Inhibition and substrate recognition – a computational approach applied to HIV protease

H.M. Vinkers[a],*, M.R. de Jonge[a], F.D. Daeyaert[a], J. Heeres[a], L.M.H. Koymans[a], J.H. van Lenthe[b], P.J. Lewi[a], H. Timmerman[c] & P.A.J. Janssen[a]

[a]*Center for Molecular Design, Janssen Pharmaceutica N.V., Antwerpsesteenweg 37, B-2350 Vosselaar, Belgium;* [b]*Theoretical Chemistry Group, Debye Institute, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands;* [c]*Leiden/Amsterdam Center for Drug Research (LACDR), Division of Medicinal Chemistry, Faculty of Chemistry, Vrije Universiteit, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands*

Summary

We have developed a computational approach in which an inhibitor's strength is determined from its interaction energy with a limited set of amino acid residues of the inhibited protein. We applied this method to HIV protease. The method uses a consensus structure built from X-ray crystallographic data. All inhibitors are docked into the consensus structure. Given that not every ligand–protein interaction causes inhibition, we implemented a genetic algorithm to determine the relevant set of residues. The algorithm optimizes the q^2 between the sum of interaction energies and the observed inhibition constants. The best possible predictive model resulting has a q^2 of 0.63. External validation by examining the predictivity for compounds not used in derivation of the model leads to a prediction accuracy between 0.9 and 1.5 $\log_{10}$ unit. Out of 198 residues in the whole protein, the best internally predictive model defines a subset of 20 residues and the best externally predictive model one of 9 residues. These residues are distributed over the subsites of the enzyme. This approach provides insight in which interactions are important for inhibiting HIV protease and it allows for quantitative prediction of inhibitor strength.

Introduction

It is estimated that around 40 million people worldwide [1] are infected with the human immunodeficiency virus (HIV), the causative agent for the acquired immunodeficiency syndrome (AIDS). A crucial enzyme of this virus is a protease; this enzyme cleaves the polyprotein products of the gag and gag-pol viral genes [2] into proteins. The proteolytic activity of this enzyme is essential for the functioning of the virus: its inhibition blocks viral maturation and infectivity, providing an opportunity for antiviral therapy. Currently, six protease-inhibiting drugs [3] are approved for treatment of HIV infection.

HIV protease is a comprehensively studied drug target. It belongs to a family of enzymes known as aspartyl proteases. In this type of enzyme the hydrolization of a peptide bond is catalyzed by two aspartic acid residues in a hydrophobic active site. HIV protease is quite specific in its substrate selection compared to other members of the protease family: it only breaks bonds between a limited number of pairs of amino acids. The enzyme consists of two identical peptide chains, each 99 amino acids long, forming a C_2-symmetrical homodimer. The homodimer has a cylindrical active site which is 23 Å long and 6–8 Å in diameter [2]. By binding to this site an inhibitor competes with the binding of the normal substrate.

Since HIV protease is a viable drug target and crystallizes relatively easy, there have been extens-

*To whom correspondence should be addressed. E-mail: mvinkers@prdbe.jnj.com

ive structure-based drug design efforts in this area. In more than one case this has been done in an iterative fashion, in which a crystal structure of an enzyme–inhibitor complex was used to design an analogue of the inhibitor, whose crystal structure in turn was resolved to guide the synthesis of new analogues. Partly as a result of this process, a wealth of structural information about HIV protease is available.

Methods to compute inhibitor strengths have also been extensively applied to HIV protease. A range of different approaches have been used; for instance FEP [4–6], LIEA [7, 8] and empirical methods such as QSAR [9], CoMFA [10, 11], 'master equation' approaches [12–14], total energy correlation [15], 'knowledge-based' methods [16–19] and methods like CScore [20] and X-Score [21] which combine different approaches. The empirical methods, often called scoring functions, are directly derived from observed inhibition strengths by correlating these strengths to components that are thought to be important for the binding process. Empirical methods have several appealing features [21], being both computationally inexpensive and conceptually tractable. Among the more common factors that are used to derive a scoring function are the van der Waals interaction energy, the Coulomb interaction energy, the hydrogen bond interaction energy, the desolvation energy, the hydrophobic energy change, the change in conformational energy upon binding and approximations for the change in entropy such as the number of flexible bonds in the ligand.

The above mentioned methods are based on the total interaction of a ligand with the protein. However, interactions with specific locations on the protein are needed in order to establish a proper inhibiting effect. In general, a molecule binding to the exterior of the protease enzyme will not cause inhibition: inhibitory action is localized. Todd et al. have found that the binding energy is not distributed uniformly throughout the binding site [22]. The interaction occurs through specific residues within protease and many other enzymes [23–25]. Kulkarni et al. [26] also found that interactions with only some of the residues contribute to the difference in activity.

The majority of scoring functions is derived from many different protein systems in order to be generally applicable. Deriving a focussed scoring function – specifically targeted to deal with one protein – obviously creates the need to derive a new one for every other protein. Another drawback is the need for ligands with known affinity enabling the deriva-

tion. Nevertheless, we expect that the better accuracy of a focussed scoring function compensates the drawbacks. We have recently used a similar type of scoring function for a different target (non-competitive inhibition of HIV Reverse Transcriptase) in a de novo drug design environment [27] with encouraging results, which shows that the method described in this article is valid for more than one system.

The computational approach in the present study derives an inhibitor's strength from its interaction energies with the relevant amino acids. Our method yields both this relevant set of residues, and the strength of inhibition for a number of inhibitors. To this end, the interaction energies of 53 HIV protease inhibitors with each of the residues of the enzyme are computed from the complex of the inhibitor and a consensus structure. The inhibitors are docked into a consensus structure to ensure that the interaction energies used to derive a model are comparable. Using one common pocket for all inhibitors also allows for application of the model to compounds of which the crystal structure has not been resolved. A best internally predictive (BIP) and best externally predictive (BEP) model are determined from the interaction energy terms employing a genetic algorithm. The residues in these models, comprising a limited set out of the 198 residues in total, are important for ligand recognition and inhibition of the viral enzyme.

Method

Construction of a consensus structure

We constructed a consensus structure from 98 X-ray crystal structures of HIV-1 proteases from the Protein Data Bank [28]. The crystal structures were stripped of inhibitor, water, solvent, and hydrogen atoms. The structures were all superimposed on 1DIF [29] – the best resolution wildtype structure – by means of the quaternion method [30] implemented by Heisterberg [31] as obtained from the Computational Chemistry List [32]. Only identical residues were considered. For instance 1GNM [33] has a V82D mutation, resulting in the exclusion of that particular residue from the superposition. Since protease is a symmetrical homodimer, the two symmetry-related superpositions were carried out for each crystal structure and the one with the lowest rmsd was retained. Subsequently, the average co-ordinates of the superimposed atoms were determined. The identifiers of the crystal structures

Table 1. Identifiers of the protease crystals used for construction of a consensus structure and their root mean square deviation from the resulting consensus structure.

PDB ID	rmsd	PDB ID	rmsd	PDB ID	rmsd
1A30	0.92	1A8G	1.02	1A8K	0.95
1A94	0.89	1A9M	0.90	1AAQ	1.06
1AID	1.52	1AJV	0.91	1AJX	0.85
1AXA	0.96	1B6J	0.65	1B6K	0.57
1B6L	0.61	1B6M	0.60	1B6N	0.65
1B6O	0.60	1B6P	0.59	1BDL	1.21
1BDQ	1.13	1BDR	1.08	1BV7	1.11
1BV9	1.12	1BWA	1.22	1BWB	1.00
1C6X	1.07	1C6Y	1.15	1C6Z	1.08
1C70	0.96	1CPI	0.68	1D4K	0.64
1D4L	0.67	1D4S	0.93	1D4Y	0.83
1DAZ	0.85	1DIF	0.47	1DMP	1.18
1DW6	0.94	1EBK	1.49	1FQX	0.90
1GNM	0.97	1GNN	0.97	1GNO	0.93
1HBV	1.05	1HIH	0.84	1HIV	0.83
1HOS	1.07	1HPO	0.85	1HPS	1.08
1HPV	1.03	1HPX	0.89	1HSG	0.81
1HTE	0.98	1HTF	0.95	1HTG	0.83
1HVC	0.77	1HVH	1.25	1HVI	0.86
1HVJ	0.93	1HVK	0.88	1HVL	0.92
1HVR	1.13	1HVS	0.90	1HWR	1.16
1HXB	0.93	1HXW	0.88	1MER	1.14
1MES	1.17	1MET	1.14	1MEU	1.14
1MTR	0.66	1ODW	0.82	1ODX	1.02
1ODY	0.94	1OHR	0.62	1PRO	0.91
1QBR	1.06	1QBS	1.16	1QBT	1.17
1QBU	1.21	1SBG	0.98	1TCX	0.82
1VIJ	1.10	1VIK	0.89	2AID	0.83
2BPV	0.88	2BPW	0.89	2BPX	0.84
2BPY	0.86	2BPZ	0.87	2UPJ	0.86
3AID	1.00	4HVP	1.01	4PHV	0.75
5HVP	0.89	7HVP	0.93	7UPJ	0.94
8HVP	0.92	9HVP	1.08		

used and their root mean square deviations from the consensus structure are given in Table 1. Hydrogens were added to the resulting average structure. The B25 aspartic acid was deprotonated [2]. The position of the hydrogens was determined by subjecting the final structure to a local geometry optimization, keeping the other atoms fixed.

Docking of the inhibitors

A total of 53 protease inhibitors for which K_i values are available was removed from their crystal structure and rigidly docked into the consensus structure.

This diverse collection of inhibitors consists of peptide analogs as well as organic ligands such as cyclic ureas. The structures of the inhibitors are depicted in Figure 1. A structural water molecule, which is in some cases involved in binding to the Ile_{50} residues, was included in the inhibitor structure. The docking algorithm consist of a combined grid-based Monte Carlo and simulated annealing method. It computes the docking energy as the sum of the van der Waals, Coulomb and hydrogen bond interactions between the ligand and the protein. Docking starts by randomly placing the ligand in a few hundred thousand orient-
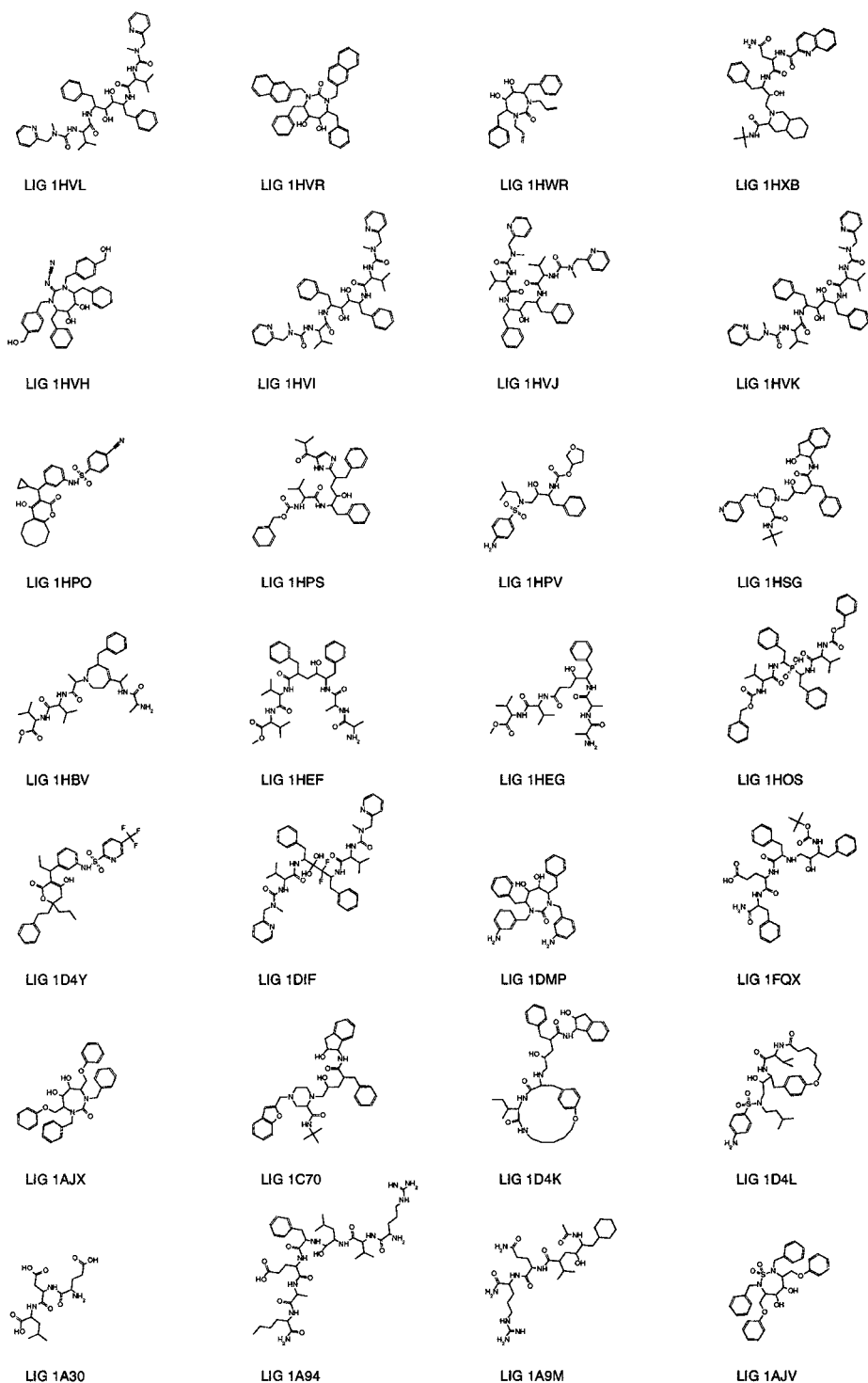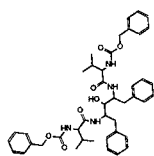
570

Figure 1. Structures of the protease inhibitors used (see Table 2). The two-dimensional co-ordinates were generated using the cactus online smiles translator [81].
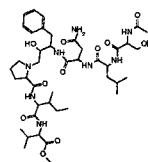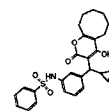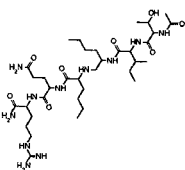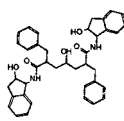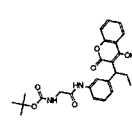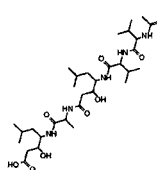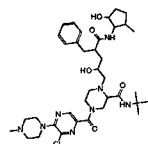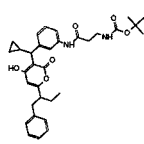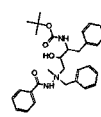
LIG 9HVP

LIG 5UPJ

LIG 6UPJ
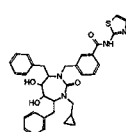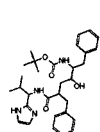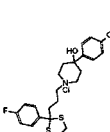
LIG 7HVP

LIG 7UPJ

LIG 4HVP
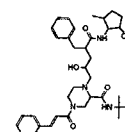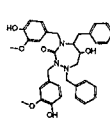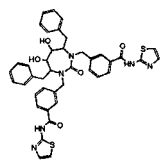
LIG 4PHV

LIG 4UPJ

LIG 5HVP
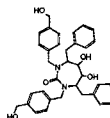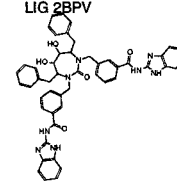
LIG 2BPY

LIG 2UPJ

LIG 3AID
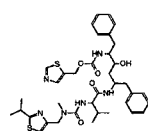
LIG 3UPJ

LIG 1QBU

LIG 1SBG
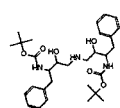
LIG 2AID

LIG 2BPV

LIG 1PRO
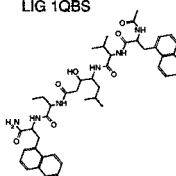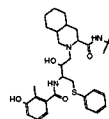
LIG 1QBR

LIG 1QBS

LIG 1QBT

LIG 1HXW

LIG 1ODW

LIG 1ODY

LIG 1OHR

Figure 1. (Continued).

Table 2. Experimentally determined pK$_i$ values of HIV-1 protease inhibitors (expressed as the negative log$_{10}$ of the inhibition constant in nanomoles) and root mean square deviation of the inhibitor atoms after docking in the consensus structure.

Identifier	pK$_i$	rmsd	Identifier	pK$_i$	rmsd
LIG 1A30 [43]	4.30	2.12	LIG 1A94 [44]	7.85	0.66
LIG 1A9M [45]	7.70	0.71	LIG 1AJV [46]	7.72	0.70
LIG 1AJX [46]	7.91	0.27	LIG 1C70 [47]	10.30	1.85
LIG 1D4K [48]	9.22	0.48	LIG 1D4L [48]	8.77	0.21
LIG 1D4Y [49]	11.10	1.14	LIG 1DIF [29]	10.66	0.80
LIG 1DMP [50]	9.47	0.47	LIG 1FQX [51]	9.82	1.00
LIG 1HBV [52]	6.37	0.39	LIG 1HEF [53]	9.00	0.59
LIG 1HEG [53]	7.74	1.01	LIG 1HOS [54]	8.55	0.85
LIG 1HPO [55]	9.22	1.71	LIG 1HPS [56]	9.22	0.57
LIG 1HPV [57]	9.22	0.81	LIG 1HSG [58]	9.42	0.43
LIG 1HVH [59]	7.96	1.05	LIG 1HVI [60]	10.08	0.56
LIG 1HVJ [60]	10.46	0.88	LIG 1HVK [60]	10.11	0.36
LIG 1HVL [60]	9.00	0.84	LIG 1HVR [61]	9.51	0.32
LIG 1HWR [62]	8.33	0.55	LIG 1HXB [63]	9.92	0.55
LIG 1HXW [64]	10.82	0.99	LIG 1ODW [65]	7.00	0.90
LIG 1ODY [66]	8.10	0.39	LIG 1OHR [67]	8.70	0.69
LIG 1PRO [68]	11.30	0.53	LIG 1QBR [50]	10.57	0.71
LIG 1QBS [69]	9.47	0.53	LIG 1QBT [50]	10.62	1.16
LIG 1QBU [50]	10.24	0.69	LIG 1SBG [70]	7.74	0.64
LIG 2AID [71]	4.82	1.39	LIG 2BPV [72]	7.67	0.34
LIG 2BPY [72]	7.40	0.79	LIG 2UPJ [73]	7.39	1.86
LIG 3AID [74]	6.86	1.47	LIG 3UPJ [73]	6.25	2.34
LIG 4HVP [75]	6.11	0.83	LIG 4PHV [76]	9.15	0.41
LIG 4UPJ [73]	6.80	0.84	LIG 5HVP [77]	7.70	0.43
LIG 5UPJ [78]	7.12	0.99	LIG 6UPJ [78]	6.32	1.76
LIG 7HVP [79]	9.62	0.36	LIG 7UPJ [55]	8.49	0.69
LIG 9HVP [80]	8.35	0.55			

ations in a soft grid: a 4–8 potential function is used in the grid computation. The best 100 initial orientations are optimized by means of simulated annealing on a less soft grid (6–12 potential) and finally with an atom-to-atom energy evaluation. Up to this point in the docking, the co-ordinates of both protein and ligand are kept rigid. Finally, the protein–ligand complexes obtained are relaxed by geometry optimization. The quality of the docking was computed as the root mean square deviation (rmsd) between the positions of the docked and original ligand atoms, after superimposing the protein atoms of the relaxed consensus structure and the original crystal structure. The smaller the rmsd, the more closely the docked orientation matches the crystallographic orientation. The rmsd of the inhibitors after docking in the consensus structure is given in Table 2.

Derivation of the model

An activity model is derived by making a selection of partial interaction energies, i.e., interactions with specific residues. The prediction of a particular selection of residues is computed from the sum of its interaction energies and the experimentally observed inhibition constants using a linear least squares fit. The experimentally determined inhibition constants are listed in Table 2. The quality of prediction is computed as the leave-one-out correlation coefficient (q^2). To compute this q^2, one inhibitor is left out, the best r^2 model is computed for the remaining n−1 inhibitors and this model is used to predict the value for the inhibitor that has been left out. When all compounds have been left out once, the value of q^2 is given by

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$
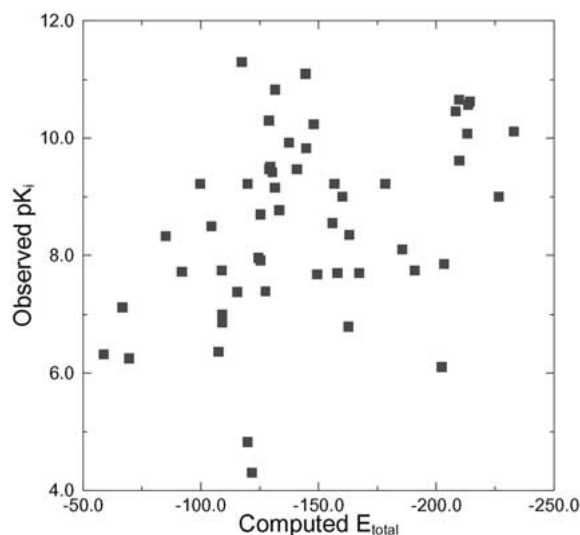
Figure 2. Observed pK$_i$ values versus total interaction energy for the 53 HIV protease inhibitors.

where y_i is the observed inhibition constant of the ith inhibitor, $\hat{y}_i$ its predicted inhibition constant not including observation i, and $\bar{y}$ the mean value for all n inhibitors.

The resulting model has the form

$$pK_i = a_1 \sum_{j=1}^{n} W_j E_j + a_0, \qquad (2)$$

where pK_i is the computed inhibition strength, a_1 and a_0 are constants determined from the fit, E_j is a contribution to the non-bonded interaction energy, and n is the number of protein residues. W_j is either zero or one, leaving these sets to be determined. Determination of the optimal set is related to the subset sum problem [34], which is known to be NP-complete. Given the dimensions of our search space, an exhaustive search would take a prohibitive amount of time. Therefore the optimal set is selected employing a genetic algorithm. In order to reduce computation time, interaction energies contributing less than 0.1 kcal/mole were excluded.

The residue selection

A genetic algorithm [35] is a simulated evolutionary process: individuals produce related offspring and survival is subjected to selective pressure according to the fitness of an individual's genes. Genetic algorithms have been shown to be efficient optimization strategies [36] in a search space that is prohibitively large for
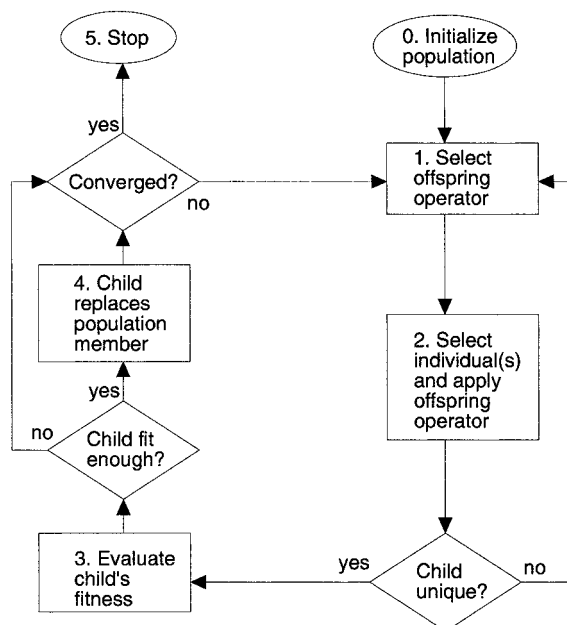

Figure 3. An overview of the steps in the genetic algorithm for feature selection.

an exhaustive search. We have implemented a genetic algorithm to search for the set of best correlating residues. The genes of an individual represent the presence or absence of a particular interaction (vector W in Equation (2) is the genome of an individual, and W_j one of its genes). The fitness of an individual is its ability to reproduce the observed pK_i values for the set of protease inhibitors, computed as the q^2 between the sum of the interaction energies for which the encoding gene is on and the observed values.

Figure 3 gives an overview of the steps in the algorithm. The acceptance decision in step 4 was taken by comparing the fitness of the child to the fitness of the parent it was created from. Doing so helps to maintain diversity in the population. If the child is fitter, it will replace one of the existing population members by roulette-wheel selection. The probability for each population member to be replaced is inversely proportional to its fitness. These probabilities are computed as

$$P_i = \frac{r_i}{\sum_{i=1}^{n} r_i}, \qquad (3)$$

where P_i is the probability for the ith individual, r_i is the rank of the individual as determined by its fitness (individuals are ranked in descending order where the first one is the fittest) and n is the population size.

The algorithm's convergence behaviour can vary in two respects: the time it takes to converge and the

actual solution it converges on. Whether the best possible solution has been found cannot be known for sure, we assume that the best solution observed in multiple trials indeed is the best possible one. The algorithm turned out to be insensitive towards the parametrization used, except for the way in which the population is initialized. If only a small fraction of genes is turned on in the starting population, the algorithm converges faster and more often on the optimal solution. Apparently, the algorithm performs better when gradually including residues that increase correlation than by finding which residues do not contribute. This gives two indications about the structure of the search space: the best solution for n residues contains the residues for the case when the n−1 residue solution is sought. Secondly, residues not in the solution rapidly destroy its fitness and hamper searching by introducing random noise on the fitness landscape.

A trial was ended if the population did not change over 1000 consecutive generations; continuing beyond this point proved not to lead to any further improvements in the solution found. The algorithm is written in ANSI C. A hundred independent trials for this problem on an SGI R14000 processor showed that convergence is achieved in 10 sec and the optimal solution is found in 98% of the cases.

Results

Best internally predictive model (BIP)

Figure 2 plots the total interaction energy versus the observed inhibition constant. The correlation coefficient r^2 is a mere 0.16 and the q^2 is 0.10. We expect to improve upon this correlation by not using all the residues.

To determine in what way the q^2 depends on the residue set size, we performed a series of runs where the set size is increased from zero to fifty. The strictly spoken undefined fitness of the zero-term model is computed by predicting each inhibitor as the average of the other inhibitors. All interactions are available for the algorithm to derive a model, only the maximum number of terms in the model is limited. Figure 4 gives a plot of the q^2 *versus* the maximum number of residues. Rather quickly a high value of q^2 is obtained, which does not change upon inclusion of more residues.

From Figure 4 it can be seen that the best internally predictive model (BIP) is reached upon inclusion of 20

terms. This model has a q^2 of 0.63 and a r^2 of 0.66. Figure 5 gives the observed versus the computed pK_i for this model. Now Equation (2) takes the following form

$$pK_i = -0.22(Arg_{A9} + Ala_{A22} + Leu_{A23} + \\ Leu_{B23} + Leu_{A24} + Leu_{B24} + Ala_{B28} + \\ Met_{A46} + Ile_{A50} + Gly_{B51} + Ile_{B54} + \\ Leu_{A76} + Leu_{B76} + Pro_{A81} + Pro_{B81} + \\ Val_{A82} + Ile_{A84} + Gly_{A86} + Gly_{B86} + \\ Arg_{B87}) + 4.9. \quad (4)$$

Note that the symmetry of the protease enzyme is broken because only one of the catalytic aspartic acids is protonated, which results in asymmetric models.

To verify that this is not a chance correlation, we performed the same genetic algorithm search with the inhibition constants randomly permuted. Figure 4 also gives the best q^2 possible for this randomized pK_i data. The q^2 obtained is the mean of five independent randomizations. It is clear that predictive models can no longer be derived, although there is the a small tendency towards higher q^2 values with increasing number of degrees of freedom. However, what cannot be seen from this plot is the magnitude of the two coefficients in the underlying model. With the real data, the model has a small intercept (a_0 in Equation (2)) and the differences in the data are explained from the difference in interaction energies. In the randomized case the models have a very small slope (a_1 in Equation (2)), meaning that all compounds are predicted to have basically the same activity. That still a q^2 of about 0.2 can be obtained this way is due to the abundance of compounds that do have an inhibition constant close to the mean.

Best externally predictive model (BEP)

As noticed by Golbraikh [37], a high value for q^2 alone is not a sufficient condition for a predictive model. It is important to validate a model by examining its ability to predict the values for an external set, i.e. a set that is not used to derive the model. To determine the external predictivity we formed ten external sets. Each set consists of seven inhibitors out of the original set, randomly chosen from each $\log_{10}$ unit of the total pK_i range. The remaining 46 inhibitors are used to derive the model. Because the accuracy in prediction of an external set depends on the particular set that is predicted, the average error for the ten sets is expressed as improvement over the zero-term prediction. Figure 6 gives the average error reduction versus the maximum
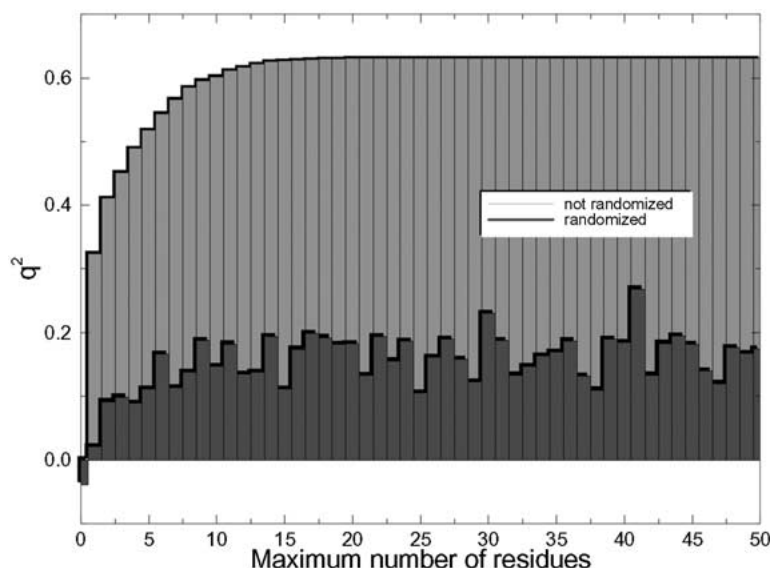
Figure 4. Best q² attainable versus the maximum number of interaction energy terms that was used to derive the model. The q² using the observed values are depicted as grey bars; the black bars give the q² after randomization of the observed values. Randomization was done independently five times, the displayed figures are averages.
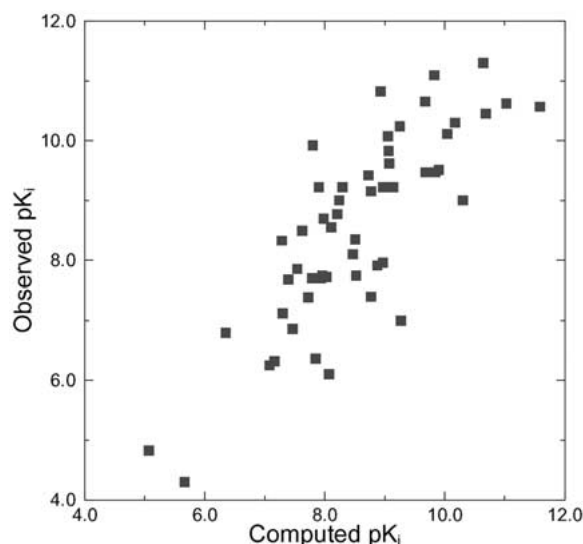

Figure 5. Observed versus computed pK$_i$ values for the 53 HIV protease inhibitors. The computed values are given by the model with the highest q² (0.63).

number of residues allowed in the model. The accuracy of the 20 residue model derived using all inhibitors in the prediction of the external sets is on average 1.2 $\log_{10}$ unit absolute deviation per compound. There appears to be a model of 9 residues that performs better at predicting the external sets. Some of the residues in the BIP model are included because they improve

the q² for the whole set. The slopes of the BIP and BEP models (Equations 4 and 5) are approximately equal, suggesting that some of the residues in the BIP model have only a minor effect on the magnitude of the computed strength.

For the best externally predictive model (BEP), Equation (2) takes the form

$$pK_i = -0.22(Leu_{B23} + Ala_{B28} + Ile_{A50} + \\ Gly_{B51} + Pro_{A81} + Pro_{B81} + \\ Val_{A82} + Ile_{A84} + Gly_{A86}) + 5.3 \quad (5)$$

The prediction for the best of the external sets is shown in Figure 7. The absolute deviation in the prediction is 0.9 $\log_{10}$ unit per compound. The prediction for the worst case is shown in Figure 8. The inhibitors in this set have an absolute deviation of 1.5 $\log_{10}$ unit per compound. The residues found in the ten external models are also present in the BIP model.

As can be seen from Figures 7 and 8 there is considerable difference in the prediction error for different external sets. We examined the compounds from the worst externally predicted set (Figure 8). This set of compounds is difficult to predict correctly regardless of the model. The pK_i for the ligand from crystal structure 4HVP is computed as 8.1, while its experimental value is 6.1. This ligand does not explicitly donate a hydrogen bond to the catalytic aspartates. Another ligand of this type, the ligand from crystal structure 1HBV, was present in one of the other
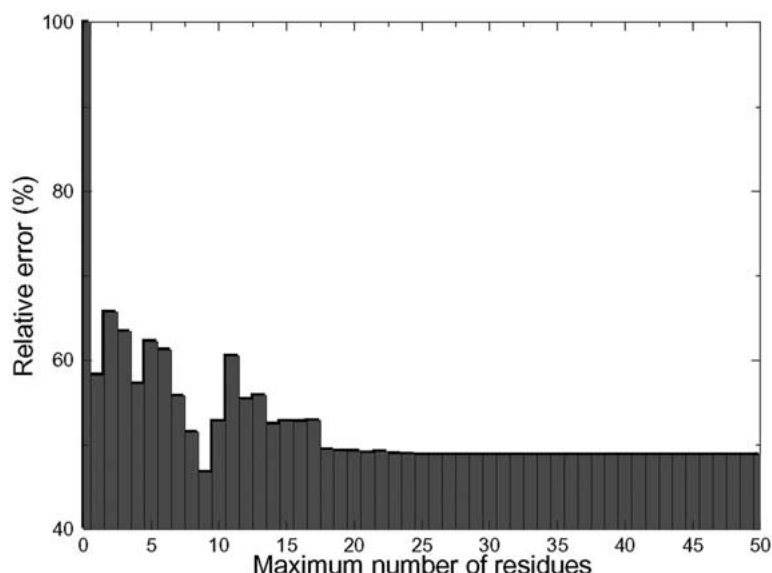
Figure 6. The relative error in prediction of an external set of seven compounds versus the maximum number of interaction energy terms that was used to derive the model. The external set is chosen randomly ten times, with observed pK_i values covering the range of activities.
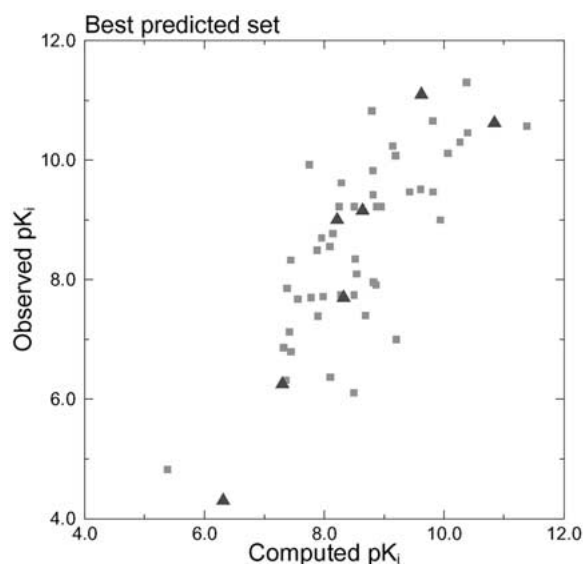
Figure 7. Observed versus computed pK_i values. From the ten external sets, the best predicted one is displayed. The inhibitors that were used to derive the model are depicted as squares, the prediction for the external set is depicted as triangles.
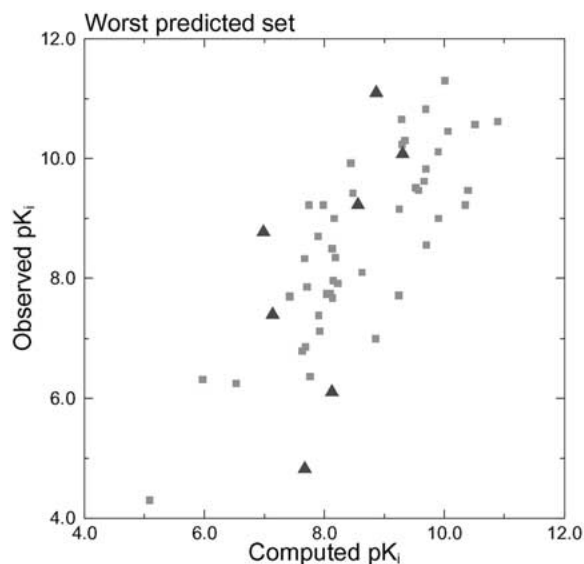
Figure 8. Observed versus computed pK_i values. From the ten external sets, the worst predicted one is displayed. The inhibitors that were used to derive the model are depicted as squares, the prediction for the external set is depicted as triangles.

external sets and also proved to be severely overestimated in strength. The interaction with the aspartates is not determining the computed figure, because the aspartates are not part of the model. Since most ligands do make a strong interaction there, it might be implicitly assumed in the model's derivation that this interaction is present. Nevertheless, the 4HVP ligand

seems difficult to predict regardless of the computational method used, as it is also mispredicted in three of the four scoring functions employed in CScore [20]. The strength for the ligand from crystal structure 2AID, which has an experimental value of 4.8, is computed at 7.7. In this crystal structure two inhibitor molecules bind to the enzyme, one in the orientation

as computed and another one next to it. We expect that the binding of this second inhibitor molecule, not considered here, contributes to the measured inhibition strength.

However, the prediction error for most compounds in an external set does not significantly change when that inhibitor is included in the derivation of the model. For example, the aforementioned ligand from 4HVP that is an outlier in the external prediction is also an outlier in the BIP model, with a computed pK_i of 8.1 (see Figure 5). Only the computed strength for the few least and most active inhibitors shows a more pronounced dependency on the model.

For comparison, we also computed the inhibition strength of the 53 compounds from our dataset with X-Score [21]. X-Score is a general scoring function and is freely available upon request from its authors. It is derived through multivariate regression analysis of 200 protein–ligand complexes. We applied X-Score to the inhibitor-consensus pocket complexes. Babel [38] was used to assign potential types to the ligand atoms. The results obtained are depicted in Figure 9. The correlation coefficient is 0.43. For this enzyme, X-Score underestimates the inhibition strength, especially for the most active compounds. Because the predictions depend on the geometry of the structures, using the original crystal structures as input will give different results. In our case these results were slightly worse (r^2 of 0.31) than the ones presented in Figure 9.

Residues in the best internally predictive and best externally predictive model

From the 198 residues of the HIV protease enzyme, 20 residues comprise the BIP model and 9 residues comprise the BEP model. These residues are not located in one single region, but are distributed over the interior of the protein. Protease has a dual character [24], consisting of both a rigid region and a more flexible region formed by two large overhanging flaps. The residues in the models are present in all regions: in the rigid region where the catalytic activity takes place, in the interconnecting loops, and in the overhanging flaps. The catalytic region is represented by three residues from the A-strand and three residues from the B-strand. One group of residues is located in the interconnecting loop composed of residues 80 through 84. This region has been proposed as one of the important determinants in substrate selection [39, 40].
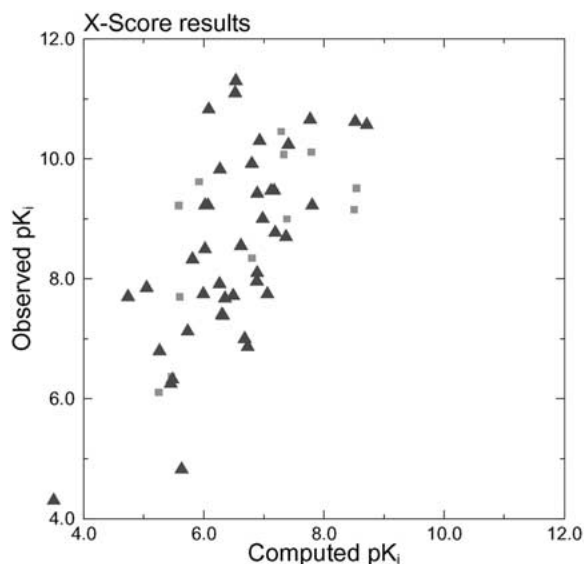

Figure 9. Observed versus computed pK$_i$ values. The inhibition strength is computed with the general scoring function X-Score [21] version 1.1, applied to the complexes that were used to derive our focussed scoring function. Squares depict the 12 inhibitors from our dataset that were also present in the X-Score dataset.

Proteases are partioned in subsites based on information about the substrate structures they process. The residues of the protein that are in contact with an amino acid of the peptide substrate define a subsite. The subsites are referred to as S_3–S_2–S_1–S_1'–S_2'–S_3', with the peptide bond or its non-scissile replacement between S_1 and S_1'. The primes refer to the C-terminal side. The definition of which residues constitute a subsite depends on the substrates used as well as the method to determine the contacts between enzyme and substrate, resulting in some variation in the definition between authors. The distribution of the BIP and BEP model residues using subsites as defined by Kulkarni [26] is depicted in Figure 10. The enzyme is coloured according to subsite, with residues from the models rendered as surfaces. As can be seen from this picture the residues are distributed over all subsites, corroborating the biochemical classification: as stated by Schechter and Berger [41] it is reasonable to expect that the subsites show specific interactions with different amino acid side-chains. The selected residues are concentrated in the inner regions (S_2–S_1–S_1'–S_2'), while the outer subsites (S_3 and S_3') contribute little. The latter subsites are less involved in substrate specificity: a wide variety of amino acid residues is tolerated here [2, 42] as compared to the more restrictive central subsites.
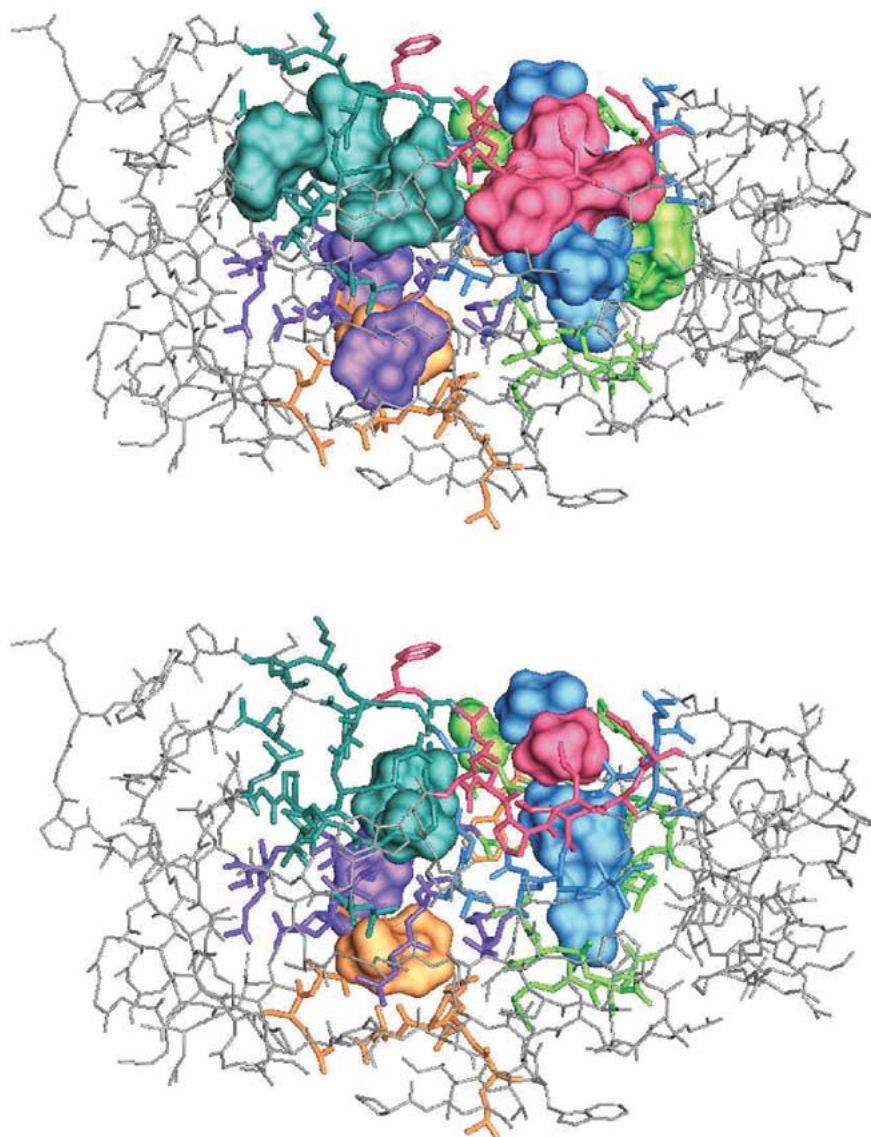
578

Figure 10. HIV protease coloured according to its subsites. The residues that define a model are rendered as surfaces. The upper enzyme shows the 20 residues from the best internally predictive model. The 9 residues from the best externally predictive model are depicted in the enzyme below.

The conserved active site residues Asp_{25}, Thr_{26} and Gly_{27} do not turn up at all in the model. The absence of the aspartic acid residues might seem surprising, because these are essential to the catalytic activity of the enzyme. The computed interactions of the inhibitors with these residues are large, especially with the deprotonated acid, and vary. Nevertheless, these interactions do not correlate with the inhibitory strength. It might be implicit in the model data that the interactions are present, but their lack of correlation can be understood from the behaviour of protease towards peptide substrates that are processed: although it is a peptide bond that interacts, well-defined sequences are processed, not arbitrary ones. The difference between a substrate and an inhibitor in the interaction with the active site aspartates is the replacement of a peptide bond by a non-hydrolyzable bond in the latter. We surmise that an inhibitor just needs a non-scissile bond and the magnitude of interaction with the aspartic acids does not contribute to its inhibitory strength.

Conclusions

We built a consensus structure for HIV protease from crystallographic data. A diverse collection of 53 protease inhibitors was docked in this consensus structure, with an average rmsd error of 0.83 for the ligand position. All inhibitors fit into the same pocket, which illustrates the relatively low flexibility of HIV protease.

We implemented a genetic algorithm to determine which residues account for the observed inhibition constants. The algorithm, that selects the set of residues whose sum of interaction energies is most predictive for the observed inhibition constants, was applied to the 53 protease complexes. The model with the best predictivity has a q^2 of 0.63, a value that is significantly higher than the q^2 of 0.10 from the total interaction energy. This best internally predictive model consists of 20 residues. The prediction of inhibition constants for external compounds leads to a model consisting of 9 residues. These residues are contained in the best internally predictive model. The external compounds are predicted within 0.9 $\log_{10}$ unit accurate for the best and 1.5 $\log_{10}$ unit for the worst set.

We found that the residues that determine protease inhibiton are located in three geometric groups: 1- around the catalytic site, 2- in the interconnecting loops, and 3- in the overhanging flaps. These groups cover all biochemically recognized subsites [41], illustrating the consistency of independent computational and biochemical approaches towards substrate recognition.

References

1. UNAIDS/WHO. AIDS epidemic update, December 2001.
2. March, D.R. and Fairlie, D.P. Designing new antiviral drugs for AIDS: HIV-1 protease and its inhibitors. Bioscience Report. R.G. Landes Company, Austin, TX, 1996.
3. Kempf, D.J., Molla, A. and Hsu, A. Protease Inhibitors as Anti-human Immunodeficiency Virus Agents, chapter 8, ASM Press, Washington, DC, 2001, pp. 147–174.
4. Ferguson, D.M., Radmer, R.J. and Kollmann, P.A., J. Med. Chem., 34 (1991) 2654.
5. Reddy, M.R., Viswanadhan, V.N. and Weinstein, J., Proc. Natl. Acad. Sci. USA, 88 (1991) 10287.
6. Cieplak, P. and Kollmann, P.A., J. Comput.-Aided Mol. Des., 7 (1993) 291.
7. Hansson, T. and Åqvist, J., Protein Eng., 8 (1995) 1137.
8. Hultén, J., Bonham, N.M., Nillroth, U., Hansson, T., Zuccarello, G., Bouzide, A., Åqvist, J., Classon, B., Danielson, U.H., Karlén, A., Kvarnström, I., Samuelsson, B. and Hallberg, A., J. Med. Chem., 40 (1997) 885.
9. Visco Jr., D.P., Pophale, R.S., Rintoul, M.D. and Faulon, J.-L., J. Mol. Graph. Model., 20 (2002) 429.
10. Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., J. Med. Chem., 36 (1993) 4152.
11. Doweyko, A.M., J. Med. Chem., 37 (1994) 1769.
12. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8 (1994) 243.
13. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118 (1996) 3959.
14. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., J. Comput.-Aided Mol. Design, 11 (1997) 425.
15. Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscount, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A., J. Med. Chem., 38 (1995) 305.
16. Verkhiver, G., Appelt, K., Freer, S.T. and Villafrance, J.E., Protein Eng., 8 (1995) 677.
17. Wallqvist, A., Jernigan, R.L. and Covell, D.G., Protein Sci., 4 (1995) 1881.
18. Muegge, I. and Martin, Y.C., J. Med. Chem., 42 (1999) 791.
19. Ishchenko, A.V. and Shakhnovich, E.I., J. Med. Chem., 45 (2002) 2770.
20. Clark, R.D., Strizhev, A., Leonard, J.M., Blake, J.F. and Matthew, J.B., J. Mol. Graph. Model., 20 (2002) 281.
21. Wang, R., Lai, L. and Wang, S., J. Comput.-Aided Mol. Design, 16 (2002) 11.
22. Todd, M.J. and Freire, E., Proteins Struct. Funct. Genet., 36 (1999) 147.
23. Luque, I. and Freire, E., Proteins Struct. Funct. Genet., S4 (2000) 63.
24. Bardi, J.S., Luque, I. and Freire, E., Biochemistry, 36 (1997) 6588.
25. Todd, M.J., Semo, N. and Freire, E., J. Mol. Biol., 283 (1998) 475.
26. Kulkarni, S.S. and Kulkarni, V.M., J. Chem. Inf. Comput. Sci., 39 (1999) 1128.
27. Vinkers, H.M., de Jonge, M.R., Daeyaert, F. F.D., Heeres, J., Koymans, L. M.H., van Lenthe, J.H., Lewi, P.J., Timmerman, H., Van Aken, K. and Janssen, P.A.J., J. Med. Chem., 46 (2003) 2765.
28. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
29. Silva, A.M., Cachau, R.E., Sham, H.L. and Erickson, J.W., J. Mol. Biol., 255 (1996) 321.
30. Kearsley, S.J., J. Comput. Chem., 11 (1990) 1187.
31. Heisterberg, D.J., Ohio Supercomputer Center, Columbus, OH, 1990.
32. http://www.ccl.net.
33. Hong, L., Treharne, A., Hartsuck, J.A., Foundling, S. and Tang, J., Biochemistry, 35 (1996) 10627.
34. LaMacchia, B.A. Basis reduction algorithms and subset sum problems. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 1991.
35. Holland, J. Adaptation in Natural and Artificial Systems. MIT Press, Cambridge, MA, 1992.
36. Wehrens, R., Pretsch, E. and Buydens, L.M.C., J. Chem. Inf. Comput. Sci., 38 (1998) 151.
37. Golbraikh, A. and Tropsha, A., J. Mol. Graph. Model., 20 (2002) 269.
38. Walters, P. and Stahl, M. Babel version 1.6, http://www.eyesopen.com/babel.

580

39. Cameron, C.E., Grinde, B., Jacques, P., Jentoft, J., Leis, J., Wlodawer, A. and Weber, I.T., J. Biol. Chem., 268 (1993) 11711.

40. Hoog, S.S., Towler, E.M., Zhao, B., Doyle, M.L., Debouck, C. and Abdel-Meguid, S.S., Biochemistry, 35 (1996) 10279.

41. Schechter, I. and Berger, A., Biochem. Biophys. Res. Commun., 27 (1967) 157.

42. Wlodawer, A. and Gustchina, A., Biochim. Biophys. Acta, 1477 (2000) 16.

43. Louis, J.M., Dyda, F., Nashed, N.T., Kimmel, A.R. and Davies, D.R., Biochemistry, 37 (1998) 2105.

44. Wu, J., Adomat, J.M., Ridky, T.W., Louis, J.M., Leis, J., Harrison, R.W. and Weber, I.T., Biochemistry, 37 (1998) 4518.

45. Hong, L., Zhang, X.-J., Foundling, S., Hartsuck, J.A. and Tang, J., FEBS Lett., 420 (1997) 11.

46. Bäckbro, K., Löwgren, S., Österlund, K., Atepo, J., Unge, T., Hultén, J., Bonham, N.M., Schaal, W., Karlén, A. and Hallberg, A., J. Med. Chem., 40 (1997) 898.

47. Munshi, S., Chen, Z., Yan, Y., Li, Y., Olsen, D.B., Schock, H.B., Galvin, B.B., Dorsey, B. and Kuo, L.C., Acta Crystallogr. D56 (2000) 381.

48. Tyndall, J. D.A., Reid, R.C., Tyssen, D.P., Jardine, D.K., Todd, B., Passmore, M., March, D.R., Pattenden, L.K., Bergman, D.A., Alewood, D., Hu, S.-H., Alewood, P.F., Birch, C.J., Martin, J.L. and Fairlie, D.P., J. Med. Chem., 43 (2000) 3495.

49. Thaisrivongs, S., Skulnick, H.I., Turner, S.R., Strohbach, J.W., Tommasi, R.A., Johnson, P.D., Aristoff, P.A., Judge, T.M., Gammill, R.B., Morris, J.K., Romines, K.R., Chrusciel, R.A., Hinshaw, R.R., Chong, K.-T., Tarpley, W.G., Poppe, S.M., Slade, D.E., Lynn, J.C., Horng, M.-M., Tomich, P.K., Seest, E.P., Dolak, L.A., Howe, W.J., Howard, G.M., Schwende, F.J., Toth, L.N., Padbury, G.E., Wilson, G.J., Shiou, L., Zipp, G.L., Wilkinson, K.F., Rush, B.D., Ruwart, M.J., Koeplinger, K.A., Zhao, Z., Cole, S., Zaya, R.M., Kakuk, T.J., Janakiraman, M.N. and Watenpaugh, K.D., J. Med. Chem., 39 (1996) 4349.

50. Jadhav, P.K., Ala, P., Woerner, F.J., Chang, C.-H., Garber, S.S., Anton, E.D. and Bacheler, L.T., J. Med. Chem., 40 (1997) 181.

51. Dohnálek, J., Hašek, J., Dušková, J., Petroková, H., Hradilek, M., Souček, M., Konvalinka, J., Brynda, J., Sedláček, J. and Fábry, M. Acta Crystallogr. D57 (2001) 472.

52. 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. and Abdel-Meguid, S.S., J. Med. Chem., 38 (1995) 3246.

53. Murthy, K. H.M., Winborne, E.L., Minnich, M.D., Culp, J.S. and Debouck, C., J. Biol. Chem., 267 (1992) 22770.

54. Abdel-Meguid, S.S., Zhao, B., Murthy, K. H.M., Winborne, E., Choi, J.-K., DesJarlais, R.L., Minnich, M.D., Culp, J.S., Debouck, C., Tomaszek Jr., T.A., Meek, T.D. and Dreyer, G.B., Biochemistry, 32 (1993) 7972.

55. Skulnick, H.I., Johnson, P.D., Aristoff, P.A., Morris, J.K., Lovasz, K.D., Howe, W.J., Watenpaugh, K.D., Janakiraman, M.N., Anderson, D.J., Reischer, R.J., Schwartz, T.M., Banitt, L.S., Tomich, P.K., Lynn, J.C., Horng, M.-M., Chong, K.-T., Hinshaw, R.R., Dolak, L.A., Seest, E.P., Schwende, F.J., Rush, B.D., Howard, G.M., Toth, L.N., Wilkinson, K.R., Kakuk, T.J., Johnson, C.W., Cole, S.L., Zaya, R.M., Zipp, G.L., Possert, P.L., Dalga, R.J., Zhong, W.-Z., Williams, M.G. and Romines, K.R., J. Med. Chem., 40 (1997) 1149.

56. Thompson, S.K., Murthy, K. H.M., Zhao, B., Winborne, E., Green, D.W., Fisher, S.M., DesJarlais, R.L., Tomaszek Jr., T.A., Meek, T.D., Gleason, J.G. and Abdel-Meguid, S.S., J. Med. Chem., 37 (1994) 3100.

57. Kim, E.E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D. and Navia, M.A., J. Am. Chem. Soc., 117 (1995) 1181.

58. Chen, Z., Li, Y., Chen, E., Hall, D.L., Darke, P.L., Culberson, C., Shafer, J.A. and Kuo, L.C., J. Biol. Chem., 269 (1994) 26344.

59. Jadhav, P.K., Woerner, F.J., Lam, P. Y.S., Hodge, N., Eyermann, C.J., Man, H.-W., Daneker, W.F., Bacheler, L.T., Rayner, M.M., Meek, J.L., Erickson-Viitanen, S., Jackson, D.A., Calabrese, J.C., Schadt, M. and Chang, C.-H., J. Med. Chem., 41 (1998) 1446.

60. Hosur, M.V., Bhat, T.N., Kempf, D.J., Baldwin, E.T., Liu, B., Gulnik, S., Wideburg, N.E., Norbeck, D.W., Appelt, K. and Erickson, J.W., J. Am. Chem. Soc., 116 (1994) 847.

61. Lam, P. Y.S., 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., Chang, C.-H., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viitanen, S., Science, 263 (1994) 380.

62. Ala, P.J., DeLoskey, R.J., Huston, E.E., Jadhav, P.K., Lam, P. Y.S., Eyermann, C.J., Hodge, C.N., Schadt, M.C., Lewandowski, F.A., Weber, P.C., McCabe, D.D., Duke, J.L. and Chang, C.-H., J. Biol. Chem., 273 (1998) 12325.

63. Krohn, A., Redshaw, S., Ritchie, J.C., Graves, B.J. and Hatada, M.H., J. Med. Chem., 34 (1991) 3340.

64. Kempf, D.J., Marsh, K.C., Denissen, J.F., McDonald, E., Vasavanonda, S., Flentge, C.A., Green, B.E., Fino, L., Park, C.H., Kong, X.-P., Wideburg, N.E., Saldivar, A., Ruiz, L., Kati, W.M., Sham, H.L., Robins, T., Stewart, K.D., Hsu, A., Plattner, J.J., Leonard, J.M. and Norbeck, D.W., Proc. Natl. Acad. Sci. USA, 92 (1995) 2484.

65. Kervinen, J., Thanki, N., Zdanov, A., Tino, J., Barrish, J., Lin, P.F., Colonno, R., Riccardi, K., Samanta, H. and Wlodawer, A., Protein Pept. Lett., 3 (1996) 399.

66. Kervinen, J., Lubkowski, J., Zdanov, A., Bhatt, D., Dunn, B.M., Hui, K.Y., Powell, D.J., Kay, J., Wlodawer, A. and Gustchina, A., Protein Sci., 7 (1998) 2314.

67. Kaldor, S.W., Kalish, V.J., Davies II, J.F., Shetty, B.V., Fritz, J.E., Appelt, K., Burgess, J.A., Campanale, K.M., Chirgadze, N.Y., Clawson, D.K., Dressman, B.A., Hatch, S.D., Khalil, D.A., Kosa, M.B., Lubbehusen, P.P., Muesing, M.A., Patick, A.K., Reich, S.H., Su, K.S. and Tatlock, J.H., J. Med. Chem., 40 (1997) 3979.

68. Sham, H.L., Zhao, C., Stewart, K.D., Betebenner, D.A., Lin, S., Park, C.H., Kong, X.-P., Rosenbrook Jr., W., Herrin, T., Madigan, D., Vasavanonda, S., Lyons, N., Molla, A., Saldivar, A., Marsh, K.C., McDonald, E., Wideburg, N.E., Denissen, J.F., Robins, T., Kempf, D.J., Plattner, J.J. and Norbeck, D.W., J. Med. Chem., 39 (1996) 392.

69. Lam, P. Y.S., Ru, Y., Jadhav, P.K., Aldrich, P.E., DeLucca, G.V., Eyermann, C.J., Chang, C.-H., Emmett, G., Holler, E.R., Daneker, W.F., Li, L., Confalone, P.N., McHugh, R.J., Han, Q., Li, R., Markwalder, J.A., Seitz, S.P., Sharpe, T.R., Bacheler, L.T., Rayner, M.M., Klabe, R.M., Shum, L., Winslow, D.L., Kornhauser, D.M., Jackson, D.A., Erickson-Viitanen, S. and Hodge, C.N., J. Med. Chem., 39 (1996) 3514.

70. Abdel-Meguid, S.S., Metcalf, B.W., Carr, T.J., Demarsh, P., DesJarlais, R.L., Fisher, S., Green, D.W., Ivanoff, L., Lambert, D.M., Murthy, K. H.M., Petteway Jr., S.R., Pitts, W.J., Tomaszek Jr., T.A., Winborne, E., Zhao, B., Dreyer, G.B. and Meek, T.D., Biochemistry, 33 (1994) 11671.

71. Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S., Furth, P.S., Ortiz de Montellano, P.R., Meng, E., Kuntz, I.D., De-

Camp, D.L., Salto, R., Rosé, J.R., Craik, C.S. and Stroud, R.M., J. Biol. Chem., 268 (1993) 15343.

72. Munshi, S., Chen, Z., Li, Y., Olsen, D.B., Fraley, M.E., Hungate, R.W. and Kuo, L.C., Acta Crystallogr. D54 (1998) 1053.

73. Thaisrivongs, S., Watenpaugh, K.D., Howe, W.J., Tomich, P.K., Dolak, L.A., Chong, K.-T., Tomich, C.-S.C., Tomasselli, A.G., Turner, S.R., Strohbach, J.W., Mulichak, A.M., Janakiraman, M.N., Moon, J.B., Lynn, J.C., Horng, M.-M., Hinshaw, R.R., Curry, K.A. and Rothrock, D.J., J. Med. Chem., 38 (1995) 3624.

74. Rutenber, E.E., McPhee, F., Kaplan, A.P., Gallion, S.L., Hogan Jr., J.C., Craik, C.S. and Stroud, R.M., Bioorg. Med. Chem., 4 (1996) 1545.

75. Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S. B.H. and Wlodawer, A., Science, 246 (1989) 1149.

76. Bone, R., Vacca, J.P., Anderson, P.S. and Holloway, M.K., J. Am. Chem. Soc., 113 (1991) 9382.

77. Fitzgerald, P. M.D., McKeever, B.M., VanMiddlesworth, J.F., Springer, J.P., Heimbach, J.C., Leu, C.-T., Herber, W.K., Dixon, R. A.F. and Darke, P.L., J. Biol. Chem., 265 (1990) 14209.

78. Romines, K.R., Watenpaugh, K.D., Tomich, P.K., Howe, W.J., Morris, J.K., Lovasz, K.D., Mulichak, A.M., Finzel, B.C., Lynn, J.C., Horng, M.-M., Schwende, F.J., Ruwart, M.J., Zipp, G.L., Chong, K.-T., Dolak, L.A., Toth, L.N., Howard, G.M., Rush, B.D., Wilkinson, K.F., Possert, P.L., Dalga, R.J. and Hinshaw, R.R., J. Med. Chem., 38 (1995) 1884.

79. Swain, A.L., Miller, M.M., Green, J., Rich, D.H., Schneider, J., Kent, S. B.H. and Wlodawer, A., Proc. Natl. Acad. Sci. USA, 87 (1990) 8805.

80. 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., Kohlbrenner, W.E., Simmer, R., Helfrich, R., Paul, D.A. and Knigge, M., Science, 249 (1990) 527.

81. http://cactus.nci.nih.gov/services/translate.