experimental data is known have been modeled three times.

Energy minimization

In the next step three different energy minimization (EM) techniques are applied to the peptide and to the α_1 and α_2 domains of the MHC. In the first case, the complexes are left as they are and immediately enter the next phase. This procedure is labeled "no minimization" (nMini). The second approach is called "pure protein minimization" (pMini). In this process, the modeled complex is energetically minimized in vacuo, using the steepest descent method of GROMACS [21]. With the third method, the complex is inserted into an artificial water bath and subsequently minimized together with the water cube. This process is called "water minimization" (wMini) (for the appropriate size of water cubes for GROMACS simulations we kindly refer the reader to our previous study [15]). Given the high throughput approach of this study we did not do any manual adjustment of side chains or hydrogen atoms. Each of these remains exactly in the position calculated by the EM. The result of this step is a set of 27801 (9267×3) modeled complexes.

Prediction of the binding affinity

The resulting complexes are evaluated using three different scoring functions. To our knowledge Fresno [22] is the only structural scoring function that has been especially designed for pMHC interactions. For this reason we decided to include Fresno despite the fact that it was developed almost a decade ago. The second scoring function that we used is XScore [23]. This is a pure consensus empirical scoring function which only focuses on the interaction between a given ligand/receptor structure without any screening or flexibility. The third method used in our study is the well-known AutoDock 4 scoring function [24].

All three scoring functions were parameterized in such a way that only the binding affinity of the given complex is predicted, without any flexibility in either ligand or receptor. The "induced fit" between the peptide and the MHC is only simulated in the threading and energy minimization steps. In a small number of cases, the scoring functions are not able to produce any significant results or simply crash during the process. This happens primarily in the nMini complexes, due to spatial clashes. For these problematic cases we defined the complex as a definite non-binder.

The result of this step is a set of 83,403 ($27{,}801 \times 3$) binding affinity predictions.

Comparison to the experimental data

In the last step the predicted scores have to be compared to the experimental binding data. Given the problem that different scoring functions produce results formulated in different units and scales, the only possible way to obtain an adequate comparison of results is to invoke a test based on rank statistics. We decided to use the area under the receiver operating characteristic (ROC) curve and the error estimation as described in the literature [25] (see "Results"). Furthermore the Spearman rank correlation coefficient [26] was used to investigate whether there is a correlation between the binding scores and the experimental data.

For each of the predicted 83,403 binding affinities a corresponding experimental IC_{50} value is available. These values classify the peptides as being either binding peptides ($IC_{50} < 500$ nM) or non-binding peptides ($IC_{50} \geq 500$ nM). This binary classification is used in combination with the predicted values to calculate the area under the ROC curve. The evaluation is performed twice: the first time all the experimental values are used (extended set), and the second time only the values with exactly determined values (core set) are part of the calculation.

In order to recognize potential differences between the areas under the ROC curves in the XScore group of the extended set, a nonparametric comparison employing the %ROC macro in SAS [27, 28] is done. This comparison is limited to XScore since the XScore function performed better in the overall setting of the study (for details see the result section below). Moreover, the inclusion of all combinations would lead to an explosion of the combinatorial space. A test including all data is performed, and all pairwise contrasts for XScore are estimated and tested. The analysis of the results is done with SAS 9.1. All p-values < 0.00139 are considered statistically significant. The critical boundary of 0.00139 is derived from the correction value for multiplicity, following Bonferroni. For the relevant number of tests, 36, the value is calculated as $0.05/36 = 0.00139$.

In addition, the Spearman rank correlation coefficient [26] was used to find whether there is a correlation between the binding scores and the experimental data. Following the practice discussed in [29], the confidence interval was calculated. For the purposes of comparing the results, the algebraic sign of the results has to be inverted, given that a lower IC_{50} value indicates stronger binding affinity while a lower pK_d score indicates weaker binding affinity. In the present study, always the positive correlation was computed.

The results of this step are 27 ($3 \times 3 \times 3$) different rocA values and 27 Spearman rank correlation values for each data set (Fig. 2).

Results

In Fig. 2 we illustrate the results. The bars in (a) represent the areas under the ROC curves that have been defined

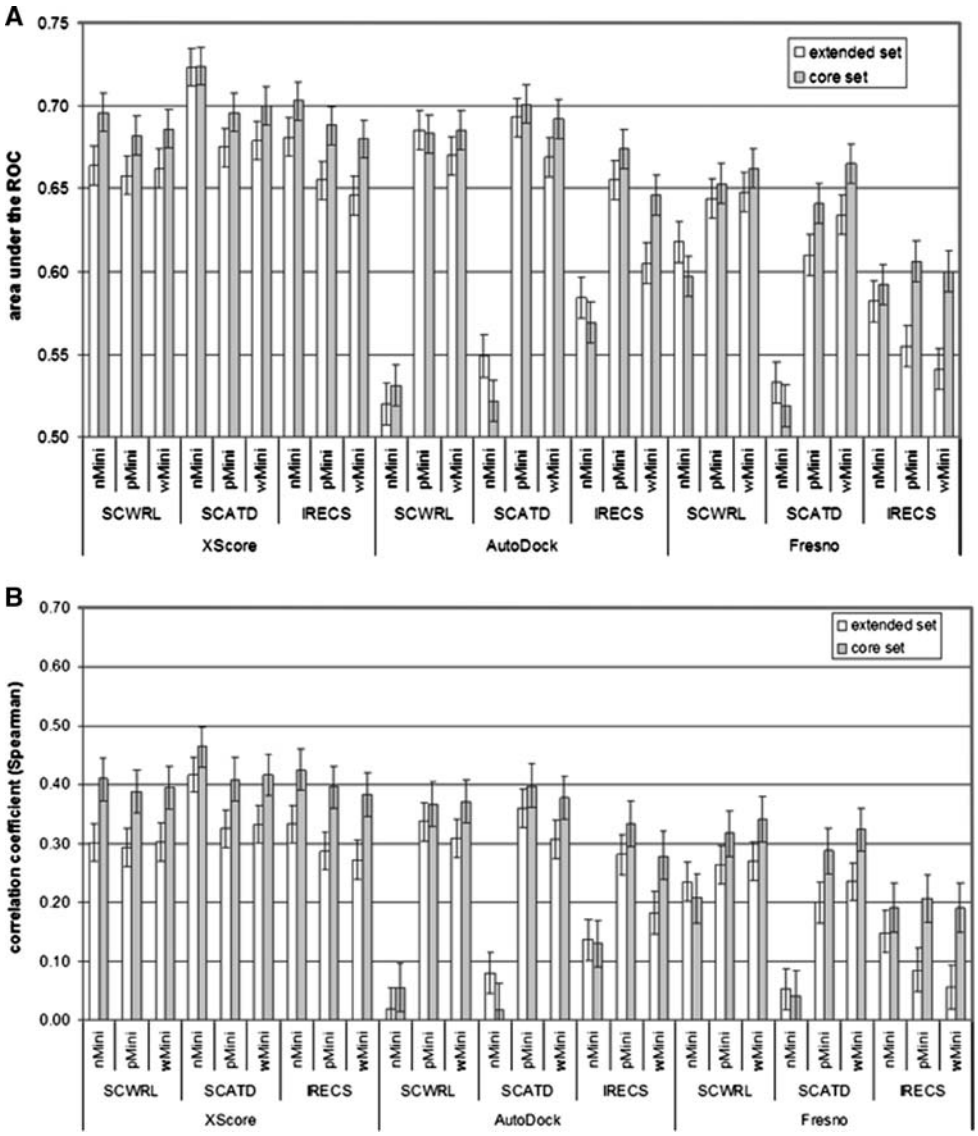

Fig 2 **a** Area under the ROC used to assess prediction quality. The values are first grouped according to the scoring method, then according to the side chain substitution tool and finally according to the energy minimization technique. The *left bar* (*empty*) always indicates the "extended set" while the *right bar* (*shaded*) shows the "core set". **b** The same methods as in **a** are compared, but the Spearman correlation coefficient is used to assess the prediction quality

assuming an experimental threshold of 500 (see the "Methods" section for details). In order to obtain a full picture of the data, we also used other thresholds. Since the results with different thresholds were similar to the results obtained with the threshold of 500, we will not include these data.

One can clearly see that the use of the Autodock scoring function in combination with an arbitrary threading tool without any kind of EM leads to almost random results (rocA $\sim$ 0.5). However, if EM is applied the results improve dramatically. This is due to the very restrictive scoring function of AutoDock which penalizes atoms that are "too close" to other atoms since such pairs add a large positive energy to the computed docking energy. This increase in positive energy occurs even if the spatial clash is not directly located in the binding groove, given that intramolecular receptor contacts are included in the binding score of AutoDock (see the AutoDock 4 scoring function publication for details [24]). In all other combinations with the scoring and threading tools, EM does not seem to show a clear tendency in its effect on the results.

The best result is achieved with SCATD in combination with XScore, without any kind of EM. This can be seen from the area under the ROC curve (Fig. 2a) as well as in the Spearman correlation coefficient (Fig. 2b). For instance, Fig. 3 shows the peptides KLFYVYYNL and SVAKCCSKT that have been modeled into the binding groove of the MHC. It is obvious that the first of the two peptides attaches well to the groove while the second peptide is not able to anchor to the groove. This optic impression is equally reflected in the experimental data and in the predicted binding affinity.

In general the results of XScore are trustworthy in combination with any of the other tools, which is why we

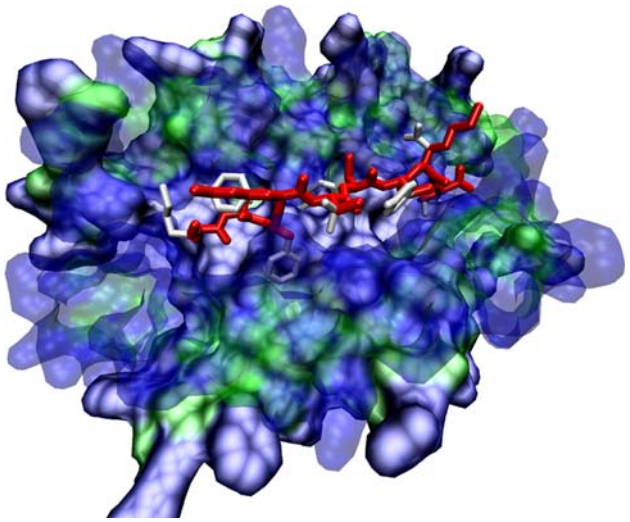

Fig 3 Peptide/MHC interaction: The peptide KLFYVYYNL is shown in *white*. It is a strong binder with an experimental IC_{50} value of 1 and a pK_d value of 9.96 predicted by XScore. The peptide SVAKCCSKT is shown in red. It is a non-binding peptide with an IC_{50} value of 36,874 and a predicted pK_d value of 6.64. The MHC-backbone is colored in *green* while the side chains are colored in *blue*. The α-helices are shown as transparent to provide insights into the binding groove

decided to perfom an additional statistical analysis. The overall statistical test for XScore shows a high statistical significance with a $p < 0.0001$, $\chi^2 = 386.1664$ and $df = 8$. Hence, at least one pairwise comparison must show a significant difference. Table 1 shows the detailed results for all pairwise comparisons. Statistically significant differences are marked. A positive estimate for the difference in rocA entails that the area under the ROC curve obtained on the basis of the method combination X is larger than the area under the ROC curve based on the method combination Y. A negative estimate indicates the opposite.

Discussion

Three side chain substitution tools, three energy minimization methods and three different scoring functions were tested on a set of 3,089 different peptides with known experimental binding affinities. Each of the tools was tested for every possible combination with the other tools. In total, 83,403 evaluations were performed ($3 \times 3 \times 3 \times 3{,}089$).

The best result by far was achieved by XScore in combination with SCATD and nMini. In our previous study we were able to show that SCATD and SCWRL can reproduce the experimentally determined X-ray structure with the lowest root mean square deviation (RMSD) [17]. On the basis of these results, the good performance of SCATD in this study could be expected. On the other hand, it is interesting to note that SCWRL does not seem to cooperate equally well with XScore, despite the fact that its calculation of the side chain positions is quite similar to the methods used by SCATD.

The general conclusion to be drawn based on the study discussed in this paper is the following. Despite the satisfactory results obtained for our data, there is still room for further improvements and innovations concerning structural methods for the prediction of binding affinities. In our case, the best combination was able to correctly predict a peptide to be either "binding" or "non-binding" with a probability of $\sim$75%. Ideally, however, probability values above 90% would be desirable. This less than optimal performance in this respect is also reflected in the Spearman correlation

Table 1 Pairwise comparison for XScore rocA results

Y \ X		Scwrl			Scatd			Irecs		
		nMini	pMini	wMini	nMini	pMini	wMini	nMini	pMini	wMini
Scwrl	nMini		-0.0070	-0.0027	0.0578*	0.0101	0.0142	0.0158*	-0.0098	-0.0193*
	pMini	0.0070		0.0042	0.0648*	0.0171*	0.0212*	0.0228*	-0.0029	-0.0124*
	wMini	0.0027	-0.0042		0.0606*	0.0129*	0.0170*	0.0186*	-0.0071	-0.0166*
Scatd	nMini	-0.0578*	-0.0648*	-0.0606*		-0.0477*	-0.0436*	-0.0420*	-0.0677*	-0.0772*
	pMini	-0.0101	-0.0171*	-0.0129*	0.0477*		0.0041*	0.0057	-0.0200*	-0.0295*
	wMini	-0.0142	-0.0212*	-0.0170*	0.0436*	-0.0041*		0.0016	-0.0241*	-0.0336*
Irecs	nMini	-0.0158*	-0.0228*	-0.0186*	0.0420*	-0.0057	-0.0016		-0.0257*	-0.0352*
	pMini	0.0098	0.0029	0.0071	0.0677*	0.0200*	0.0241*	0.0257*		-0.0095*
	wMini	0.0193*	0.0124*	0.0166*	0.0772*	0.0295*	0.0336*	0.0352*	0.0095*	

* significant difference (p<0.00139)
method X significantly worse than method Y
method X significantly better than method Y

Evaluation for the purposes of finding the optimal combination of procedures for threading and minimization. Each cell of the table gives the following information: The estimated rocA difference between one specific pairing of threading and minimization X (e.g., ScwrlpMini) and a second specific pairing Y (e.g., ScatdpMini) is computed. The *-symbol indicates a significant difference in rocA. The whole evaluation was based on the extended set and XScore, since the latter proved superior with regard to the other scoring functions

coefficients (Fig. 2b), where only values below 0.5 could be obtained. It seems obvious that these results could easily be improved if the methods under discussion were specifically calibrated for the interaction of peptide/MHC, or if the structures could be adjusted manually. In the long run, however, the ultimate goal must be to develop general methods which are not restricted to a certain type of complex but could accommodate all types of complexes. The objective is thus to develop scoring functions as well as methods for spatial arrangement which are applicable to a very diverse set of molecules.

From the present work the following general recommendations can be deduced: (1) When using the pure AutoDock scoring function, EM is necessary. (2) In comparison to other scoring functions, XScore works best with arbitrary side chain substitution tools and different EMs. Specifically, XScore results were always better than random in our study. (3) Fresno—although it has been developed for pMHC—seems to require manual adjustment of the positions of the polar hydrogen atoms to be able to yield good results. However, this was not a valid possibility in our high-throughput approach. (4) Only the core set with exactly determined experimental values should be used for further testing, since affinities defined over inequalities (e.g., "$IC_{50} > 5{,}000$ nM") will always yield results with a certain degree of uncertainty.

Acknowledgment This work is supported in part by the Austrian Grid Project of the Austrian Ministry of Education, Science and Culture (contract no. GZ 4003/2-VI/4c/2004).

Conflict of interest The authors declare no conflict of interests. None of the authors is author or co-author of any tool that was discussed and evaluated in this paper.

References

1. Saxova P, Buus S, Brunak S, Kesmir C (2003) Predicting proteasomal cleavage sites: a comparison of available methods. Int Immunol 15(7):781–787. doi:10.1093/intimm/dxg084
2. Korber B, LaBute M, Yusim K (2006) Immunoinformatics comes of age. PLOS Comput Biol 2(6):e71. doi:10.1371/journal.pcbi.0020071
3. Tsurui H, Takahashi T (2007) Prediction of T-cell epitope. J Pharmacol Sci 105(4):299–316. doi:10.1254/jphs.CR0070056
4. Sousa SF, Fernades P, Ramos MJ (2006) Protein-ligand docking current status and future challanges. Proteins 65:15–26. doi:10.1002/prot.21082
5. Gowthaman U, Agrewala JN (2008) In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J Proteome Res 7(1):154–163. doi:10.1021/pr070527b
6. Miller PJ, Pazy Y, Conti B, Riddle D, Appella E, Collins EJ (2007) Single MHC mutation eliminates enthalpy associated with T cell receptor binding. J Mol Biol 373(2):315–327. doi:10.1016/j.jmb.2007.07.028
7. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG (1991) Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296. doi:10.1038/351290a0
8. Kjer-Nielsen L, Clements CS, Purcell AW et al (2003) A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity. Immunity 18(1):53–64. doi:10.1016/S1074-7613(02)00513-7
9. Bergman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242. doi:10.1093/nar/28.1.235
10. Peters B, Bui HH, Frankild S et al (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLOS Comput Biol 2(6):e65. doi:10.1371/journal.pcbi.0020065
11. Rognan D, Zimmermann N, Jung G, Folkers G (1992) Molecular dynamics study of a complex between the human histocompatibility antigen HLA-A2 and the IMP58–66 nonapeptide from influenza virus matrix protein. Eur J Biochem 208(1):101–113. doi:10.1111/j.1432-1033.1992.tb17163.x
12. Zoete V, Michielin O (2007) Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: Application to the TCR-p-MHC complex. Proteins 67(4):1026–1047. doi:10.1002/prot.21395
13. Gregoire C, Lin SY, Mazza G, Rebai N, Luescher IF, Malissen B (1996) Covalent assembly of a soluble T cell receptor-peptide-major histocompatibility class I complex. Proc Natl Acad Sci USA 93(14):7184–7189. doi:10.1073/pnas.93.14.7184
14. Toh H, Kamikawaji N, Tana T, Muta S, Sasazuki T, Kuhara S (2000) Magnitude of structural changes of the T-cell receptor binding regions determine the strength of T-cell antagonism: molecular dynamics simulations of HLA-DR4 (DRB1*0405) complexed with analogue peptide. Protein Eng 13(6):423–429. doi:10.1093/protein/13.6.423
15. Omasits U, Knapp B, Neumann M et al (2008) Analysis of key parameters for molecular dynamics of pMHC molecules. Mol Simul 34:781–793. doi:10.1080/08927020802256298
16. Wan S, Coveney P, Flower DR (2004) Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: can the alpha3 and beta2 m domains be neglected? J Comput Chem 25(15):1803–1813. doi:10.1002/jcc.20100
17. Knapp B, Omasits U, Schreiner W (2008) Side chain substitution benchmark for peptide/MHC interaction. Protein Sci 17(6):977–982. doi:10.1110/ps.073402508
18. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12(9):2001–2014. doi:10.1110/ps.03154503
19. Xu J (2005) Rapid side-chain prediction via tree decomposition. RECOMB 3500:423–439
20. Hartmann C, Antes I, Lengauer T (2007) IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci 16(7):1294–1307. doi:10.1110/ps.062658307
21. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7:306–317
22. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V (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(22):4650–4658. doi:10.1021/jm9910775
23. Wang R, Lai L, Wang S (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des 16:11–26. doi:10.1023/A:1016357811882
24. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semi-empirical free energy force field with charge-based desolvation. J Comput Chem 28(6):1145–1152. doi:10.1002/jcc.20634

25. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
26. Spearman C (1904) The proof, measurement of association between two things. By C. Spearman, 1904. Am J Psychol 100(3–4): 441–471
27. Roc-macro (2008) Nonparametric comparison of areas under correlated ROC curves. SAS website 2008 July 16. Available from http://support.sas.com/kb/25/017.html. Cited 2008 Jul 16
28. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3): 837–845. doi:10.2307/2531595
29. Kates L, Petzoldt T proto (2007) An R Package for Prototype Programming. http://cran.r-project.org/web/packages/proto/. Accessed 2 Oct 2008