Research Article

Prediction of Peptide Binding to Major Histocompatibility II Receptors with Molecular Mechanics and Semi-Empirical Quantum Mechanics Methods

Sarah Aldulaijan and James A. Platts

School of Chemistry, Cardiff University, Park Place, Cardiff CF10 3AT, UK.

Received on November 25, 2011; Accepted on January 23, 2012; Published on February 15, 2012

Correspondence should be addressed to James A. Platts; Phone: +44 2920 874950, Fax: +44 2920 874030, Email: platts@cf.ac.uk

Abstract

Methods for prediction of the binding of peptides to major histocompatibility complex (MHC) II receptors are examined, using literature values of IC_{50} as a benchmark. Two sets of IC_{50} data for closely structurally related peptides based on hen egg lysozyme (HEL) and myelin basic protein (MBP) are reported first. This shows that methods based on both molecular mechanics and semi-empirical quantum mechanics can predict binding with good-to-reasonable accuracy, as long as a suitable method for estimation of solvation effects is included. A more diverse set of 22 peptides bound to HLA-DR1 provides a tougher test of such methods, especially since no crystal structure is available for these peptide-MHC complexes. We therefore use sequence based methods such as SYFPEITHI and SVMHC to generate possible binding poses, using a consensus approach to determine the most likely anchor residues, which are then mapped onto the crystal structure of an unrelated peptide bound to the same receptor. This analysis shows that the MM/GBVI method performs particularly well, as does the AMBER94 forcefield with Born solvation model. Indeed, MM/GBVI can be used as an alternative to sequence based methods in generating binding poses, leading to still better accuracy.

Introduction

Major Histocompatibility Complex (MHC) molecules are an important class of receptor in the immune system of all vertebrates: in humans they are termed human leukocyte antigens (HLA). Their role is to bind peptides presented to cell surfaces, hence allowing recognition of self or non-self and stimulating appropriate immune response in the case of non-self. MHC receptors are generally separated into class I and class II. Both have a single peptide binding site, which in class I is made up of a single amino-acid chain, whereas in class II the active site is located at the junction between two chains (Mantzourani *et al.* 2005, Wearsch & Cresswell 2008). Incorrect recognition of self peptides as being non-self is implicated in a number of auto-immune diseases such as multiple sclerosis and rheumatoid arthritis. The exact mechanism of this is not known but the concept of "molecular mimicry", in which certain self-peptide sequences are sufficiently similar to non-self sequences to induce immune attack on the body, has been proposed. Prediction of the key binding event between peptide and MHC is therefore desirable, both in understanding the origin of these debilitating diseases and in design of new therapies to treat them. Figure 1 shows a peptide bound to MHC II, taken from PDB entry 1YMM.

In order to understand the way that a peptide or drug interacts with its receptor to affect the biological system in the body's cells, we must concentrate on the interactions between the drug and the receptor (Mantzourani *et al.* 2008, Meyer *et al.* 2003, Zhao & Truhlar 2007). In most cases, the most significant interactions between drugs and their biological receptors are non-covalent (Cerny & Hobza 2007). Although typically weaker than covalent interactions, collectively they exert important influence in many properties of biomacromolecules, for example they are well known to affect the structure of proteins, DNA and RNA (Eistner *et al.* 2001, Grimme 2004, Jurecka *et al.* 2006a, b).

Accurate and efficient theoretical description of non-covalent interactions is an intense and ongoing area of research (Hobza *et al.* 1997, McNamara &

Hillier 2007, Sharma *et al.* 2008). *Ab initio* and density functional theory (DFT) methods can give quantitative accuracy, but are not generally applicable to large systems such as those of interest here. Semi-empirical methods offer speed and simplicity, making them appropriate for study of large systems (Anisimov *et al.* 2011, Eistner *et al.* 2001, McNamara & Hillier 2007, MRocha *et al.* 2006, Tuttle & Thiel 2008), but typically perform poorly for non-covalent interactions, especially stacking (Eistner *et al.* 2001). Addition of a dispersion correction term improves performance: AM1-D and PM3-D give errors of 1.1 and 1.2 kcal/mol, respectively, across a wide range of interactions (McNamara & Hillier 2007). More recent developments in semi-empirical methods include RM1 (Puzyn *et al.* 2008, Rocha *et al.* 2006, Stewart 2007) and PM6 (Stewart 2007), which encompasses many more elements within self-consistent set of parameters, and performs well for many classes of compound. PM6-DH2 is a further development of PM6 to include corrections for the dispersion and H-bond interactions (Korth *et al.* 2010b, Rezac *et al.* 2009). This method succeeds in calculating hydrogen bond energies with accuracy close to DFT-D approach, but is three orders of magnitude faster (Korth 2010, Korth *et al.* 2010a). The applicability of semi-empirical methods to large systems is further enhanced by the MOZYME method, using localized molecular orbital instead of the standard SCF procedure, implemented in current versions of MOPAC (http:/OpenMOPAC.net). By using the MOZYME method, studying large systems such as entire drug-receptor complexes is feasible.

Atomistic force field, or molecular mechanics (MM), methods are widely used in simulation of biological systems by reducing the essentials of systems of interest to simple mathematical forms. Non-covalent interactions are typically treated by a combination of point charges, to account for electrostatics, and Lennard-Jones potentials, for dispersive and repulsive interactions. More than a decade ago, Hobza *et al.* (1997) showed that the force field of Cornell et al (often referred to as AMBER) best reproduced *ab initio* data for interaction of DNA base pairs. More recently, Paton and Goodman showed that the OPLS-AA force field performs well for binding energy prediction of both hydrogen bonding and dispersion-bound complexes (Paton & Goodman 2009).

Because peptide-receptor interactions always occur in biological solvent (Klamt 1994), and in order to estimate the interaction energies for these complexes in appropriate ways, solvent must be considered in calculations. Calculating interaction energies for large biological complexes in solvent by computational methods is a challenging task (Klamt 1994). Many approaches have been tested to estimate the effect of the solvent in these interactions (Klamt 1994). Conductor-like Screening Model (COSMO) is widely used to model solvents, especially water (Anisimov & Cavasotto 2011, Klamt 1994, Klamt & Schuurmann 1993). This method depends on generation of a conducting surface at vdW distance in order to calculate the dielectric screening charges and energies (Klamt 1994).

The generalized Born model/surface area approach (GB/SA) is another method used to calculate binding free energy, developed by Still et al (Labute 2008a, Qiu *et al.* 1997, Still, *et al.* 1990), and is widely used in calculating free energy of binding for ligand-receptor complexes (Anisimov & Cavasotto 2011, Zoete *et al.* 2010, Zoete and Michielin 2007). In this method, cavitation energy depends on molecular surface area, while relative solvation of separated ligand and receptor compared to their complex is estimated from a generalization of the Born model. When combined with MM methods for calculation of electrostatic and van der Waals interactions, these are referred to as MM-GB/SA methods. The GB/VI (generalized Born/

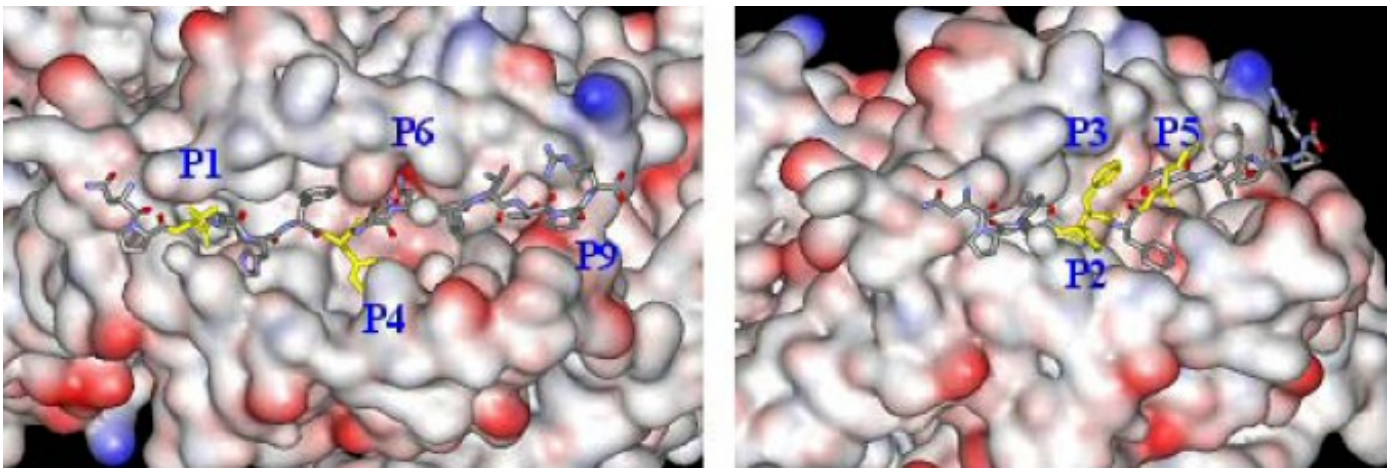

Figure 1. Two views of an epitope of myelin basic protein bound to an HLA receptor. P1 and P4 are the main "anchor" residues, while P3 and P5 are contacts to T-cell receptors.

volume integral) model, implemented in recent versions of MOE software (http://www.chemcomp.com), is similar to GB/SA in most respects, but calculates the cavitation energy in as an integral over molecular volume rather than surface area (Labute 2008a, b). MM/GB-VI is a fast and promising method to calculate the interaction energy in solvent. There are many advantages of using this method, the dielectric constant of the solvent is estimated based on the atoms (http://www.chemcomp.com/) present in the specific complex under study, rather than on an idealised values. In addition, this method yields an estimate of binding free energy, unlike all other methods used here that give only interaction energies. The change in entropy on binding is not explicitly included in MM/GB-VI: it has previously been shown that although entropy is essential in calculating absolute binding free energy (Zoete & Michielin 2007) it is not essential for estimating the relative binding free energy (Gohlke *et al.* 2003; Wang & Kollman 2000), a conclusion we would like to test for the flexible peptide ligands used in this study.

In a previous study (Aldulaijan & Platts 2010), several approximate methods were tested against correlated *ab initio* calculations for their ability to predict the energy of interaction between amino acids, focussing on the interaction of a MBP peptide with its MHC -II receptor. It was found that the semi-empirical RM1 approach with additional correction for dispersion effects gives the best reproduction of *ab initio* data, with a mean unsigned error of a little more than 1 kcal/mol over almost 50 interactions after optimisation of the global scaling factor. Performance is similar for several other parameterisations of semi-empirical theory, with RM1 chosen for its slightly better results. The atomistic forcefield OLPS-AA also shows promise, with a mean error slightly greater than 2 kcal/mol.

The speed of semi-empirical methods allows examination of the interaction energies of larger models of the peptide than single amino acids, especially when coupled with the MOZYME method. Therefore, we sought biological data to compare against the computational methods, in order to choose the most suitable method for predicting peptide-receptor interactions. IC_{50} data, *i.e.* the concentration required to inhibit 50% of binding of natural peptide in competitive binding, are widely used in such cases (Harrison *et al.* 1997). Although it is possible to convert IC_{50} to inhibition constant (Ki), which is directly related to binding free energy, using the Cheng-Prusoff equation (Cheng & Prusoff 1973), we were not able to perform this conversion for the peptide-MHC II complexes under study, due to lack of information about ligand and receptor concentrations in literature data. IC_{50} values are sensitive to conditions such as the temperature and solvent (www.bdbiosciences.com; Tajkhorshid 1998), it is therefore preferable to choose sets of IC_{50} data measured in a consistent manner in the same laboratory. We have therefore concentrated on several sets of peptide-MHCII receptor complexes with IC_{50} values measured in the same conditions, and with related X-ray structures published, and used these as tests of possible methods for prediction of peptide-MHCII binding using a variety of statistical techniques. We employed many of the methods discussed above (molecular mechanics methods OPLS-AA, AMBER94, and MM/GB-VI and semi-empirical RM1-D and PM6-DH2) to examine in detail the interaction of three sets of peptides with Major Histocompatibility Complex (MHC) class II receptor, and to compare calculated binding energies to available IC_{50} data.

Data Sets and Computational Methods

The first set studied is derived from hen egg lysozyme (HEL), and is based on the complex of 12 amino acids (MKRHGLDNYRGY) with MHC class II mouse I-Ag7 (Harrison *et al.* 1997). The X-ray structure of the peptide-receptor complex was taken from PDB entry 1F3J. IC_{50} data has been reported for analogues of the HEL peptide, in which one or more N-terminal and/or C-terminal residues are truncated to reveal the key residues for binding. IC_{50} values of 1000nM or more are denoted non-binders, and IC_{50} less than 1000nM are binders (Harrison *et al.* 1997). This set therefore contains 5 binders and 5 non-binders.

The second set studied is based on a complex of myelin basic protein (MBP)-derived peptide with HLA DRB1*1501(Harrison *et al.* 1997, Krogsgaard *et al.* 2000). It contains fourteen amino acids (ENPVVHFFKNIVTP; Harrison *et al.* 1997, Krogsgaard *et al.* 2000), and the relevant X-ray structure was taken from PDB entry 1BX2. In this set, each amino acid is replaced in turn by Ala and values of IC_{50} measured (Krogsgaard *et al.* 2000). In this set, all the IC_{50} values show stable interactions according to the 1000 nM cutoff used above, and many interactions have the same value of IC_{50}. However, two peptides have rather higher IC_{50} values, in which Val89 and Phe92 are replaced by Ala. Both amino acids are known to form strong interactions, in pocket 1 and pocket 4 of the binding site, respectively (Aldulaijan & Platts 2010, Harrison *et al.* 1997; Krogsgaard *et al.* 2000) and by replacing these amino acids with Ala, the binding affinity of the peptides decreased (Krogsgaard *et al.* 2000).

A third was taken from Southwood *et al*'s study (1998), and contains 22 peptides with much more diverse sequences than the first two sets interact-

ing with HLA DRB1*0101. In this case, X-ray structures of complexes are not available. Instead, manual docking was performed by mutating amino acids to the relevant sequence in MOE, using the X-ray structure of human class II MHC protein HLa-DR1 in complex with the tight binding peptide A2 103-117, PDB code 1AQD (Murthy & Stern 1997) as a template. In order to guide this procedure, possible amino acids that could act as "anchors" within binding pockets were identified by means of sequence-based prediction methods SYFPEITHI and SVMHC, as well as the algorithm set out by Southwood *et al.* (1998).

SYFPEITHI is a databank and prediction algorithm for peptide-MHC binding, and contains a large range of ligands and peptide motifs, used to predict the peptide binding with MHC receptor (http:/ alkaid001.atspace.com; Jalkanen, *et al.* 2004, Rommensee *et al.* 1999) based on published motifs of amino acids and anchor positions. It calculates a score to identify the amino acid as anchor, auxiliary anchor, preferred residues or if the amino acid has "negative effect on the binding ability" (Jalkanen *et al.* 2004). SVMHC is a prediction server for MHC class I and II, used to test the ability of peptides to bind with different MHC alleles, and to find the best "binders in a protein sequence" (Donnes & Kohlbacher 2006). According to Donnes and Elofsson, the performance of SVMHC and SYFPEITHI for six MHC types common between these methods are compared (Donnes & Kohlbacher 2006), with SVMHC giving 95% correct predictions and 91% for SYFPEITHI (Donnes & Elofsson 2002). The final sequence-based prediction method used is the algorithm set out by Southwood *et al. (*1998), which is specific for the DRB1*0101 allele MHC class II. Each residue has value based on its position on the receptor, encoded into an in-house awk program to evaluate the most likely binding sites of the peptide based on these values.

The X-ray crystallographic coordinates were obtained from PDB entry 1BX2 for MBP peptide and 1F3J for HEL peptide (http://www.chemcomp.com; Harrison *et al.* 1997, Labute 2008a; MOE). For the Southwood data set, three prediction methods (SYFPEITHI, SVMHC and Southwood) were used to identify the best peptide anchors that fit in the receptor pockets. We choose the best alignment of peptide in the receptor as a consensus of these methods, and only this alignment was used in further study.

Coordinates were loaded into MOE, and protonated according to typical protonation states. All hydrogen positions were optimised using the AMBER94 forcefield, with heavy atoms fixed at their X-ray positions. MOE program was used to calculate interaction energies using OPLSAA and AMBER94 force fields with dielectric constant 1 (vacuum), 4, 20 and 78.4 (water) (Krogsgaard *et al.* 2000, Mantzourani *et al.* 2005). MOE was used to calculate interaction energies with the Born solvation model, and also binding free energies with the MM/GBVI method. For this method, the dielectric constant is estimated according to the atoms present in the receptor, and a constrained energy minimization performed for ligand atoms (Klamt & Schuurmann 1993).

MOPAC was used to carry out semi-empirical calculations, using the RM1-D tested in our previous study (Aldulaijan & Platts 2010) and also the recent PM6-DH2 method, incorporating corrections for both dispersion and hydrogen bonding (Korth *et al.* 2010b, Rezac *et al.* 2009). For larger systems we used MOZYME keyword to accelerate the calculations (Wearsch & Cresswell 2008). COSMO was used to estimate the effect on a solvent (http:/ OpenMOPAC.net; Stewart 2009), with the same values for dielectric constant noted above (Southwood *et al.* 1998) and NSPA (number of geometrical segments per atom; Labute 2008a) equal to 122.

Several statistical tests were used to investigate the suitability of different theoretical methods for prediction of peptide-MHCII bonding, using published IC_{50} values as a test. Specifically, we employed the standard Pearson R^2 value against the negative log of IC_{50} values, Spearman's rank correlation coefficient (http://www.wessa.net/rankcorr.wasp), and area under relative operating characteristic (ROC) curves by using the ROCkit package (Dorfman & Alf 1969, Metz *et al.* 1998). In each case, a value of 1.0 indicates the ideal of perfect prediction.

Results and Discussions

Table 1 reports IC_{50} and interaction energies from OPLS-AA, AMBER94, MM/GB-VI and RM1-D for the series of peptides based on HEL. According to Tsai (2002), the value 4 of the dielectric constant is suitable to be used in protein interaction, and so was employed here. Sequential removal of one to three residues from the C-terminus of the native peptide increases IC_{50}, a trend that is reflected in interaction energies from all methods considered. In contrast, removal of the N-terminal methionine residue actually increases potency: two of the four methods reflect this in increased binding, and the remaining two methods show only very small change in interaction energy. The shortest sequence, KRHGLDNY, is by some distance the least potent peptide in this data set, and again all methods predict weak binding for this peptide. From the GB-VI results, we can see that the binding energy is approximately additive: for example, removal of M from the

N-terminus of the peptide reduces binding energy by *ca*. 1 kcal/mol independently of the other residues present. Similarly, simultaneous removal of both M and Y from N- and C-termini reduces binding by 10.5 kcal/mol, a value that is very close to the sum of individual values (1.0 and 9.4 kcal/mol, respectively).

Statistical measures across the entire data set clarify the differences in methods. Plotting $\log(1/IC_{50})$ against interaction energy gives some correlation for all methods, but noticeably superior performance for MM/GBVI over others considered (Figure 2). The pattern is similar, but not as clear cut, when considering rank correlation, whether using raw or averaged data. ROC data shows that MM/GBVI and AMBER94/Born are able to unambiguously separate binders from non-binders with no false positive or negatives, whereas PM6-DH2/COSMO, RM1-D/COSMO and OPLS-AA/Born cannot. However, even those methods give high values, indicating that their predictive ability remains rather good.

Table 2 reports similar data for the data set consisting of peptides based on MBP. In this case, all but two peptides are quite strongly bound to the receptor, and also exhibit very low IC_{50} values. The two exceptions to this are for mutation of Val89 and Phe92, which are well-known to be important as "anchor residues": mutation of these into alanine significantly increases IC_{50} values. All methods considered predict that the F92A mutation is particularly weakly bound. The instability of the F92A mutant is most marked with the forcefield method: OPLS-AA predicts that this peptide is not bound at all to the receptor. In contrast, the semi-empirical methods succeed in predicting the relatively weak binding of the V89A mutant, whereas with force field methods this mutant does not stand out as being more weakly bound than other peptides.

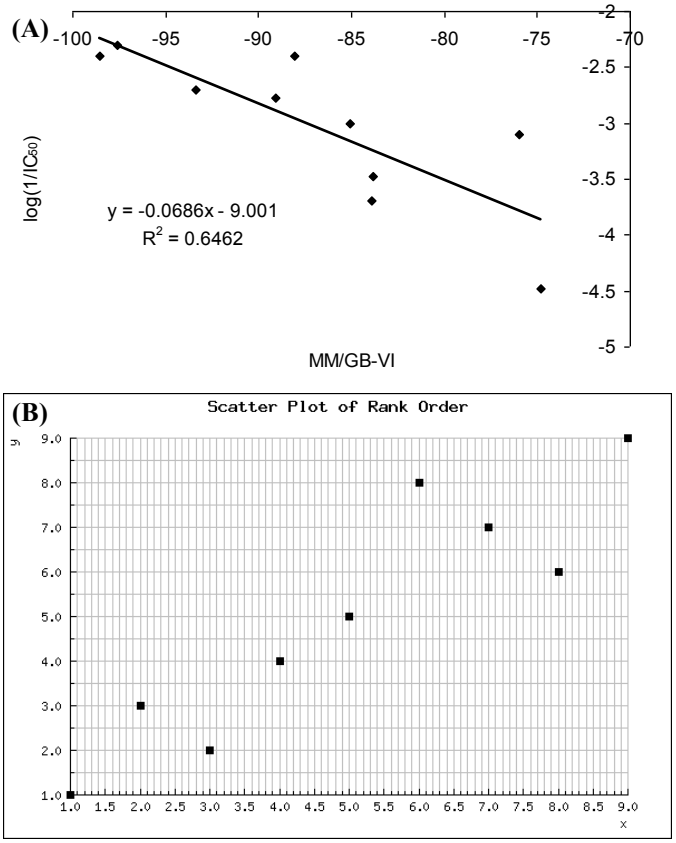

Figure 2. (A) Linear and (B) rank correlations from MM/GB-VI data for HEL data set. ROC curve not shown due to perfect prediction.

Table 1: IC_{50} values (nM) and interaction energies (kcal/mol) for HEL along with R^2, rank correlation, and ROC area for each method.

Peptide	IC_{50}	MM/ GBVI	RM1-D/ COSMO	PM6-DH2/ COSMO	OPLS-AA/ Born	AMBER94/ Born
MKRHGLDNYRGY	**250**	**-98.55**	**-214.89**	**-347.72**	**-129.53**	**-130.69**
MKRHGLDNYRG	600	-89.11	-205.87	-344.21	-119.61	-120.62
MKRHGLDNYR	1000	-85.05	-188.91	-312.20	-112.30	-111.85
MKRHGLDNY	1250	-75.98	-145.17	-293.10	-99.28	-95.17
KRHGLDNYRGY	200	-97.57	-209.74	-337.18	-138.25	-135.53
KRHGLDNYRG	**250**	**-88.03**	**-200.92**	**-334.00**	**-128.33**	**-125.49**
KRHGLDNYR	5000	-83.94	-183.75	-302.20	-121.03	-116.70
KRHGLDNY	30000	-74.83	-140.01	-282.69	-108.02	-100.01
RHGLDNYRGY	500	-93.37	-150.54	-303.55	-125.39	-120.58
RHGLDNYRG	3000	-83.83	-141.75	-300.10	-115.47	-110.54
R^2		0.65	0.46	0.65	0.43	0.55
Rank correlation		0.88	0.82	0.63	0.80	0.86
*(IC50-average) rank correlation		0.92	0.90	0.82	0.73	0.82
ROC area (cutoff 1000 nM)		1.00	0.93	0.96	0.96	1.00

* Rank correlation for the set after taking the average values for peptides in **bold** (IC_{50} = 250).

The R^2 statistic indicates reasonable correlation between IC_{50} and RM1-D interaction energy, a slightly worse correlation with MM/GBVI data, and poor correlations with OPLS-AA and AMBER94 data. Application of the rank correlation statistic is not straightforward for this data set, since four peptides have IC_{50} = 4 nM and a further four have IC_{50} = 10 nM. Therefore, we took the average energies for the peptides with IC_{50} = 4 nM and the average energies for the peptides with IC_{50} = 10 nM and used these averages on the calculation of the rank correlation of this set. The standard cutoff of 1000 nM to distinguish binders from non-binders for ROC analysis is inappropriate in this case.

While the results for HEL and MBP data sets is encouraging, the structural similarities and restricted range of IC_{50} data (particularly for MBP) mean that more stringent tests are required before we can reach any conclusions on the suitability of the methods examined. For this, we employed Southwood *et all's* (1998) set of 22 structurally diverse peptides bound with IC_{50} values ranging from below 2 to over 2000 nM. Initially, SYFPEITHI and SVMHC prediction servers, along with our own implementation of Southwood *et al*'s algorithm, were used to identify the best alignment for each peptide. This alignment was then constructed by manual mutation of PDB structure 1AQD in MOE, and energy minimized. The peptide on 1AQD PDB structure contains 14 residues, with the fourth residue located in pocket 1 of the receptor. So, in order to mutate this peptide to Southwood's peptides, we located the anchor residue in pocket 1 and then mutated the rest of the original peptide from 1AQD to that employed by Southwood *et al*. By using this technique we included the core residues of the peptides (located on pocket 1 to pocket 9 of the receptor, the importantly binders residues) and some more residues of the peptides to our calculations, but we missed few residues from each peptide on the set. On table 3 and 4, we underline the residues included in our calculations and identify the residue in pocket 1 in bold red.

These structures were then used to examine the performance of the methods discussed above in predicting binding energy for this more challenging set of data (Table 3). As in other data sets considered above, all methods clearly identify the peptide with the highest IC_{50} value, namely 27.415, as being particularly weakly bound. Indeed, MM/GBVI predicts this peptide not to be bound at all to the receptor. Across the entire set, statistical measures show promising performance for MM/GBVI and AMBER94/Born methods, with rather worse performance for OPLS-AA/Born and PM6-DH2/COSMO, and poor results from RM1-D/COSMO. The MM/GBVI R^2 value of 0.54 is more than 99.9% significant, and corresponds to a standard error for estimate of IC_{50} of 0.64 nM. The rank correlation coefficient of 0.78 indicates that this method puts almost 80% of peptides in the correct rank order. For ROC results, we used a cutoff of 50 nM to distinguish binders from non-binders, resulting in 11 peptides in each category, thereby giving a balanced test of predictions. The area under the ROC curve of 0.93 found using MM/GBVI is highly encouraging, indicating that very few false positives/negatives result from this approach. In contrast, the value of 0.62 for

Table 2: IC_{50} values (nM) and interaction energies (kcal/mol) for MBP along with R^2, rank correlation, and ROC area for each method.

Peptide	IC_{50}	MM/ GBVI	RM1-D/ COSMO	PM6-DH2/ COSMO	OPLS-AA/ Born	AMBER94/ Born
EAPVVHFFKNIVTP	7	-71.09	-20.38	-124.32	-12.86	-14.50
ENAVVHFFKNIVTP	**10**	**-70.00**	**-18.01**	**-125.35**	**-9.24**	**-15.18**
ENPAVHFFKNIVTP	**10**	**-68.85**	**-19.83**	**-127.46**	**-10.72**	**-18.02**
ENPVAHFFKNIVTP	50	-67.19	-15.66	-118.41	-6.64	-14.23
ENPVVAFFKNIVTP	**10**	**-65.74**	**-17.54**	**-124.17**	**-8.68**	**-13.08**
ENPVVHAFKNIVTP	**10**	**-67.28**	**-18.19**	**-124.50**	**-6.28**	**-13.20**
ENPVVHFAKNIVTP	199	-63.90	-13.41	-117.60	+0.05	-5.55
ENPVVHFFKAIVTP	4	-68.45	-17.22	-113.15	-29.96	-37.80
ENPVVHFFKNAVTP	4	-70.95	-20.91	-128.37	-14.70	-23.28
ENPVVHFFKNIATP	4	-69.25	-18.76	-127.31	-9.29	-16.37
ENPVVHFFKNIVAP	4	-69.35	-21.79	-128.18	-29.98	-34.72
R^2		0.57	0.69	0.22	0.46	0.45
*(IC50-average) rank correlation		1.00	0.90	0.60	1.00	0.90

* Rank correlation for the set after taking the average energies for peptides with IC_{50} = 4 nM (underlined) and for peptides with IC_{50} = 10 nM (**bold**).

Table 3: IC_{50} values (nM) and interaction energies (kcal/mol) for Southwood data set along with R^2, rank correlation, and ROC area for each method.

Peptide No.	Sequence*	IC_{50}	MM/ GBVI	RM1-D/ COSMO	PM6-DH2/ COSMO	OPLS-AA/ Born	AM-BER94/ Born
1188.34	HN**W**VNHAVPLAMKLI	14	-40.62	-85.53	-152.33	-143.45	-126.40
1188.16	KSK**Y**KLATSVLAGLL	3.7	-49.04	-182.02	-246.12	-138.83	-139.35
1136.47	THHY**F**VDLIGGAMLSL	2.2	-57.48	-26.04	-101.91	-145.89	-143.04
1188.32	GLA**Y**KFVVPGAATPY	3.1	-42.75	-105.60	-172.91	-129.11	-120.34
1136.16	LTSQF**F**LPALPVFTWL	1.6	-53.28	-71.09	-143.94	-148.22	-138.43
27.415	NVKYLV**I**VFLIFFDL	2011	+17.46	-4.54	-71.56	-93.76	-95.87
27.403	**L**VNLLIFHINGKIIK	78	-13.19	-77.91	-127.79	-168.31	-128.50
1136.21	IPQE**W**KPAITVKVLPA	2.2	-36.32	-130.90	-209.19	-132.11	-117.41
1136.28	LAAII**F**LFGPPTALRS	0.23	-53.05	-90.94	-161.75	-140.97	-135.16
1136.11	VVFPAS**F**FIKLPIILA	0.89	-59.32	-84.21	-146.19	-155.07	-137.81
1136.14	FATCF**L**IPLTSQFFLP	5.3	-24.52	-63.29	-136.09	-133.68	-131.21
1188.13	AGL**L**GNVSTVLLGGV	116	-28.96	-86.62	-149.38	-115.74	-102.50
1136.24	NLSNV**L**ATITTGVLDI	182	-25.61	-27.96	-96.12	-113.74	-96.22
1136.12	**I**KLPIILAFATCFLIP	105	40.92	-118.30	-150.80	-107.73	-98.42
27.392	SSV**F**NVVNSSIGLIM	41	-38.79	-51.92	-123.90	-133.96	-123.70
27.417	**V**KNVIGPFMKAVCVE	56	-53.73	-100.38	-152.69	-128.45	-126.36
1136.55	QEID**P**LSYNYIPVNSN	65	-11.14	-7.50	-78.95	-119.45	-102.80
1136.71	EPQGST**Y**AASSATSVD	5.1	-58.73	-16.20	-96.59	-127.66	-113.00
1136.38	**S**SIIFGAFPSLHSGCC	70	-8.49	-33.79	-85.15	-90.11	-85.43
27.388	**M**RKLAILSVSSFLFV	50	-13.22	-73.82	-125.30	-143.71	-128.80
1136.59.01a	**R**VYQEPQVSPPQRAET	130	+29.36	-28.23	-86.59	-94.42	-110.26
1136.46	LWWST**M**YLTHHYFVDL	68	-9.91	-106.31	-190.35	-135.71	-128.45
R^2			0.54	0.14	0.23	0.36	0.48
Rank correlation			0.78	0.29	0.48	0.66	0.74
ROC area			0.93	0.62	0.75	0.79	0.87

* The underlined residues are the residues which we included in our calculations and the residues on bold red are the residue which located on pocket one of the receptor.

RM1-D is only slightly higher than random.

Because of the encouraging performance of MM/GBVI, we then explored whether this method could be used to predict alignment of peptides within the receptor, rather than relying on purely sequence-based methods. To do this, numerous potential binding poses were generated with SYFPEITHI and SVMHC algorithms, and the one with the most negative MM/GBVI interaction energy selected. In 20 of the 22 cases, this agreed with the results from sequence-based prediction methods, but for two peptides (nos. 1136.14 and 1136.16) a lower energy alternative was found from this analysis. For the 1136.14, MM/GBVI predicts Thr as the anchor residue instead of Leu, and for 1136.16, the MM/GBVI predicts Gln as the anchor residue instead of Phe. This is illustrated in Figure 3 for 1136.16: as might be expected, sequence-based predictions place Phe in the hydrophobic environment of pocket 1. However, this leads to placement of Ala into pocket 4, Pro in pocket 6 and Thr in pocket 9, none of which are particularly favourable for binding. With Gln as the residue in pocket 1, a hydrogen bond can form to the side-chain carbonyl (Figure 3, bottom left). In addition, this alignment places Leu in pocket 4, Ala in pocket 6 and Val in Pocket 9, all of which contribute to favourable binding. Comparison of Tables 3 and 4 shows that the second alignment has almost 10 kcal/mol greater binding energy, despite the apparent anomaly of a having relatively polar residue in the hydrophobic pocket 1.

Using the new values for these two peptides improves all statistical tests slightly, as shown in Table 4. MM/GBVI data shows small increases in R^2 and rank correlation coefficient, with plots corresponding to these data shown in Figure 4. The area under the ROC curve increases from 0.93 to 0.96, again illustrated in Figure 4. The statistics from other methods are barely affected by this change. Thus, we conclude that MM/GBVI interaction energies are a useful addition to sequence-only methods of prediction of peptide -MHC-II binding alignments.

Conclusion

We have tested several methods to calculate the interaction energy for peptide-MHC-II complexes for three separate data sets, using IC_{50} data to evaluate the accuracy of each theoretical method. We show that MM/GBVI approach is a promising way to calculate the free energy for peptide-receptor complexes, with reliable performance for all three data sets as measured by three distinct statistical tests. For two data sets where peptides are closely related, HEL and MBP, excellent performance is evident from these statistics, with

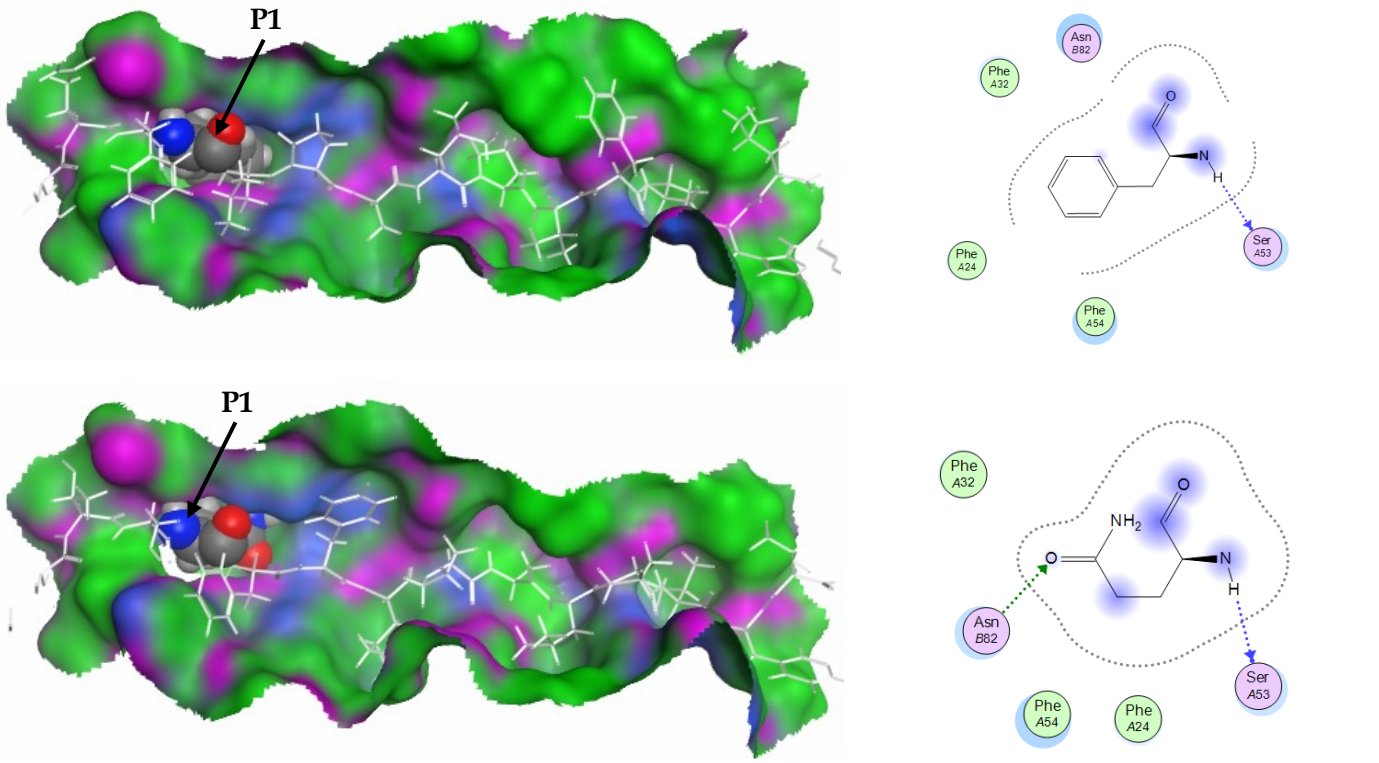

Figure 3 3D and 2D ligand interaction views of two possible alignments of peptide 1136.16 in HLA-DR1. Top: Phe in pocket 1; Bottom: Gln in pocket 1. On the left, the MHC receptor is shown as a continuous surface, the residue in pocket 1 as space-filling CPK spheres, and the remainder of the peptide as white wireframe. On the right, blue-shading of the peptide residue indicates exposed atoms.

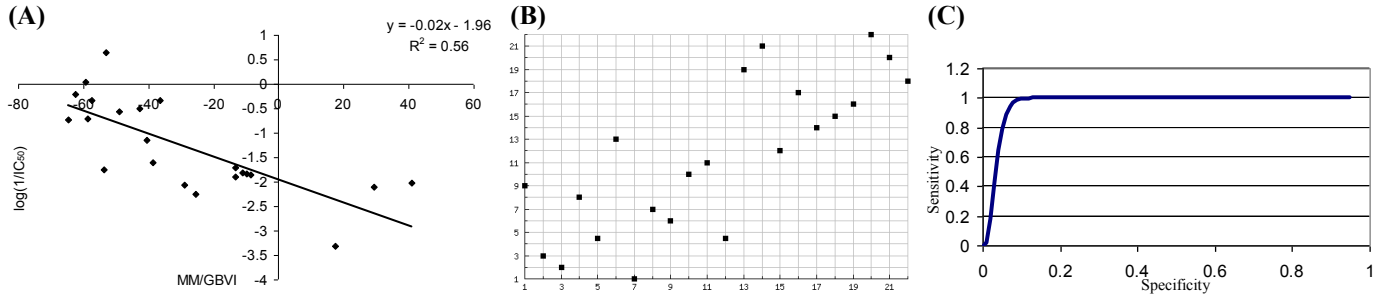

Figure 4 (A) Linear correlation, (B) rank correlation and (C) ROC curve from MM/GBVI data for Southwood data set.

strongly significant correlation between interaction energy and log(1/IC_{50}) good or perfect ranking of activity, and no false negatives/positives. AMBER94 with a Born model of solvation performs almost as well, while OPLS-AA/Born and RM1-D/COSMO give rather worse performance. MM/GBVI also performs well for the more diverse set of peptides contained in the Southwood data set despite the lack of entropy contributions to these calculations, apparently confirming that such contributions are not required in evaluation of relative binding free energies even for ligands as flexible as peptides.

We also show that this method can be used to predict the anchor residues that reside in receptor binding pockets, and that this approach gives slight improvement in statistics over purely sequence-based

Table 4: IC_{50} values (nM) and interaction energies (kcal/mol) for Southwood data set from MM/GBVI alignment along with R^2, rank correlation, and ROC area for each method.[a]

Peptide No.	Sequence*	IC50	MM/ GBVI	RM1-D/ COSMO	OPLS-AA/ Born	AMBER94/ Born
1188.34	HN**W**VNHAVPLAMKLI	14	-40.62	-85.53	-143.45	-126.40
1188.16	KSK**Y**KLATSVLAGLL	3.7	-49.04	-182.02	-138.83	-139.35
1136.47	THHY**F**VDLIGGAMLSL	2.2	-57.48	-26.04	-145.89	-143.04
1188.32	GLA**Y**KFVVPGAATPY	3.1	-42.75	-105.60	-129.11	-120.34
1136.16	LTS**Q**FFLPALPVFTWL	**1.6**	**-62.43**	**-68.14**	**-146.77**	**-131.54**
27.415	NVKYLV**I**VFLIFFDL	2011	17.46	-4.54	-93.76	-95.87
27.403	**L**VNLLIFHINGKIIK	78	-13.19	-77.91	-168.31	-128.50
1136.21	IPQE**W**KPAITVKVLPA	2.2	-36.32	-130.90	-132.11	-117.41
1136.28	LAAII**F**LFGPPTALRS	0.23	-53.05	-90.94	-140.97	-135.16
1136.11	VVFPAS**F**FIKLPIILA	0.89	-59.32	-84.21	-155.07	-137.81
1136.14	FA**T**CFLIPLTSQFFLP	**5.3**	**-64.73**	**-66.80**	**-140.92**	**-130.17**
1188.13	AGL**L**GNVSTVLLGGV	116	-28.96	-86.62	-115.74	-102.50
1136.24	NLSNV**L**ATITTGVLDI	182	-25.61	-27.96	-113.74	-96.22
1136.12	**I**KLPIILAFATCFLIP	105	40.92	-118.30	-107.73	-98.42
27.392	SSV**F**NVVNSSIGLIM	41	-38.79	-51.92	-133.96	-123.70
27.417	**V**KNVIGPFMKAVCVE	56	-53.73	-100.38	-128.45	-126.36
1136.55	QEID**P**LSYNYIPVNSN	65	-11.14	-7.50	-119.45	-102.80
1136.71	EPQGST**Y**AASSATSVD	5.1	-58.73	-16.20	-127.66	-113.00
1136.38	**S**SIIFGAFPSLHSGCC	70	-8.49	-33.79	-90.11	-85.43
27.388	**M**RKLAILSVSSFLFV	50	-13.22	-73.82	-143.71	-128.80
1136.59.01a	**R**VYQEPQVSPPQRAET	130	29.36	-28.23	-94.42	-110.26
1136.46	LWWST**M**YLTHHYFVDL	68	-9.91	-106.31	-135.71	-128.45
R^2			0.56	0.14	0.36	0.47
Rank correlation			0.79	0.29	0.66	0.74
ROC area			0.96	0.62	0.80	0.87

[a] Alignments that differ from Table 3 shown in bold.

* The underlined residues are the residues which we included in our calculations and the residues on bold red are the residue which located on pocket one of the receptor.

prediction methods such as SYFPEITHI or SVMHC. The accuracy of the MM/GBVI approach may stem from the fact that the dielectric constant employed is estimated from the atoms present in the specific complex under study, rather than on an idealised value, or from the use of constrained optimisation that allows ligand and some receptor flexibility while keeping the overall binding mode fixed. Of course, both peptide ligand and protein receptor are flexible objects, such that the single snapshots used here can only be approximations of the entire binding event. We are currently exploring the use of molecular dynamics to calculate MM-GB/SA averaged over multiple snapshots, and will report the results in a future publication. For now, we have shown that the MM/GBVI approach can deliver reasonable predictions of peptide-MHC binding in a matter of a few seconds on a desktop computer.

Acknowledgements

SA thanks the Saudi Arabian government for funding. JAP is grateful to the Leverhulme Trust for a Research Fellowship.

References

Aldulaijan S & Platts JA 2010 Theoretical prediction of a peptide binding to major histocompatibility complex II *Journal of Molecular Graphics and Modelling* **29** 240-245

Anisimov V, Ziemys A, Kizhake S, Yuan Z, Natarajan A & Cavasotto C 2011 Computational and experimental studies of the interaction between phospho-peptides and the C-terminal domain of BRCA1. *Journal of Computer-Aided Molecular Design* **25** 1071-1084.

Anisimov VM & Cavasotto CN 2011 Quantum mechanical binding free energy calculation for phosphopeptide inhibitors of the Lck SH2 domain. *Journal of Computational Chemistry* **32** 2254-2263.

Cerny J & Hobza P 2007 Non-covalent interactions in biomacromolecules *Physical Chemistry Chemical Physics* **9** 5291-5303

Cheng Y & Prusoff W 1973 Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharmacol* **22** 3099-3108.

Donnes P & Elofsson A 2002 Prediction of MHC class I binding peptides using SVMHC. *bioinformatics* **3**.

Donnes P & Kohlbacher O 2006 SVMHCL a server for prediction for MHC-binding peptides. *Nucleic acids research* **34** web server issue.

Dorfman DD & Alf E 1969 Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals - rating method data. *Journal of Mathematical Psychology* **6** 487-496.

Eistner M, Hobza P, Frauenheim T, Suhai S & Kaxiras E 2001 Hydrogen bonding and stacking interactions of nucleic base pairs: A density-functional-theory based treatment. *Journal of Chemical physics* **114** 5149-5155.

Gohlke H, Kiel C & Case DA 2003 Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *Journal of Molecular Biology* **330** 891 -913.

Grimme S 2004 Accurate description of van der waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* **25** 1463-1473.

Harrison LC, Honeyman MC, Termbleau S, Gregori S, Gallazzi F, Augstein P, Brusic V, Hammer J & Adorini L 1997 A Peptide-binding Motif for I-Ag7, the Class II Major Histocompatibility Complex (MHC) Molecule of NOD and Biozzi AB/H Mice. *Journal of Experimental Medicine* **185** 1013-1021.

Hobza P, Kabelac M, Sponer J, Mejzlik P & Vondrasek J 1997 Performance of empirical potentials (AMBER, CFF95, CVFF, CHARMM, OPLS, POLTEV), semiempirical quantum chemical methods (AM1, MNDO/M, PM3), and ab initio Hartree-Fock method for interaction of DNA bases: Comparison with nonempirical beyond Hartree-Fock results *Journal of Computational Chemistry* **18** 1136-1150

Jalkanen KJ, Elstner M & Suhai S 2004 Amino acids and small peptides as building blocks for proteins: comparative theoretical and spectroscopic studies *Journal of Molecular Structure: THEOCHEM* **675** 61-77.

Jurecka P, Cerny J, Hobza P & Salahub D 2006a Density functional theory augmented with empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with Ab initio quantum mechanics calculations. *Journal of Computational Chemistry* **28** 555-569.

Jurecka P, Sponer J, Cerny J & Hobza P 2006b Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies for small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **8** 1985-1993.

Klamt A 1994 Conductor-like Screening Model for Real Solvent: A New Approach to the Quantitative Calculation of Solvation Phenomena. *Journal of Physical chemistry* **99** 2224-2235.

Klamt A & Schuurmann G 1993 COSMO: A new approch to dielectric screening in solvent with explicit experessions for the screening energy and its gradient. *Journal of Chemical Society Perkin Transactions* **2** 799-805.

Korth M 2010 Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields *Journal of Chemical Theory and Computation* **6** 3808-3816.

Korth M, Pitonak M, Rezac J & Hobza P 2010a A Transferable H-bonding correction for semiempirical quantum-chemical methods. *Journal of chemical theory and computation* **6** 344-352.

Korth M, Pitonak M, Rezac J & Hobza P 2010b A Transferable H-bonding correction for semiempirical quantum-chemical methods. *Journal of chemical theory and computation* **6** 344-352.

Krogsgaard M, Wucherpfenning KW, Canella B, Hansen

BE, Svejgraad A, Pyrdol J, Ditzel H, Raine C, Engberg J & Fugger L 2000 Visualization of Myelin Basic Protein (MBP) T Cell Epitopes in Multiple Sclerosis using a Monoclonal Antibody Specfic for the Human Histocompatibility Leukocyte Antigen (HILA)-DR2-MBP 85-99 Complex. *Journal of Experimental Medicine* **191** 1395-1412.

Labute P 2008a The Generalized Born/Volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area. *Journal of Computational Chemistry* **29** 1693-1698.

Labute P 2008b The Generalized Born/Volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area. *Journal of computaional chemistry* **29** 1693-1698.

Mantzourani E, Laimou D, Matsoukas MT & Tselios T 2008 Peptides as Therapeutic Agents or Drug Leads for Autoimmune, Hormone Dependent and Cardiovascular Diseases *Anti-Inflammatory & Anti-Allergy Agents in Medicinal Chemistry* **7** 294-306

Mantzourani ED, Mavromoustakos TM, Platts JA, Matsoukas JM & Tselios TV 2005 Structural requirements for binding of Myelin Basic Protein (MBP) peptides to MHC II: Effects on immune regulation. *Current Medicinal Chemistry* **12** 1521-1535

McNamara JP & Hillier IH 2007 Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules *Physical Chemistry Chemical Physics* **9** 2362-2370

Metz CE, Herman BA & Shen J-H 1998 Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Stat. Med. Journals* **17** 1033-1053.

Meyer EA, Castellano RK & Francois Diederich 2003 Interactions with Aromatic Rings in Chemical and Biological Recoqnition. *Angewandte chemie* **42** 1210-1250

Murthy VL & Stern LJ 1997 The class II MHC protein HLA -DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding. *Structure* **5** 1385-1397.

Paton RS & Goodman JM 2009 Hydrogen Bonding and pi-stacking: How reliable are Force Fields? A critical evaluation of force field descriptions of non-bonded interactions *Journal of Chemical Information and Modelling* **49** 944-955.

Puzyn T, Suzuki N, Heranczyk M & Rak J 2008 Calculation of Quantum-Mechanical Descriptor for QSPR at the DFT level: Is it Necessary? *Journal of chemical information and modeling* **48** 1174-1180.

Qiu D, Shenkin PS, Hollinger FP & Still WC 1997 The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *Journal of physical chemistry A* **101** 3005-3014.

Rezac J, Fanfrlik J, Salahub D & Hobza P 2009 Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-bonding Correction Terms Reliably describes Various Types of Noncovalent Complexes. *Journal of Chemical Theory and Computation* **5** 1749-1760.

Rocha GB, Freire RO, Simas AM & Stewart JJP 2006 RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, and I. *Journal of Computational Chemistry* **27** 1101-1111.

Rommensee H-G, Bachmann J, Emmerich NP, Bachor OA & Stevanovic S 1999 SYFPEITHI: database for MHC ligands and peptides motifs. *Immunogenetics.* **50** 213-219.

Sharma R, McNamara JP, Raju RK, Vincent MA, Hillier IH & Morgado CA 2008 The interaction of carbohydrates and amino acids with aromatic systems studied by density functional and semi-empirical molecular orbital calculations with dispersion corrections *Physical Chemistry Chemical Physics* **10** 2767-2774

Southwood S, Sidney J, Kondo A, Guercio M-Fd, Appella E, Hoffman S, Kubo RT, Chesnut RW, Gery HM & Sette A 1998 Several common HLA-DR types share largely overlapping peptide binding repertoires. *The Journal of Immunology* **160** 3363-3373.

Stewart JJP 2007 Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular modeling* **13** 1173-1213.

Still WC, Tempczyk A, Hawley RC & Hendrickson T 1990 Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **112** 6127-6129.

Tajkhorshid EJ, K. J; Suhai, S 1998 Structure and Vibrational Spectra of the Zwitterion l-Alanine in the Presence of Explicit Water Molecules:? A Density Functional Analysis *Journal of Physical chemistry B* **102** 5899-5913.

Tsai CS 2002 *An introduction to computational biochemistry.*

Tuttle T & Thiel W 2008 OMx-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application. *Physical Chemistry Chemical Physics* **10** 2159-2166.

Wang W & Kollman PA 2000 Free Energy Calculations on Dimer Stability of the HIV Protease using Molecular Dynamics and a Continuum Solvent Model. *Journal of Molecular Biology* **303** 567-582.

Wearsch PA & Cresswell P 2008 The quality control of MHC class I peptide loading *Current Opinion in Cell Biology* **20** 624-631

Zhao Y & Truhlar D 2007 Density functionals for noncovalent interaction energies of biological importance *Journal of Chemical Theory and Computation* **3** 289-300

Zoete V, Irving M & Michielin O 2010 MM-GBSA binding free energy decomposition and T cell receptor engineering. *Journal of Molecular Recognition.* **23** 142-152.

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** 1026-1047.