APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening

Violeta I. Pérez-Nueno, Obdulia Rabal, José I. Borrell, and Jordi Teixidó*

Grup d'Enginyeria Molecular, Institut Químic de Sarriá (IQS), Universitat Ramon Llull, Via Augusta 390, 08017 Barcelona, Spain

Received February 4, 2009

A new interaction fingerprint (IF) called APIF (atom-pairs-based interaction fingerprint) has been developed for postprocessing protein−ligand docking results. Unlike other existing fingerprints which employ absolute locations of individual interactions, APIF considers the relative positions of pairs of interacting atoms. Docking-based virtual screening was performed with GOLD using the crystal structures of trypsin, rhinovirus, HIV protease, carboxypeptidase, and estrogen receptor-α as targets. A score derived from the similarity of the bit strings for each docking solution to that of a known reference binding mode was obtained. Comparisons between APIF, GoldScore function, and standard interaction fingerprint (CHIF) scores were performed using enrichment plots. Superior recovery rates were observed in the IF score cases. Comparable results were achieved by using either of the two interaction fingerprints, substantially improving GoldScore function enrichment factors. Binding mode analyses were also carried out in order to study the best method for selecting conformations with a binding mode similar to that of the reference crystallized complex. These showed that the first conformations retrieved by interaction fingerprint scores had a more similar binding mode to the reference complex than those retrieved by the GoldScore function.

INTRODUCTION

Interaction fingerprints (IFs) have been developed to enhance the representation and analysis of three-dimensional protein−ligand interactions.[1−3] In particular, they have proven to be very useful in docking output postprocessing as a virtual screening (VS) filter and for binding mode detection.[4] These methods have been developed in order to overcome the known deficiencies in identifying accurately the conformations with closest binding modes to the X-ray structures.[5,6]

IFs encode the 3D protein−ligand contacts in bit strings of a length derived from the number of residues/atoms in the target protein binding cavity. Typically, each bit denotes either the presence (1) or absence (0) of a particular interaction such as a hydrogen bond or hydrophobic or van der Waals contact.

The different interaction fingerprint implementations vary depending on the bit string definition and the type of interactions considered. The initial proposal of Deng and co-workers[1,4] operated at the residue level and considered hydrophobic and hydrogen-bond contacts. Following this idea, Kelly and Mancera[2] transferred the initial concept based on residues to a new one based on atoms for hydrogen-bond sites. Moreover, these authors introduced the concept of weighting the importance of the detected interactions. Recent atom-based IF advances were developed by Mpamhanga and co-workers.[3] These authors encoded hydrogen bonds and hydrophobic contacts in a fingerprint of length equal to the number of heavy atoms in the binding site.

In this paper, we present a new fingerprint called APIF (atom-pairs-based interaction fingerprint) that encodes ligand−protein binding modes in a bit string based on the concept of atom pairs. This approach is widely used in the context of fragment-based similarity searches.[7] APIF encodes ranges of distances between two receptor−ligand interaction points. Each observed distance increases a count in an associated seven-range bin. Depending on the combination of the type of contacts, the corresponding bit is set on.

The three IF approaches previously reported encode ligand−protein interaction information in an absolute manner; that is, a contact is expected or not at a concrete atom or residue of the protein sequence. On the other hand, APIF considers the relative pairwise position of the interacting atoms rather than their absolute locations. Thus, from our viewpoint, the main novelty of this approach is that our IF encodes the conserved distance between two receptor−ligand interactions rather than requiring a specific atom or residue of the protein.

The performance of this new fingerprint was validated through docking-based VS using both enrichment plots and binding mode analyses. Enrichment results obtained with APIF were compared with those retrieved with the GoldScore function and an in-house implementation of the CHIF fingerprint of Mpamhanga et al. Inspection of the binding modes for the poses selected by these three criteria was carried out in order to analyze their ability to retrieve the closest binding modes to the crystallographic structures within the first top-ranked conformations.

METHODS

Case Studies: Protein and Databases Preparation. To evaluate our approach, we decided to dock several different

* Corresponding author phone:+34-93-267.20.00; fax: +34-93-205.62.66; e-mail: j.teixido@iqs.url.es.

10.1021/ci900043r CCC: $40.75
Published on Web 04/13/2009

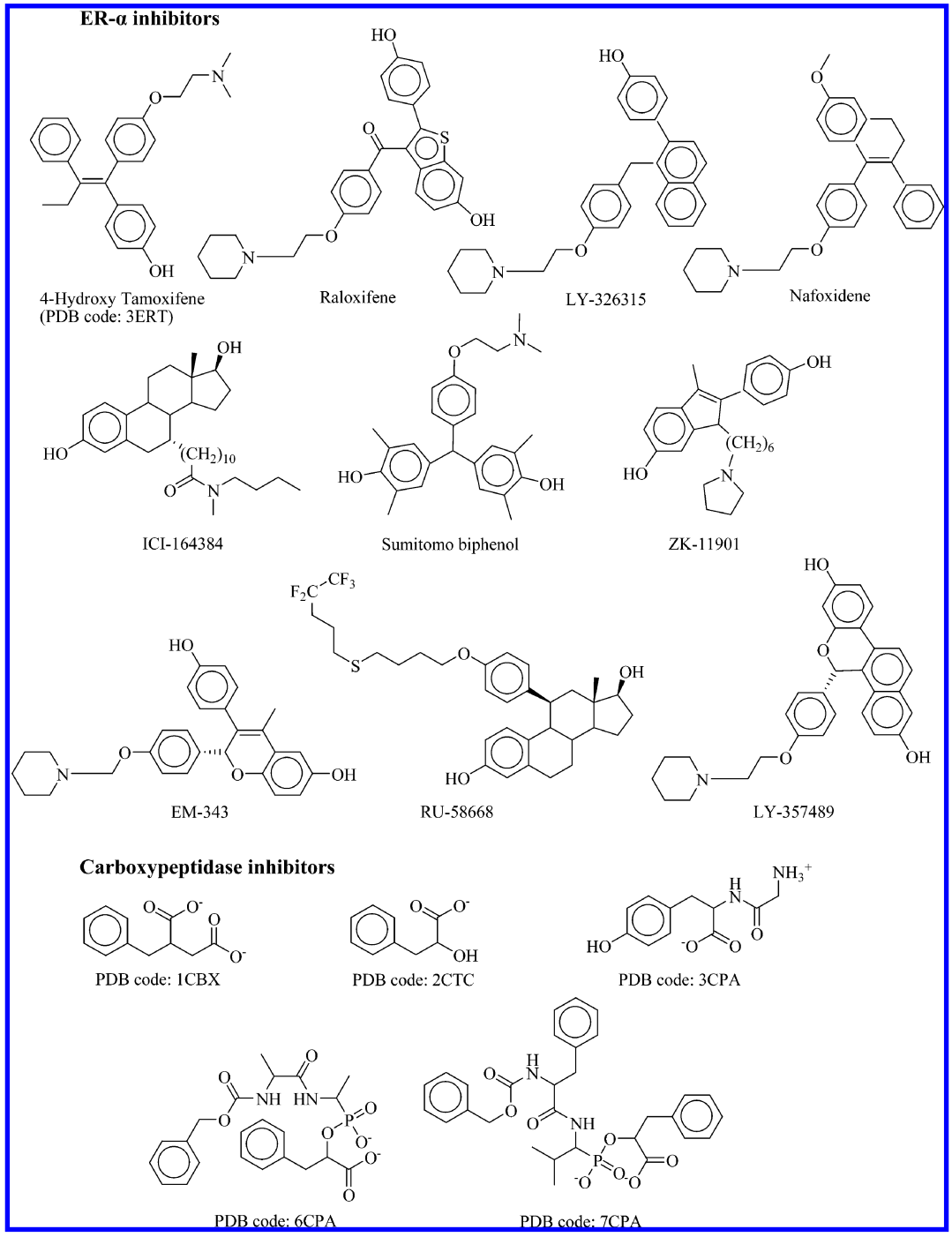

Figure 1. Known ER-α and carboxypeptidase active inhibitors in the virtual screening data sets.

Table 1. Reference Complexes of Trypsin, Rhinovirus, HIV Protease, Carboxypeptidase and ER-α Targets Used in Docking and IFs Virtual Screening

target	complexed ligand	PDB code	resolution/Å
trypsin	benzamidine inhibitor	3PTB	1.7
rhinovirus	5-(7-(4-(4,5-dihydro-2-oxazolyl)phenoxy)heptyl)-3-methyl isoxazole	2R04	3
HIV protease	*N*,*N*-bis(2-hydroxy-1-indanyl)-2,6-diphenylmethyl-4-hydroxy-1,7-heptandiamide	4PHV	2.1
carboxypeptidase	α-hydroxy-β-phenyl-propionic acid	2CTC	1.4
ER-α	4-hydroxy tamoxifene	3ERT	1.9

experimentally determined X-ray cocrystal structures and known inhibitors. Thus, the set of known inhibitors was collected from the FlexS-77 data set[8] for the trypsin, rhinovirus, HIV protease, and carboxypeptidase targets. The "Bissantz active set"[9] was used to compile the inhibitors for the estrogen receptor-α (ER-α) target for comparison purposes with Mpamhanga et al.'s work.[3] The structures of these compounds are shown in Figures 1−3. For each target of these sets, the complex with the best crystallographic resolution was selected as the reference for docking. These

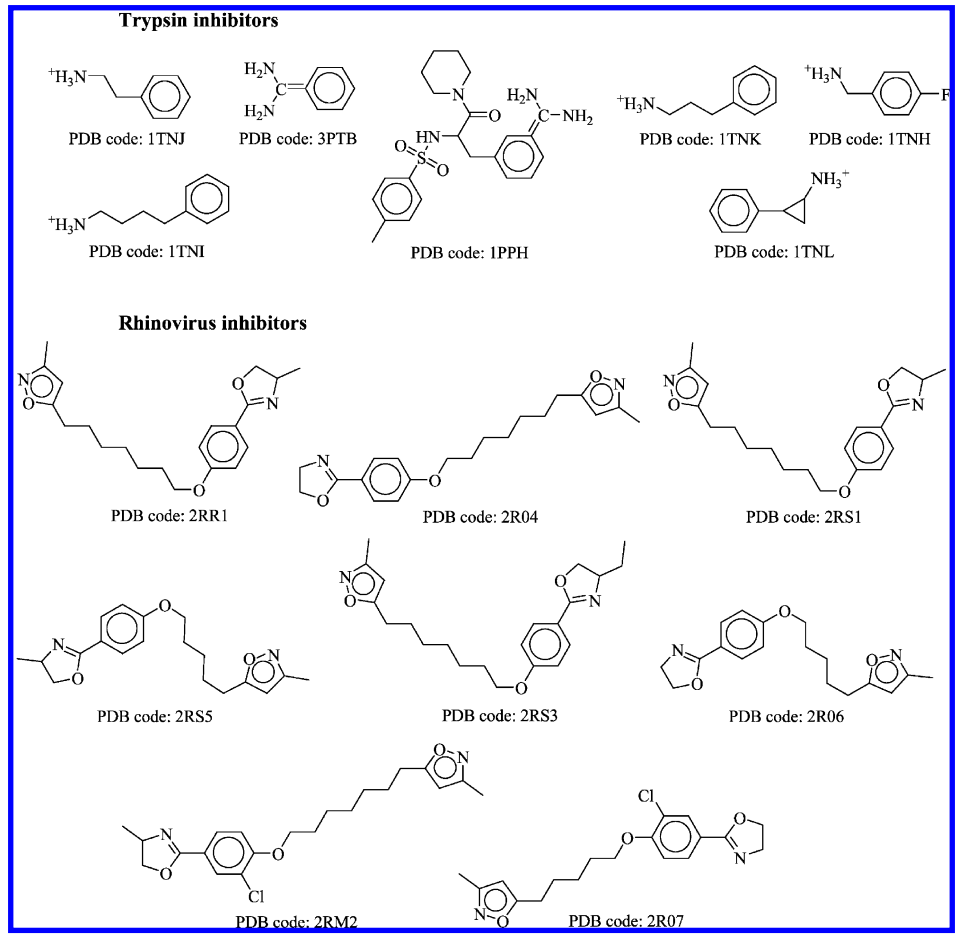

Figure 2. Known trypsin and rhinovirus active inhibitors in the virtual screening data sets.

reference complexes and their corresponding PDB entries are listed in Table 1. For each target complex, the ligand binding site was defined from the bound ligand using a radius cutoff of 10 Å. Bound waters were removed from the binding sites, and the receptors were protonated at pH 7.

In order to perform the virtual screening, the known actives were combined with presumed inactive compounds from the Maybridge Screening Collection database[9] in such a way that several 1D properties calculated by MOE[11] were similar to those of the active compounds (molecular weight, number of rotatable single bonds, number of hydrogen-bond acceptor atoms, number of hydrogen-bond donor atoms, octanol−water partition coefficient, and number of hydrophobic atoms).[12] Table 2 shows the average and standard deviations of these properties for the data sets used. It can be seen that they are quite similar for the active and inactive pools.

We would like to remark that, for APIF and CHIF (in-house implementation) comparison purposes, the ER-α inactive pool of 490 compounds differs from that of Mpamhanga et al.[3] However, it is worth mentioning that Mpamhanga et al. repeated their calculation for three different inactive pools without finding significant differences, so it is not expected that our modified inactive pool will have much influence in reproducing results.

Docking Methodology. All of the data set compounds were docked into the aforementioned protein structures using GOLD.[13] In the GOLD runs, the ligand binding site was limited to all protein atoms within 10 Å from the centroid of the binding residues.[14−26] The GOLD cavity-detection algorithm was enabled in order to confine the calculation to concave regions in the vicinity of the binding site. A total of 100 docking runs per experiment (conformations) were performed, with each run consisting of a maximum of 100 000 genetic algorithm (GA) operations. All other GA parameters used default values. Cutoff distances of 2.5 Å for hydrogen bonds and 4.0 Å for nonbonded contacts were set. In each study, all of the ligand poses generated were retained for subsequent binding mode analyses. The GoldScore function was used for scoring the docked conformations as the first criterion for VS ranking and for the subsequent respective enrichment plots.

Construction of APIF. The algorithm to generate the APIF was implemented in the MOE SVL language.[11] First, given a complex, the active site is defined using a radius value (10 Å in the present study). Second, the interactions between the protein and the ligand are detected using the function pro_Contacts, as implemented in MOE: hydrogen bonds are defined following the Stickle446 function, and hydrophobic contacts are determined using a cutoff of 4.5 Å. Depending on the type of interaction, both the atoms of the protein (P) and the ligand (L) are labeled as hydrogen-bond donor, hydrogen-bond acceptor, or hydrophobic. This results in six possible types: acceptor-L, acceptor-P, donor-L, donor-P, hydrophobic-L, and hydrophobic-P. Third, all possible pairwise protein−ligand interactions are detected and classified depending on one of the six possible combinations of pairs of interaction contacts. The six possible types are listed in Figure 4a. For each pairwise interaction detected in a complex, the distance between the two receptor atoms (d_1) and the distance between the two ligand atoms (d_2) are measured. Figure 4b shows this process. Each

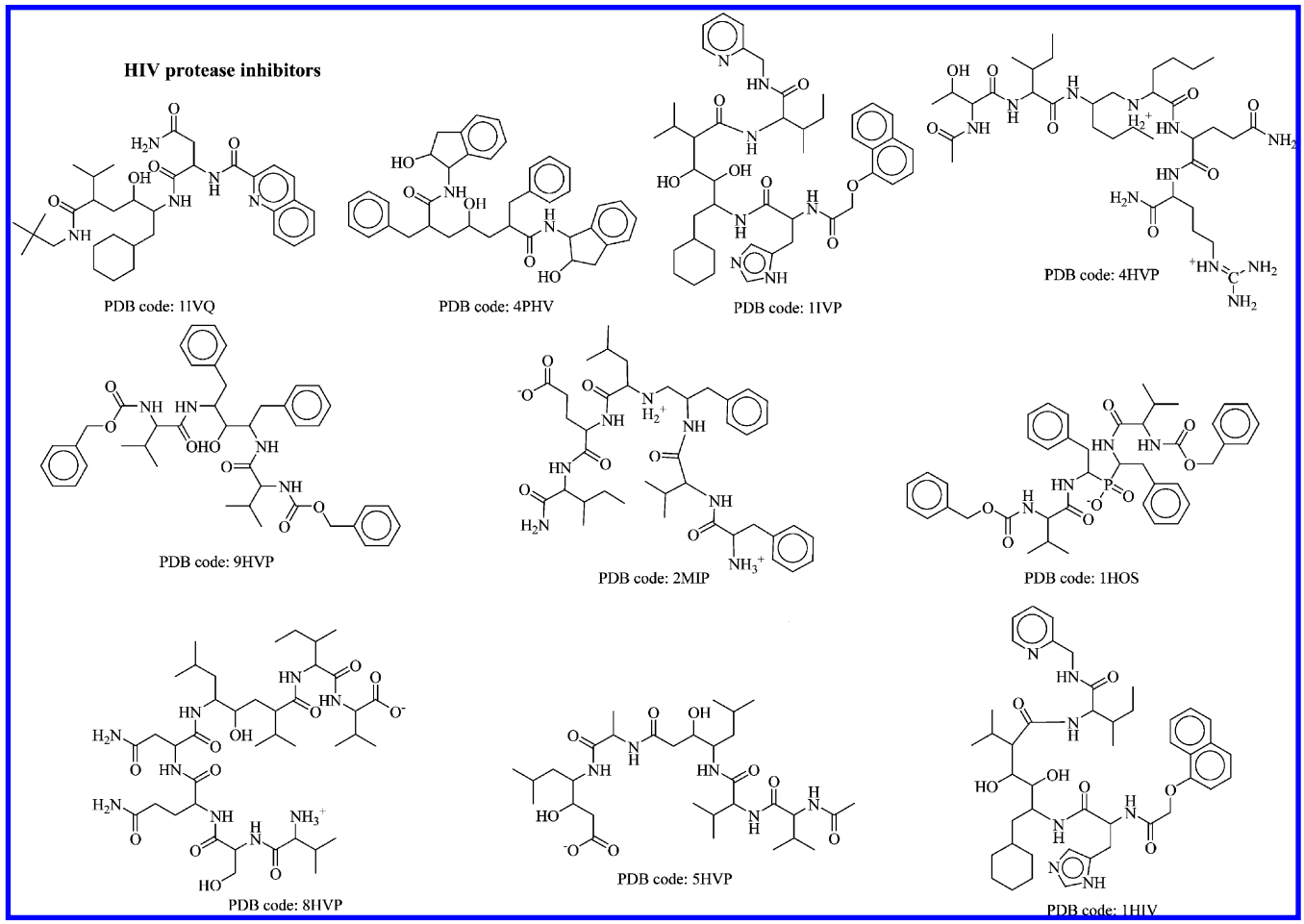

Figure 3. Known HIV protease active inhibitors in the virtual screening data set.

observed distance increases a counter within a bin divided into seven ranges, taken from Mason,[27] which correspond to distance ranges of (Å) 0−2.5, 2.5−4, 4−6, 6−9, 9−13, 13−18, and >18. The two distances taken together define a single bit in a string of 49 bits (enumerated from zero) according to eq 1:

$$\text{bit position} = \text{bin}(d_2) + 7 \cdot \text{bin}(d_1) \quad (1)$$

Fourth, the final fingerprint length corresponds to the number of possible combinations of pairs of interaction contacts (six) and the dimension of the binning partition scheme for the ligand distances (seven) and the receptor distances (seven). In this way, the total fingerprint is composed of $6 \times 7 \times 7 = 294$ bits. Figure 4 illustrates the APIF, and Figure 5 illustrates this encoding system. Both raw fingerprints and normalized ones between 0 and 1 were constructed.

In-House CHIF Fingerprint Implementation. Here, the CHIF fingerprint was also implemented in SVL. We followed the CHIF design description from Mpamhanga et al.,[3] although some differences arise in the function used to determine protein contacts. In our case, and as for APIF, the MOE pro_Contacts function was used. This function uses a different hydrogen-bond distance threshold and different atom-type definitions (donor, acceptor, and hydrophobic) from those of Mpamhanga et al.[3] We also fixed a radius of 10 Å to define the binding site.

Virtual Screening Protocol. After docking the inhibitors against their corresponding target, CHIF and APIF fingerprints were calculated for all of the retrieved poses. Similarly, reference CHIF and APIF fingerprints were directly generated from the crystallographic reference complexes. Then, a similarity search was performed between the fingerprints derived from each docked conformation and the reference fingerprints. Two scoring systems were used to evaluate the conformations for each ligand and to rank the screened list:

- Traditional similarity coefficients:[28] Euclidean distance, Manhattan distance, Tanimoto coefficient, and simple matching coefficient. These similarity values can be calculated using eqs 2−5 (below), where A and B denote the numbers of bit sets in the two IFs that are being compared and C denotes the number of bits in common. This scoring system will subsequently be called SCORE1, specifying in each case the particular coefficient used (Euclidean, Manhattan, Tanimoto, or simple matching).

$$\text{Euclidean distance} = \sqrt{A + B - 2C} \quad (2)$$

$$\text{Manhattan distance} = A + B - 2C \quad (3)$$

$$\text{Tanimoto coeffcient} = \frac{C}{A + B - C} \quad (4)$$

$$\text{simple matching coefficient} = C \quad (5)$$

- Following Mpamhanga's work,[3] a second kind of score was calculated, resulting from the multiplication of the value obtained from the GoldScore function for each solution by the similarity coefficient (in this case, we only considered Tanimoto and simple matching). This scoring system will

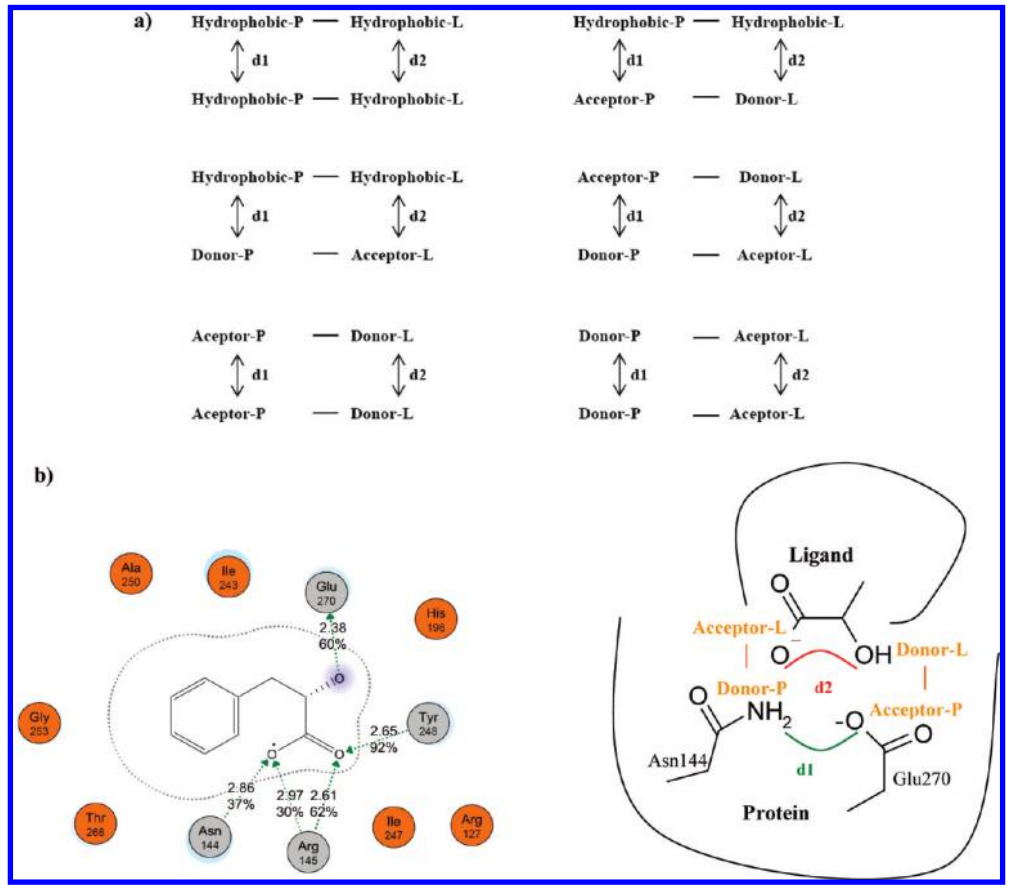

Figure 4. Atom-pairs-based interaction fingerprint (APIF). (a) Six possible combinations of pairs of interactions defining a set of 49 bits (seven bits for a total of seven distances). (b) Codification of the pairwise interactions from the distances measured between the two protein (d_1) and the two ligand (d_2) interacting atoms. This specific case shows the carboxypeptidase in complex with an α-hydroxy-β-phenyl-propionic acid. For a pair of interactions detected, for example, a hydrogen bond between ND2 of Asn144 and a negatively charged oxygen of the α-hydroxy-β-phenyl-propionic acid and a hydrogen bond between OE2 of Glu270 and an oxygen of the α-hydroxy-β-phenyl-propionic acid, the interacting distances between the two protein atoms (d_1) and the two ligand atoms (d_2) are measured.

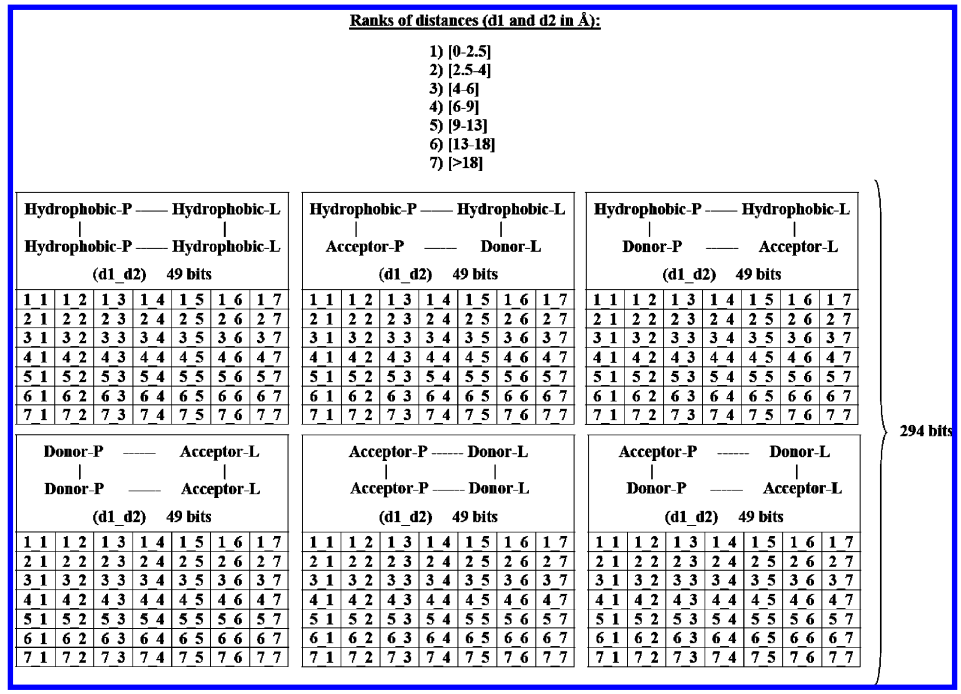

Figure 5. Atom-pairs-based interaction fingerprint encoding system. The six possible combinations of pairs of interaction contacts are each represented with 49 bits. For each pairwise interaction detected in a complex, the distance between the two receptor atoms (d_1) and the two ligand atoms (d_2) is measured. The results are clustered into seven distance ranges (Å): 0−2.5, 2.5−4, 4−6, 6−9, 9−13, 13−18, and >18. Thus, the total fingerprint length is always $6 \times 7 \times 7 = 294$ bits.

Table 2. Summary of the 1D Physico−Chemical Properties of Active and Inactive Molecules in the Trypsin, Rhinovirus, HIV Protease, Carboxypeptidase, and ER-α Screening Databases[a]

trypsin	weight	b_1rotN	a_acc	a_don
7 actives	174.8 (113.6)	2.9 (2.5)	0.4 (1.1)	0.1 (0.4)
467 inactives	238.7 (53.4)	1.9 (1.2)	0.2 (0.7)	0.1 (0.4)
rhinovirus	weight	b_1rotN	a_acc	a_don
8 actives	351.1 (23.8)	9.4 (1.2)	3.0 (0.0)	0.0 (0.0)
498 inactives	356.2 (20.4)	5.4 (1.0)	3.1 (0.6)	0.2 (0.5)
HIV protease	weight	b_1rotN	a_acc	a_don
10 actives	740.2 (88.9)	20.7 (5.8)	7.0 (2.4)	6.5 (2.0)
489 inactives	609.2 (55.2)	8.4 (3.9)	4.9 (2.3)	1.4 (1.5)
carboxypeptidase	weight	b_1rotN	a_acc	a_don
5 actives	333.3 (183.7)	7.6 (4.7)	1.8 (1.3)	1.4 (0.9)
272 inactives	281.3 (43.0)	5.5 (0.8)	2.6 (0.5)	1.5 (0.7)
ER-α	weight	b_1rotN	a_acc	a_don
10 actives	458.8 (67.1)	11.3 (4.3)	3.6 (0.8)	1.6 (0.7)
490 inactives	465.1 (16.2)	8.5 (3.1)	4.2 (1.8)	1.0 (0.9)

[a] This table shows the average and standard deviations (in parentheses) of the following properties: weight (molecular weight), b_1rotN (number of rotatable single bonds), a_acc (number of hydrogen-bond acceptor atoms), and a_don (number of hydrogen-bond donor atoms).

Table 3. ER-α Enrichment Factor Values for the First 2%, 5%, and 10% of the Screened Database

estrogen receptor-α	2%	5%	10%
GOLDSCORE (docking)	15	10	7
CHIF-SCORE1-TANIMOTO	15	10	6
CHIF-SCORE1-SIMPLE_MATCHING	25	12	**8**
CHIF-SCORE1-EUCLIDEAN	15	10	6
CHIF-SCORE1-MANHATTAN	15	10	6
CHIF-SCORE2-TANIMOTO	25	12	7
CHIF-SCORE2-SIMPLE_MATCHING	**30**	**14**	**8**
APIF-SCORE1-TANIMOTO	10	4	4
APIF-SCORE1-SIMPLE_MATCHING	0	6	4
APIF-SCORE1-EUCLIDEAN	10	4	2
APIF-SCORE1-MANHATTAN	0	4	3
APIF-SCORE2-TANIMOTO	20	10	**8**
APIF-SCORE2-SIMPLE_MATCHING	20	10	7
NORMALIZED_APIF-SCORE1-TANIMOTO	0	2	2
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	10	6	5
NORMALIZED_APIF-SCORE1-EUCLIDEAN	0	4	2
NORMALIZED_APIF-SCORE1-MANHATTAN	0	0	2
NORMALIZED_APIF-SCORE2-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	20	**14**	**8**

subsequently be called SCORE2, specifying in each case the particular coefficient used (Tanimoto or simple matching).

$$\text{SCORE2} = \text{SCORE1} \times \text{GoldScore function} \quad (6)$$

Finally, the VS was analyzed in terms of enrichment plots using the three criteria: GoldScore function, SCORE1, and SCORE2.

Binding Mode Analyses. IFs provide a good method for analyzing the protein−ligand interactions and optimizing the resulting docking poses. Several studies have been made for analyzing the binding modes obtained from IFs' selected conformations.[1−4,29,30] As many docking validation studies have shown, scoring functions (such as GoldScore) do not always identify within the first-ranked conformations the cocrystallized binding modes. This also happens with the poses selected using IF-based similarity scores. However, since IFs take into account experimental data, it is reasonable

Table 4. Trypsin Enrichment Factor Values for the First 2%, 5%, and 10% of the Screened Database

trypsin	2%	5%	10%
GOLDSCORE (docking)	7	3	1
CHIF-SCORE1-TANIMOTO	0	0	0
CHIF-SCORE1-SIMPLE_MATCHING	0	6	3
CHIF-SCORE1-EUCLIDEAN	0	0	0
CHIF-SCORE1-MANHATTAN	0	0	0
CHIF-SCORE2-TANIMOTO	7	6	4
CHIF-SCORE2-SIMPLE_MATCHING	14	6	**7**
APIF-SCORE1-TANIMOTO	14	**14**	**7**
APIF-SCORE1-SIMPLE_MATCHING	14	9	**7**
APIF-SCORE1-EUCLIDEAN	14	6	3
APIF-SCORE1-MANHATTAN	14	6	4
APIF-SCORE2-TANIMOTO	**29**	**14**	**7**
APIF-SCORE2-SIMPLE_MATCHING	21	**14**	**7**
NORMALIZED_APIF-SCORE1-TANIMOTO	14	11	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	14	11	6
NORMALIZED_APIF-SCORE1-EUCLIDEAN	14	6	3
NORMALIZED_APIF-SCORE1-MANHATTAN	7	3	1
NORMALIZED_APIF-SCORE2-TANIMOTO	**29**	**14**	**7**
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	21	11	**7**

Table 5. Rhinovirus Enrichment Factor Values for the First 2%, 5%, and 10% of the Screened Database

rhinovirus	2%	5%	10%
GOLDSCORE (docking)	6	8	5
CHIF-SCORE1-TANIMOTO	13	5	4
CHIF-SCORE1-SIMPLE_MATCHING	19	8	4
CHIF-SCORE1-EUCLIDEAN	13	5	5
CHIF-SCORE1-MANHATTAN	13	5	4
CHIF-SCORE2-TANIMOTO	13	8	4
CHIF-SCORE2-SIMPLE_MATCHING	13	5	4
APIF-SCORE1-TANIMOTO	6	8	5
APIF-SCORE1-SIMPLE_MATCHING	0	5	3
APIF-SCORE1-EUCLIDEAN	13	8	5
APIF-SCORE1-MANHATTAN	6	5	4
APIF-SCORE2-TANIMOTO	19	10	6
APIF-SCORE2-SIMPLE_MATCHING	6	8	4
NORMALIZED_APIF-SCORE1-TANIMOTO	13	8	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	6	5	5
NORMALIZED_APIF-SCORE1-EUCLIDEAN	6	10	6
NORMALIZED_APIF-SCORE1-MANHATTAN	0	3	4
NORMALIZED_APIF-SCORE2-TANIMOTO	**31**	**13**	8
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	13	10	**9**

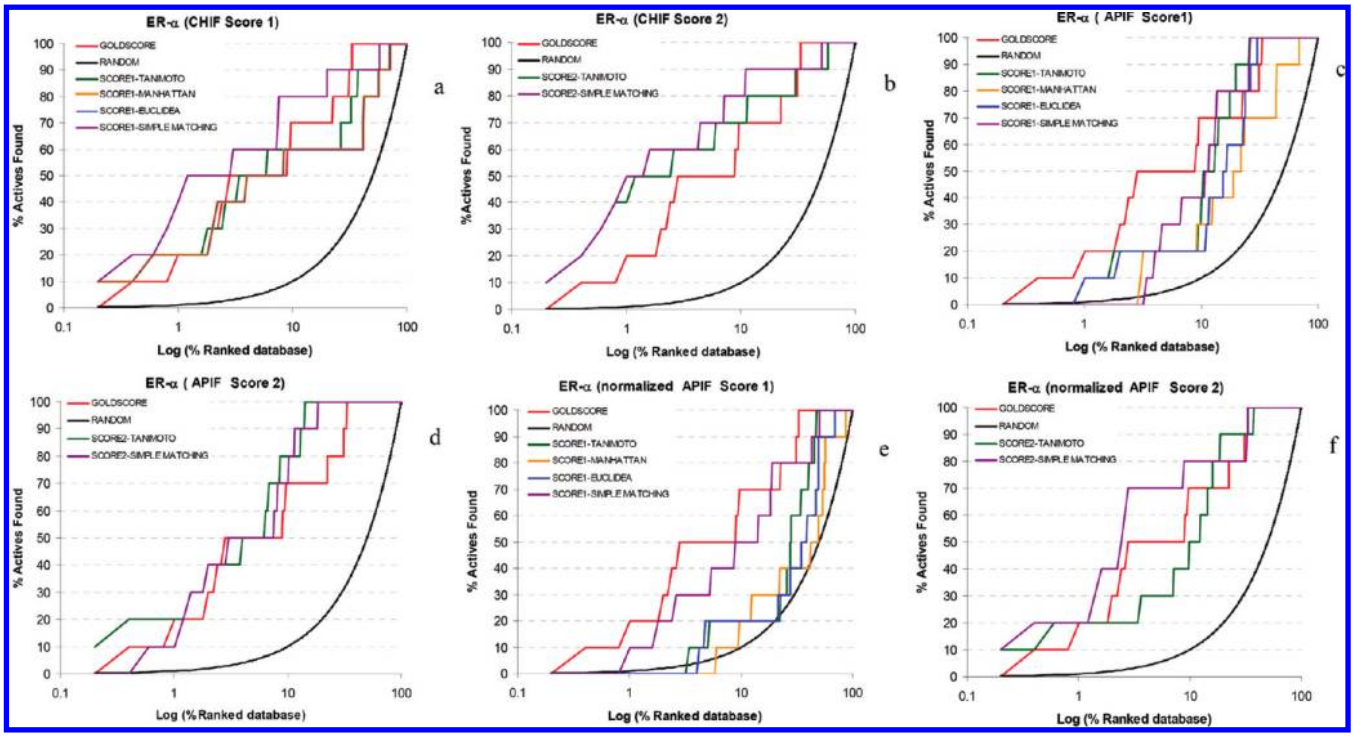

Figure 6. ER-α enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) normalized APIF and SCORE1, and (f) normalized APIF and SCORE2. The different similarity scores used correspond to simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green), and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The *x* axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the *y* axis.

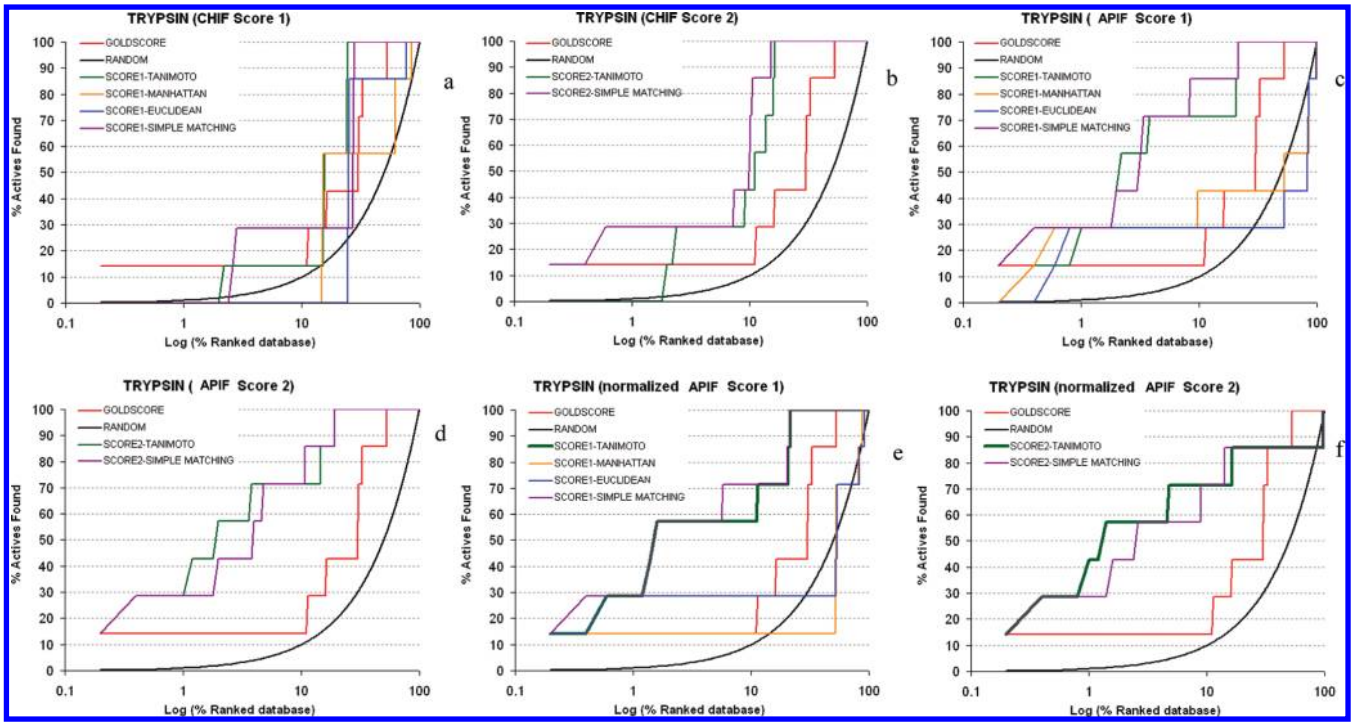

Figure 7. Trypsin enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) normalized APIF and SCORE1, (f) and normalized APIF and SCORE2. The different similarity scores used correspond to simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green), and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The *x* axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the *y* axis.

to suppose that they can select closer poses to the experimental crystallographic complex than using only docking scoring functions. In this work, we compared the abilities of the GoldScore function and APIF-based and CHIF-based similarity fingerprints to retrieve the closest binding modes to the crystallographic structures within the first top-ranked conformations. Binding modes of four out of the five targets used in the enrichment studies (trypsin, rhinovirus, HIV protease, and carboxypeptidase) were analyzed. For each target, the root-mean-square deviation (rmsd) between the

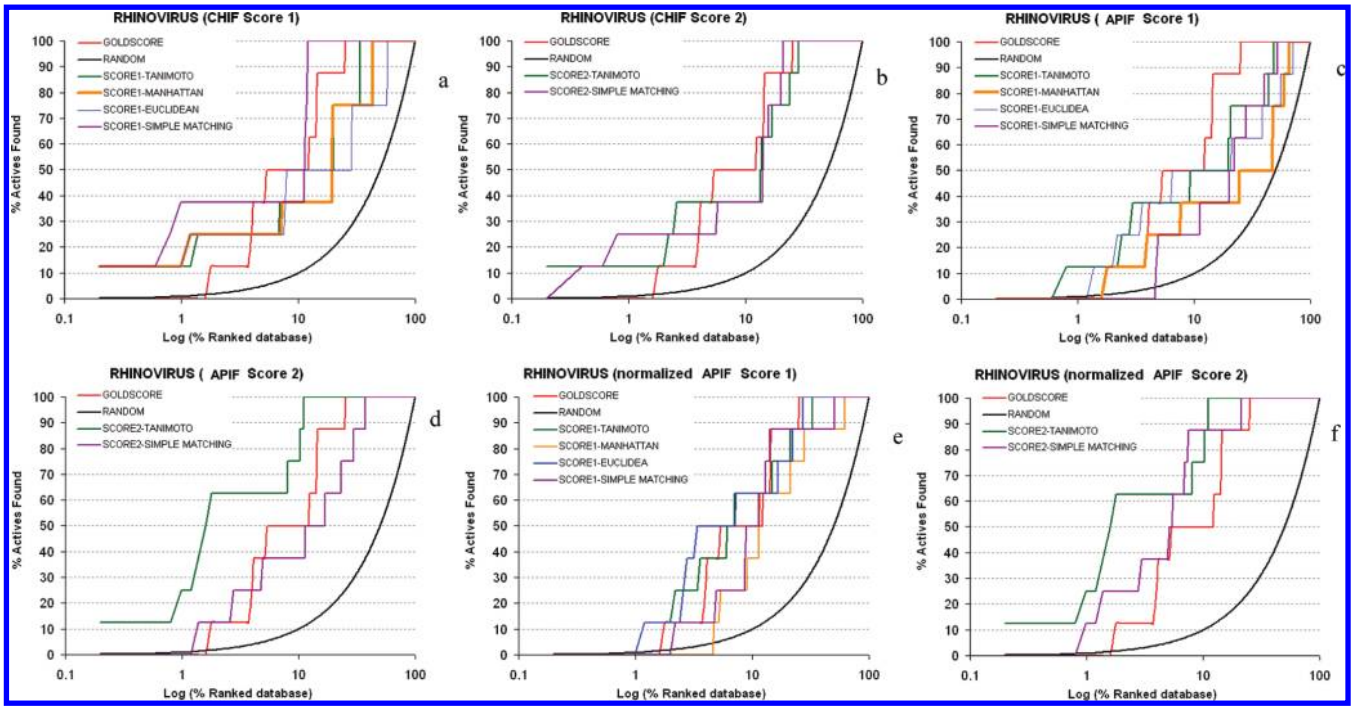

Figure 8. Rhinovirus enrichment plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) normalized APIF and SCORE1, and (f) normalized APIF and SCORE2. The different similarity scores used correspond to simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green), and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The *x* axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the *y* axis.

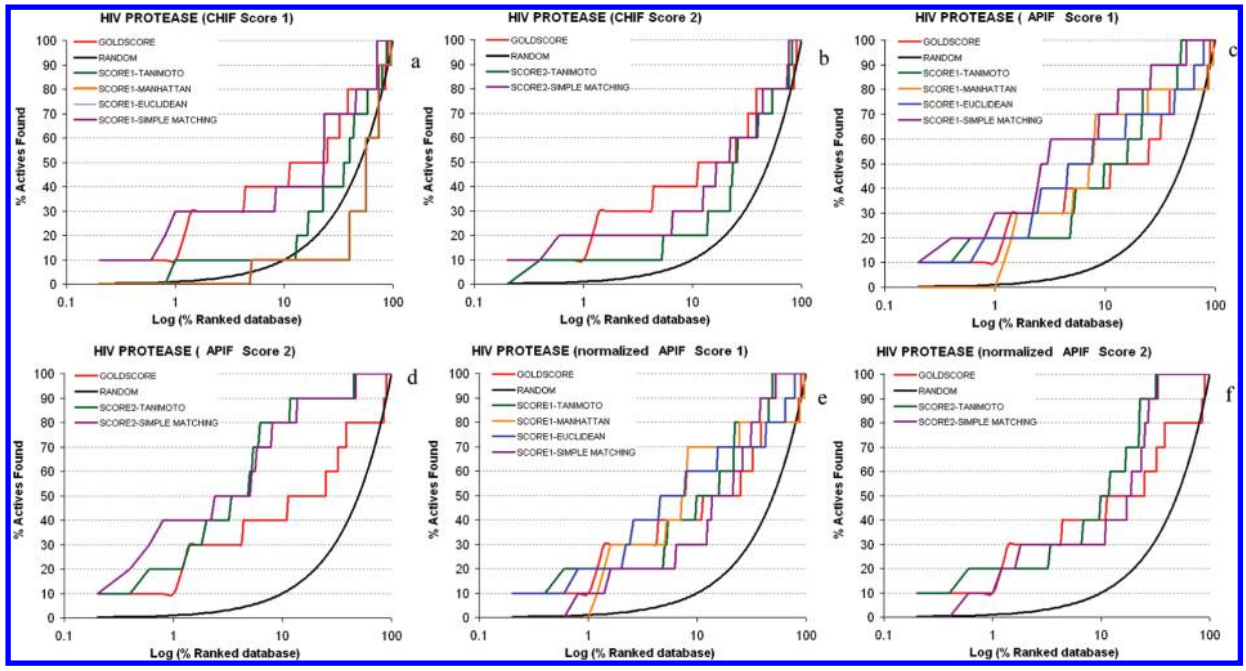

Figure 9. HIV protease plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) normalized APIF and SCORE1, and (f) normalized APIF and SCORE2. The different similarity scores used correspond to simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green), and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The *x* axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the *y* axis.

crystallographic reference complex and each docked pose was calculated. Results were analyzed using three different plots:

- Graph 1 (rmsd from crystallographic binding mode versus CHIF-based Tanimoto or APIF-based Tanimoto similarity coefficient). Plotting rms deviations from the X-ray pose versus the similarity of IFs expressed by a Tanimoto coefficient, calculated from APIF and CHIF fingerprints, generated by the X-ray and the predicted docking pose.

- Graph 2 (rmsd from crystallographic binding mode versus GoldScore rank or APIF-based Tanimoto rank or CHIF-based Tanimoto rank). Plotting rms deviations from

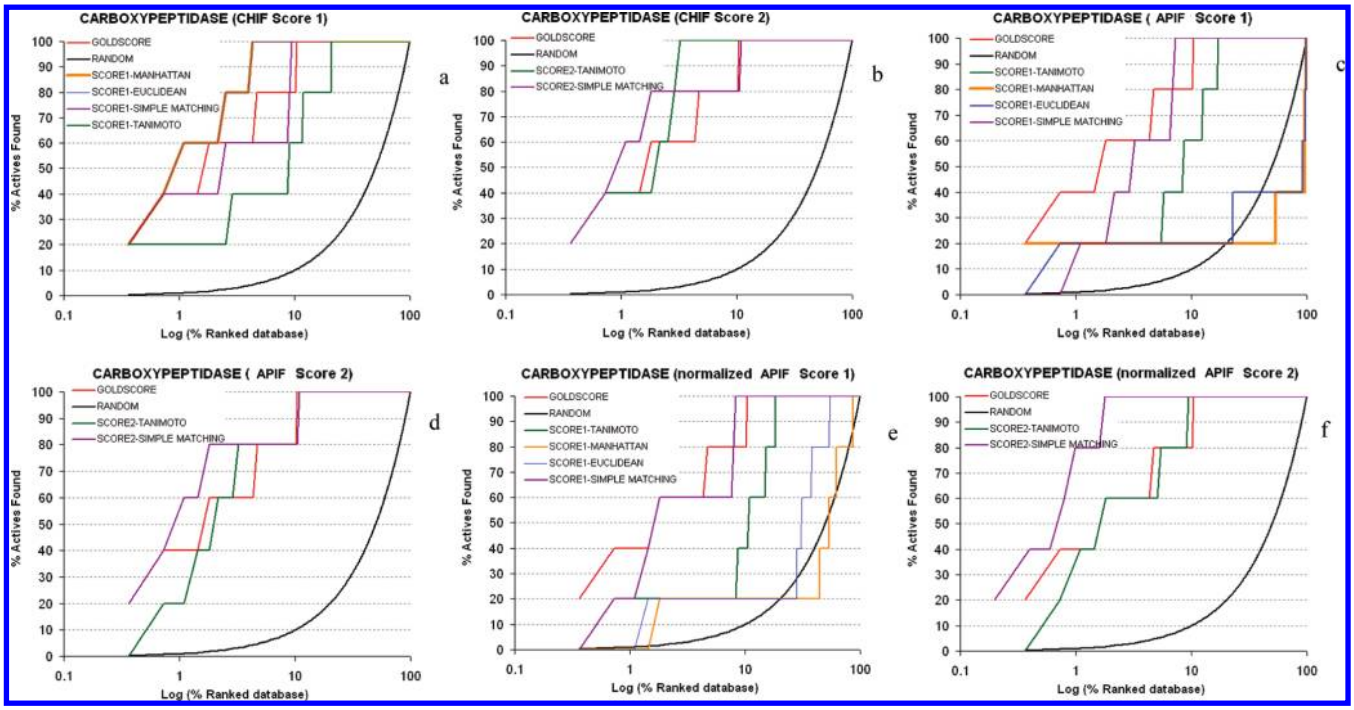

Figure 10. Carboxypeptidase plots obtained using (a) CHIF in-house implementation and SCORE1, (b) CHIF in-house implementation and SCORE2, (c) APIF and SCORE1, (d) APIF and SCORE2, (e) normalized APIF and SCORE1, and (f) normalized APIF and SCORE2. The different similarity scores used correspond to simple matching (purple), Euclidean distance (blue), Tanimoto coefficient (green), and Manhattan distance (yellow). The enrichment plot obtained using GoldScore is shown in red and a random screening in black. The *x* axis is the logarithm of the percent of the database screened, plotted against the percent recovery of known active compounds on the *y* axis.

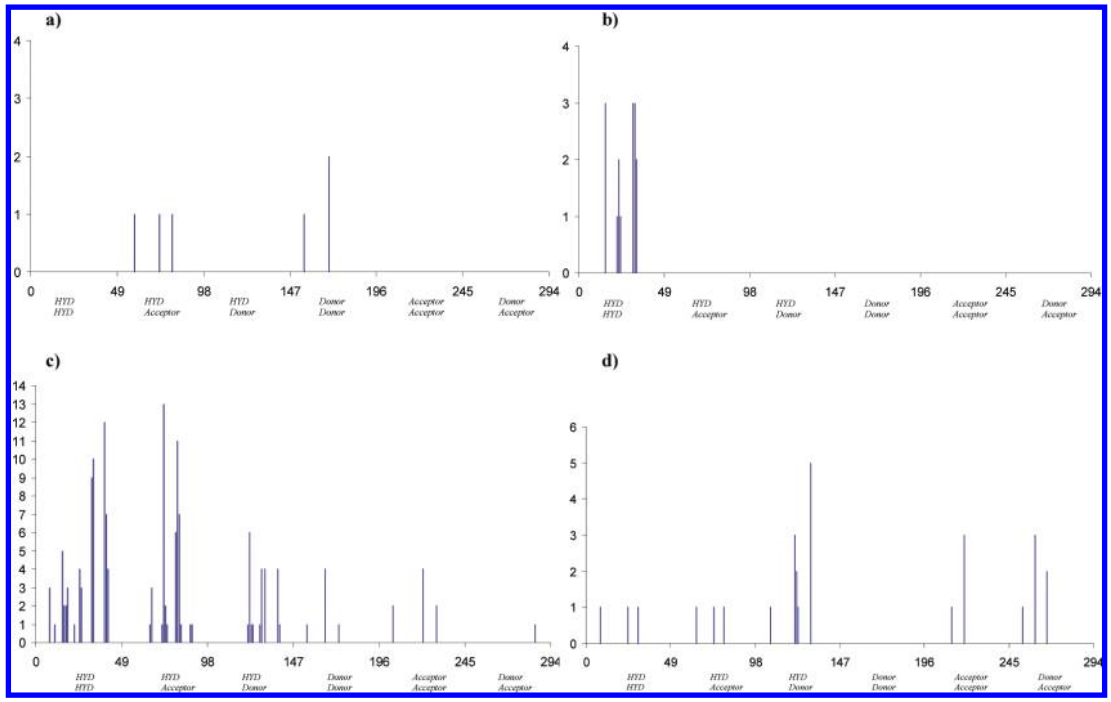

Figure 11. Correlation diagram of the APIF fingerprint for (a) trypsin, (b) rhinovirus, (c) HIV protease, and (d) carboxypeptidase complexes. The total fingerprint length is 294 bits, divided into six atom pair contacts: *HYD HYD*, referring to hydrophobic−hydrophobic protein/ligand interactions; *HYD Acceptor*, referring to hydrophobic−acceptor protein/ligand interactions; *HYD Donor*, referring to hydrophobic−donor protein/ligand interactions; *Donor Donor*, referring to donor−donor protein/ligand interactions; *Acceptor Acceptor*, referring to acceptor−acceptor protein/ligand interactions; and *Donor Acceptor*, referring to donor−acceptor protein/ligand interactions.

X-ray pose versus ranked active ligand docked conformations according to the GoldScore function or APIF-based and CHIF-based similarity Tanimoto scores.

- Graph 3 (% cases predicted within 2 Å rmsd versus binding mode rank). A comparison of the effects of using GoldScore function, and APIF and CHIF IFs to postprocess the docking-generated poses on the likelihood of identifying the crystallographic binding mode within the active docked conformations obtained.

RESULTS AND DISCUSSION

Performance of APIF in Virtual Screening: Database Enrichment. Here, we present the VS enrichment plots for the ER-α, trypsin, rhinovirus, HIV protease, and car-

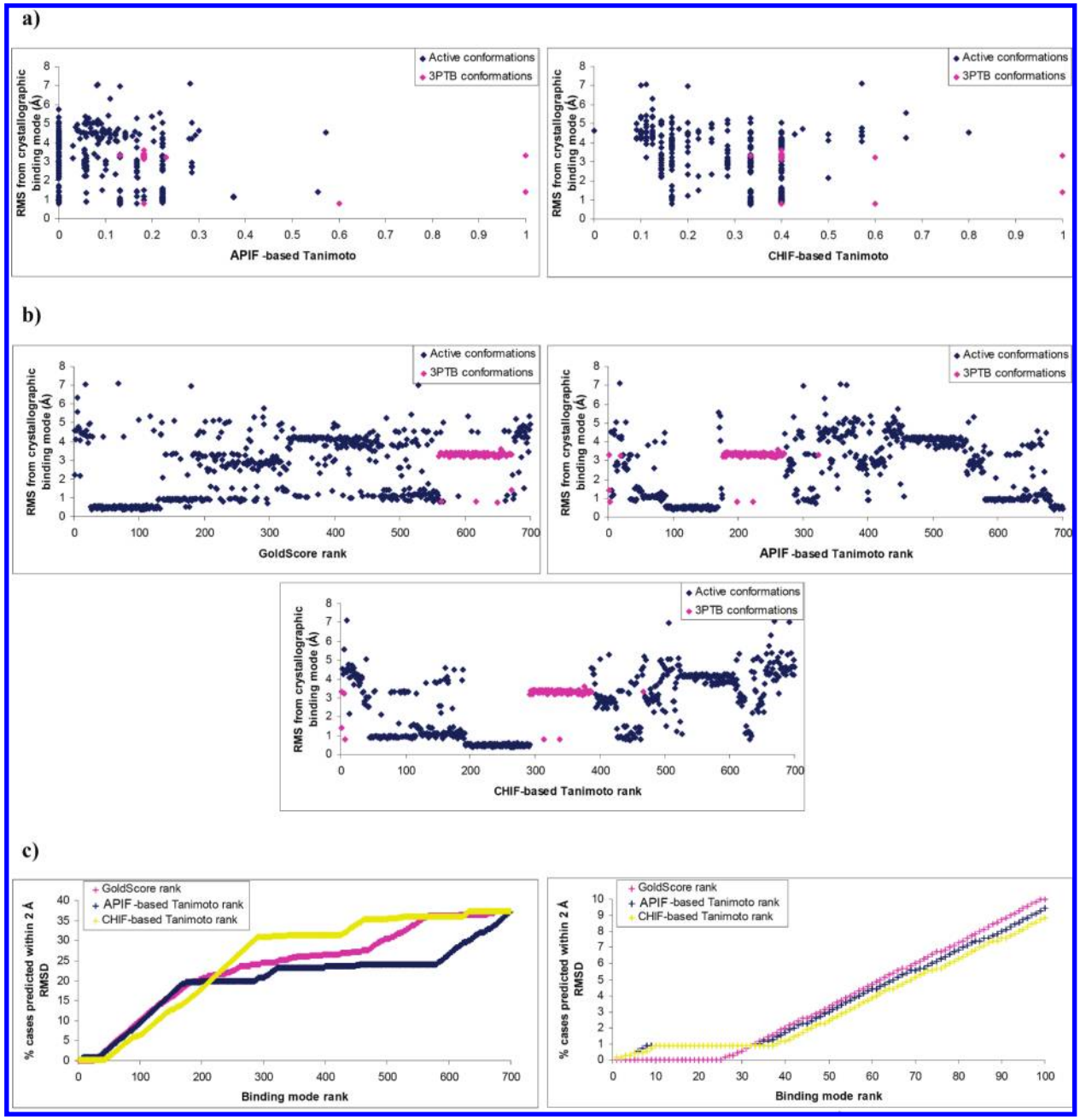

Figure 12. Trypsin binding mode analyses. (a) The rmsd from crystallographic binding mode (Å) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of seven active ligands. (b) The rmsd from crystallographic binding mode (Å) vs GoldScore rank (left), APIF-based Tanimoto rank (right), and CHIF-based Tanimoto rank (center). (c) Fraction of cases (%) predicted within 2 Å rmsd vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 700 active docked conformations obtained (left) and the top 100 ranked solutions (right).

boxypeptidase targets calculated using the GoldScore function and the CHIF, APIF, and normalized APIF-based similarity criteria. For the similarity, both the SCORE1 and SCORE2 metrics were used. In order to enhance the first part of the plots, the *x* axis uses a logarithmic scale for the percent of the database screened plotted against the percent recovery of known active compounds. We also list the enrichment factor values (EFs) for the first 2%, 5%, and 10% of the screened databases.

ER-α recovery plots (Figure 6) show that our in-house CHIF implementation (Figure 6a and b) reproduces the results reported for the same case study analyzed in Mpamhanga et al.'s work.[3] Therefore, we have validated our CHIF implementation in spite of the previously mentioned differences in determining protein contacts. Regarding the APIF (Figure 6c and d) and normalized APIF (Figure 6e and f) results, although APIF is able to retrieve compounds over a random selection, the enrichment obtained with the similarity scores is lower than the enrichment achieved using the GoldScore function (Figure 6c and e). The combination of the similarity and energetic criteria (SCORE2) achieves higher performance than that obtained only with the energetic criterion (Figure 6d and f). However, even considering SCORE2, APIF does not achieve the high performance achieved by CHIF in the first 1−2% of screened database, although it does at higher percentages. Regarding normalization, no well-defined tendency can be found. Whereas for

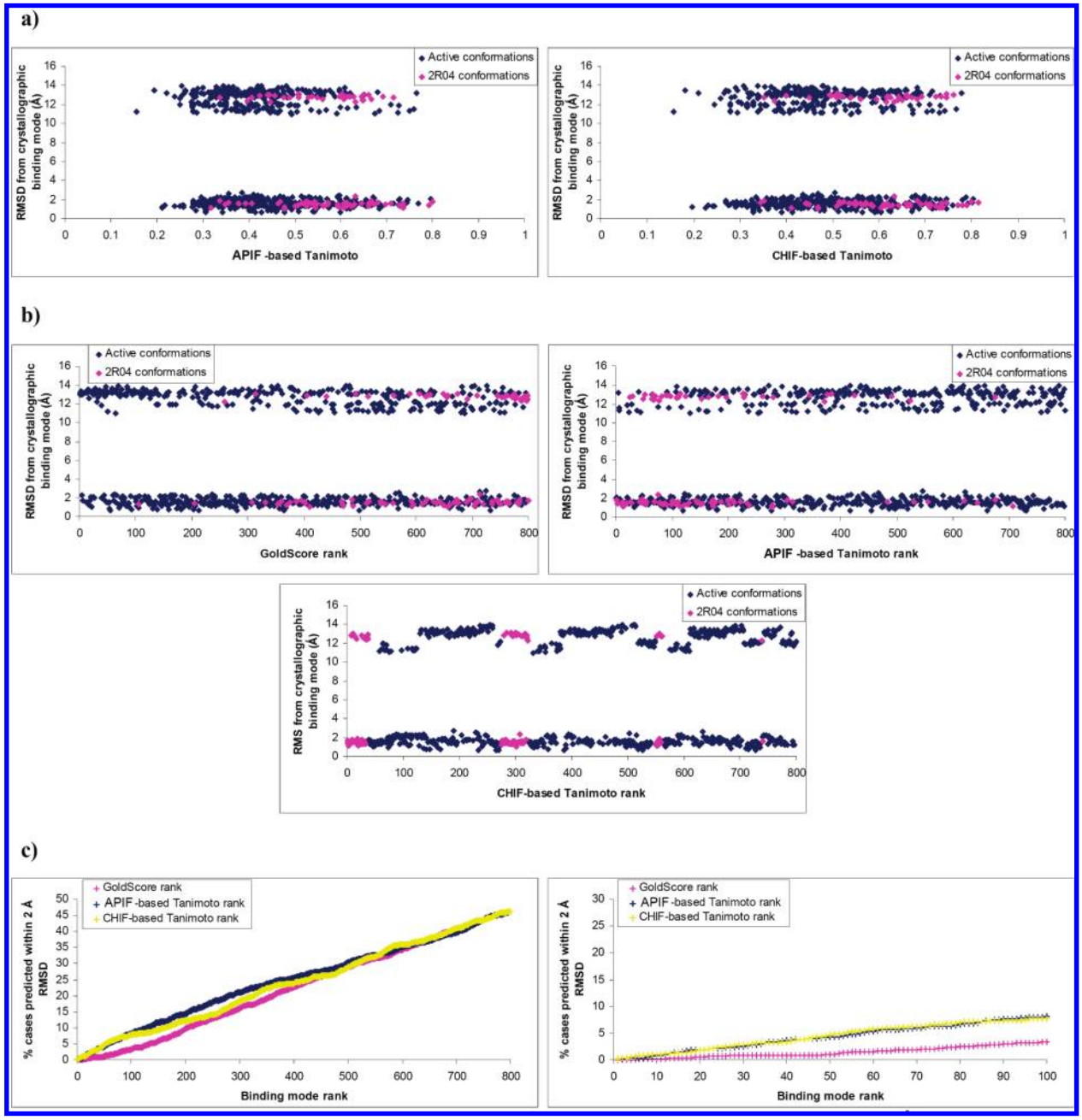

Figure 13. Rhinovirus binding mode analyses. (a) The rmsd from crystallographic binding mode (Å) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of eight active ligands. (b) The rmsd from crystallographic binding mode (Å) vs GoldScore rank (left), APIF-based Tanimoto rank (right), and CHIF-based Tanimoto rank (center). (c) Fraction of cases (%) predicted within 2 Å rmsd vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 800 active docked conformations obtained (left) and the top 100 ranked solutions (right).

SCORE1 the normalization gives worse results (Figure 6c and e), for SCORE2 it has a positive effect (Figure 6d and f).

For ER-α, the EFs for the first 2%, 5%, and 10% of the screened database are shown in Table 3. The maximum theoretical value for the EF is 50 (500/10). The maximum value found for each percentage is shown in bold. It can be observed that CHIF with SCORE2 gives the optimum result. APIF and normalized APIF with SCORE1 are not able to discriminate between active and inactive compounds better than the docking energetic criterion (GoldScore), although their behavior improves in combination with the energetic criterion.

Results for trypsin (Figure 7) show that docking does not perform so well in this case. The enrichment obtained with the similarity scores is higher than the enrichment calculated using GoldScore for APIF (Figure 7c−f), which achieves higher performance than CHIF (Figure 7a and b). In both cases, the combination of the similarity and energetic criteria (SCORE2) achieves higher performance than that obtained only with the energetic criterion (Figure 7b, d, and f). APIF normalization does not improve the results in this case (Figure 7e and f). Table 4 shows the EFs obtained at the first percentages of the database screened. The maximum theoretical value for the EF is 67.7 (474/7). The maximum value found for each percentage is shown in bold.

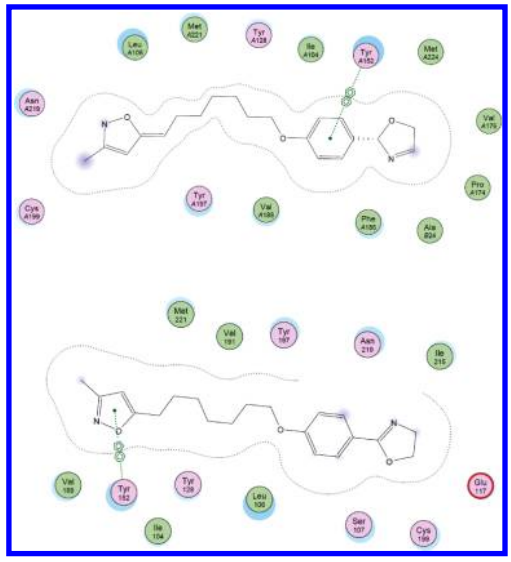

Figure 14. Rhinovirus binding modes. Two binding modes found for rhinovirus conformations interacting with the same protein atoms.

Table 6. HIV Protease Enrichment Factor Values for the First 2%, 5%, and 10% of the Screened Database

	HIV protease		
	2%	5%	10%
GOLDSCORE (docking)	15	8	4
CHIF-SCORE1-TANIMOTO	5	2	1
CHIF-SCORE1-SIMPLE_MATCHING	15	6	4
CHIF-SCORE1-EUCLIDEAN	0	2	1
CHIF-SCORE1-MANHATTAN	0	2	1
CHIF-SCORE2-TANIMOTO	5	2	2
CHIF-SCORE2-SIMPLE_MATCHING	10	4	3
APIF-SCORE1-TANIMOTO	15	10	7
APIF-SCORE1-SIMPLE_MATCHING	15	**12**	7
APIF-SCORE1-EUCLIDEAN	15	8	4
APIF-SCORE1-MANHATTAN	10	6	4
APIF-SCORE2-TANIMOTO	**20**	**12**	**8**
APIF-SCORE2-SIMPLE_MATCHING	**20**	10	**8**
NORMALIZED_APIF-SCORE1-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	10	4	3
NORMALIZED_APIF-SCORE1-EUCLIDEAN	10	10	6
NORMALIZED_APIF-SCORE1-MANHATTAN	15	8	7
NORMALIZED_APIF-SCORE2-TANIMOTO	10	6	5
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	15	6	3

Results for rhinovirus (Figure 8) show that the enrichments given by the APIF-based and CHIF-based similarity scores are higher than those obtained by docking. The CHIF fingerprint achieves higher performance than APIF in all SCORE1 metrics (Figures 8a, c, and e) and simple matching SCORE2 (Figure 8b). APIF achieves higher performance than CHIF in Tanimoto SCORE2 (Figure 8d) and normalized APIF Tanimoto SCORE1 and SCORE2 (Figure 8e and f). Table 5 shows in detail some EF values. In this case, the maximum theoretical EF is 63.2 (506/8).

Results for HIV protease (Figure 9) show that the enrichments obtained using GoldScore are similar to those from the APIF-based similarity scores (Figure 9c−f) and higher than those obtained from the CHIF-based similarity scores (Figure 9a and b). APIF (Figure 9c and d) achieves higher performance than CHIF (Figure 9a and b). In both cases, the simple matching score gives the best results. In this case, APIF normalization does not improve the APIF results for both SCORE1 and SCORE2 (Figure 9e and f). Table 6 shows some EF values. The maximum theoretical value for the EF is 49.9 (499/10).

Table 7. Carboxypeptidase Enrichment Factor Values for the First 2%, 5%, and 10% of the Screened Database

	carboxypeptidase		
	2%	5%	10%
GOLDSCORE (docking)	30	16	10
CHIF-SCORE1-TANIMOTO	10	8	6
CHIF-SCORE1-SIMPLE_MATCHING	20	12	**10**
CHIF-SCORE1-EUCLIDEAN	30	**20**	**10**
CHIF-SCORE1-MANHATTAN	30	**20**	**10**
CHIF-SCORE2-TANIMOTO	30	**20**	**10**
CHIF-SCORE2-SIMPLE_MATCHING	40	16	**10**
APIF-SCORE1-TANIMOTO	10	4	6
APIF-SCORE1-SIMPLE_MATCHING	20	12	**10**
APIF-SCORE1-EUCLIDEAN	10	4	2
APIF-SCORE1-MANHATTAN	10	4	2
APIF-SCORE2-TANIMOTO	30	16	8
APIF-SCORE2-SIMPLE_MATCHING	30	12	**10**
NORMALIZED_APIF-SCORE1-TANIMOTO	10	4	6
NORMALIZED_APIF-SCORE1-SIMPLE_MATCHING	30	12	**10**
NORMALIZED_APIF-SCORE1-EUCLIDEAN	10	4	2
NORMALIZED_APIF-SCORE1-MANHATTAN	10	4	2
NORMALIZED_APIF-SCORE2-TANIMOTO	30	12	**10**
NORMALIZED_APIF-SCORE2-SIMPLE_MATCHING	**50**	**20**	**10**

Results for carboxypeptidase (Figure 10) show that the enrichment obtained using GoldScore is similar to the enrichment performed by similarity coefficients. CHIF (Figure 10a and b) achieves higher performance than APIF (Figure 10c and d). CHIF-based Manhattan and Euclidean SCORE1 and CHIF-based Tanimoto and simple matching SCORE2 perform better than the GoldScore function (Figure 10a and b). Regarding the combination of the similarity and energetic criteria (SCORE2), simple matching gives the best results in all cases. Regarding normalization, the normalized APIF gives similar results for both SCORE1 and SCORE2 to those of APIF, except for the simple matching score (Figure 10e and f), which improves results. Table 7 shows the enrichment values for the first percentages of the database screened. The maximum theoretical value for the EF is 55.4 (277/5).

Finally, we show the correlation diagram of the APIF fingerprint for trypsin, rhinovirus, HIV protease, and carboxypeptidase reference complexes (Figure 11). The number of contacts found for each target and the type of protein−ligand interactions are shown: hydrophobic−hydrophobic (*HYD HYD*), hydrophobic−acceptor (*HYD Acceptor*), hydrophobic−donor (*HYD Donor*), donor−donor (*Donor Donor*), acceptor−acceptor (*Acceptor Acceptor*), and donor−acceptor (*Donor Acceptor*).

Summarizing database enrichment analyses, CHIF obtains the best EF values for ER-α, rhinovirus, and carboxypeptidase. For trypsin and HIV protease, APIF achieves the best EFs. Moreover, rhinovirus and carboxypeptidase APIF

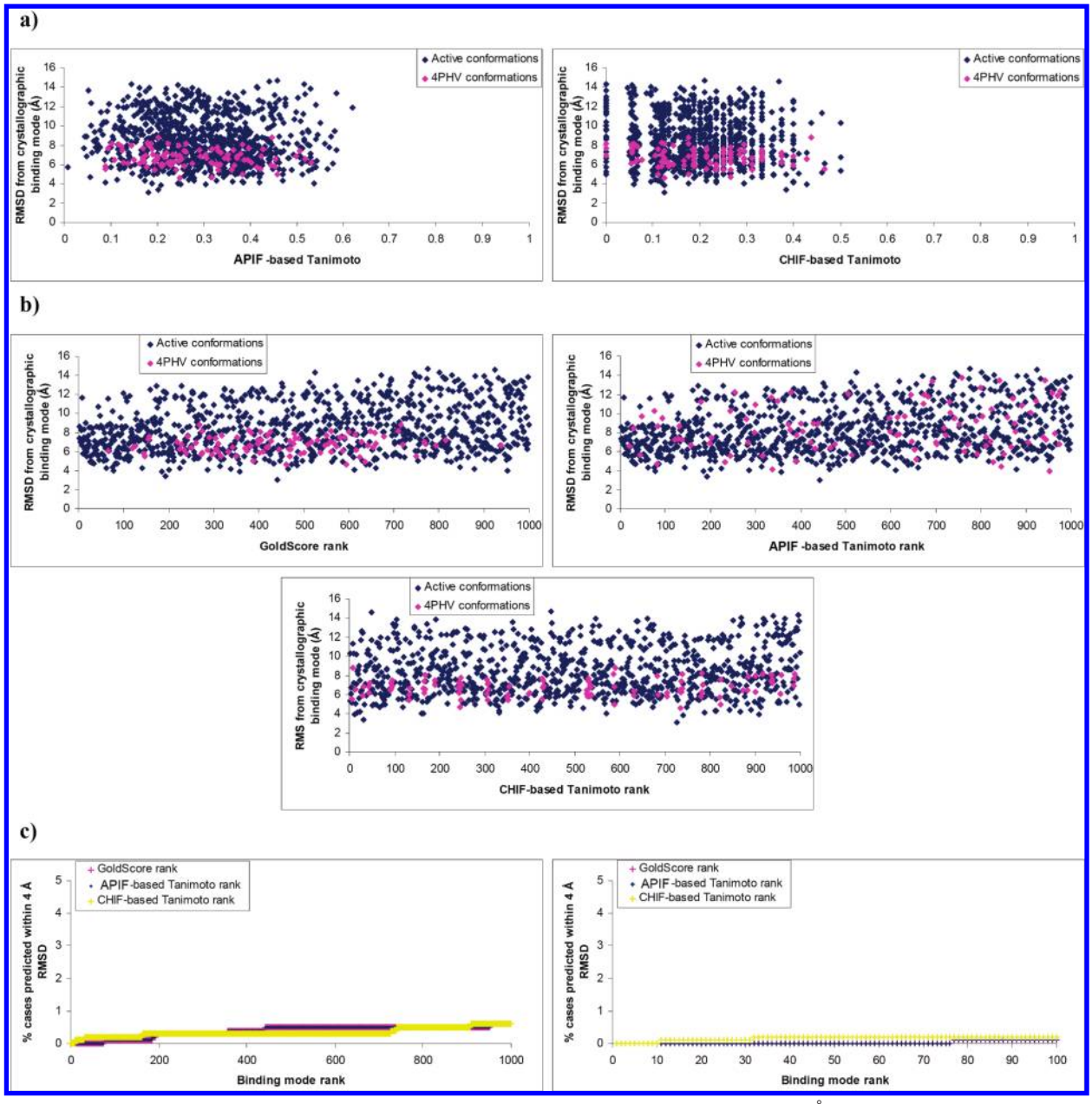

Figure 15. HIV protease binding mode analyses. (a) The rmsd from crystallographic binding mode (Å) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of 10 active ligands. (b) The rmsd from crystallographic binding mode (Å) vs GoldScore rank (left), APIF-based Tanimoto rank (right), and CHIF-based Tanimoto rank (center). (c) Fraction of cases (%) predicted within 2 Å rmsd vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 1000 active docked conformations obtained (left) and the top 100 ranked solutions (right).

SCORE2 and normalized APIF SCORE2, respectively, improve CHIF EF results. Furthermore, the APIF Tanimoto and Euclidean similarity scores always return good enrichments even though they do not always achieve the best results. Generally, the combination of the similarity and energetic criteria (SCORE2) achieves higher enrichments than those obtained only with the energetic criterion, except for failed docked conformations or bad scoring function behavior (Tables 3−7).

APIF Recognition of The Binding Mode. Here, we present the binding mode analyses for trypsin, rhinovirus, HIV protease, and carboxypeptidase first-ranked conformations according to the previously described criteria (Figures 12−15). For each target, the docked conformations corresponding to the reference crystallographic ligand are shown in pink, whereas those corresponding to the rest of the active compounds of the set are shown in blue.

Results for the trypsin target (Figure 12) show a non-well-defined tendency to associate high Tanimoto scores with low rmsd values from the crystal structure (Figure 12a), although this tendency is clearer for the conformations corresponding to the complexed ligand (PDB code: 3PTB). It can be seen that docking performs randomly (broad range of rmsd values), but rather well for some conformations (rmsd < 2 Å). Moreover, IFs capture the basic interactions for the lowest rmsd conformations (Tanimoto score values = 1 for 3PTB conformations). Lower rmsd conformations from the crystallographic binding mode are found in the top CHIF and APIF hitlist ranking positions, especially for CHIF top-ranked conformations (Figure 12b). Ligand conformations

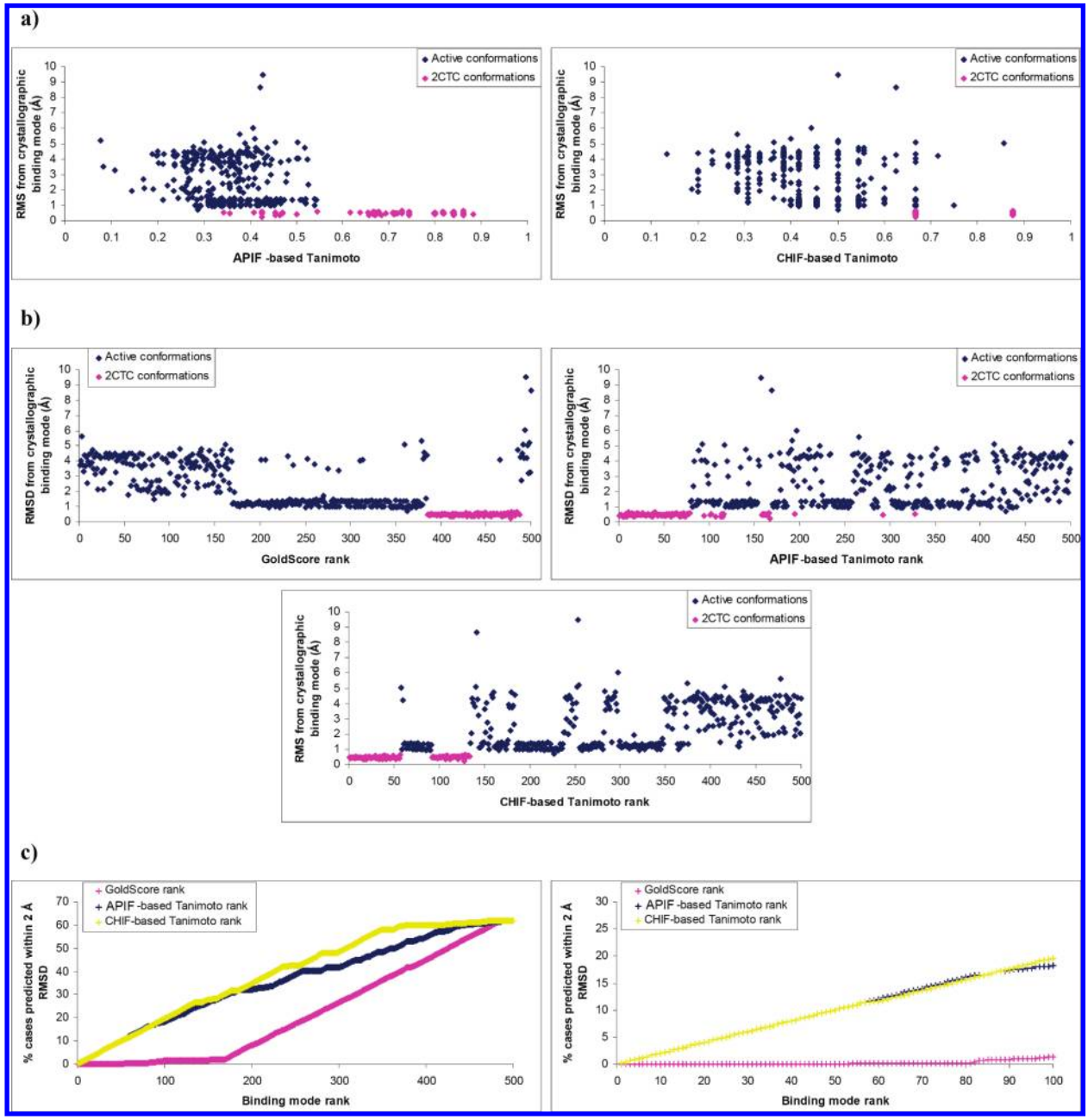

Figure 16. Carboxypeptidase binding mode analyses. (a) The rmsd from crystallographic binding mode (Å) versus APIF-based Tanimoto (left) and CHIF-based Tanimoto (right) for a set of five active ligands. (b) The rmsd from crystallographic binding mode (Å) vs GoldScore rank (left), APIF-based Tanimoto rank (right), and CHIF-based Tanimoto rank (center). (c) Fraction of cases (%) predicted within 2 Å rmsd vs binding mode rank for GoldScore (pink curve), APIF (blue curve), and CHIF (yellow curve) for the 500 active docked conformations obtained (left) and the top 100 ranked solutions (right).

corresponding to the complexed compound are found in the first ranking positions for CHIF and APIF ranking lists and in the last positions for the GoldScore function (Figure 12b). The first 35 CHIF and APIF top-ranked ligand conformations have lower rmsd from the crystallographic binding mode than the first top-ranked GoldScore conformations. For the subsequent ranked conformations, the APIF-based Tanimoto score and GoldScore function give better results (Figure 12c).

Results for the rhinovirus target (Figure 13) show no tendency to associate high Tanimoto scores with low rmsd values from the crystal structure because multiple conformations with the same protein interacting points but different binding modes are found (Figure 14). Two groups of ligand conformations are found, one with low rmsd from the crystallographic binding mode (between 0 and 2 Å) and the other with higher rmsd values (between 11 and 14 Å). Moreover, the conformations corresponding to the complexed ligand (PDB code: 2R04) show high Tanimoto score values (Figure 13a), but not exceeding 0.8. These conformations with rmsd < 2 Å do not achieve Tanimoto score values of 1 due to the fact that a hydrophobic interaction present in the crystal reference complex is changed to a hydrogen-bond contact, and a new hydrophobic interaction is created between the ligand and a neighboring residue to the crystallographic interacting one. Lower rmsd from the crystallographic binding mode ranked conformations alternate with higher rmsd ranked conformations, according to the two binding modes found (Figure 13b). Ligand conformations corresponding to the complexed compound are found in the first ranking positions for CHIF and APIF ranking lists and

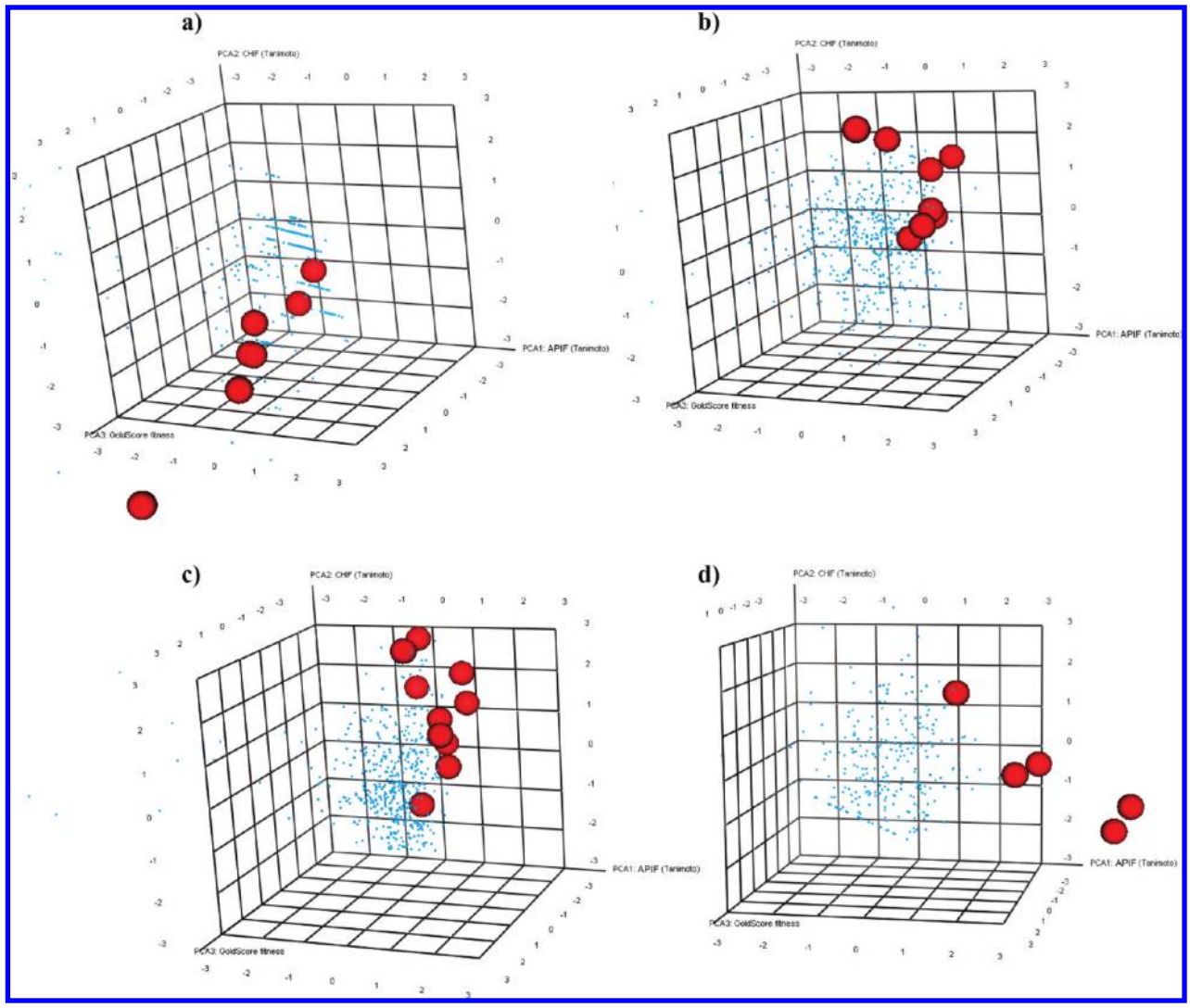

Figure 17. PCA analyses of trypsin, rhinovirus, HIV protease, and carboxypeptidase compound databases, showing the separation of active compounds' IFs into specific eigenvector spaces. (a) Trypsin compounds database PCA analysis. (b) Rhinovirus compounds database PCA analysis. (c) HIV protease compounds database PCA analysis. (d) Carboxypeptidase compounds database PCA analysis. In all cases, PCA axes correspond to APIF-based Tanimoto score, CHIF-based Tanimoto score, and GoldScore function. Active compounds are shown in red ball-and-stick and inactive compounds in blue stick representations.

in the last positions for the GoldScore function (Figure 13b). The first top-ranked CHIF and APIF conformations show lower rmsd from the crystallographic binding mode than the first-ranked GoldScore function conformations (Figure 13c).

Results for the HIV protease target (Figure 15) show that the docking procedure performs poorly in this case. The rmsd values obtained are generally high, that is, over 4 Å from the crystallographic binding mode, and the similarity between the docked conformations and the reference complex is always lower than 0.5 (Figure 15a). Both for the GoldScore function and for CHIF and APIF IFs, the ligand conformations corresponding to the complexed compound (PDB code: 4PHV) are found randomly along the ranking lists (Figure 15b). Given that the docking procedure cannot find the experimental binding mode, the IF calculation from the docked poses achieves poor results too (Figure 15c).

Results for the carboxypeptidase target (Figure 16) show that docking performs well (high number of conformations within rmsd < 2 Å). Higher Tanimoto similarity scores correspond to lower rmsd from the crystal structure values (Figure 16a). Moreover, this tendency is emphasized for the ligand conformations corresponding to the complexed compound (PDB code: 2CTC). Ligand conformations with the lowest rmsd from the crystallographic binding mode are found in the first IF ranking positions. However, they are found in the last positions for GoldScore (Figure 16b). The first top-ranking IF ligand conformations show lower rmsd from the crystallographic binding mode than the first GoldScore ranked conformations (Figure 16c).

In order to visualize binding modes, a PCA analysis was performed. Figure 17 shows three-dimensional PCA plots for trypsin, rhinovirus, HIV protease, and carboxypeptidase complexes. These plots show that active molecules are located in a different region than the inactive compounds. Corroborating the above-mentioned binding mode results, trypsin and carboxypeptidase seem to best recognize the binding mode closest to the crystallographic structures for the first top-ranked conformations, restricting active molecules to a specific region of space far from inactive compounds in the PCA plot.

Summarizing binding mode analyses, carboxypeptidase and trypsin show the best tendency to associate high Tanimoto scores with low rmsd values from the crystal structure. Rhinovirus and HIV protease do not follow this

tendency because docking is not able to find the correct binding mode conformations. However, all results show that the lowest rmsd conformations are found in the first CHIF and APIF top-ranked positions and in the subsequent ranked GoldScore positions (Figures 12c, 13c, 14c, and 15c). Therefore, IFs provide a better method for identifying low rmsd conformations from the crystallographic binding mode than only using a docking scoring function.

CONCLUSION

The analyses in this study indicate that our new interaction fingerprint (APIF) yields satisfactory results, often comparable to our CHIF implementation, and it improves the GoldScore results, inasmuch as our enrichment plots exhibit good recognition of the known actives. Overall, this study shows that APIF has proven to be suitable for ranking and filtering virtual screening docking results. However, the quality of the EFs obtained by APIF scoring strongly depends on docking success. Our results show that, if docking is successful, as in the trypsin and carboxypeptidase cases, then APIF scoring retrieves good enrichments, substantially improving the results obtained when using only a docking scoring function. Using APIF is thus a good way to select poses or virtual hits that satisfy a defined ligand−protein interaction reference, which will be useful for receptor-based prospective virtual screening.

ACKNOWLEDGMENT

We thank Dave Ritchie for proof-reading the manuscript. V.I.P.N. thanks the Generalitat de Catalunya−DURSI for a grant within the Formació de Personal Investigador (2008FI) Program. This work was supported by The TV3 Marathon Foundation (AIDS-2001) promoted by the Catalan Radio and Television Corporation (Corporació Catalana de Ràdio i Televisió, CCRTV) and the Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01).

REFERENCES AND NOTES

(1) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
(2) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.
(3) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686–698.
(4) Chuaqui, C.; Deng, Z.; Singh, J. Interaction Profiles of Protein Kinase-Inhibitor Complexes and Their Application to Virtual Screening. *J. Med. Chem.* **2005**, *48*, 121–133.
(5) Warren, G. L.; Andrews, C. V.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(6) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 151–166.
(7) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definitions and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(8) FlexS-77 data set collected by C. Lemmen, G. Klebe, and M. Böhm, first published in: Lemmen, C.; Lengauer, T; Klebe, G. FlexS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.
(9) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.
(10) *Maybride Bringing Life to Drug Discovery*; Maybridge Databases Autumn 2005; Fisher Scientific International: England, 2005.
(11) *MOE (Molecular Operating Environment)*, 2006.08 release; Chemical Computing Group, Inc.: Montreal, Canada, 2004.
(12) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.
(13) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609–623.
(14) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and its Complexes with Inhibitors. *Acta Crystallogr., Sect. B* **1983**, *39*, 480–490.
(15) Renatus, R.; Bode, W.; Huber, R.; Stürzebecher, J.; Stubbs, M. T. Structural and Functional Analyses of Benzamidine-Based Inhibitors in Complex with Trypsin: Implications for the Inhibition of Factor Xa, tPA, and Urokinase. *J. Med. Chem.* **1998**, *41*, 5445–5456.
(16) Böhm, M.; Stürzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42*, 458–477.
(17) Badger, J.; Minor, I.; Oliveira, M. A.; Smith, T. J.; Rossmann, M. G. Structural analysis of antiviral agents that interact with the capsid of human rhinoviruses. *Proteins* **1989**, *6*, 1–19.
(18) Matthews, D. A.; Dragovich, P. S.; Webber, S. E.; Fuhrman, S. A.; Patick, A. K.; Zalman, L. S.; Hendrickson, T. F.; Love, R. A.; Prins, T. J.; Marakovits, J. T.; Zhou, R.; Tikhe, J.; Ford, C. E.; Meador, J. W.; Ferre, R. A.; Brown, E. L.; Binford, S. L.; Brothers, M. A.; DeLisle, D. M.; Worland, S. T. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 11000–11007.
(19) Bone, R.; Vacca, J. P.; Anderson, P. S.; Holloway, M. K. X-Ray Crystal Structure of the HIV Protease Complex with L-700,417, an Inhibitor with Pseudo C2 Symmetry. *J. Am. Chem. Soc.* **1991**, *113*, 9382–9384.
(20) Specker, E.; Böttcher, J.; Brass, S.; Heine, A.; Lilie, H.; Schoop, A.; Müller, G.; Griebenow, N.; Klebe, G. Unexpected Novel Binding Mode of Pyrrolidine-Based Aspartyl Protease Inhibitors: Design, Synthesis and Crystal Structure in Complex with HIV Protease. *ChemMedChem* **2006**, *1*, 106–117.
(21) Specker, E.; Böttcher, J.; Lilie, H.; Heine, A.; Schoop, A.; Müller, G.; Griebenow, N.; Klebe, G. An Old Target Revisited: Two New Privileged Skeletons and an Unexpected Binding Mode For HIV-Protease Inhibitors. *Angew. Chem., Int. Ed.* **2005**, *44*, 3140–3144.
(22) Teplyakov, A.; Wilson, K. S.; Orioli, P.; Mangani, S. High-resolution structure of the complex between carboxypeptidase A and L-phenyl lactate. *Acta Crystallogr. Sect., D* **1993**, *49*, 534–540.
(23) Rees, D. C.; Lipscomb, W. N. Binding of ligands to the active site of carboxypeptidase A. *Proc. Natl Acad. Sci. U. S. A.* **1981**, *78*, 5455–5459.
(24) Kim, H.; Lipscomb, W. N. Crystal Structure of the Complex of Carboxypeptidase A with a Strongly Bound Phosphonate in a New Crystalline Form: Comparison with Structures of Other Complexes. *Biochemistry* **1990**, *29*, 5546–5555.
(25) Christianson, D. W.; Lipscomb, W. N. Binding of a possible transition state analogue to the active site of carboxypeptidase A. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 6840–6844.
(26) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927–937.
(27) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(28) Willet, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900–908.
(29) Hert, J.; Willet, P.; Wilton, D. J. Comparison of Fingrprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(30) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.

CI900043R