Knowledge-Based Interaction Fingerprint Scoring: A Simple Method for Improving the Effectiveness of Fast Scoring Functions

Chidochangu P. Mpamhanga,[†,‡] Beining Chen,*[,†] Iain M. McLay,[§] and Peter Willett[‡]

Department of Chemistry, University of Sheffield, Sheffield S3 7HF, United Kingdom,
Department of Information Studies, University of Sheffield, Sheffield S1 4DP, United Kingdom, and
GlaxoSmithKline Medicines Research Centre, Computational and Structural Sciences, Stevenage,
Hertfordshire SG1 2NY, United Kingdom

Received September 23, 2005

A new method for the postprocessing of docking outputs has been developed, based on encoding putative 3D binding modes (docking solutions) as ligand−protein interactions into simple bit strings, a method analogous to the structural interaction fingerprint. Instead of employing traditional scoring functions, the method uses a series of new, knowledge-based scores derived from the similarity of the bit strings for each docking solution to that of a known reference binding mode. A GOLD docking study was carried out using the Bissantz estrogen receptor antagonist set along with the new scoring method. Superior recovery rates, with up to 2-fold enrichments, were observed when the new knowledge-based scoring was compared to the GOLD fitness score. In addition, top ranking sets of molecules (actives and potential actives or decoys) were structurally diverse with low molecular weights and structural complexities. Principal component analysis and clustering of the fingerprints permits the easy separation of active from inactive binding modes and the visualization of diverse binding modes.

1. INTRODUCTION

Current methods for docking small ligands into protein targets employ a variety of approaches including genetic or evolutionary programming, simulated annealing, Monte Carlo, distance geometry, shape matching, and incremental construction. By and large, these methods appear to be relatively successful at predicting known ligand poses.[3,4] However, although docking is reasonably successful (in that it can return poses close to the X-ray structures), there is increasing agreement that the scoring schemes in current use often fail to predict accurately the measured ligand affinity for the receptor.

There are three common classes of fast scoring functions: empirical-, knowledge-, and force-field-based models. Each is derived from an analysis of experimental data (which may be incomplete and inconsistent) such as measured binding constants, descriptors derived from the analysis of ligand/protein X-ray complexes, or physicochemical experimental data.[5] The difficulties of scoring are further aggravated by our current inability to correctly apportion the hydrophobic and desolvation energy terms. These failures can be moderated, to some extent, by using data fusion methods, that is, the combination of two or more scoring functions into a more effective consensus score.[6,7] More complex approaches that have been studied to address the limitations of current systems include a variety of "knowledge-based" or "target-biased" approaches that impose constraints based on ligand or receptor pharmacophores thought to be required for activity[8−10] and methods based on free energy grids.[11]

The main objective of the work reported here was to develop a generally applicable structure-based scoring methodology which uses knowledge of the binding interactions gained from the analysis of available protein−ligand crystal structures. To achieve this, interactions between a ligand and its receptor were identified from available X-ray structures for reference molecules and from the predicted binding modes for docked libraries of compounds. The information gained was encoded into simple bit strings or *interaction fingerprints* (IFs). A variety of similarity coefficients were then employed to rank the solutions or poses obtained from ligand docking by evaluating the similarity between each bit string and a reference bit string. The similarity measures or coefficients employed were adapted from the field of ligand-based virtual screening and chemoinformatics.[12−21] Deng and co-workers[1] have previously reported a technique for residue-based interaction fingerprinting that is based on an analysis of interaction profiles from kinase structures and that was aimed at understanding the basis of inhibitor selectivity. This technique has recently been extended by Kelly and Mancera[33] to encompass the use of atom-based interaction fingerprints for clustering and retrieving binding modes. The IF method described here employs a similar system for encoding interactions, but the motivation of this work was the need to find simple, all-purpose knowledge-based scoring strategies for use in studies where at least one co-crystallized X-ray structure is available. The basic assumption underpinning our approach is not that all novel potential leads will have the same binding modes as the reference compound but that they will share many similar

[*] Corresponding author E-mail: B.Chen@shef.ac.uk.
[†] Department of Chemistry, The University of Sheffield.
[‡] Department of Information Studies, The University of Sheffield.
[§] GlaxoSmithKline.

10.1021/ci050420d CCC: $33.50
Published on Web 02/02/2006

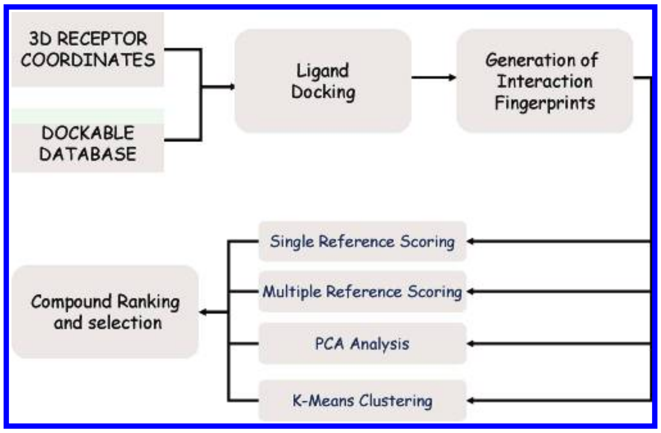

Figure 1. General workflow proposed for interaction-fingerprint-based scoring. The interaction fingerprint can also be used for separating active compounds from inactive by clustering and PCA analysis.

interaction patterns and will therefore show high similarity to the reference compound(s).

Methods exploring the use of single reference crystal structures for knowledge-based scoring were implemented and evaluated using the well-known Bissantz[2] inhibitor set of the α estrogen receptor (ER). The methodology was extended to encompass multiple reference crystal structures using methods analogous to work on ligand-based similarity virtual screening discussed by Hert et al.[22] Essentially, multiple reference scoring can be viewed as a simplified learning algorithm in which the scoring is improved as each new reference IF becomes available. In addition to scoring docked structures, it is possible to use IFs to explore the relationships between ligand binding modes by clustering[16,17] or principal components analysis (PCA).[31]

2. METHODS

The workflow proposed in Figure 1 shows that the procedure starts with the customary receptor site and database preparation, followed by docking a "reference" compound and analysis of its docking accuracy, then finally docking and scoring the compound database in a virtual screening mode using both the traditional (GOLD fitness score) and novel (IF-based) scoring methods.

2.1. Computational Tools. The tools used for this study included the Daylight toolkit[23] for database preparation, CONCORD[24,25] for 3D structure generation, and the SYBYL tool case[25] for visualization and preparation of the docking site. For virtual screening, the GOLD version 2.2[26] PVM Linux version was run on a 47-processor Linux system. In-house SPL, Perl, and C scripts were developed for the generation of interaction fingerprints. Finally, the UNITY[25] database analysis tools were utilized for structural analysis of the top-ranked sets of compounds.

2.2. Protein Site Preparation. To facilitate the docking and scoring protocol, we selected the same reference X-ray structure for the ER (PDB entry 3ert) as that used in the Bissantz study. This is a co-crystallized structure of the receptor with 4-hydroxy-tamoxifene (4HT).[27] In addition, one other X-ray structure, raloxifene[28] (1err), was selected and docked to demonstrate multiple reference scoring. Protein preparation included extracting the ligands and the water molecules from the active site. Once extracted, the 3D ligand conformations were stored for use in the docking evaluation steps (virtual crystallography). The 4HT site was used as the principle docking site for this study.

2.3. Preparation of 3D Databases. A set of 490 druglike molecules was selected at random from the GlaxoSmithKline (GSK) preferred supplier's compound library (SOBAX)[7] and represented in the SMILES format. This data set was preprocessed to produce a dockable library, after ascribing realistic protonation states to potential ionizable groups (amines, amidines, carboxylic acids, and other acidic isosteric forms), and then the 3D structures were generated using CONCORD. This data set of 490 compounds was used as the background molecule set (all assumed to be inactive) and was seeded with the "Bissantz active set" consisting of 10 high-affinity ERα antagonists to make up a database of 500 compounds; this data set was used for all the analyses documented in this study. Two more compound databases were designed and docked alongside the SOBAX set; these included seeding the same 10 known antagonists into 1000 background "druglike" molecules made available online by Rognan et al.[28] (Rognan set) and a randomly selected set of 500 compounds from the MAO data set[30] (MAO set).

2.4. Ligand Docking. GOLD version 2.2 was used to dock both the reference structures and the compound libraries described immediately above. The coordinates of a point deep in the binding site were located manually during protein site preparation, and the GOLD flood-fill algorithm was used to generate a solvent-accessible binding site from all of the atoms within a 10 Å radius of the selected point. For each of the 10 independent genetic algorithm runs, a default maximum of 100 000 genetic operations was performed, using the default operator weights and a population size of 100 chromosomes.

2.5. Generation of the Interaction Fingerprints. A simple algorithm was designed to generate three different types of interaction fingerprints, called CIF, HIF, and CHIF. First, this involved determining the atomic coordinates of the receptor site and then those of each GOLD-predicted solution for the compounds. Second, two pairwise interatomic parameters [distance (d) and H-bond angle (a)] were measured. And last, for all acceptor−polar hydrogen pairs, if d was less than or equal to 3.0 Å and a was less than 90°, then a hydrogen bond was detected and a bit representing the receptor atom on a linear bit string was set to 1 (thus generating a HIF). If d was within van der Waals (vdW) close contact constraints and was between non-hydrogen bonding atoms, a close contact was detected and the corresponding receptor atom bit set for the CIF. Close contacts are defined as any interatomic distance (d) shorter than the sum of the vdW radii of any non-hydrogen-bonding atomic pair. A third type of fingerprint, the CHIF, is built by an amalgamation of the HIFs and CIFs. In our implementation, the IF is built from a fixed-length bit string as long as the number of heavy atoms in the binding site (see Figure 2 for an illustration). Hydrogen-bond acceptors are defined as any heteroatom with potential lone pairs; these included the following SYBYL mol2 atom types: N.ar, N.1-3, N.pl3, S.3, S.2, S.O, O.2, O.3, and P.3. Hydrogen-bond donor atoms include any heteroatoms bonded to hydrogen atoms: N.ar, N.1, N.2, N.3, N.pl3, S.2, S.3, S.O, O.2, O.3, and P.3.

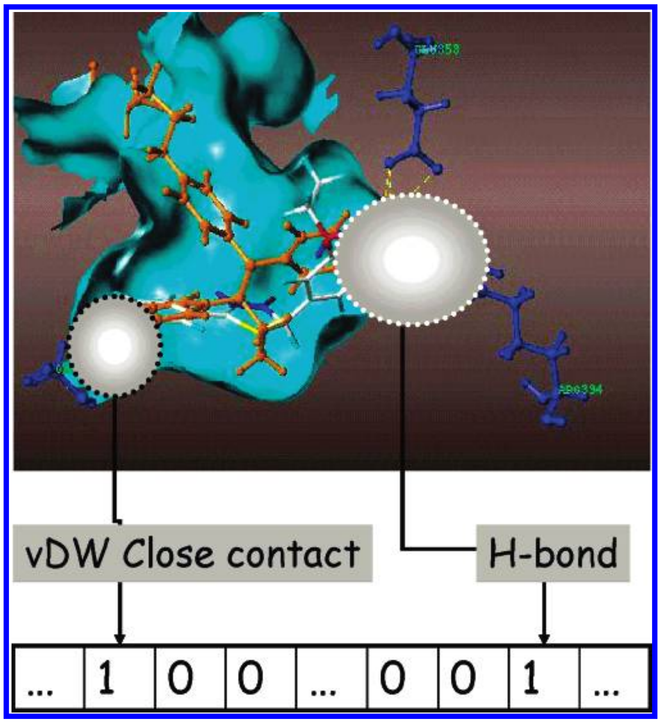

Figure 2. Fingerprint generated based on binding modes or pairwise interactions (H-bonds and vdW) formed between the docking proposed ligand conformation and a receptor. The length of the IF is equal to the number of heavy atoms in the binding site.

This simple technique was used to encode the binding modes of all the conformations generated for the reference ligands and all the docked database compounds. Each pose from the ensemble of GOLD-predicted solutions for the reference ligand was then compared to its relevant X-ray structure and the best root-mean-square deviation (RMSD) pose used to generate the reference fingerprint. Once identified, the best reference compound pose was then used for similarity-based scoring of the rest of the compounds in the docked compound library.

2.6. Single Reference Knowledge-Based Scoring. One of the most important aspects of this work is that it enables the development of new scoring functions that incorporate knowledge derived from X-ray complexes. We believe that our new method will avoid the tendency of current scoring functions to favor high molecular weight and complex compounds and, consequently, will rank compounds highly for the correct reasons.[9] Single reference scoring involves a very straightforward method of comparing each docking solution IF to the reference IF(s) using well-known similarity coefficients for comparing binary fingerprints. Assuming that active compounds will adopt highly similar IFs, then one is able to separate the actives from the inactives.

We refer to the first series of scoring methods as the binding mode similarity-based scoring functions (BMSs). This is simply the use of the values obtained from the similarity coefficient to rank all docked compounds. This allows us to avoid the use of the "traditional" energy estimation-based fast scoring functions. In this study, we employed the following simple similarity coefficients:[21] the Tanimoto coefficient, Euclidean distance, and simple matching coefficient. The similarity values can be calculated using the equations given below (eqs 1−3), where A and B denote the numbers of bits set in the two IFs that are being compared (X_A and X_B) and C denotes the number of bits in common (eq 4).

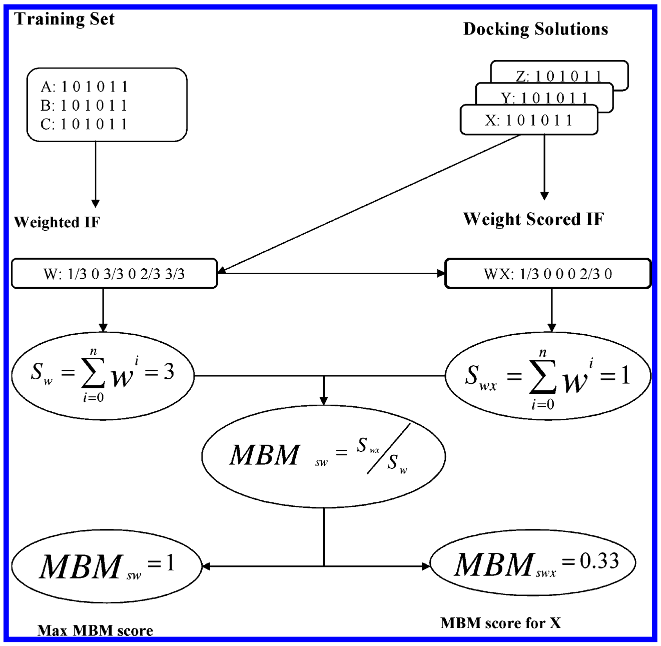

Figure 3. Simple procedure for computing the multiple binding mode interaction fingerprint score (MBM).

$$\text{Tanimoto coefficient} = \frac{C}{A + B - C} \qquad (1)$$

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

$$\text{Simple matching} = C \qquad (3)$$

where

$$A = \sum_{j=1}^{n} X_{jA},\ B = \sum_{j=1}^{n} X_{jB},\ \text{and } C = \sum_{j=1}^{n} X_{jA}X_{jB} \qquad (4)$$

2.7. Binding Knowledge Modified or Biased GOLD Scoring. A second series of single reference scoring methods attempts not to avoid fast scoring completely but to include it in such a manner as to penalize those compounds which cannot assume the reference structure binding mode as represented by the IF. To achieve this, we multiplied the value obtained from the GOLD fitness score, for each solution, with the BMS value to give us a new score referred to as the binding mode biased (BMB) score. Thus, a greater score is obtained for any compound that can assume a similar binding mode to that of the X-ray or reference compound. The new scores are calculated as shown in eq 5. Notice that the general form of this equation is modified (eq 6) to take into account the nature of the Euclidean distance value, which increases with dissimilarity [by multiplying the Goldscore (Gold fitness score] with $n -$ BMS, where n is the number of bits in the IF.

$$\text{BMB}_{\text{similarity}} = \text{Goldscore} \times \text{BMS} \qquad (5)$$

$$\text{BMB}_{\text{Euclidean}} = \text{Goldscore} \times (n - \text{BMS}) \qquad (6)$$

Along with the BMB series of scoring functions, we also implemented a method to filter compounds that could not adopt the expected binding mode(s) after docking; this is

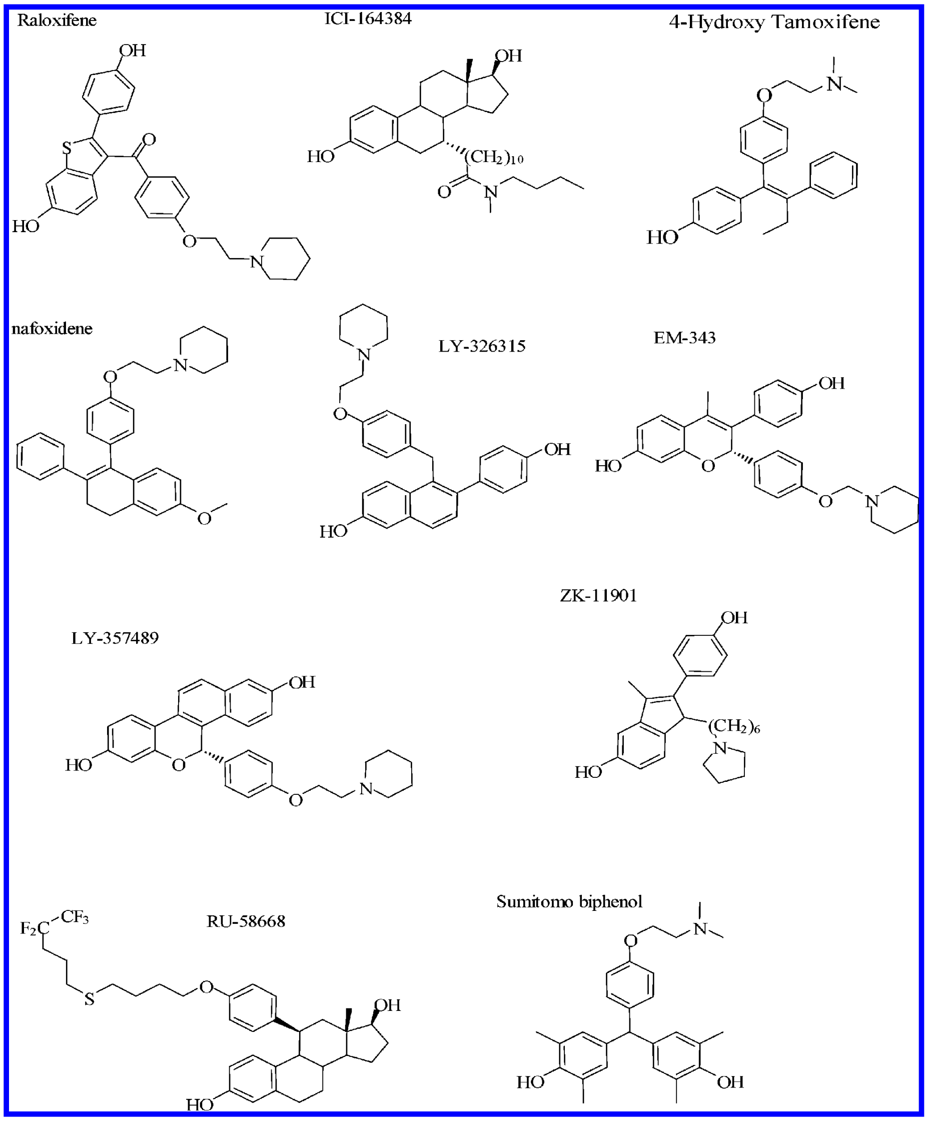

Figure 4. ER-receptor α antagonist set seeded into a set of 490 randomly selected druglike compounds to form the docking library. 4-Hydoxy tamoxifene is the reference structure co-crystallized to the docking site (PDB entry 3ert).

referred to as binding mode filtering (BMF). Using the simple matching coefficient, we assigned a BMF score of 1 to compounds that had BMS values greater than or equal to 1, and the BMF score was 0 otherwise. This filter score was multiplied by the GOLD fitness value, thus effectively filtering from the ranked database all of the compounds that could not form the desired interactions (i.e., binding mode filtering).

2.8. Multiple Reference Scoring. Multiple binding mode (MBM) scoring is an extension of the binding mode scoring methods described in the previous section. Here, we attempt to capture and score our compounds on the basis of the knowledge stored in not just one reference structure but in as many X-ray structures as are available. It is known that scoring functions often vary in their ability to separate active from inactive compounds across different chemical series,[7] and a variety of methods have been suggested to overcome this problem (one logical approach is to use multiple docking runs for each X-ray structure). To this end, our method presents us with a novel and easy system of incorporating all available X-ray data into a single novel scoring scheme. The step-by-step computation of the scoring scheme developed to incorporate multiple binding modes is illustrated in Figure 3. First, a multiple binding mode IF or a frequency-weighted fingerprint is generated for all the co-crystallized reference structures (training set). Then, for each docked compound, an IF is produced and compared to the weighted fingerprint by multiplying all the coded interactions with the frequency-based weights allocated for each bit position. In this way, an MBM score can be obtained and used to rank the docked libraries. As highlighted in the previous section, this score can also be extended to calculate a modified GOLD fitness score by the substitution of BMS with MBM in eq 5 to obtain a similarity-biased score.

2.9. *k*-Means Clustering and PCA for the Analysis of the Interaction Fingerprints. *k*-Means clustering and PCA were used for cluster analysis and automated pose visualization of the docking results to illustrate how useful the interaction fingerprint method can be for the researcher.

Cluster analysis involves the automated grouping of comparable objects (in this case, molecular binding modes as represented by the CHIF fingerprint). However, cluster analysis is not only for grouping objects on the basis of intragroup similarity but is also useful for understanding intergroup similarities. For simplicity, we elected to use a nonhierarchical *k*-Means-based clustering algorithm for this

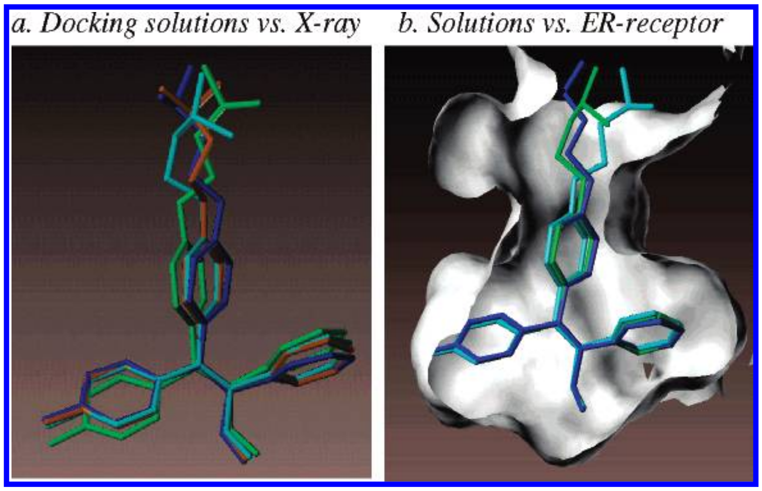

Figure 5. ER-receptor α-antagonist 4HT docking poses. The X-ray structure (orange) compared to the GOLD-returned ensemble. Poses 3 (dark blue), 2 (cyan), and 1 (green) with RMSD values of 0.42, 0.69, and 0.76 and Gold fitness scores of 67.43, 69.69, and 68.95, respectively.

study.[13,16] We estimated the required number of clusters (c_{required}) by using a simplified coarse optimization method. First, this involved rapidly clustering the fingerprint set 10 times using the estimated number of clusters $c_{\mathrm{estimated}}$ where $10 \leq c_{\mathrm{estimated}} \leq 100$ and $c/10 = N$. Second, this was followed by selecting c_{required} manually such that the clusters obtained contained as few singletons as possible. The squared Euclidean distance was used as the measure of dissimilarity between the IFs for the cluster analysis. We realize that the manual selection of c_{required} can prejudice the final outcome, so we also used PCA to verify the results obtained.

A simple nonmathematical working definition and treatment of PCA is available in a tutorial by Smith.[31] In this case, PCA will be treated as simply a method for the automatic identification of patterns or relationships in data with high dimensionality. The advantage of PCA is that, once the major contributing components (principal components) have been found, then the data can be compressed from the higher dimensional spaces into lower dimensions without the loss of information. In our case, each object or binding mode is represented by a binary vector of length equal to the number of potential receptor interaction points (i.e., the number of heavy atoms within the binding site). This becomes a multidimensional problem because each potential interacting atom forms a variable x, where $x \in (0, 1)$ and the number of dimensions are equal to the length of the bit string. The PCA studies were carried out using Spotfire.[32]

3. RESULTS AND DISCUSSION

The validation of the new, knowledge-based scoring schemes was carried out using data from a well-known docking case study.[2] The ER receptor was used as the docking site for a 500-compound library seeded with 10 known antagonists (the SOBAX set). We must also highlight here that two other libraries were also used for the docking study, and the results obtained showed trends similar to those obtained using the SOBAX set. This was satisfactory as it illustrated that the enrichments obtained for the SOBAX database are not attributable to the specific difference between the background molecules and actives.

3.1. ER Receptor. As noted by Bissantz et al., the ER receptor provides us with a suitable, well-studied target for docking, and it is, thus, a good test for our new methodology: high-resolution X-ray co-crystallized structures are available in the PDB, and there are 10 known, high-affinity active α antagonists (Figure 4), thus fulfilling the basic requirements for a retrospective docking and scoring study. The binding site was constructed from the ER receptor co-crystallized to 4HT (PDB entry 3ert). To analyze the effect of using more than one reference structure for scoring (MBM scoring), another X-ray structure was obtained from the PDB (entry 1ERR), which is the ER receptor co-crystallized with raloxifene.

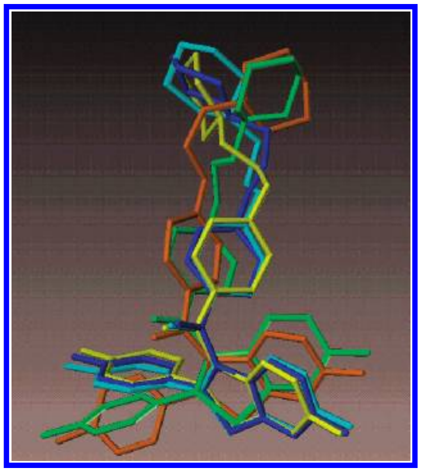

Figure 6. ER-receptor α-antagonist raloxifene docking poses. The X-ray structure (orange) compared to the GOLD-returned ensemble. Poses 1 (green), 2 (cyan), 3 (blue), and 4 (yellow) with RMSD values of 1.59, 2.02, 2.08, and 2.13 and Gold fitness scores of 61.58, 69.17, 65.13, and 66.76, respectively.

As discussed in the Bissantz study, GOLD docking is very effective at returning solutions (conformations or poses) close (in terms of the RMSD) to the X-ray conformation for this target. Our results show that each of the top three poses returned for the reference structure (4HT) had very reasonable RMSD values (0.42, 0.69, and 0.76, when compared to the X-ray structure) and GOLD fitness scores of 67.43, 69.69, and 68.95, respectively. The overlay of the ensemble returned is shown along with the MOLCAD surface of the binding site in Figure 5. It is also encouraging to note that even when we docked raloxifene into the (ERT site) surrogate site we

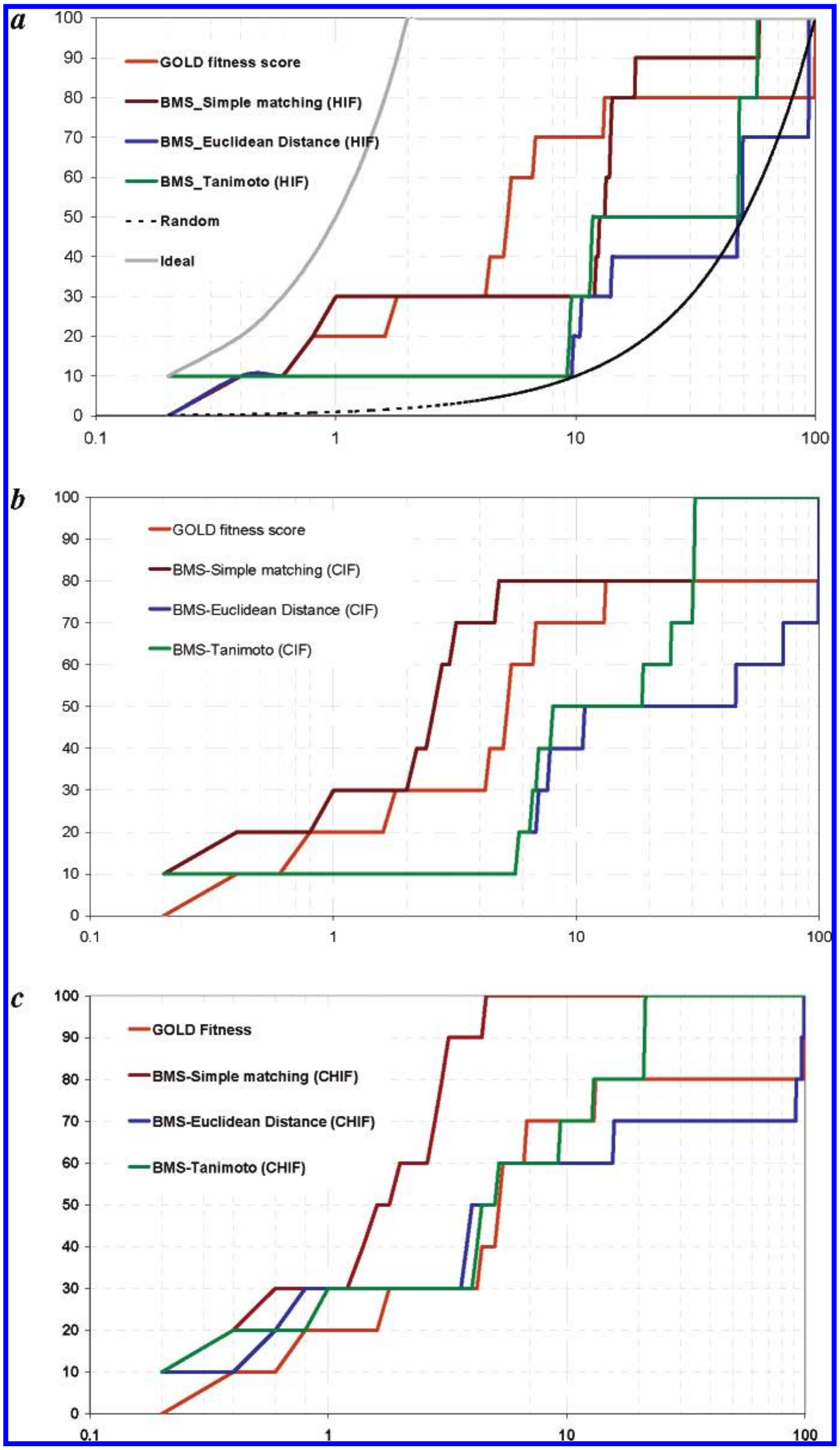

Figure 7. Percentage rate of recall of active ER antagonists plotted as a function of the percentage-ranked database for all new binding-mode-similarity-based scoring functions compared to Gold fitness score. The graphs a, b, and c correspond to the three fingerprint types HIF, CIF, and CHIF, respectively. The *x* axis is the logarithmic scale of the percent of the database screened plotted against the percent recovery of known active compounds.

obtained impressive results. The RMSD values for the four returned conformations were 1.59, 2.02, 2.08, and 2.13 (we regard RMSD values less than 2.50 to be of acceptable quality), with GOLD fitness scores of 61.58, 69.17, 65.13, and 66.76, respectively (Figure 6).

The best GOLD fitness score values do not always coincide with the best RMSD values, see the scores achieved by the "best poses" for both 4HT and raloxifene, Figures 1 and 2. This presents a fundamental problem for conventional scoring methods because, for the most part, the selection of the "best" pose for each molecule is based on ranking the docking solutions using the fast scoring functions, and poorer conformations sometimes achieve quite high scores. A subtle advantage of the IF methodology is that scoring is biased toward the known binding mode and therefore avoids this problem. This means each “pose” scored highly using the IF similarity-based method should be preferred, even if its energy score is lower than its “ensemble siblings” from the results of a docking study.

Once the best RMSD conformation for the reference structure(s) has been determined, the IFs (HIF, CIF, and CHIF) are generated for the pose. We chose to use the returned pose with the best RMSD value to generate the reference IF, since this pose represents the best solution that the docking engine would have been able to reproduce and, hence, the most reasonable starting point for comparison with other virtually predicted poses. An innovative user may also choose to use knowledge gained from other experiments such as site-directed mutagenesis or intuition to edit the reference fingerprint.

The effectiveness of the scoring methods was assessed using cumulative recall (recovery) plots for purposes of analysis, visualization, quantification, and comparison.[7] We define recall (R) as the ratio of recovered actives or hits to the total number active compounds in the database; R was multiplied by 100 and plotted as a function of the screened percentage of the database, to yield the cumulative recall plot for each scoring function. In practice, only the top 1−5% of a ranked database is of interest since in silico screening is primarily used to identify small sets (numbering in their tens to thousands from large databases) of compounds for further in vitro screening. The short list is usually obtained from the virtual screening of in-house compound libraries (typically numbering in the hundreds of thousands or millions of compounds) or external “druglike” libraries. We used the logarithmic scale for the independent variable axis or x axis (% database) to focus attention on the top-ranked fractions of the database (Figure 7).

3.2. Single Reference Similarity-Based Scoring. *3.2.1. Binding Mode Similarity-Based Scoring Functions (BMS).* The resulting recall plots for this set of scoring schemes are shown in Figure 7, where it will be seen that the CIFs (Figure 7b) perform far better than do the HIFs (Figure 7a). This may be attributable to the fact that the ER-receptor ligand binding is primarily driven by hydrophobic interactions (and therefore this may be different for other protein targets). The figure also shows that the CHIF, an amalgamated form of the two fingerprints types, seems to exhibit the best recovery rates (Figure 7c). Indeed, the CHIF BMS results (simple matching, Tanimoto coefficient, and Euclidean distance) are significantly better than conventional GOLD fitness scoring. At 5% of the ranked database, we see recovery values of 100%, 50%, and 50%, respectively, compared to 40% for the GOLD score (Table 1). A key advantage of the BMS scoring method is that it successfully and completely avoids the need to use fast scoring functions.

A further advantage observed for simple matching and Tanimoto coefficient BMS scoring is that the recovery of the full set (minus the reference molecule) of αER antagonists is achieved at much higher rank positions (4.8% and 21.6%, respectively) than with the GOLD fitness score. CHIF, hence, appears to provide a successful way of effectively encoding X-ray crystallographic knowledge and, thus, forms a viable alternative to other reported “HIF-like” methods that are based mainly on the hydrogen-bond interactions.[1,33] We propose that the inclusion of just the vdW close contacts as defined in our methods may improve on such approaches.

Table 1. ER-Receptor Antagonists Recall for the Top 5% Ranked for All Scoring Methods

scoring scheme	HIF (%)	CIF (%)	CHIF (%)
Gold fitness score	40	40	40
BMS_Simple matching	30	50	80
BMS_Euclidean distance	10	10	50
BMS_Tanimoto coefficient	10	10	90
BMB_Simple Matching (weighted filter)	50	80	80
BMB_Euclidean distance	30	50	60
BMB_Tanimoto coefficient	20	10	70
BMS_filtered GOLD score	70	80	70
MBM_wScore			100
MBM_wGOLD_score			80

3.2.2. Binding Mode Similarity-Biased GOLD Scoring Functions (BMB). As previously discussed, we also explored the effect of combining the interaction knowledge with an existing fast scoring function, by multiplying the value obtained from the GOLD fitness score by the various BMS-based scores (eqs 5 and 6). The scoring methods derived in this way can be separated into two main sets: the binding mode biased scores (BMB scores) shown in Figure 8 and the BMB filtered scores shown in Figure 9, where the latter represent an attempt to penalize compounds on the basis of the absence of reference binding mode interactions.

All the BMB scoring methods (simple matching, Tanimoto coefficient, and Euclidean distance) exhibit significant improvements on the conventional GOLD fitness scores. At a 5% fraction of the ranked database, the new methods have recall figures of 80, 70, and 60%, respectively, compared with 40% for GOLD. The results for the HIFs and CIFs are included here for completeness since the discussion will focus on the CHIF results. The BMB IF-based scoring methods present a new, practical, and effective way of enhancing the performance of traditional scoring functions. Filtered scoring also appears to have a similar effect, since 70% of the active compounds are recovered in the top 5% of the ranked database (Figures 8 and 9).

3.2.3. Multiple Reference Binding Mode Knowledge-Based Scoring (MBM). We can, as discussed previously, extend the scoring method to include not just a single reference interaction pattern (derived from one X-ray compound) but also multiple interaction modes, by using a weighted fingerprint method. For the ER receptor, we used two available X-ray co-crystallized structures (4-hydroxy tamoxifene and raloxifene). As we know, the docking of these two compounds was very successful with RMSD values for the best GOLD poses of 0.42 and 1.59, respectively. These best poses were used to generate CHIFs, and a weighted fingerprint was computed from these two CHIFs as detailed in the methodology section.

We recognize that more sophisticated methods have been employed by other researchers for the multiple reference scoring of ligand-based fingerprints.[22] However, we consider our simple multiple binding mode weighted fingerprint score (MBM_wScore) sufficient to demonstrate the effectiveness of the method (Figure 3). The frequency-based weights generated for the weighted fingerprint used in MBM scoring are based only on the positive interactions (bits in the fingerprint) and ignore negative interactions: this is important

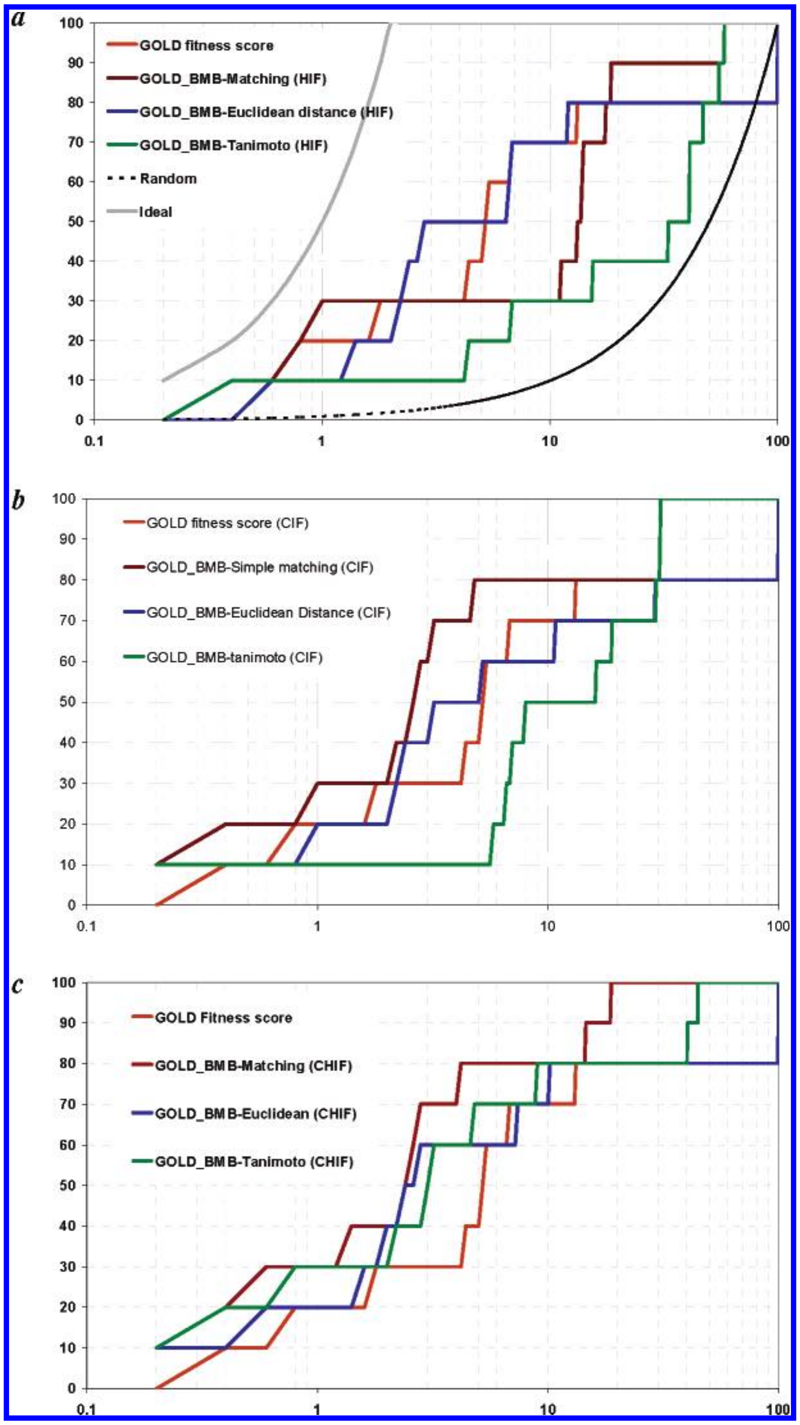

Figure 8. Recovery plots for the GOLD interaction knowledge biased scoring methods (BMB scores). The graphs a, b, and c correspond to the three fingerprint types HIF, CIF, and CHIF, respectively. The *x* axis is the logarithmic scale of the percent of the database screened plotted against the percent recovery of known active compounds.

because by their very nature X-ray structures do not capture negative interactions.

The results obtained are very encouraging (Figure 10), with all of the actives being recovered in the top 3% of the screened database using the MBM_wScore. When this score is incorporated into the fast scoring function (MBM_wGOLD_Score), better results are attained than with the GOLD fitness score, with 80% of the actives being recovered by MBM_wGOLD in the top 5% of the screened database. Weighted scoring recovers all actives much earlier in the ranking than does GOLD, suggesting that for those compounds for which GOLD fails MBM scoring is relatively more successful; for example, using MBM, we recovered 100% of the compounds at 27% of the ranked database (Figure 10).

One of the reasons why the ER receptor forms a successful site for docking is that only minor conformational changes are observed in the X-ray structures when different ligands are bound.[2,27,28] A quick macroconformational analysis reveals very small changes between the 4HT- and raloxifene-bound (3ERT and 1ERR) protein structures for all residues lining the binding pocket (weighted RMSD values of 1.44, 1.53, and 1.94 for Cα, backbone, and non-hydrogen atoms). Differences are observed in the microenvironment of the

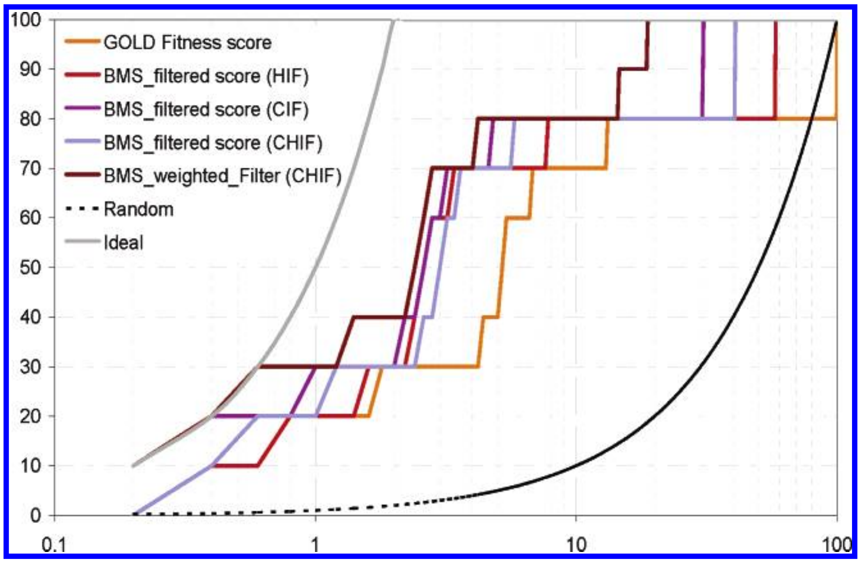

Figure 9. Recovery plots for the GOLD interaction filtered scores (BMS_Filtered scores) compared to the Gold fitness score. The *x* axis is the logarithmic scale of the percent of the database screened plotted against the percent recovery of known active compounds.

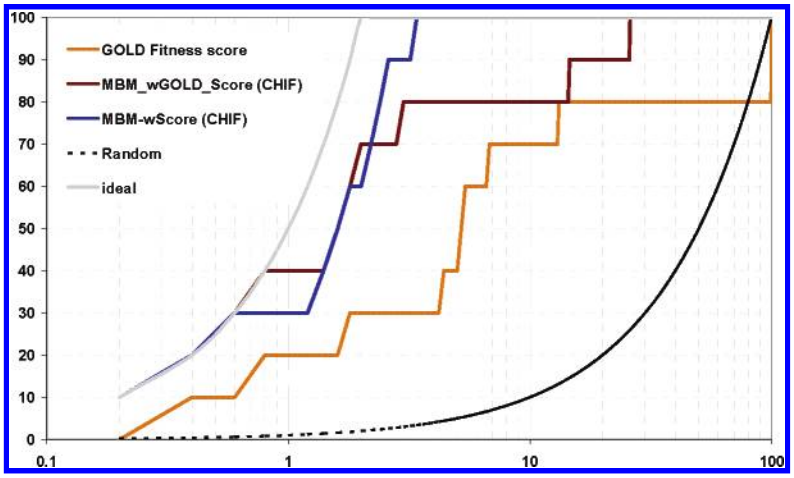

Figure 10. Recovery plots for multiple reference-weighted interaction scores (MBM_wScore). The *x* axis is the logarithmic scale of the percent of the database screened plotted against the percent recovery of known active compounds.

binding sites, but these are minor. Current docking algorithms tend to fail for targets that show large conformational changes for different ligands, as a consequence of the "rigid receptor" limitation.

MBM scoring provides a way of using multiple reference structures but does suffer from some possible limitations, in particular, a predisposition to miss novel binding modes especially when the method is confronted by instances where no new information is gained from considering additional X-ray structures. This is a problem common to pharmacophore-based methods. However, because we have decoupled docking from the scoring, our technique allows a user to filter out compounds that rank highly because of fingerprint similarity before focusing on the remaining, potentially more interesting modes. In addition, our approach to scoring can be useful when extending docking, from its usual use in lead discovery (virtual screening of large, heterogeneous databases), to the more upstream utility of lead optimization (virtual screening of smaller, more focused libraries), where the focus will be on docking-focused libraries in pursuit of compounds which can adopt predetermined binding modes after docking.

Another limitation is that, although the interactions identified may have a high frequency due to their predominance in the available set of X-ray structures, this may not be related to the energy contributions of such interactions. It may be interesting, in the future, to investigate how the inclusion of the energy contributions will influence the results of MBM IF scoring. However, as discussed elsewhere, apportioning energy contributions is not trivial, and the results thus far suggest that our new method offers a relatively economical and effective way of scoring without the need for energy apportioning.

Although new, our methods are analogous to those employed by Verdonk et al.,[9] who achieved similar results using pharmacophore filtering to screen for low-affinity and -molecular-weight compounds for CDK2, and mirror efforts by Fradera et al.,[10] who successfully incorporated a pharmacophore and ligand similarity measure to direct their docking using DOCK.[34] Other related work includes the knowledge-based scoring function developed by Feher et al.[8] and the encoding of pharmacophoric points into "pharmacophore multiplets" or fingerprints[35] (which provides an implicit encoding of binding modes). However, when

Table 2. Structural Analysis of the Top 10% Ranked Structures Returned by the Scoring Functions

	GOLD	BMS_Euc	BMS_Tan	BMS_Mat	random
Distribution of Rotatable Bonds (Top 10% Rank)					
min per structure	2	1	1	1	0
max per structure	11	9	9	11	15
mean per structure	6.55	5.18	4.84	5.40	5.25
std deviation from mean	1.81	1.95	1.92	2.25	2.0
Distribution of Heavy Atoms (Top 10% Set)					
min per structure	22	14	19	14	11
max per structure	42	34	37	34	42
mean per structure	28.21	24.34	28.46	25.13	26.16
std deviation from mean	3.62	4.32	4.94	4.46	4.74
Distribution of Heteroatoms (Top 10% Set)					
min per structure	3	2	2	2	1
max per structure	10	10	10	11	11
mean per structure	6.61	5.94	5.83	6.20	5.91
deviation from mean	1.71	1.72	1.73	1.85	1.81

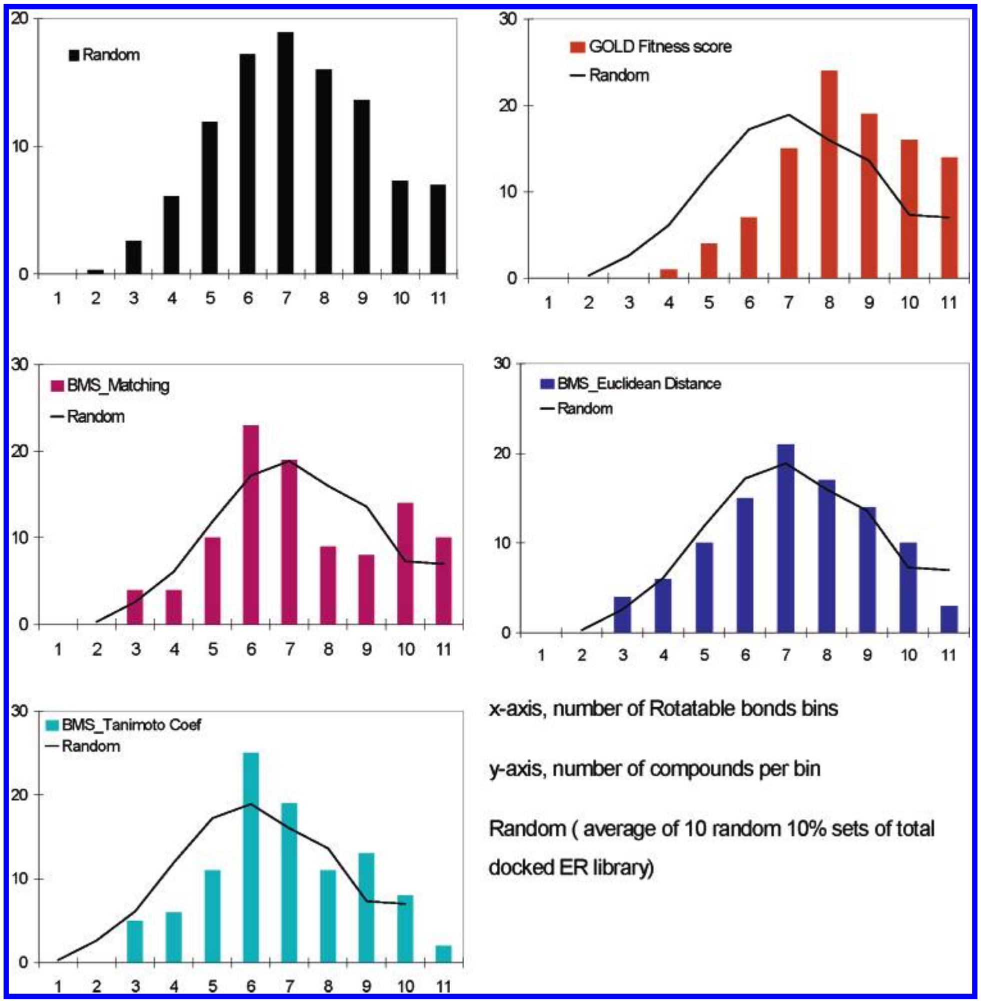

Figure 11. Distribution of rotatable bonds in the top 10% ranked sets for the scoring functions compared to randomly selected sets.

compared to all of these, our method provides an easy-to-implement, -interpret, and -adapt (transferable to any docking program) and less computationally demanding generally applicable post-docking scoring tool.

It is important that any new method should be tested on a variety of targets in order to establish its robustness. Current work involves more difficult protein targets (PDE4, a data set introduced by Mpamhanga et al.,[7] and the Bissantz thymidine kinase set), and preliminary results show very encouraging trends for these targets (to be published). However, it would not be surprising if poorer results are obtained for less “dockable” targets, that is, targets with shallow, less well-defined binding pockets, since the scoring method is highly dependent on the initial docking success of the reference molecules.

3.3. Structural Analysis of Top Ranking Compounds. Besides the common problems of fast scoring methods that have been highlighted previously, it is also well-known that scoring functions are inclined to favor molecules that are highly complex and have large molecular weights.[9] A structural analysis of the top 10% ranked molecules (actives as well as the decoys) for each scoring function was

Table 3. Structural Self-Similarity Analysis of the Top 10% Ranked Structures Returned by Scoring Functions

scoring function	mean self "Tanimoto" similarity	std deviation of the mean
GOLD	0.57	0.12
BMS_Euc	0.53	0.11
BMS_Tan	0.54	0.14
BMS_Mat	0.59	0.13
random	0.68	0.13

performed in order to demonstrate that the new scoring methods (BMS) can avoid this problem. The numbers of rotatable bonds (a measure of flexibility), heteroatoms (a measure of complexity), and heavy atoms (the molecular weight measure) were determined for the top-ranking sets, and a comparison was made of the differences between the top 10% recovered by the GOLD fitness score and by each of the BMS scoring functions.

The results showed that, on average, the number of rotatable bonds per structure is higher (6.6 per structure) for the GOLD fitness score than for the whole docked library (5.3). Conversely, the new scoring functions all show averages that are not greater than those of the original library (Table 2). The histograms in Figure 11 show that not only is the average number of rotatable bonds lower for the new scoring function but also more compounds for the GOLD fitness score have more than five rotatable bonds (73% of the top 10% scoring structures have greater than five rotatable bonds). In contrast to this, the new scoring methods show lower proportions of recovered molecules with numbers of rotatable bonds greater than five, specifically, only 44%, 41%, and 34%, respectively, for Euclidean distance, simple matching, and Tanimoto coefficient BMS scores. For comparison, only 44% of the compounds in the parent library have more than five rotatable bonds.

The results in Table 2 are encouraging, since it would appear that the new methods favor smaller, less flexible compounds. The average number of heavy atoms for molecules in the top 10% GOLD-fitness-score-recovered sets is higher than for the new methods; thus, the minimum number of heavy atoms per structure is significantly higher, 22 for GOLD, when compared to those of all the new scoring methods (14, 19, and 14 for Euclidean distance, Tanimoto coefficient, and simple matching BMS scores, respectively). Finally, there are no significant variations between the average numbers of heteroatoms for the recovered sets compared to the whole database: hence, the average molecule recovered by the scoring is not more complex than a randomly selected one.

Last, to fulfill one of the key requirements of any virtual screening protocol, we measured the diversity of all of the compounds retrieved in the top 10% of the ranking, since a good virtual screening method will not only retrieve many active molecules but will also retrieve sets of compounds that have a high diversity (or low "self-similarity"). This is particularly useful if the method is to be used for prospective work where the focus of the project will be to shortlist small sets containing novel compounds or chemical series to enable scaffold hopping. To evaluate this, we calculated the Unity fingerprint-based mean Tanimoto self-similarity index for the top-ranked 10% for each new scoring function and for the GOLD fitness score (see Table 3). The results we obtained were very promising, as the mean self-similarity values for the new scoring functions were all below 0.6 and as there is no significant difference between the Tanimoto similarity values using the new scoring functions and using the GOLD fitness function. The new methods, hence, retain this most desirable aspect of docking and scoring.

3.4. Visualizing Binding Modes (*k*-Means Cluster Analysis and PCA). The workflow scheme in Figure 1 allows our method to be applied for the automated visualization of putative binding modes using cluster analysis. After clustering the CHIFs, one could trace the lead clusters [clusters where the known actives or reference compound(s) fall]: this will enable the straightforward shortlisting of

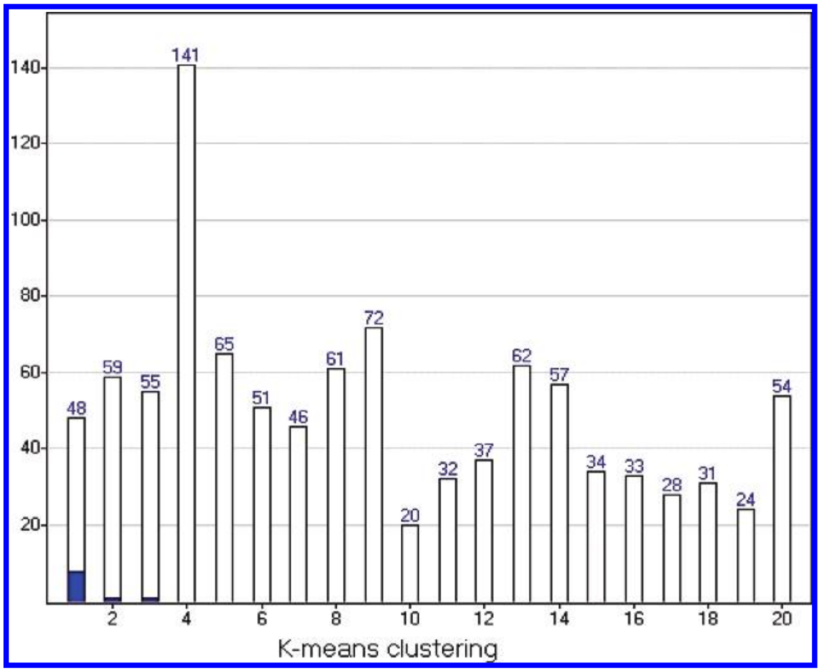

Figure 12. The *k*-Means cluster analysis of the interaction fingerprints (CHIFs) showing the partition of active compounds (in blue) into specific clusters; the *x* axis presents the cluster identifiers and the *y* axis the number of interaction fingerprints or poses clustered.

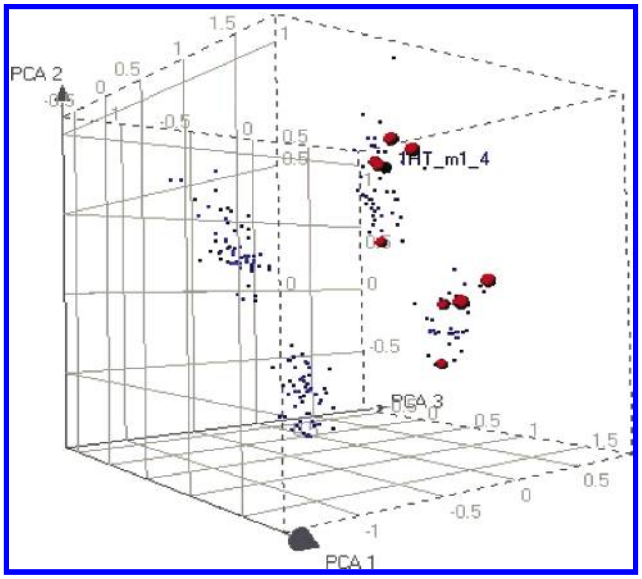

Figure 13. PCA of CHIFs showing the separation of active compounds' IFs into specific eigenvector spaces.

compounds and binding modes as it would allow the user to focus attention on just the lead clusters. This approach can be especially valuable for compound selection in large-scale docking projects, which can yield top-scoring sets containing tens of thousands of compounds, but manual visualization of individual binding modes can still be a problem even with more manageable numbers of compounds. The results (Figure 12) indicate that clustering can group all active compounds into just a few lead clusters. A PCA analysis was performed in order to verify this observation, resulting in a three-dimensional PCA plot showing the 10 active molecules restricted to only two regions of space (Figure 13). It is apparent that both PCA and cluster analysis can, if used properly, be useful for automated visualization and shortlisting of compounds based on their IFs.

4. CONCLUSION

This paper introduces an approach, for post-docking processing, which highlights the importance of knowledge-based scoring, and which shows that fast scoring does not need to be based purely on estimating the interaction energy but can draw on the knowledge implicit in X-ray co-crystallized structures.

The benefits of the approach have been demonstrated using the well-known Bissantz ER-receptor test set. The scores are generated from the similarity-based comparison of interaction fingerprints (binding modes) of putative docking poses to one or more reference structures. In this paper, we have used three frequently used similarity coefficients [Tanimoto coefficient, squared Euclidean distance (Euclidean distance), and simple matching]. We have proposed a simple and viable work flow which integrates the new scoring functions into a system that can be applied with any docking algorithm for virtual screening. The new methods demonstrate significant improvements when compared to the GOLD score, a typical and effective fast scoring function. It is also clear from our investigations that the CHIF outperforms its component fingerprint types, HIF and CIF, when used for scoring for this particular protein target.

The method can also be modified to form multiple reference scoring schemes, which attempt to integrate into one score the latent knowledge stored up in multiple X-ray structures. Multiple reference scoring appears to outperform all single reference scoring functions; even so, still better results might be obtained for targets with more X-ray data (Pde4b, on going work). We have also used the IFs to enable the quick shortlisting and automated visualization of putative binding modes using clustering and PCA.

In addition, we believe that IF scoring can enable researchers to apply docking tools for more upstream drug discovery research projects, where the aim would be to suggest new leads for old targets and even for lead optimization. It is encouraging to see that fingerprint scoring, though favoring only compounds that form predetermined binding modes, is still able to suggest highly diverse sets of potentially active hits, as indicated by the low Tanimoto self-similarity values obtained for the top-ranking sets studied here. Finally, structural analysis reveals that our methods appear to favor less-complicated and smaller molecules than does the GOLD fitness scoring.

ACKNOWLEDGMENT

Special gratitude goes to Dr. Val Gillet and Dr. Roger Mutter (Sheffield University) for their advice and proof reading. We are also grateful to Dr. Andrew Leach and his department at GSK for providing computational facilities. Last but not least our thanks go to Dr. Giampolo Bravi and Daniel Ormsby for help with the initial SPL implementation of the fingerprinting program. This work is funded by The Department of Health (U.K., DH007/0102).

REFERENCES AND NOTES

(1) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SiFT): A Novel Method for Analysing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(2) 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.

(3) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods On Systems Of Pharmaceutical Relevance. *Proteins* **2004**, *56*, 235−249.

(4) 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.

(5) Bohm, H.; Stahl, M. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Indiana University−Purdue University and Wiley-VCH: Hoboken, NJ, 2002; Vol. 18, Chapter 2, pp 41−87.

(6) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method For Obtaining Improved Hit Rates From Docking Databases Of Three-Dimensional Structures Into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(7) Mpamhanga, C. P.; Chen, B.; Mclay, I. M.; Ormsby, D.; Lindvall M. K. A Retrospective Docking Study Of PDE4B Ligands And An Analysis Of The Behaviour Of Selected Scoring Functions. *J. Chem. Inf. Model.* **2005**, *45*, 1061−1074.

(8) Feher, M.; Deretey, E.; Roy, S. BHB: A Simple Knowledge-Based Scoring Function to Improve the Efficiency of Database Screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1316−1327.

(9) Verdonk, M. L. B.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein−Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(10) Fradera, X.; Knegtel, R. M. A.; Mestres, J. Similarity-Driven Flexible Ligand Docking. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 623−636.

(11) Bernacki, K.; Kalyanaraman, C.; Jacobson, M. P. Virtual Ligand Screening against Escherichia coli Dihydrofolate Reductase: Improving Docking Enrichment Using Physics-Based Methods. *J. Biomol. Screening* [Online] **2005**, *10*, 675−681.

(12) Carhart, R. S.; Smith, D. H.; Venkataraghava, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(13) Willett, P. W.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.

(14) Fisanick, W. C.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.

(15) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.

(16) Downs, G. M. W.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(17) Brown, R. D. M. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(18) Kearsley, S. K. S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheriden, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(19) Thorner, D. A. W.; Willett, P.; Wright, P. M. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.

(20) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 559−614.

(21) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(22) Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(23) Daylight. http://www.daylight.com (accessed Sept 2005).

(24) Taylor, R.; Mullier, G. W.; Sexton, G. J. Automation of Conformational Analysis and other Molecular Modeling Calculations. *J. Mol. Graphics* **1992**, *10*, 152−160.

(25) Tripos. http://www.tripos.com (accessed Sept 2005).

(26) Jones, G. W.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(27) 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.

(28) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* **1997**, *389*, 753−758.

(29) Rognan dataset. www.bioinfo-pharma.u-strasbg.fr/download/random1000.mol/ (accessed Sept 2005).

(30) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(31) Smith, L. I. A Tutorial on Principal Component Analysis. www.cs.otago.ac.nz/student_tutorials/ (accessed Sept 2005).

(32) Spotfire. http://www.spotfire.com (accessed Sept 2005).

(33) 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.

(34) Ewing, T., Kuntz, I. D. Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening. *J. Comput. Chem.* **1997**, *18*, 1175−189.

(35) Mason, J. S.; Beno, B. R. Library Design Using BCUT Chemistry-Space Descriptors and Multiple Four-Point Pharmacophore Fingerprints: Simultaneous Optimization and Structure-Based Diversity. *J. Mol. Graphics Modell.* **2000**, *18*, 438−451.

CI050420D