J Comput Aided Mol Des (2008) 22:287–297
DOI 10.1007/s10822-008-9186-7

Docking and multivariate methods to explore HIV-1 drug-resistance: a comparative analysis

Anna Maria Almerico · Marco Tutone · Antonino Lauria

Received: 2 August 2007 / Accepted: 23 January 2008 / Published online: 14 February 2008

Abstract In this paper we describe a comparative analysis between multivariate and docking methods in the study of the drug resistance to the reverse transcriptase and the protease inhibitors. In our early papers we developed a simple but efficient method to evaluate the features of compounds that are less likely to trigger resistance or are effective against mutant HIV strains, using the multivariate statistical procedures PCA and DA. In the attempt to create a more solid background for the prediction of susceptibility or resistance, we carried out a comparative analysis between our previous multivariate approach and molecular docking study. The intent of this paper is not only to find further support to the results obtained by the combined use of PCA and DA, but also to evidence the structural features, in terms of molecular descriptors, similarity, and energetic contributions, derived from docking, which can account for the arising of drug-resistance against mutant strains.

Keywords HIV-1 · NNRTIs · Molecular docking · Multivariate analysis · Mutation · PIs · Resistance

Abbreviations

NNRTIs	Non-nucleoside reverse transcriptase inhibitors
PIs	Protease inhibitors
PR	Protease
RT	Reverse transcriptase

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9186-7) contains supplementary material, which is available to authorized users.

A. M. Almerico (✉) · M. Tutone · A. Lauria
Dipartimento Farmacochimico, Tossicologico e Biologico, Università degli Studi di Palermo, Via Archirafi 32, 90123 Palermo, Italy
e-mail: almerico@unipa.it

Introduction

The treatment regimens for the human immunodeficiency virus type-1 (HIV-1) may result in dramatic suppression of viral replication in infected individuals. The therapy has included both HIV protease (PR) and reverse transcriptase (RT) enzyme inhibitors. When viral replication is incompletely suppressed, drug-resistant variants emerge through the accumulation of mutations in the HIV-1 RT o PR genes. The arise of these drug-resistant mutations during treatment have significantly affected patient management and the long-term effectiveness of antiretroviral therapy [1, 2]. Evaluating the effects of these mutations has become an important factor in developing treatment strategies for infected patients [3, 4]. Several methodologies have been developed to either phenotypically or genotypically determine HIV-1 drug susceptibility [5, 6].

Phenotypic tests are long-lasting but expensive, on the other side, genotypic testing for resistance is a relatively rapid and inexpensive method to identify PR and RT amino acid substitutions leading to drug resistance [7, 8]. Genotyping is recommended for use in clinical practice [9–11]. However, rules-based interpretation systems are retrospectives in nature and must be frequently updated to accommodate new mutational patterns and new antiretrovirals, so the interpretation of genotyping information is difficult when complex mutational patterns and large numbers of polymorphisms interact to cause resistance, cross or reversal resistance [11].

Several works have reported attempts to develop computational methodologies for molecular modeling of HIV-1 enzyme-inhibitor complexes, and to build QSAR models for activity prediction [12–14]. However, these computational methods have not been developed to predict resistance to RT and PR variants. Only few works have been published with

the aim to consider phenotypically HIV-1 PR inhibitors resistance [15, 16], on the basis of the available data of the FDA approved drugs. But, to date, comparative analysis between multivariate models and docking results was never used to disclose the characteristics which can influence the susceptibility or resistance of both NNRTI and PI against the most common mutant strains of RT and PR. Previously, we have developed a simple but efficient method to evaluate, on the basis of physico-chemical descriptors and structural similarity, the features of compounds that are less likely to trigger resistance or are effective against mutant HIV strains, describing the use of multivariate statistical procedures, PCA (Principal Component Analysis) and DA (Discriminant Analysis), as tools to explore the inhibitory activity of NNRTIs and PIs classes against HIV-1 viruses (wild type RT and more frequent mutants, Y181C, V106A, K103N, L100I; wild type PR and more frequent single mutants, V82A, V82F, I84V) [17, 18].

This proposed procedure can be considered as a reliable method for the activity prediction of inhibitors, for which the data against mutant strains have not been reported. The approach could be used as a sufficiently good and fast discriminator to evaluate preliminarily the probable resistance to newer synthesized compounds before it is actually verified in biological tests. In the attempt to create a more solid background for the susceptibility or resistance prediction of an inhibitor against a mutant strain, we carried out a comparative analysis between these multivariate results and a molecular docking studies, performed on the same sets of compounds already considered and PDB structures of the enzymes (RT and PR) which present the mutations reported in our works [17, 18]. The molecular structure of a drug can not prevent the appearance of a mutation, due to the high rate replication of the virus and to the lack of repairing enzymes. Thus the intent of this paper is not only to confirm the results obtained by the combined use of PCA and DA, but also to evidence the structural characteristics, in terms of molecular descriptors, similarity, and energetic contributions derived from docking, which can justify the arising of drug-resistance or explain the activity against a mutant strain.

Materials and methods

Dataset

A total of 55 NNRTIs and 51 PIs (peptidomimetic and non peptide compounds) was retrieved from the database of National Institute of Allergy and Infectious Diseases (NIAID) [19] and from literature, the inhibitors structures and the full reference list is reported in the supplementary information (SI-F1, SI-F2, SI-T1, and SI-T2). As previously reported, experimental data of activity for both wild type and mutant strains of RT and PR were selected. The phenotypic results were classified as resistant (R) or susceptible (S) based on the fold change cut-off values previously calculated [17, 18].

Protein structures selection

The X-ray crystal structures of HIV-1 enzyme-inhibitor complexes used for the binding energy determination were obtained from the Brookhaven Protein Data Bank (PDB) [20]. First, the structures of the two native proteins and structures which present single mutation were selected. The choice of crystallographic structures, presenting such characteristics, makes easier to identify the differences between the native and the mutated active site and, therefore, the features of the inhibitors which can modulate the activity. It was impossible to select suitable structure for L100I and V106A mutants due to lack of single mutant structures in the PDB.

Only few structure files of single mutant enzyme were available, and, in order to retain the homogeneity in the conformations of the proteins, in both cases, structures with the same ligand bound in the wild type strain and in the mutant strains were chosen (Table 1). For RT, two crystal structures, with efavirenz or with nevirapine, as ligands, were selected. Due to a higher resolution, the structures with efavirenz were preferred: 1FK9 for wt, 1FKO for K103N, and 1JKH for Y181C. Instead for PR, two wild type proteins

Table 1 Retrieved structures of HIV-1 RT and PR in the Protein Data Bank (selected in italic)

Ligand	Mutation	PDB ID	Resolution (Å)
Reverse transcriptase			
Efavirenz	K103N	1IKV	3.00
Efavirenz	K103N	*1FKO*	2.90
Nevirapine	K103N	1FKP	2.90
PNU142721	K103N	1IKX	2.80
Efavirenz	Y181C	*1JKH*	2.50
TNK651	Y181C	1JLA	2.50
Nevirapine	Y181C	1JLB	3.00
Efavirenz	Wild Type	*1FK9*	2.50
Protease			
A77003	V82A	*1HVS*	2.25
DMP450	I84V	1MER	1.90
DMP323	I84V	*1MES*	1.90
DMP323	V82F	*1MET*	1.90
A77003	Wild Type	*1HVI*	1.80
DMP323	Wild Type	*1QBS*	1.80

were selected due to the presence of two different ligands in the mutant strains (1QBS and 1HVI for wt, 1HVS for V82A, 1MET for V82F, and 1MES for I84V).

Proteins and ligands setup

Preparation of inhibitor structures was carried out as previously reported [17, 18, 21]. In the case of protein structures, missing hydrogens are added to the templates and the water molecules solved by X-ray crystallography are removed from the proteins. The cavity is explored to identify and select the protein active site. To this aim, the Ligandfit software, included in the modeling environment of Cerius^2 by Accelrys, was used [22]. Ligandfit involves the use of a flood-filling algorithm which begins by first constructing a rectangular grid with an user-defined spacing. In these investigation, the grid spacing was set as default (0.5 Å). Each grid point was then classified as either an occupied or free point. Occupied grid points are those that lie within the contact distance of the nearest protein atom. The contact distance is set equal to the radius of the protein atom. The radius of each protein heavy atom is set at 2.5 Å, while the radius for protein hydrogen atoms is set to 2.0 Å. Grid points lying outside contact distance are free (unoccupied). An "eraser", which determines the opening size of the site, removes then the free grid points lying outside the protein and it retains the entire found cavity which contains a specified number of free points. The opening size was set to 5.0 Å and the cut-off size to 100 free points (Protein Shape mode, PS). Ligandfit provides another tool for constructing the site which is based on the presence of a known ligand pose (Docked Ligand mode, DL). In this case, all free grid points that lie within the radius of any ligand atom are determined. The radius of both types of atoms was set as in PS mode. In this work both types of cavity detection were employed to characterize the binding site sizes.

Docking settings

Docking calculations were carried out employing Monte Carlo method for the ligands conformational search. During the search, bond lengths and bond angles were fixed, only the rotatable bonds were allowed to rotate freely. The number of trials for the Monte Carlo search was fixed to 10,000 and the maximum number of conformers saved was set to 10, and each conformer is classified different from another if RMSD (root mean square deviation) value is up to 1.50. In the calculation of ligands internal energy, the electrostatic energy beyond the van der Waals energy (CFF1.02 force field) was also included. Both contributes were computed using the 9-6 Lennard-Jones functions. For the docking energy calculation, the softened Lennard-Jones potential, with a trilinear interpolation and a rigid body minimization of the nonbonded interaction energy between the ligands and the protein, was used [23]. The final docked energy was calculated as the sum of the intermolecular energy and the internal energy of ligand (all the details of the docking setting are included in SI-T3).

Scoring docked ligands

Ligandfit protocol employs scoring functions to prioritize docked ligand relative to one another and to predict binding affinities. The pool of available scoring functions comprises LigScore [23], Ludi [24], PLP [25, 26], and PMF [27]. LigScore is a fast, simple scoring function for predicting protein-ligand binding affinities. Three descriptors are used to calculate LigScore, as shown in Eq. 1.

$$LigScore = pK_i = vdW + (C + pol) + Totpol^2 \quad (1)$$

where vdW is in kcal/mol, C + pol is in Å^2, Totpol^2 is in Å^4

Totpol^2 represents surface descriptors, but while C+pol is a count of the buried polar surface area between protein and ligand involving attractive interactions, the first involves attractive and repulsive protein-ligand interaction. Similar equation is used for LigScore depending on the vdW descriptor being calculated by means CFF or Dreiding force field. Two kinds of LigScore (LIGSC1 and LIGSC2) functions can be employed: they differ on the considered types of protein-ligand interaction. At variance with LIGSC1, LIGSC2, besides using the rule based polarity assignment, includes a third descriptor Burypol^2, which is the squared sum of the buried polar surface area of the protein as well as the ligand molecule.

PLP is a simple scoring function, in arbitrary units of energy, which has been shown to correlate well with protein-ligand binding affinities (Eq. 2).

$$pK_i = -PLP \quad (2)$$

In the PLP1 function, each non-hydrogen ligand or non-hydrogen receptor atom is assigned a PLP atom type. All hydrogen atoms are excluded from the PLP function. There are four PLP atom types:

1. Hydrogen bond (H-bond) donor; 2. H-bond acceptor; 3. As both H-bond donor and acceptor; and 4. Non-polar. There are two types of pairwise interactions in PLP1, namely H-bond and steric. The two interactions are described by the same functional form, but with different parameters. PLP1 score is the sum of the function values of all pairwise interactions in a receptor/ligand complex.

In the PLP2 function, PLP atom typing remains the same as in PLP1, but in addition, a PLP atomic radius is

assigned to each atom, except for hydrogen. There are three different radii: small (1.4 Å for F and metal ions); medium (1.8 Å for C, O, and N); large (2.2 Å for S, P, Cl, and Br).

There are three types of pairwise interactions in PLP1, namely H-bond, dispersion and repulsion. There are two types of functional forms, for H-bond and dispersion interactions, which have the same function form, but with different parameters. A scaling factor is used for H-bond and repulsion terms based on the angle of receptor/ligand atoms involved. The PLP2 score is the sum of the function values of all pairwise interactions in a receptor/ligand complex.

PMF is defined as the sum of the free energies of interaction over all interatomic pairs of the protein-ligand complex, including folding entropy of protein and ligand in the complex conformation, solvation terms and terms that compensate for the loss of conformational freedom of the ligand upon complex conformation (Eq. 3). The PMF scores are reported in arbitrary energy units.

$$\Delta G_{bind} = \frac{PMF}{\varepsilon} \quad (3)$$

Ludi score (kcal/mol) is a sum of five contributions: 1. ideal H-bonds; 2. perturbed ionic interactions; 3. lipophilic interactions; 4. freezing of freedom internal degrees; 5. loss of translational and rotational entropy of the ligand (Eq. 4).

$$Ludi = k\Delta G_{bind} \quad (4)$$

Results and discussion

On the structures of RT and PR, selected from the PDB, the two site search modes were employed according to the material and methods section, and the results are reported in Table 2.

As known, the size of the RT active site is quite small, and it is 10 Å away from the active site of polymerase. An analysis of the active site size calculated by both methods revealed that there is only a little increase when the PS mode was used, demonstrating that the ligand usually occupies quite completely the active site (Fig. 1). The most interesting result is the decrease of the mutant strain binding site size with respect to the wild type one. In the case of K103N mutation, the distance of the ligand from the from key amino group in the mutated amino acid increases from 3.59 Å to 4.23 Å. On the contrary, there is a little decrease of the distance from Lys101 for the bound inhibitor. But, as overall effect, there is a reduction of the mutated enzyme active site because of the presence of two additional H-bonds (Leu234-His235, His235-Pro236) respect the wild type protein. In the case of Y181C, as well, a reduction in the calculated active site was found, although the side chain of Cys is less bulky than the Tyr side chain. This change leads to a distance increase of the ligand from the Cys residue, but allows the formation of six new H-bonds (Pro95-Trp229, Leu100-Lys101, Lys101-Lys103, Lys103-Pro236, Leu234-His235, His235-Pro236), which produces a sensible decrease of the active site size.

Table 2 Size of the Ligandfit calculated active site and ratio between the sizes of mutant and wild type active sites

	PDB ID	Free points		Strain
		DL	PS	
RT	1FK9	1716	1979	Wild Type
	1FKO	1688	1856	K103N
	1JKH	1558	1774	Y181C
PR	1HVI	6934	1663	Wild Type
	1QBS	4153	2014	Wild Type
	1HVS	6954	1453	V82A
	1MES	4167	1997	I84V
	1MET	4029	3167	V82F

	Strain	Mutant/WT	Ratio	
			DL	PS
RT	Y181C	1JKH/1FK9	0.910	0.900
	K103N	1FKO/1FK9	0.980	0.937
PR	I84V	1MES/1QBS	1.003	0.990
	V82F	1MET/1QBS	0.970	1.570
	V82A	1HVS/1HVI	1.003	0.870

The calculated active sites of PR in the two different modes show a neat difference. In DL mode, the number of free points, that characterize the binding pocket, is much higher than the retrieved number in PS mode. Such a result can be rationalized considering that the active site, formed only by the amino acids Arg8 and Asp29, is rather open, exceeding 5.0 Å, so that in PS search mode, the algorithm does not recognize it as a part of the three-dimensional space to be included into the active site. Another sharp difference can be found between the structure of the proteins that bind DMP323 (1QBS, 1MES, 1MET) and A77003 (1HVI, 1HVS), as shown in Figs. 2 and 3. The two different ligands have really different structures, in terms of molecular mass (DMP323 593.01 Da, A77003 827.42 Da), accessible surface area (DMP323 550.25 $Å^2$, A77003 690.88 $Å^2$), molecular volume (DMP323 463.26 $Å^3$, A77003 653.44 $Å^3$). This furnishes a reasonable explanation for the high number of free-points retrieved in the calculated active site from the two structures and justify the selection of two different wt proteins related to the ligand bound to the mutated protein.

In the case of PR, data obtained by PS mode resulted not suitable enough to carry on the docking calculation, because both the size of the calculated active site is too small to set peptidomimetic/non-peptide inhibitors and the

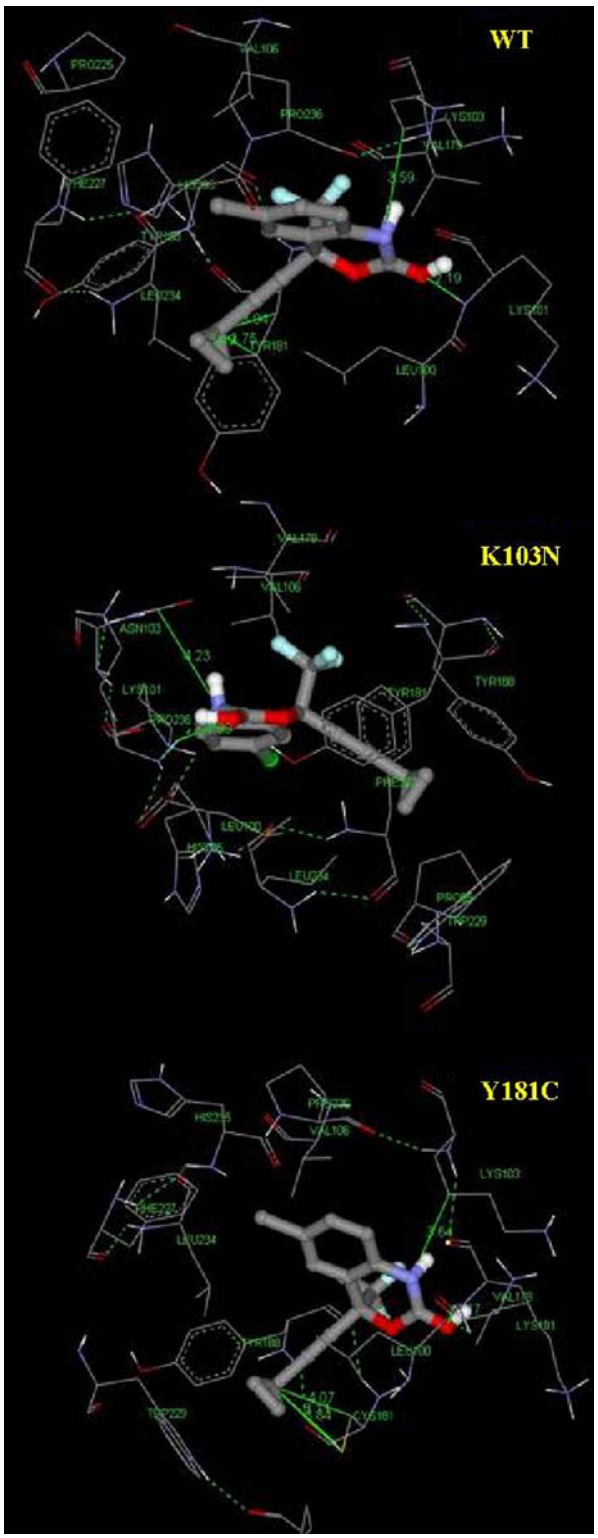

Fig. 1 NNRTIs binding sites as found in crystal structures with efavirenz bound

ratio mutant/wild type can result overestimated: in the case of V82F an increment of 50% of the size for the mutant was found, which of course is not useful. Therefore, for the docking calculation in the case of protease, we used only the DL method for the determination of the active sites. From a structural point of view, the mutation V82A determines a reduced lipophilicity in the pocket and the loss of H-bond interactions with the residue Asp25'. But the most relevant structural difference is the loss of three H-bonds with Thr80 and Pro81, due to the changed amino acid and to the increment of the active site size. The mutation V82F also involves an increase of the lipophilicity in the pocket, joint to the reduction of the active site. In this case, in fact, the distances of the ligand from key residues (Val/Phe) decrease, from 3.78–4.30 Å to 3.57 Å. Also, the distances from Asp25/25′ is reduced, whereas the number and the type of H-bond are unchanged. As overall result, this mutation involves variation of steric and lipophilic interactions near the amino acid Phe82. Mutation I84V determines only a slight increase in lipophilicity. The distances of the ligand from the mutated amino acid is partially unaltered, considering the presence only of one additional methyl group, while the distances from Asp25/25′ are shorter than those found in the case of wild type. The number of H-bonds in the core of the mutated active site is higher than in the wt one. The region between Thr80, Pro81 and Val82 is the most interested by this change and the direct consequence of the new H-bonds presence is the formation of an additional interaction between Gly49 and Ile50.

The docking scores of the complexes were then analyzed. All the scoring functions, as available in Ligandfit, were considered in order to evaluate all the possible energetic contributions of the protein-ligand interaction. Usually when a docking study is carried out, the conformer with the best scoring value is taken into account as the entity which better represents the virtual interaction ligand-protein, but, we decided to utilize 10 scoring outputs (the best of 10,000 trials), in each case. This way of treating the docking results can improve the statistical validity, since usually in these calculations the experimentally known binding mode of an inhibitor is present within a given range of the best docking solutions.

To better manipulate the huge amount of data obtained by the docking scores, for the comparative analysis, we decided to consider the maximum value among the selected 10 scores, the minimum value and the mean value of all the scoring obtained for every inhibitor studied. The algebraic difference of the scores ($Q = Score\ mutant - Score\ wild\ type$) was calculated as the difference between the score obtained in the case of the mutant protein and the score calculated on the wild type protein, with the aim of classifying as S or R all the ligand molecules. In fact if $Q > 0$ the hypothesis that the docked compound is potentially active against the mutant strain (S) could be considered, if $Q < 0$ the compound could be classified as less active against the

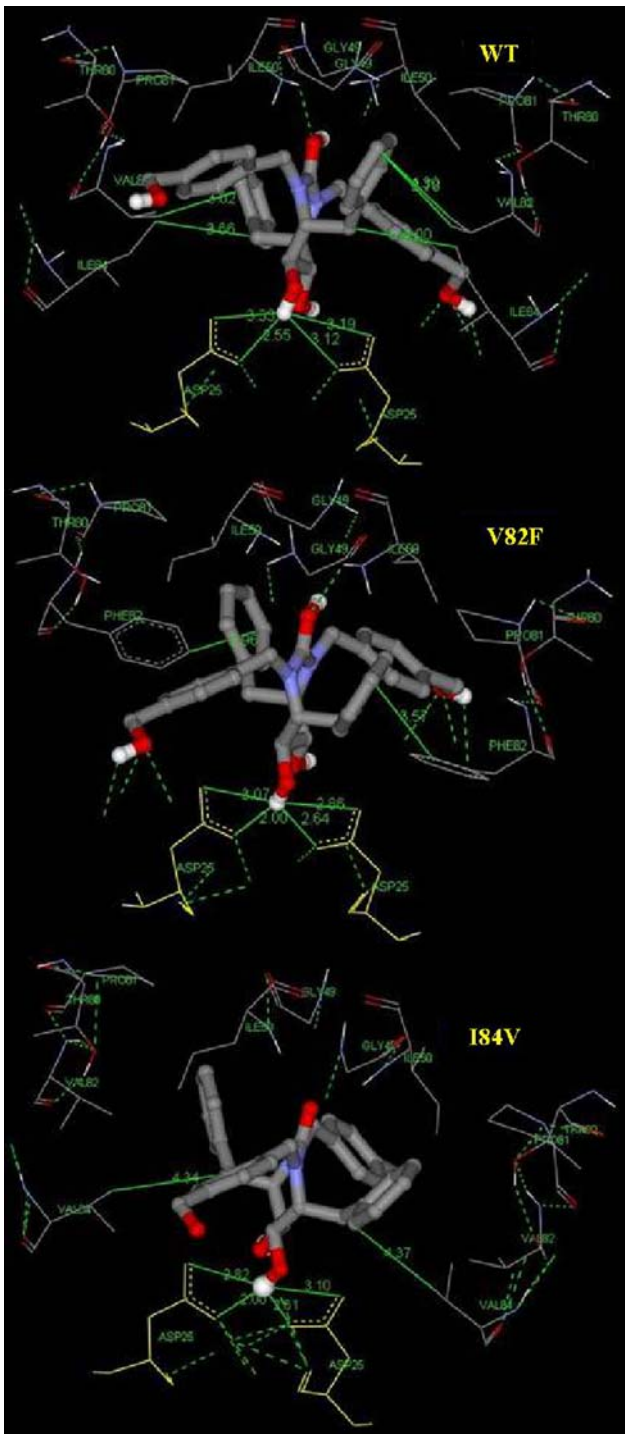

Fig. 2 PIs binding sites as found in crystal structures with DMP323 bound

mutant (R) with respect to the native protein. Every scoring function was separately considered and the compounds are definitively assigned as S or R when at least 2 of 3 scoring values (Q_{max}, Q_{min}, Q_{mean}) resulted concordant. This way of scores interpretation has the advantage to compare the descriptors which define the scoring functions with the

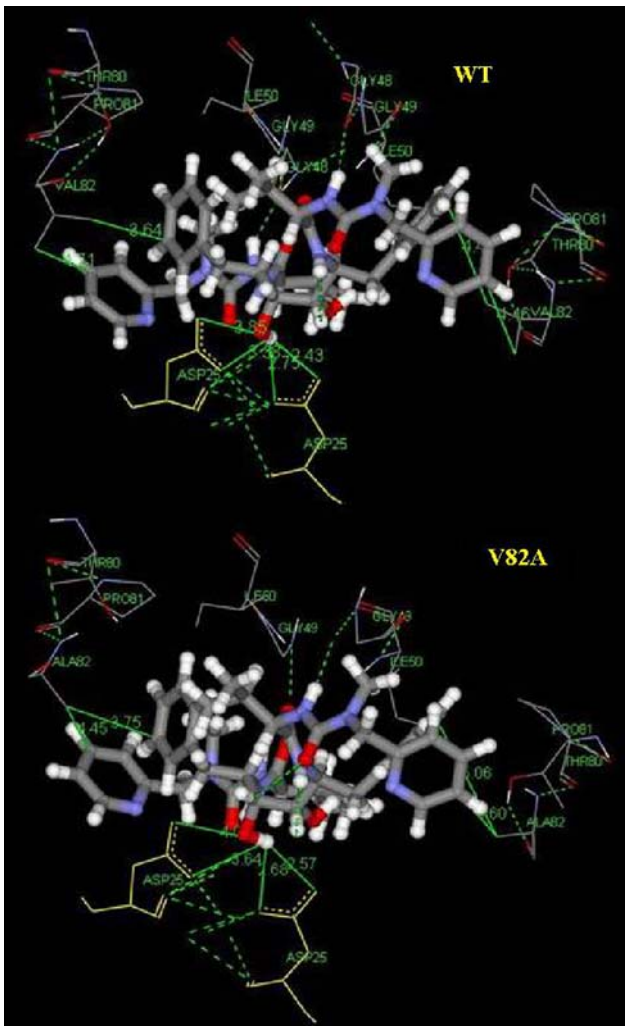

Fig. 3 PIs binding sites as found in crystal structures with A77003 bound

descriptors involved in the multivariate analysis. The assignments of all the compounds to S or R classes are reported in the supporting information (Tables SI-T4 and SI-T5), whereas the percentage of correct assignment to class S or R for all compounds with known activity is reported in Fig. 4.

Comparative analysis of multivariate/docking approaches for inhibitors, R/S classified on the basis experimental data

Analysing, first separately, the results obtained by using the PCA/DA and the docking approach for the compounds with known experimental activity data, it is possible to evidence that in the case of NNRTIs the percentage of correct classification in the PCA/DA approach is found in the range 65.0–80.0% (Fig. 4a), whereas for the docking approach this percentage reaches 60.0–65.0% (Fig. 4b). In

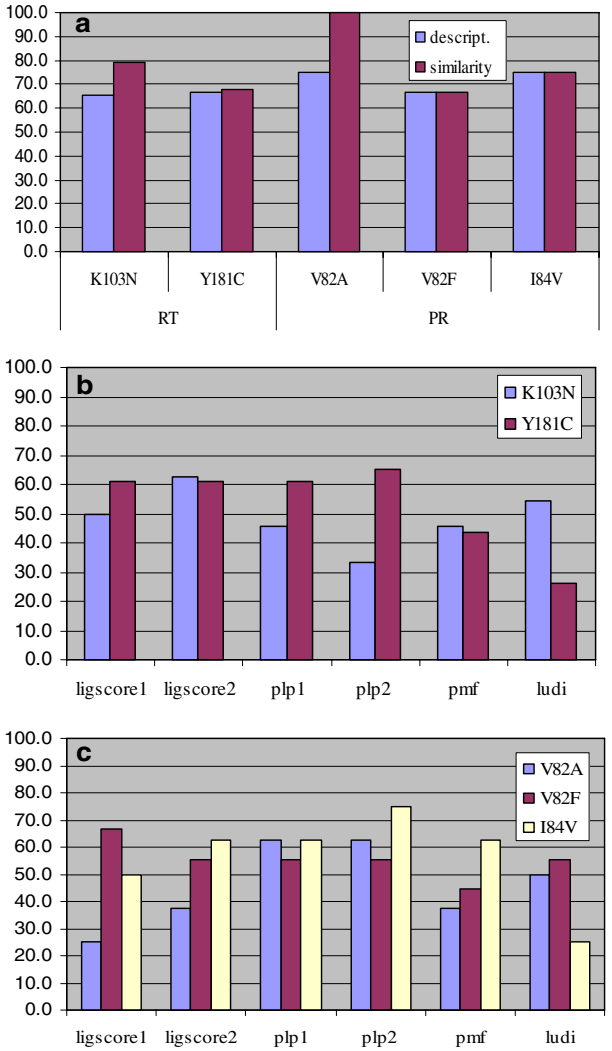

Fig. 4 Percentage of correct class assignments for inhibitors with literature data: (**a**) in the case of PCA/DA approach; (**b**) for NNRTIs in the case of PS docking approach; (**c**) for PIs in the case of DL docking approach

the case of drugs in clinical use, the results obtained by using the descriptors seems to give information with a higher structural contents (lower number of misclassified compounds) than when the similarity index was considered, despite the fact that this last furnished higher percentages of correct assignment. Analogous considerations are suggested by the docking results.

Also in the case of PIs, by using the PCA/DA method, the results obtained considering only compounds with known activity data showed a higher number of correct prediction if compared with the docking data (Fig. 4a vs. c). This number is generally higher when the similarity index was used.

A comparison of the two different approaches testifies that PCA/DA performed better with respect to docking. However, the latter allows to make assignment as S or R for all the compounds, while with our first approach, it was possible only to classify a significant number of compounds, but not the whole set.

Comparative analysis of the multivariate/docking approaches for all the inhibitors

The comparison of the results obtained by the multivariate analysis and the docking approach was then extended to all the compounds studied herein and predicted by us as belonging to S or R classes. The percentage of corresponding assignment, achieved by both methods, is depicted in Fig. 5, whereas the degree of correspondence between the different docking scores, examined in the study, is reported in the supporting information (Tables SI-T6 and SI-F3).

In the case of the RT inhibitors, it is possible to evidence that LigScore 1 and 2 gave the higher correspondence for both the mutant strain considered (Fig. 5a). Analogous results were also obtained with Ludi when the mutation K103N is taken into account. These results are not surprising considering which score is better related to the descriptors giving higher contribution when the PCA/DA studies were carried out: LigScore 1 and 2 are functions defined by energetic contribution of steric and electronic features and this is of particular interest because the descriptors which had a major importance in the PCA were ellipsoidal volume, surface area, variables which clearly reflect the importance of steric approach to the binding pocket. The electronic features which have the greater

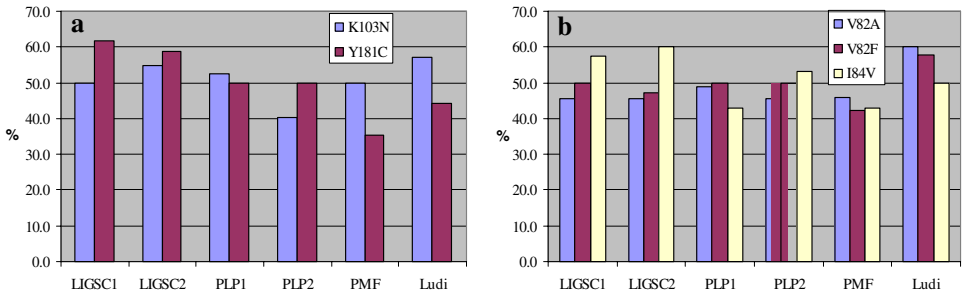

Fig. 5 Percentage of corresponding assignments between multivariate analysis (descriptors) and docking scores: (**a**) for PS site search mode (NNRTIs); (**b**) for DL site search mode (PIs)

weight in the PCs were E_{LUMO}, E_{HOMO}, total dipole moment, reflecting thus the importance of π-stacking during the drug-receptor interaction.

In the case of the PR, the results reported in Fig. 5b show that Ludi is the scoring function which provides the higher number of corresponding assignments for the mutant V82A and V82F and this is noteworthy because it is possible to gain similar information from the multivariate analysis. In fact, the Ludi function includes energetic contribution from ideal H-bond, perturbed ionic interactions, lipophilic interaction and rotational entropy, all physico-chemical properties which had the major weight in the PCs. This reflects the importance of the requirement for a high degree of flexibility to achieve suitable interactions, and underlines the importance of lipophilicity in the binding mode, due to the presence of hydrophobic pocket S/S′ which allows the correct accommodation of the inhibitor. PCs are well expressed in terms of total dipole μ and number of H-bond acceptors, together with log P: this shows as the electronic and lipophilic features, but mainly, the ability to form hydrogen bond represents a fundamental element for the inhibition process. In the case of mutant I84V, the higher percentage of corresponding assignments is provided by LigScore 1 and 2, in agreement with the results of multivariate analysis in which the included PCs are also well expressed in terms accessible surface area and hydrophobic interactions.

All these findings can be better appreciated after an analysis of the spatial arrangement of the ligands docked at the binding sites, that can evidence if the type of interactions, which the compounds classified respectively S or R establish with the binding pocket of the mutant enzymes, takes into account the modifications occurred in the active site of the protein. In Figs. 6 and 7 the superimpositions of the RT and PR inhibitor structures, classified consistently as S or R, docked at the suitable enzymes, are reported.

The mutation in position 103 of RT is not directly involved in the interaction with the ligands. The narrowing of the active site, due to the H-bonds between the residues 234–235–236, as previously evidenced, provides an easier access to small compounds. This result demonstrates that all the molecules, classified as R, present long side chain and bulky groups, generally substituted phenyl ring, or are characterized by large moieties of three or more rings (Fig. 6a). But, the presence of a large lipophilic moiety seems to be fundamental to interact with the hydrophobic pocket created by Tyr181, Tyr188, and Trp229. On the contrary, the inhibitors, classified as S against this mutant, have generally a reduced size with respect to the compounds classified as resistant: all of them are able to assume the butterfly conformation, which was already demonstrated to be the suitable accommodation of the NNRTIs for the inhibitory activity (Fig. 6b). In fact polycyclic structures

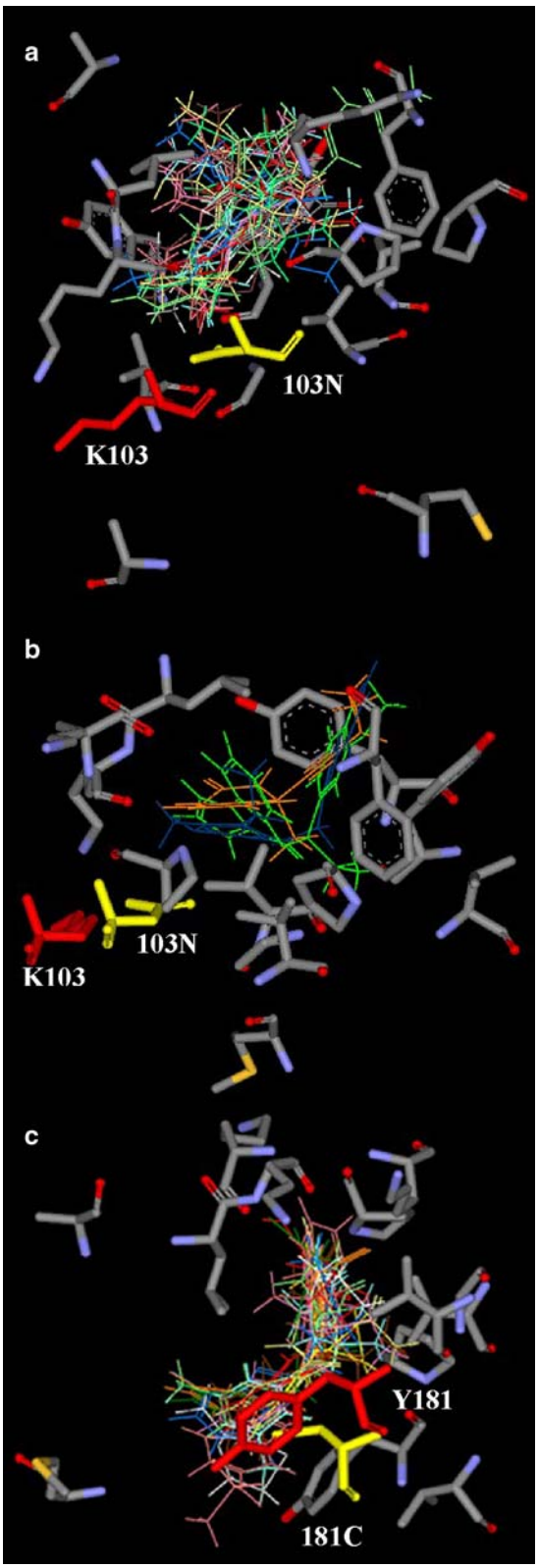

Fig. 6 NNRTIs docked at the binding site, WT (red amino acid) and mutant RT (yellow amino acid): (**a**) compound classified as R for the K103N mutation; (**b**) compounds classified as S for the K103N mutation; (**c**) compound classified as R for the Y181C mutation

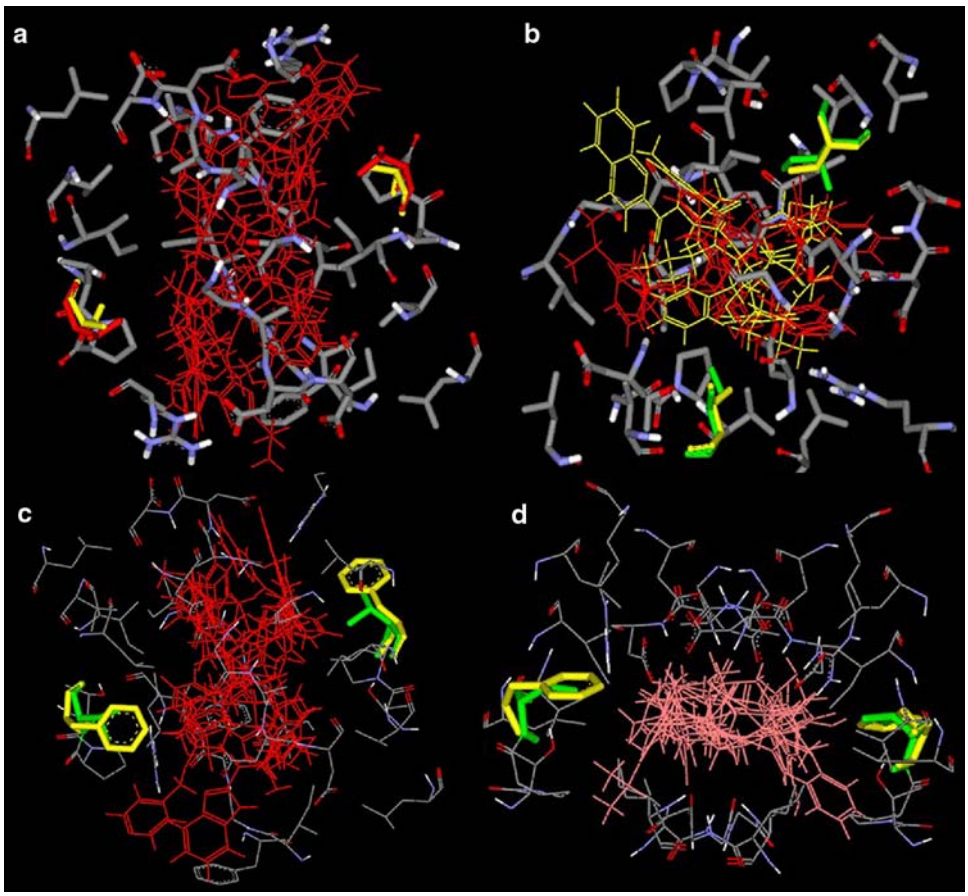

Fig. 7 PIs docked at the binding site, WT [green (red in **a**) amino acid], and mutant PR (yellow amino acid): (**a**) compound classified as R for the V82A mutation; (**b**) compound classified as R (red) and as S (yellow) for the I84V mutation; (**c**) compound classified as R for the V82F mutation; (**d**) compounds classified as S for the V82F mutation

such as calanolide or costatolide compounds, possessing a four rings system, are classified as resistant since they are not able to interact with the triad Tyr181, Tyr188, and Trp229, or to assume a flexible conformation. Delavirdine, although presenting all these requirements of flexibility and lipophilicity, is classified as R because the large number of bumps with residues that do not allow a perfect interaction with key portions of the active site. Other compounds, such as UC781, although allowed to assume a butterfly conformation, are too folded, so that they partially loose the hydrophobic interaction with amino acids 100, 106, 179, 188, 227, 235. For the mutation Y181C, the higher information content was obtained for the compound classified as resistant. As previously confirmed by the calculated size of the binding pocket, it is possible to notice a reduced cavity, but more importantly the lack of the tyrosine aromatic ring involves a reduced ability of the inhibitors to establish the π-stacking interactions (Fig. 6c). All the molecules (Nevirapine, Delavirdine and inhibitors with similar characteristics) bear an aromatic ring oriented towards this amino acid. Consequently the loss of this π-stacking produces a lengthening of the distance from the lipophilic portion Phe227, His 235, and Pro236. The presence of a spacer, between the aromatic ring which provides π-stacking and the lipophilic moiety which interact with the three residues, could guarantee favorable interaction, overcoming thus the drug-resistance induced by the mutation.

In the case of the PR, the mutation V82A leads to a decrease in the lipophilicity of the binding pocket, so that the molecules classified as resistant, independently by their peptidomimetic or non-peptide structure, bear large lipophilic moieties which reduce their binding capability (Fig. 4a). In fact, they present a lipophilic moiety which faces the mutated amino acid, but the reduced lipophilicity of the pocket influences the anchorage of the inhibitor. Since all these inhibitors represent state transition analogs of the protease substrate, the OH group, responsible of the interaction with the Asp25/25′ and therefore of the inhibitory activity, is found very far away from the reacting site. Inhibitors classified as S do not present such a lipophilic portion, so that the distances from Asp25/25′ remain unchanged. Also in the case of I84V (Fig. 4b) the decreased lipophilicity, following the mutation, determines a less effective anchorage in the pocket by large aromatic moieties. Moreover, the inhibitors classified as R make H-bond with Ile50, and this increases the distances of the pharmacofore OH group from the aspartate residues, whereas compounds classified as S do not show this H-bond with

Ile50 and do maintain the optimal distance aspartates-OH. The opposite consideration can be made in the case of the mutation V82F. The narrowing of the subsite S/S′, due to the presence of the phenyl ring of Phe, involves for bulky molecules an unsuitable positioning of the OH, while molecules coherently classified as susceptible (Fig. 4c, d) show a suitable volume and lipophilicity (number of aromatic rings sufficient enough to interact but not to generate steric bumps), so that the key reacting OH group maintains a correct distance from Asp25/25′.

Conclusions

Following our development of some multivariate statistical procedures (PCA and DA) to evaluate, on the basis of physico-chemical descriptors and structural similarity, the features of compounds to which mutant HIV strains are susceptible or are less likely to trigger resistance, we now extended the study including the use of docking procedure. We started this work searching in the PDB the crystallographic structures of the single mutant strains of RT and PR. Using all the available Ligandfit utilities, the binding pocket for both the wild type enzymes and the mutants was investigated. The docking scores of the complexes were calculated by using six scoring functions (LigScore1 and 2, PLP1 and 2, PMF, Ludi) which give different information on the basis of the different energetic contribution involved in the function. Then the algebraic difference between of the scores (Q) was calculated as the difference between the score obtained in the case of the mutant protein and the one calculated on the wild type protein, with the aim of classifying as S or R all the ligands.

In the case of RT mutant strain, the comparative analysis demonstrated that LigScore 1 and 2 gave the higher percentage of corresponding assignments with the multivariate ones. This is in agreement with the descriptors which were shown to have a major importance in the PCA (ellipsoidal volume, surface area, E_{LUMO}, E_{HOMO}, total dipole moment). In fact, these scoring functions are defined by energetic contribution of steric and electronic features. Moreover, in the case of the NNRTIs, it was possible to evidence that the compound classified as R have a larger size with respect to the molecules classified as S, and the π-stacking interaction is of fundamental importance in the modulation of the activity.

In the case of the PIs, it could be noticed that, depending on the mutant strain, the presence of bulky lipophilic moieties could influence the behaviour of the inhibitor as resistant or susceptible. Ludi was evidenced as the scoring function which provided better results for the mutant V82A and V82F, an interesting finding because analogous information was obtained from the multivariate analysis. Ludi function includes energetic contribution related to the physico-chemical properties which have the major weight in the PCs. This reflects the need of a high degree of flexibility to achieve suitable interactions, underlines the importance of lipophilicity in the binding mode, due to the presence of hydrophobic pocket S/S′, which allows the correct accommodation of the inhibitor. The presence of electronic and lipophilic features, but, mainly, the ability to form H-bonds represents a fundamental element for the inhibition process.

The challenge of comparative analysis between docking and multivariate methods to explore HIV-1 drug-resistance absolutely showed an interesting qualitative valiance considering the use of scoring functions the value of which is derived by the sum of defined molecular descriptors.

The chemometric and docking techniques used in these papers to study the HIV-1 inhibitors revealed to be of particular interest and could provide guidelines for the synthesis of new compounds with suitable structural features to maximize the drug-receptor interaction and consequently the biological activity, also in the case of mutant enzymes. Moreover, the possibility to discern how multivariate analysis and docking complement each other in discriminating the same set of inhibitors might give clues to the study of drug resistance in other biochemical systems.

Supplementary material

The online version of this article contains supplementary material, which is available to authorized users. The structures of NNRTIs and PIs included in the study are reported as Figures SI-F1 and SI-F2. The Tables SI-T1 to SI-T5 list complete data sets, and classification of inhibitors.

References

1. Hammer S, Vaida F, Bennett K, Holohan M, Sheiner L, Eron J, Wheat L, Mitsuyasu R, Gulick R, Valentine F, Aberg J, Rogers M, Karol C, Saah A, Lewis R, Bessen L, Brosgart C, DeGruttola V, Mellors J (2002) J Am Med Assoc 288:169
2. Gallant J (2000) J Am Med Assoc 283:1329
3. Carpenter C, Cooper D, Fischl M, Gatell J, Gazzard B, Hammer S, Hirsch M, Jacobsen D, Katzenstein D, Montaner J, Richman D, Saag M, Schechter M, Schooley R, Thompson M, Vella S, Yeni P, Volberding P (2000) J Am Med Assoc 283:381
4. Hirsch M, Richman D (2000) J Am Med Assoc 284:1649
5. Rusconi S, La Seta Catamancio S, Sheridan F, Parker D (2000) J Clin Virol 19:135
6. Holodniy M, Katzenstein D, Winters M, Montoya J, Shafer R, Kozal M, Ragni M, Merigan T (1993) J Acquir Immune Defic Syndr 6:366
7. Schuurman R, Demeter L, Reichelderfer P, Tijnagel J, de Groot T, Boucher C (1999) J Clin Microbiol 37:2291

8. Race E, Gilbert S, Sheldon J, Rose J, Moffatt A, Sitbon G, Dissanayeke S, Cammack N, Duncan I (1998) AIDS 12:1465
9. Shafer R (2002) Clin Microbiol Rev 15:247
10. Schinazi R, Larder B, Mellors J (2000) Int Antiv News 8:65
11. Baxter J, Mayers D, Wentworth D, Neaton J, Hoover M, Winters M, Mannheimer S, Thompson M, Abrams D, Brizz B (2000) AIDS 14:F83
12. Holloway M, Wai J, Halgren T, Fitzgerald P, Vacca J, Dorsey B, Levin R, Thompson W, Chen L, deSolms S (1995) J Med Chem 38:305
13. Nair A, Jayatilleke P, Wang X, Miertus S, Welsh W (2002) J Med Chem 45:973
14. Perez C, Pastor M, Ortiz A, Gago F (1998) J Med Chem 41:836
15. Shenderovich M, Kagan R, Heseltine P, Ramnarayan K (2003) Protein Sci 12:1706
16. Jenwitheesuk E, Samudrala R (2005) Antivir Ther 10:157
17. Almerico AM, Lauria A, Tutone M, Diana P, Barraja P, Montalbano A, Cirrincione G, Dattolo G (2003) QSAR Comb Sci 22:984
18. Almerico AM, Tutone M, Lauria A, Diana P, Barraja P, Montalbano A, Cirrincione G, Dattolo G (2006) J Chem Inf Model 46:168
19. http://www.niaid.nih.gov/daids/dtpdb
20. http://www.rcsb.org/pdb
21. All of the calculations were run on a Silicon Graphics Indigo II workstation using the software TSAR 3.2 (Tools for Structure Activity Relationships), VAMP 6.0, and ASP 3.2 (Automated Similarity Packages) (Oxford Molecular-Accelrys). Molecular descriptors were derived according to the method and assumptions reported in the TSAR 3.2 Reference Guide, Oxford Molecular Limited (1998)
22. Ligandfit User Manual, Accelrys Inc (2003)
23. Venkatachalam C, Jiang X, Oldfield T, Waldman M (2003) J Mol Graph Model 21:289
24. Böhm H (1998) J Comput Aided Mol Des 12:309
25. Verkhivker G, Bouzida D, Gehlhaar D, Rejto P, Arthurs S, Colson A, Freer S, Larson V, Luty B, Marrone T, Rose P (2000) J Comput Aided Mol Des 14:731
26. Gehlhaar D, Verkhivker G, Rejto P, Sherman C, Fogel D, Fogel L, Freer S (1995) Chem Biol 2:317
27. Muegge I, Martin Y (1999) J Med Chem 42:791