SCORING PEPTIDE(MIMETIC)-PROTEIN INTERACTIONS

E.E. Moret*, M.C. van Wijk, A.S. Kostense, M.B. Gillies

Department of Medicinal Chemistry, Utrecht Institute for Pharmaceutical Sciences, Faculty of Pharmacy, PO Box 80082, 3508 TB Utrecht, The Netherlands

ABSTRACT. Methods for affinity prediction do not perform well in the case of peptide-protein binding. Here we investigate the cause for this failure. We conclude that the affinity of a small set of complexes can be described well by sidechain steric interactions and the number of buried polar groups, but only if outliers are removed. In the PLS models, particularly the hydrogen bonding of the peptide backbone plays an anomalous role, explaining the inability of general scoring methods to predict peptide-protein binding.

Although structural biology has helped tremendously in understanding and rationalising drug action, it should be noted that the realm of molecular structure is still far removed from in vitro pharmacology and even further from in vivo experiments. We would like to predict the affinity of a drug for a receptor solely based on the structure of its complex. Such a prediction of affinity is hampered by the fact that we tend to look at single ligands complexed to single targets. We neglect the events before and after this complexation. We typically neglect multimolecule effects, solvent, hydrophobicity, induced fit, entropy and differences between the crystal and physiological state.

Unfortunately scoring is the Achilles heel of modelling. The development of reliable and efficient general-purpose scoring methods shows few signs of imminent breakthrough, a situation unchanged from that described in the review of Ajay and Murcko.[1] Techniques based rigorously on the principles of statistical thermodynamics,[2] such as Free Energy Perturbation,[3] require intractably long-running simulations even for relatively simple

systems. A search for less computationally demanding methods has led to a number of approaches relying on use of static properties of the ligand-receptor complex, rather than ensemble averages derived from simulations. The relationship between these properties and affinity is described either by regression techniques or a master-equation which is intended to include terms for the most important contributions to the net free energy change upon binding.

The SCORE method of Böhm[4,5] is a general purpose regression-based technique. Noteworthy characteristics are terms which are intended to account for entropic contributions to affinity due to loss of degrees of freedom on binding, and hydrophobic effects. Expressions of a similar form are used in some master equation approaches,[6] while further approaches rely on the analysis of interaction energies derived from molecular mechanics calculations.[7] Despite efforts in this direction as seen in SCORE, most of these techniques appear to be inadequate in their treatment of entropic effects, and more specifically solvation and molecular flexibility. Nevertheless, there are some notable successes, at least for protein-protein interaction[8] and ligand-protein binding.[9]

Many processes in biology and immunology are based on the interaction of peptides with proteins. These interactions often play a role in infectious and chronic diseases. Peptides, however, are not normally used for therapy, because of their instability. Peptidomimetic compounds have been successfully introduced as their more stable counterparts. All five HIV-1 protease inhibitors currently registered (at the end of 1999) are peptidomimetics. Despite their obvious importance, affinity prediction of peptides and peptidomimetic compounds is difficult. Peptide behaviour is not simply intermediate between small compounds (molecular weight below 500) and proteins (polypeptides of more than 20 amino acids). Small compounds and proteins have well defined conformations. Peptides however, may have too many polar groups and rotatable bonds to be described sucessfully by regular scoring functions. Furthermore, a scoring function that performs well for a variety of compounds differing enormously in binding affinity, may not be able to simply order a series of analogous peptides.

We will investigate the usefulness of some popular scoring methods in peptide-protein interaction. We accept in advance the fact that a general scoring function for small

ligands, peptide-like ligands and protein ligands is not feasible with our current understanding of molecular recognition. Our aim is not to find a generally applicable scoring function, but rather to find out why peptides are so badly predicted. To do this, we selected some peptide(mimetic) protein complexes from the Protein Data Bank (PDB),[10] in which we have a scientific interest. We will investigate five peptides that bind to a Major Histocompatibility Complex (MHC) class I HLA-A2 molecule and 15 peptidomimetics that bind to Human Immunodeficiency Virus (HIV) 1 protease. This data set may not be sufficient for development of a general scoring function. However, this set might enable us to investigate:

- peptides as well as peptidomimetics
- binding to a groove on the outside as well as on the inside of a protein
- binding to a transport protein as well as to a proteolytic enzyme
- high and moderate affinity binding
- binding with and without a structural water molecule

We will compute energetic parameters with molecular mechanics and continuum electrostatics, and we will tabulate descriptive parameters or properties that could favour or disfavour interaction. We will also decompose the interaction in sidechain and backbone effects. Many of the procedures have been described before, as reviewed by Oprea[11] and Knegtel and Grootenhuis.[12] We will add an energy decomposition to ligand sidechain and backbone atoms and an enumeration of buried and exposed polar groups.

Results

The most important step is the data preparation. The most subjective step is the neutralisation of Asp B25 of the HIV protease molecules for all complexes. It is quite well possible that some ligands have induced other pK_a shifts in the protease, leading to other protonation states. The coordinate minimisation consists of a stage where the newly added hydrogen atoms were minimised and a stage where the entire ligand as well as all sidechains of nearby protein residues were minimised. Using the energies from the latter complexes led to slightly better regression models than the energies from the hydrogen_only minimised models.

Table 1. Experimental (Kcal/mol) and predicted affinities of the complexes

PDB-code	*Ligand*	ΔG_{exp}	ΔG_{pred1}	ΔG_{pred2}	*Ludi*
1hhg[13]	GP-9	-8.90[6]	-11.19	-10.14	1472
1hhh[13]	HBV-10	-11.60[6]	-7.53	-10.47	1617
1hhi[13]	FLU-9	-11.20[6]	-11.13	-10.89	1478
1hhj[13]	RT-9	-9.00[6]	-9.12	-9.17	1985
1hhk[13]	TAX-9	-10.90[6]	-11.78	-11.87	1370
4hvp[14]	MVT101	-8.33[15]	-10.29		1297
5hvp[16]	Ace-Pep	-10.50[15]	-10.20	-11.24	1388
7hvp[17]	JG365	-13.12[15]	-10.61	-12.04	1274
8hvp[18]	U85548e	-11.61[19]	-9.34	-10.82	1049
9hvp[20]	A74704	-11.38[15]	-12.82	-13.34	1185
1aaq[21]	PSI	-11.45[15]	-8.58	-11.63	1196
1hbv[22]	SB203238	-8.68[15]	-13.50		907
1hpv[23]	VX478	-12.57[15]	-11.84	-12.34	923
1htf[24]	GR126045	-11.04[15]	-12.22	-10.88	1069
1htg[24]	GR137615	-13.19[15]	-13.13	-12.89	1175
1hvi[25]	A77003	-13.74[15]	-13.12	-13.72	1343
1hvj[25]	A78791	-14.26[15]	-12.83	-13.25	1348
1hvk[25]	A79928	-13.79[15]	-13.00	-13.42	1412
1hvl[25]	A76889	-12.27[15]	-13.05	-13.49	1280
1hvr[26]	XK263	-12.97[15]	-14.86	-13.66	1034

The starting point of this study was the fact that established scoring methods performed badly. Table 1 lists the 20 complexes studied. The log K_i values were rewritten as ΔG_{exp}-values (third column) and relate to the Ludi Score (sixth column) with the following statistics: r^2=0.039, n=20, s=1.79, F=0.74 without obvious outliers. It is clear that the most established method and score function does not predict well for this data set.

Table 2 lists the energy and property parameters used in this study. Leave-one-out PLS analyses were performed on different training sets of 18 compounds and the resulting model was used to predict the 2 complexes that were left out entirely, as explained in the methods section. These externally predicted values are also listed in Table 1 in the fourth column as ΔG_{pred1}.The statistical data of these 10 runs are listed in Table 3. A differing number of latent

Table 2 Explanatory energies and properties tested in the PLS models

Parameter	*Explanation*	*Program*
Ei_c_s	sidechain Coulomb interaction energy	Insight/Discover(MSI)
Ei_v_s	sidechain van der Waals interaction energy	Insight/Discover

Ei_s	sidechain non-bonded interaction energy	Insight/Discover
Ei_c	Coulomb interaction energy	Insight/Discover
Ei_v	van der Waals interaction energy	Insight/Discover
Ei	non-bonded interaction energy	Insight/Discover
Ei_l	ligand bonded energy	Insight/Discover
Eh_hb	intermolecular hydrogen bond energy	HINT(EduSoft)
Eh_ab	intermolecular acid/base energy	HINT
Eh_hf	intermolecular hydrophobic energy	HINT
Eh_bb	intermolecular base/base energy	HINT
Eh_hp	intermolecular hydrophobic/polar energy	HINT
Eh	intermolecular energy	HINT
Ed_c	Coulombic energy	Delphi[27]
Ed_crf	corrected reaction field energy (solvation)	Delphi
Ed	summed energy	manually
L_hb	Ludi hydrogen bridge score	Insight/Ludi(MSI)
L_li	Ludi lipophilic contact score	Insight/Ludi
L_sb	Ludi salt bridge score	Insight/Ludi
L_ro	Ludi rotatable groups score	Insight/Ludi
Ludi	Ludi score	Insight/Ludi
MS	total buried surface area	MS[28]
HB_pos	number of possible ligand hydrogen bonds	manually
HB_sat	number of hydrogen bonds formed	Insight/Ludi
HB_bur	number of buried polar groups	Whatif[29]
HB_exp	number of exposed ligand polar groups	manually
MW	ligand molecular weight	Sybyl(Tripos)
Nat	number of ligand atoms	Sybyl
RES	resolution of X-ray structure	PDB
Random	random numbers	Sybyl

variables, the optimum number of components in column 6, seems to point at an instable overall model, since the model depends on which compounds are left out. Comparing the ΔG_{exp} to the predictions leads to a relation with r^2=0.138, n=20, s=1.69, F=2.88. There are two notable outliers, complexes 4hvp and 1hbv.

Table 3 Statistical parameters of the PLS analyses on all complexes

Run	*Omitted*	*# of xv*	*# comp.*	q^2	*Optim. # comp.*	r^2	*s*	*F*
1	1hhg,1aaq	18	10	0.44	4	0.94	0.50	49.32
2	1hhh,1hbv	18	10	0.68	6	0.99	0.23	160.47
3	1hhi,1hpv	18	10	0.26	2	0.66	1.16	14.51
4	1hhj,1htf	18	10	0.18	1	0.54	1.23	18.98
5	1hhk,1htg	18	10	0.24	2	0.65	1.15	13.97
6	4hvp,1hvi	18	10	0.17	2	0.64	1.04	13.35
7	5hvp,1hvj	18	10	0.19	2	0.64	1.12	13.08
8	7hvp,1hvk	18	10	0.28	2	0.66	1.08	14.52
9	8hvp,1hvl	18	10	0.34	2	0.71	1.07	18.48
10	9hvp,1hvr	18	10	0.36	2	0.71	1.07	17.94

Table 4 Statistical parameters of the PLS analyses after outlier removal

Run	*Omitted*	*# of xv*	*# comp.*	q^2	*Optim. # comp.*	r^2	*s*	*F*
1	1hhg,1aaq	16	10	0.61	3	0.88	0.54	30.01
2	1hhh,1hpv	16	10	0.73	3	0.91	0.53	41.85

3	1hhi,1htf	16	10	0.61	3	0.89	0.61	30.69
4	1hhj,1htg	16	10	0.56	3	0.86	0.58	25.13
5	1hhk,1hvi	16	10	0.66	3	0.89	0.56	33.00
6	5hvp,1hvj	16	10	0.63	3	0.88	0.57	28.55
7	7hvp,1hvk	16	10	0.60	3	0.90	0.54	33.93
8	8hvp,1hvl	16	10	0.71	3	0.93	0.48	52.99
9	9hvp,1hvr	16	10	0.84	3	0.95	0.41	70.80

After removal of these outliers, the entire procedure was re-done. The predicted values of these runs are listed in the fifth column in Table 1 as ΔG_{pred2}. The correlation of ΔG_{exp} with these predicted values is acceptable with r^2=0.674, n=18, s=0.90 and F=33.09. The statistical results for these last PLS runs are listed in Table 4. All runs now point at an optimum number of 3 latent variables.

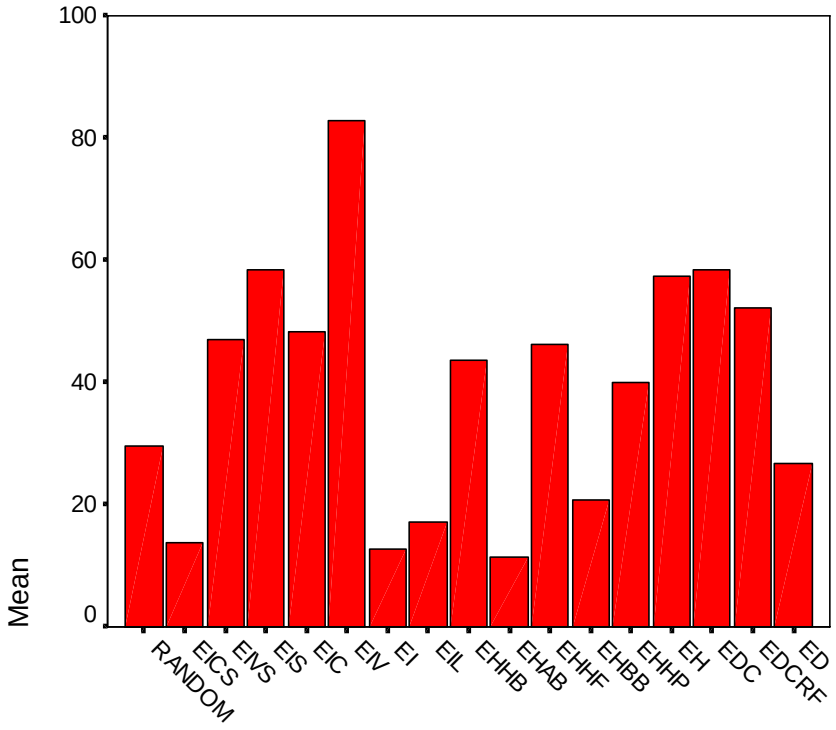

Figure 1 The mean (% times 10) of the fractional contributions of the energy parameters. Nine PLS models were derived for 18 compounds, as shown in Table 4. Sixteen out of thirty parameters in the models are related to energy and are shown in this Figure. If all would contribute equally to the model, a mean value of 0.33% (or 33 in this Figure) would result, as seen for e.g. the random numbers.

In order to investigate which parameters principally determine the PLS model, the mean (expressed as per thousand) of their contributions in the 9 PLS runs is shown in Figures 1 and 2, for the energies and properties, respectively.

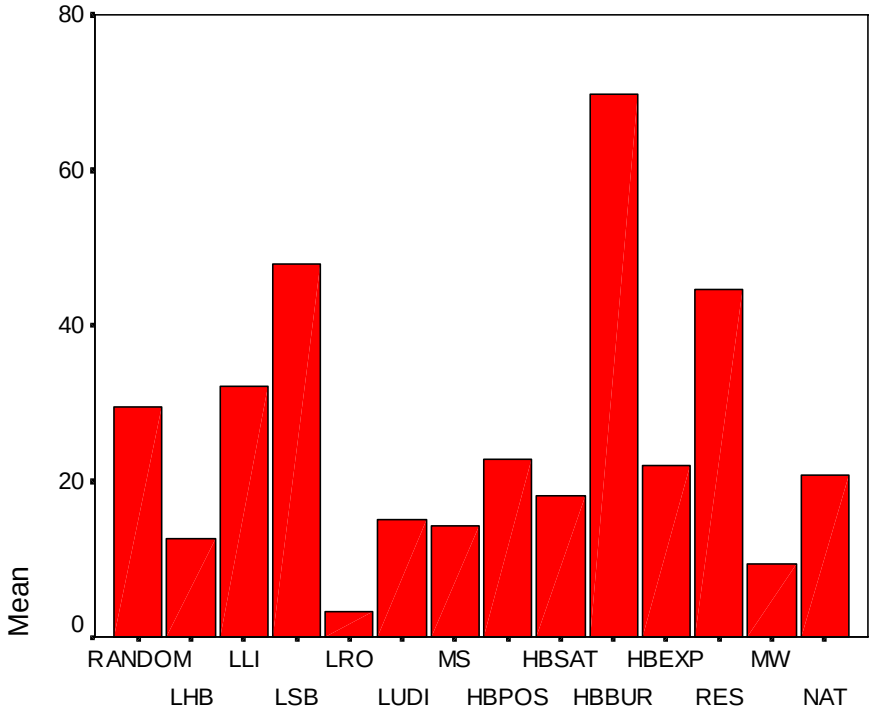

Figure 2 The mean (% times 10) of the fractional contributions of the property parameters. Nine PLS models were derived for 18 compounds, as shown in Table 4. Thirteen out of thirty parameters in the models are related to properties and are shown in this Figure. If all would contribute equally to the model, a mean value of 0.33% (or 33 in this Figure) would result, as seen for e.g. the random numbers.

Only four properties outperform the random numbers: Ludi lipophilic and salt bridge scores, the number of buried polar groups and the X-ray structure resolution. On the other hand, 10 energy parameters contribute more to the model than random numbers: four non-bonded interaction energies, four HINT intermolecular energies and the energies from the continuum electrostatics.

There is extensive intercorrelation between some of these variables, however, as can be seen in Figure 3. Figure 3 shows the (varimax) rotated loadings of the variables that outperform the random variable, plotted with the first two principal components.

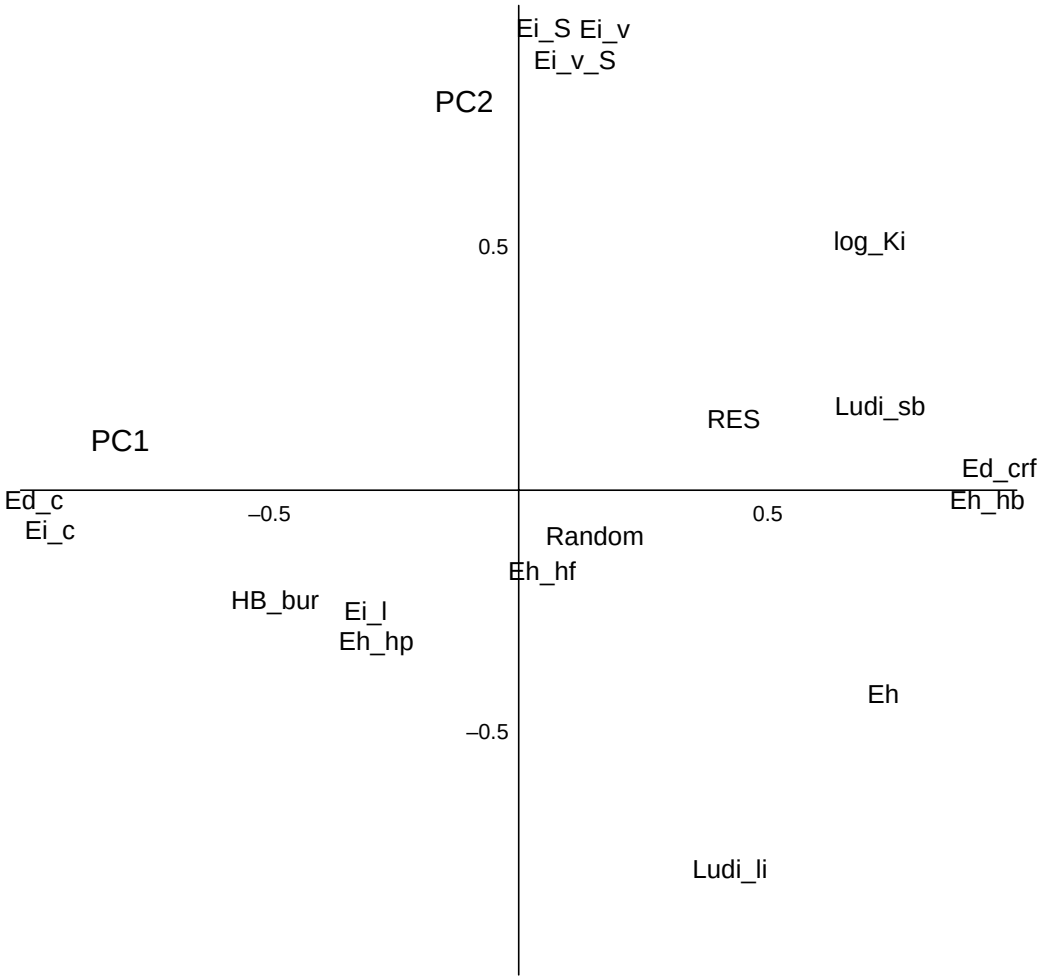

Figure 3 Varimax rotated loadings of PC1 and PC2. Fourteen principal components were extracted to replace the highly intercorrelated fourteen parameters that outperformed the random numbers in the PLS analyses (Figures 1 and 2). In order to help in interpreting these principal components, their axes have been rotated to increase correlation with the original parameters. The loadings resemble correlation coefficients; a value of 0.99 means that PC1 is correlated strongly to the Delphi solvation energy Ed_crf.

The first principal component (44% of the total variance) relates to electrostatic effects like the Coulombic interaction energy, the HINT hydrogen bond energy and the Delphi Coulombic and solvation energy. The second component (24% of the total variance) relates to steric effects like the van der Waals interaction energy between ligand and protein as well as between ligand sidechains and protein and the entire non-bonded interaction energy between the ligand sidechains and the protein atoms. The third significant component (12% of the total variance) is related mainly to the HINT hydrophobic intermolecular energy Eh_hf. PCA elegantly recognises the three basic interaction types, well-known in Hansch analysis. It is especially noteworthy that the HINT hydrophobic energy does not score in PC1 and PC2 and renders unique information to the data set. A similar finding was reported by Wei et al.[30]

Discussion

The Ludi Score routine fails to predict the affinities of the peptides and peptidomimetic compounds for their proteins. Especially the MHC binding peptides are predicted to bind better than they actually do. The Ludi Score equals 100 times the estimated log K_a, so the RT-9 peptide is predicted to bind to HLA-A2 with an association constant of almost 10^{20}, as can be seen in Table 1. An enormous number of hydrogen bonds and lipophilic contacts are responsible for this. Morgan et al.[31] postulated in the case of thermolysin inhibitors that polar groups that hydrogen bond with solvent before binding, and with a protein after binding, do not have a net contribution to the affinity. The observation that hydrogen bonds do not contribute to our PLS models, may indicate that for peptides and peptidomimetic compounds most hydrogen bonds do not contribute to affinity due to desolvation penalties. The number of buried polar groups, however, is the most important parameter in our models, as can be seen in Figure 2. This again points to a considerable desolvation cost for the peptide polar groups. We postulate that empirical scoring techniques that enumerate hydrogen bonds will fail to give good predictions for peptide binding, which is related to the avoidance of unfulfilled hydrogen bonds.

Another scoring method using a master equation and continuum electrostatics[6] failed to order the MHC binding peptides. In our results we observe that the Discover Coulombic interaction energy, the Delphi Coulombic energy and the Delphi solvation energy are almost

completely correlated. We observe, as did Froloff,[6] that the sum of Coulombic energy and solvation energy, i.e. the total electrostatic energy, opposes binding. Here, the Coulombic interaction energy, has negative (attracting) values, but contributes to the PLS equation with a negative slope. This implies that Coulombic interactions decrease binding. This may be an explanation for failures in predicting peptide binding with molecular mechanics energies, which are not decomposed. Coulombic and van der Waals contributions should be considered separately. Like Knegtel and Grootenhuis[12] we scaled down the electrostatics using a distance dependent dielectric constant. Holloway[32] uses a constant dielectric constant of 1 and came to the conclusion that a subdivision of the molecular mechanics energy in Coulomb and van der Waals effects did not improve their relationships. We think, however, that this subdivision is needed for peptides. We do not see any contribution of the sidechain Coulombic interaction energy (meaning the Coulombic interaction energy of the sidechains of the ligand and the entire protein), suggesting that the adverse relation is mainly determined by the peptide backbone atoms. This may be an artifact of the minimisation, where almost all possible hydrogen bonds are optimised, leading to Coulombic interaction energies, that suggest interaction. The negative sign in the PLS equation, however, indicates that these optimised hydrogen bonds are not as stable as when they were fully solvated. Holloway[32] was able to use the total interaction energy as a sole indicator of the K_i values of HIV protease inhibitors. Despite differences in force field and dielectric masking, we believe that it is the difference in data sets, and especially our inclusion of the MHC binding peptides, which underlies this apparent discrepancy.

The latent variables and PLS models became consistent only after removal of two outliers. Removing two out of the already modest 20 compounds, may seem unwise. Given the uncertainty in their X-ray structures and affinity data, we believe that this removal is essential. The crystals of MVT-101 and HIV protease (4hvp) have been re-analysed which led to quite a different set of coordinates,[33] that have not yet been deposited at the PDB. The other outlier (1hbv), the complex of HIV protease with the reduced amide peptidomimetic SB203238, cannot be as easily explained. It is possible that the reduced amide induces a different protonation at the catalytic aspartates of HIV protease. So, we neutralised both aspartic acids and re-computed the energies. The complex remained an outlier, and it now was the only compound with positive Coulombic energy as computed with Delphi. We are currently trying to predict pK_a values in this complex, based on continuum electrostatics. In

their study of 13 HIV-protease complexes, Bardi et al.[34] predicted the same 1hbv complex worst using another scoring method. If 10% of the current small data set is such an obvious outlier, we are warned against the use of fully automated procedures on large data sets for scoring method development. All complexes should be manually inspected, and completed for missing atoms where necessary. We feel that no scoring methods can be compared if different starting structures, atom types and charges are used. An annotated database of protonated complexes with association constants, pertinent to the crystallisation conditions, would be most welcome. Such a database would indeed allow a comparison between scoring methods, as long as it is representative and non-redundant.

In the PLS models the van der Waals interactions between the ligand sidechain and the protein, and the number of buried polar groups dominate. Interestingly, these parameters are not encountered in published scoring functions. Despite the fact that we tested 30 parameters, we are likely to have missed some important phenomena. As our understanding of molecular recognition increases, we accept that buried hydrogen bonds contribute more to binding than the hydrogen bonds that are solvent exposed. We also see an increasing number of reports stating that amino-aromatic, aromatic-aromatic and the polar CH-pi interactions and CH hydrogen bonds should not be neglected. Umezawa et al.[35] conclude for the same HLA-A2 peptide complexes as studied here, that there are many CH-pi interactions. We are also currently investigating CH pi interactions in peptide protein interactions. It is possible that polar effects can explain thermodynamic differences between seemingly similar hydrophobic sidechains, as reported recently by Davies et al.[36] Gilmer et al.[37] reported that by replacing isoleucine by leucine in the Ace-pYEEIE phosphopeptide the affinity for the GST Src SH3-SH2 domain dropped 3-fold. It is known that this residue resides in a hydrophobic pocket in the SH2 domain. The parameters we have used in this analysis would not be able to predict such a difference between these seemingly identical residues. An in-depth analysis of the polar component of interactions that were formerly believed to be purely hydrophobic, is in order.

Another feature that we did not fully include is the induced fit of ligand and protein. Since it is impossible to completely search the conformational space of large peptides, we have decided not to compute a ligand free energy, despite a report by Nicklaus et al.,[38] showing high conformational strain for protein bound ligands. Recent work by Boström et

al.[39] suggested that ligand minimisation in vacuum and in solvation models can lead to electrostatic collapse, in the case of relatively polar ligands. We could expect the same thing to happen in our peptide data set and decided to only use the bonded (internal) energy of the ligand, minimised in the complex, as a measure of conformational strain. Such an absolute energy is of course dependent on the force field. This conformational strain did not contribute to the PLS models. The fate of a ligand before complex formation remains a blind spot in structural biology.

Concluding, we have observed that peptides and their mimetics are indeed hard to score with general partitioning methods, like the empirical regression methods[5] and the master equation methods.[6] Especially the hydrogen bonds of the peptide backbone show anomalous behaviour, compared to hydrogen bonds between small ligands and proteins. Furthermore, we have observed that it is almost impossible to determine the protonation and charge of the complexes, in a manner which reflects the experimental conditions of the affinity measurement with any certainty. Discrepancies in X-ray coordinates and discrepancies in the measurement of association constants introduces a standard deviation that may well be as large as the standard error of 1 log unit that seems to be the current asymptotic barrier for most scoring methods (see e.g. Knegtel and Grootenhuis).[12] Ofcourse, the X-ray structure resolution is a limiting factor. Our observation that the resolution participates significantly to the PLS models might be a chance factor or the result of the fact that the MHC complex structures are in general of lower resolution than the HIV protease complex structures, so that the resolution discriminates between both types of proteins.

A complete conformational analysis of peptides is impossible. Scoring the generated conformations in peptide-protein complexes is difficult, as long as we do not understand, for example, the difference in behaviour between leucine and isoleucine. We must conclude that at this moment, we can not reliably score (nor dock) peptides and peptidomimetics to proteins without user bias. Even non-partitioning methods, like free energy perturbation molecular dynamics, do not include polarisation terms and polar effects of hydrophobic groups, and they compute the conformational ensemble of one molecular complex only, not allowing for partial protonation, oligomerization or allosteric influences. Still, experimental uncertainties may outweigh our current computational inadequacies.

Methods

Data collection

The coordinates from the following complexes were downloaded from the PDB:[10] 1hhg, 1hhh, 1hhi, 1hhj, 1hhk, 4hvp, 5hvp, 7hvp, 8hvp, 9hvp, 1aaq, 1hbv, 1hpv, 1htf, 1htg, 1hvi, 1hvj, 1hvk, 1hvl, 1hvr. K_i values were taken from Froloff et al.[6] for the five MHC peptide complexes, and from Eldridge et al.[15] for the HVP complexes, and checked against the original references. The K_i value of 8hvp was reported by Lin et al.[19] All values were rewritten as ΔG_{exp}=-RTlnK_i in Kcal/mol at a temperature of 298 K.

Data preparation

The complexes were prepared in InsightII 98.0 (MSI, San Diego, USA). Water molecules were removed (except for the groove water in 14 HIV protease complexes) and missing sidechains and hydrogen atoms were added, consistent with the dominant state at a pH value of 6. Asp B25 was protonated in all HIV protease complexes. The chains were capped with charged ends. With Discover 98.0 (MSI, San Diego, USA), the CFF91 forcefield, and conjugate gradients optimisation to below a gradient of 0.01 Kcal/mol, first the hydrogen positions were optimised and then the entire ligand together with the sidechains of all amino acids with an atom within 5 Å of the ligand. The cell multipole method and a distance dependent dielectric constant (1*r) were used for the non-bonded interactions.

Molecular mechanics energies

An analysis of the non-bonded interaction energy between ligand and protein was performed on the minimised complex with a cut-off radius of 99 Å and a distance dependent dielectric of (4*r) to reduce the electrostatic dominance in vacuum. We extracted the interaction energy of the entire ligand with the entire protein and of only the ligand sidechains with the entire protein (the sidechain atoms were defined manually in the case of peptidomimetic ligands). As a measure of ligand conformational strain, we extracted the internal bonded energy of the ligand after complex minimisation. Hint 2.27I[40,41] was used to compute hydrophobic and polar interactions using the intermolecular HintTable and a cut-off radius of 6.00 Å and a van der Waals limit of 0.90 Å. A 1 Å grid was computed with a border space of 5 Å and an exponential hydropathic function of exp(-1r). No steric term was added because the minimised complex was used. The amino acids were partitioned by

dictionary, with only polar hydrogen atoms and inferred solvent conditions. The heteroatomic ligands of HIV proteases were partitioned by calculation with a polar_proximity via_bonds.

Continuum electrostatics energies

The complexes were converted to PDB-type files and CFF91 charges and PARSE3[42] sizes were given to all atoms. With qdiffxs (Delphi 3.0[27]) and a grid size of 65, the Coulomb and solvation (corrected reaction field) energies were computed of the ligand, the protein and the complex. The dielectric constants were 2 within solute and 80 outside. The energy of the complex minus the energies of the ligand and the protein were used in the statistical analyses after conversion from kt to Kcal/mol.

Property collection and calculation

With the MS programme[28] from the QCPE (nr. 429) and the autoMS script (part of the Dock 4 suite of programmes),[43] we computed the buried surface area (TOTAL AREA in the output file). With the Ludi programme (MSI, San Diego, USA) we computed the Ludi score (the SCORE programme of 1994 was used),[4] which consists of contributions from hydrogen bridges and salt bridges, as well as from the lipophilic contact area and the number of rotatable groups in the ligand. We manually counted the maximum number of possible hydrogen bonds a ligand could make. The number of hydrogen bonds that were actually made was generated by Ludi in the previous stage. With Whatif,[29] version 19970813-1517, we computed the hydrogen bond network (HNQCHK command)[44] for the complex and the protein_only and thus obtained the buried polar groups of both protein and ligand. We manually computed the exposed polar groups of the ligand, which could be neutralised by surrounding solvent (this is the maximum number of hydrogen bonds that a ligand could make minus the hydrogen bonds that it does make and the buried polar atoms of the ligand). Further simple properties are the resolution of the X-ray structure, the number of ligand atoms and the molecular weight as well as a random number (generated with Sybyl).

Statistical analysis

The statistical analyses were performed with Sybyl 6.5 (Tripos, St. Louis, USA) and SPSS 6.5. We performed the following strategy, since it was immediately obvious that the models depended heavily on the selection of test set and training set. We selected all

compounds minus number 1 and 11 for a PLS run with 10 components and 18 crossvalidations. We then computed a model with the optimum number of components and without crossvalidations. We predicted compounds 1 and 11 with this model. Then, we selected all compounds minus number 2 and 20 and performed the same procedure to predict them. Then we performed a linear regression of the ΔG_{exp}-values and the predictions. After removing two obvious outliers (see Results and Discussion) we performed the same strategy, selecting all compounds minus 1 and 10, 2 and 11, 3 and 12 etc… again on a leave-one-out basis. So, the internal validation is leave-one-out. The external validation is leave-two-out. A principal component analysis was also performed with Sybyl 6.5 using factor analysis with all components and Varimax rotation.

Acknowledgements

The authors would like to acknowledge Drs C. Murray (1aaq), M. Miller (4hvp) and J. Trylska for helpful discussions regarding their work.

References

1)Ajay; Murcko, M. A. *J Med Chem* **1995**, *38*, 4953-67.
2)Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys J* **1997**, *72*, 1047-69.
3)Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564-8.
4)Böhm, H. J. *J Comput Aided Mol Des* **1994**, *8*, 243-56.
5)Böhm, H. J. *J Comput Aided Mol Des* **1998**, *12*, 309-23.
6)Froloff, N.; Windemuth, A.; Honig, B. *Protein Sci* **1997**, *6*, 1293-301.
7)Wade, R. C.; Ortiz, A. R.; Gago, F. *Perspectives in Drug Discovery and Design* **1998**, *9/10/11*, 19-34.
8)Strynadka, N. C.; Eisenstein, M.; Katchalski-Katzir, E.; Shoichet, B. K.; Kuntz, I. D.; Abagyan, R.; Totrov, M.; Janin, J.; Cherfils, J.; Zimmerman, F.; Olson, A.; Duncan, B.; Rao, M.; Jackson, R.; Sternberg, M.; James, M. N. *Nat Struct Biol* **1996**, *3*, 233-9.
9)Murray, C. W.; Auton, T. R.; Eldridge, M. D. *J Comput Aided Mol Des* **1998**, *12*, 503-19.
10)Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J Mol Biol* **1977**, *112*, 535-42.

11)Oprea, T. I.; Marshall, G. R. *Perspectives in Drug Discovery and Design* **1998**, *9/10/11*, 35-61.

12)Knegtel, R. M. A.; Grootenhuis, P. D. J. *Perspectives in Drug Discovery and Design* **1998**, *9/10/11*, 99-114.

13)Madden, D. R.; Garboczi, D. N.; Wiley, D. C. *Cell* **1993**, *75*, 693-708.

14)Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B.; Wlodawer, A. *Science* **1989**, *246*, 1149-52.

15)Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J Comput Aided Mol Des* **1997**, *11*, 425-45.

16)Fitzgerald, P. M.; McKeever, B. M.; VanMiddlesworth, J. F.; Springer, J. P.; Heimbach, J. C.; Leu, C. T.; Herber, W. K.; Dixon, R. A.; Darke, P. L. *J Biol Chem* **1990**, *265*, 14209-19.

17)Swain, A. L.; Miller, M. M.; Green, J.; Rich, D. H.; Schneider, J.; Kent, S. B.; Wlodawer, A. *Proc Natl Acad Sci U S A* **1990**, *87*, 8805-9.

18)Jaskolski, M.; Tomasselli, A. G.; Sawyer, T. K.; Staples, D. G.; Heinrikson, R. L.; Schneider, J.; Kent, S. B.; Wlodawer, A. *Biochemistry* **1991**, *30*, 1600-9.

19)Lin, Y.; Lin, X.; Hong, L.; Foundling, S.; Heinrikson, R. L.; Thaisrivongs, S.; Leelamanit, W.; Raterman, D.; Shah, M.; Dunn, B. M.; et al. *Biochemistry* **1995**, *34*, 1143-52.

20)Erickson, J.; Neidhart, D. J.; VanDrie, J.; Kempf, D. J.; Wang, X. C.; Norbeck, D. W.; Plattner, J. J.; Rittenhouse, J. W.; Turon, M.; Wideburg, N.; et al. *Science* **1990**, *249*, 527-33.

21)Dreyer, G. B.; Lambert, D. M.; Meek, T. D.; Carr, T. J.; Tomaszek, T. A., Jr.; Fernandez, A. V.; Bartus, H.; Cacciavillani, E.; Hassell, A. M.; Minnich, M.; et al. *Biochemistry* **1992**, *31*, 6646-59.

22)Hoog, S. S.; Zhao, B.; Winborne, E.; Fisher, S.; Green, D. W.; DesJarlais, R. L.; Newlander, K. A.; Callahan, J. F.; Moore, M. L.; Huffman, W. F.; et al. *J Med Chem* **1995**, *38*, 3246-52.

23)Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. *J. Am. Chem. Soc.* **1995**, *117*, 1181-2.

24)Jhoti, H.; Singh, O. M.; Weir, M. P.; Cooke, R.; Murray-Rust, P.; Wonacott, A. *Biochemistry* **1994**, *33*, 8417-27.

25)Hosur, M. V.; Bhat, T. N.; Kempf, D. J.; Baldwin, E. T.; Liu, B.; Gulnik, S.; Wideburg, N. E.; Norbeck, D. W.; Appelt, K.; Erickson, J. W. *J. Am. Chem. Soc.* **1994**, *116*, 847-55.

26)Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; et al. *Science* **1994**, *263*, 380-4.

27)Gilson, M. K.; Honig, B. *Proteins* **1988**, *4*, 7-18.

28)Connolly, M. L. *Science* **1983**, *221*, 709-13.

29)Vriend, G. *J Mol Graph* **1990**, *8*, 52-6, 29.

30)Wei, D. T.; Meadows, J. C.; Kellogg, G. E. *Med Chem Res* **1997**, *7*, 259-270.

31)Morgan, B. P.; Scholtz, J. M.; Ballinger, M. D.; Zipkin, I. D.; Bartlett, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 297-307.

32)Holloway, M. K. *Perspectives in Drug Discovery and Design* **1998**, *9/10/11*, 63-84.

33)Miller, M.; Geller, M.; Gribskov, M.; Kent, S. B. *Proteins* **1997**, *27*, 184-94.

34)Bardi, J. S.; Luque, I.; Freire, E. *Biochemistry* **1997**, *36*, 6588-96.

35)Umezawa, Y.; Nishio, M. *Bioorg Med Chem* **1998**, *6*, 2507-15.

36)Davies, T. G.; Hubbard, R. E.; Tame, J. R. *Protein Sci* **1999**, *8*, 1432-44.

37)Gilmer, T.; Rodriguez, M.; Jordan, S.; Crosby, R.; Alligood, K.; Green, M.; Kimery, M.; Wagner, C.; Kinder, D.; Charifson, P.; et al. *J Biol Chem* **1994**, *269*, 31711-9.

38)Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. *Bioorg Med Chem* **1995**, *3*, 411-28.

39)Boström, J.; Norrby, P. O.; Liljefors, T. *J Comput Aided Mol Des* **1998**, *12*, 383-96.

40)Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J Comput Aided Mol Des* **1991**, *5*, 545-52.

41)Meng, E. C.; Kuntz, I. D.; Abraham, D. J.; Kellogg, G. E. *J Comput Aided Mol Des* **1994**, *8*, 299-306.

42)Sitkoff, D.; Sharp, K. A.; Honig, B. *J Phys Chem* **1994**, *98*, 1978-88.

43)Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J Mol Biol* **1982**, *161*, 269-88.

44)Hooft, R. W.; Sander, C.; Vriend, G. *Proteins* **1996**, *26*, 363-76.