Expanded Interaction Fingerprint Method for Analyzing Ligand Binding Modes in Docking and Structure-Based Drug Design

Matthew D. Kelly[‡] and Ricardo L. Mancera*[,§]

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K., and De Novo Pharmaceuticals, Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, U.K.

Received April 16, 2004

An expanded interaction fingerprint method has been developed for analyzing the binding modes of ligands in docking and structure-based design methods. Taking the basic premise of representing a ligand in terms of a binary string that denotes its interactions with a target protein, we have expanded the method to include additional interaction-specific information. By considering the hydrogen-bonding strength and/or accessibility of the hydrogen bonding groups within a binding site as well as their geometric arrangement we aim to provide a better representation of a ligand-protein interaction. These expanded methods have been applied to the postprocessing of binding poses generated in a docking study for 220 different proteins and to the analysis of ligands generated by an automated ligand-generation algorithm for the anthrax oedema factor. In the docking study, the application of the interaction fingerprint method as a postprocessing tool resulted in an increased success rate in identifying the crystallographic binding mode. In the analysis of the ligands generated for the anthrax oedema factor, the incorporation of additional interaction-specific information resulted in a more intuitive and comprehensive analysis of automated ligand-generation output.

1. INTRODUCTION

The application of in silico structure-based drug-design methods, such as ligand-protein docking,[1−4] virtual high-throughput screening,[5,6] and automated ligand generation,[7,8] to the drug discovery process invariably produces a large number of potential lead candidates. These prospective ligands need to be filtered in order to reduce their number for more precise and labor-intensive studies. Standard implementations of the above methods enlist a scoring function in order to rank the predicted ligand-protein interactions to aid the identification of those candidates possessing a greater chance of success in vitro.[1,9] However, these scoring functions alone are insufficient for the accurate computation of ligand-protein interaction energies and hence the unambiguous identification of true binding modes, most likely due to the approximations required in their calculation. The performance of these scoring functions can nonetheless be generally improved with the inclusion of postprocessing tools such as clustering based upon structural similarity.[10,11]

Clustering molecules based upon similarity requires some quantitative measure (descriptor) of the similarity between two molecules. The first measures of chemical similarity counted the number of common fragment substructures for a pair of molecules.[12,13] Fragment-based similarity searches have been refined and enhanced since these original methods were first published; for example, the subsimilarity searching method[14] is an example of a local similarity search method that identifies molecules containing a substructure similar to the target structure (or substructure). A more recent investigation reported that encoding molecular fragments as atom sequences, i.e., the sequence of all atoms in the shortest bond path between two atoms with the terminal atoms included, was a particularly effective method for identifying active compounds based on their similarity to a probe molecule.[15]

In addition to these common substructure approaches to similarity searching, other kinds of descriptors have been developed and implemented. These include physicochemical property descriptors,[16−19] topological indices,[20,21] and pharmacophore points.[22] A detailed and comprehensive review of the many descriptors applied to chemical similarity searching is provided by Willett.[23] Different descriptors better represent different intrinsic properties of a set of molecules. For example, the physicochemical property ClogP (partition coefficient) provides a useful representation of hydrophobic properties.[16,21] On the other hand, assigning pharmacophore points such as hydrogen-bond donors or acceptors onto the atoms of a ligand proves useful for comparing ligands whose binding interactions are largely determined by hydrogen bonding.[18]

Following a selection of the descriptor or combination of them to represent a ligand, the information must be encoded in a suitable manner. The most common approach is to use a fixed length binary string (bitstring).[23] In this representation, each bit can denote either the presence (1) or absence (0) of a particular discrete feature or property. Alternatively, in the case of continuous variables, a range of possible values may be divided into bins with each bin represented by a bit in the final bitstring. Following the generation of a set of bitstrings, each representing a different ligand, the similarity between them can be computed with one of the many available similarity indices. The relative advantages and disadvantages of these different methods have been the subject of a number of comprehensive reviews,[15,23−25] with

* Corresponding author e-mail: rlm1003@cam.ac.uk.
[‡] University of Cambridge.
[§] De Novo Pharmaceuticals.

10.1021/ci049870g CCC: $27.50
Published on Web 10/01/2004

the Tanimoto coefficient being by far the most popular method.

Traditionally, similarity methods have focused on representing a molecule based upon its own structural and chemical composition. However, Deng and co-workers[10] recently described a novel approach to representing the properties of a ligand. As opposed to calculating the properties of a ligand from the perspective of its own structural and chemical components, the Structural Interaction Fingerprint (SIFt) method represents a ligand by the interactions it forms in the binding site of a protein. While this requires knowledge of the binding mode of each of the ligands with a common protein, a basic implementation of the method has proven effective in the postprocessing of output from a docking study and as a filter during a virtual chemical library screening process. Using seven bits per binding-site residue to represent seven different types of interaction, the SIFt method encoded a ligand-protein interaction into a binary string. Clustering the resultant strings for a set of ligands for a particular protein allowed the ligands to be grouped into similar binding modes, aiding the identification of the crystallographically observed binding mode. The SIFt method also demonstrated a good database enrichment performance in a virtual library screen for p38 inhibitors, outperforming the scoring functions ChemScore[26] and PMF.[27]

With the objective of developing a method for postprocessing in silico docking and automated ligand generation data, we have created a method that, while sharing the basic premise of the above-mentioned SIFt method, extends it to investigate the effect of encoding more interaction-specific information into the bitstring. By representing the interactions at the atomic-level as opposed to the residue level and including measures of the strength of the interactions or their geometric grouping, we are able to better describe a ligand-protein interaction and produce a more informative analysis of docking and ligand generation data.

Classifying automated ligand generation output on the basis of their binding modes with a target protein would allow both the identification of ligands sharing similar binding modes to known active compounds and the filtering out of those ligands demonstrating unfavorable binding modes. This would increase the likelihood of identifying further active compounds. Furthermore, such an approach would be well suited for use in high-throughput virtual screening methods.

2. METHODS

Generation of Bitstrings. The observed interactions between ligand and protein were encoded into a binary string of fixed length. The number of bits is dependent on both the total number of interaction groups (sitepoints) present within the binding pocket of the target protein, and the representation method is employed. Our initial investigations have been limited to hydrogen-bonding interactions in order to analyze the effects of encoding various hydrogen bonding interaction-specific information. However, this method can be readily expanded to include hydrophobic and/or steric sitepoints to deliver a fuller description of all kinds of ligand-protein interactions.

The first step in the implementation of any of the representation methods examined was an analysis of the binding site of the protein. The extent of the binding site was determined by selecting all protein atoms within a 6.0 Å cutoff of the ligand in its crystallographic binding mode. The initial analysis of the binding site identified all potential interaction points (e.g. hydrogen-bonding sitepoints) enabling the determination of the total length of the binary string. In the first and most basic representation investigated, the length of the bitstring is equal to the total number of potential interaction points with each bit representing either the presence (1) or absence (0) of an interaction with that hydrogen-bonding sitepoint by a particular ligand. Therefore, for a given set of ligands that bind to a particular protein, an array of binary strings is generated (one for each ligand). In each of these strings, the position of the bit representing a particular hydrogen-bonding interaction is the same, allowing the strings to be directly compared with a similarity index such as the Tanimoto coefficient.

In addition to this simple representation, a number of additional methods were investigated. One such approach was the incorporation of a measure of importance of each of the potential hydrogen-bonding sitepoints in the binding site. We used the recently introduced strength-weighted accessible-probability score (SWAPS)[28] to determine the relative importance of the sitepoints within a binding pocket. For each hydrogen-bonding sitepoint, the SWAPS is calculated as the product of the potential strength of the hydrogen bond that may be formed and the solvent accessible probability score (SAPS). For the strength component, the relative density (d_{rel}) of the hydrogen bonding group, as calculated in the IsoStar database of the Cambridge Crystallographic Data Centre (CCDC),[29] was used. The d_{rel} value represents the propensity of a hydrogen-bonding group to form close, and therefore strong, contacts with a probe group (i.e. a hydrogen-bond donor or acceptor). The other component of SWAPS, the SAPS value, represents the hydrogen-bonding-specific accessibility of the sitepoint.[28] This is calculated by determining the accessibility of binned regions surrounding the atom. Each of the bins has an associated weight that represents the likelihood of observing a complementary hydrogen-bonding atom in that position. The greater the likelihood, the greater the weight is, and, therefore, the greater the contribution of that bin to the final SAPS value. The end result is a value that represents the accessibility of a sitepoint based upon the availability of those regions surrounding the atom that are more favorable for the positioning of a complementary hydrogen-bonding atom.

To include this SWAPS value in the binary representation, each sitepoint was represented not by a single bit but by a range of bits dependent on the value of SWAPS. For example, a sitepoint with the maximal SWAPS score of 4.9 (a fully accessible oxygen of a carboxylic acid group), would be represented by 10 bits, with either all the bits set to one if the sitepoint is satisfied by the ligand or all set to zero if not. However, a sitepoint with a submaximal SWAPS score would be represented by less than 10 bits, with the exact number directly proportional to its fraction of the maximum SWAPS value. The rationale behind this modification of the simple "one-bit-per-sitepoint" method is that if two ligands interact with a common high-scoring sitepoint (i.e. one that has the potential to form a strong interaction), then these ligands should be deemed to have a more similar binding mode than if instead they have a low-scoring sitepoint in

common (i.e. one that has the potential to form only weak interactions). In other words, the underlying assumption is that a strong hydrogen-bonding group is likely to provide a greater contribution to the overall binding energy and specificity of the ligand than a weak, less specific hydrogen-bonding group, i.e., stronger hydrogen-bonding groups are more likely to determine the binding mode of a ligand.

Another modification of the simple "one-bit-per-sitepoint" method was the inclusion of a simple measure for the actual strength of the hydrogen bond formed. For this purpose we have implemented the basic scoring function of Bohm[30] (Equation 1), in which the calculated strength of the hydrogen bond is dependent upon the deviation from "ideal" distance and angle criteria (which give a maximum hydrogen bond energy of −4.7 kJ/mol):

$$f(\Delta R,\Delta\alpha) = f1(\Delta R) f2(\Delta\alpha) \tag{1}$$

$$f1(\Delta R) = \begin{cases} 1 & \Delta R \le 0.2 \text{ Å} \\ 1 - (\Delta R - 0.2)/0.4 & \Delta R \le 0.6 \text{ Å} \\ 0 & \Delta R > 0.6 \text{ Å} \end{cases} \tag{2}$$

$$\Delta\alpha \le 30°$$

$$\Delta\alpha \le 80°$$

$$\Delta\alpha > 80°$$

$$f2(\Delta\alpha) = \begin{cases} 1 \\ 1 - (\Delta\alpha - 30)/50 \\ 0 \end{cases} \tag{3}$$

where ΔR is the deviation of the H···O/N hydrogen-bond length from its ideal value of 1.9 Å and $\Delta\alpha$ is the deviation of the hydrogen-bond angle $\angle_{\text{N/O–H}\cdots\text{O/N}}$ from the ideal value of 180°. This function tolerates small deviations of up to 0.2 Å and 30° from the ideal geometry in order to account for small uncertainties in the crystal structure. While this scoring function does not account for the preferences of certain functional groups to form angles different than 180° (i.e. carbonyl···H−N/O bond), it does provide a useful approximation of the strength of a hydrogen bond.

To encode this information into the binary string, each potential interaction point in the binding site was again represented by 10 bits, with the number of bits turned on for each satisfied interaction point proportional to the score resulting from the deviation of the hydrogen bond geometry from the ideal. The rationale behind this modification of the bitstring is that the formation of a near "ideal" hydrogen bond between a ligand and the protein provides a greater contribution to the interaction (and, thus, the binding mode) with the protein than the formation of a hydrogen bond on the limits of acceptable geometric criteria.

The final modification to the simple "one-bit-per-sitepoint" method was the inclusion of a sitepoint clustering method, where each string is divided into segments, with each segment representing a geometrically clustered group of sitepoints. The purpose of this modification was to assign a greater difference (i.e. reduce their similarity) to a pair of ligands that differ by the addition or relocation of an interaction to a previously unoccupied subregion (i.e. sitepoint cluster) of the pocket (see Figure 1). In other words, we aimed at reducing the effect of small changes in the binding mode of a ligand when "switching" between closely

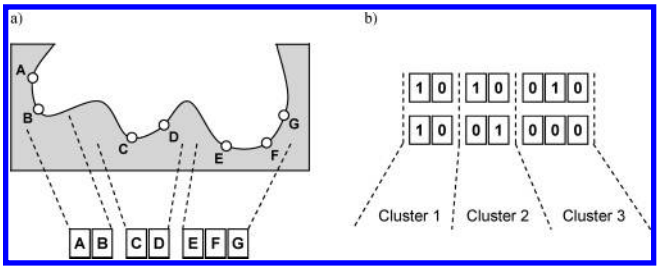

Figure 1. This schematic representation of the cluster-based similarity method demonstrates (a) the creation of a three-segmented bitstring representing three distinct clusters of sitepoints within the binding pocket. For the example bitstrings shown (b); cluster 1 shows no change, cluster 2 shows an intracluster change, and cluster 3 shows an intercluster change. For segments 1 and 2 the similarity is calculated as-per-normal, for segment 3, the similarity is calculated and then multiplied by a weighting factor to reduce the computed similarity.

neighboring alternative sitepoints, so that the associated binding modes are deemed more similar. Our modification achieves this in practice by increasing the relative contribution of *inter*cluster differences as compared to *intra*cluster variations. The geometric clustering of the sitepoints was performed using a hierarchical agglomerative method implementing average linkage.[31] The endpoint for the clustering was determined using the rule that the maximum distance between the centroid of a cluster and any of its members is less than half the average distance between all existing centroids.[32] Once the segmented strings are generated, the similarity between any two strings is calculated not by comparing the entire length of the two strings but by calculating a similarity value for each of the corresponding segments. The similarity was calculated by either the Tanimoto coefficient or the modified Tanimoto coefficient (further details below). To incorporate the increased effect of intercluster differences on the final distance metric relative to intracluster differences, the similarity calculated between any two segments representing an intercluster difference was multiplied by a weighting factor (between 0 and 1) to effectively reduce the calculated similarity. The mean of the similarities between each of the segments, following any necessary weight adjustments, was then subtracted from unity to produce the final distance measure.

Clustering of Bitstrings. In addition to investigating the effects of modifying the interaction properties encoded into the binary string, two different methods for measuring the similarity between two bitstrings were also compared. As stated earlier, the most commonly implemented measure of similarity between two binary strings is the Tanimoto coefficient (T_1)[33]

$$T_1 = \frac{c}{a + b - c} \tag{4}$$

where for any two binary strings **A** and **B**, a is the number of "on" bits in **A**, b is the number of bits on in **B**, and c is the number of bits on in both **A** and **B**. While this method has proved its usefulness in many applications, it has been shown to fall prey to often unwanted, size-dependent effects since it only considers the "on" bits in its computation.[24,33] To combat this drawback the modified Tanimoto coefficient[34] (*MT*) was developed

$$MT = \alpha T_1 + (1 - \alpha)T_o \quad (5)$$

where

$$\alpha = \frac{2 - p}{3} \quad (6)$$

and where

$$p = \frac{a + b}{2n} \quad (7)$$

where a and b are as for the Tanimoto coefficient and n is the total number of bits in **A** (or **B** as equal). T_0 represents the Tanimoto complement

$$T_0 = \frac{d}{a + b - 2c + d} \quad (8)$$

where a, b, and c are for the Tanimoto coefficient and d is the number of "off" bits in both **A** and **B**. By considering both "on" bits (through computing the Tanimoto coefficient, T_1 (Equation 4)) and "off" bits (through computing the Tanimoto complement coefficient, T_0 (Equation 8)) and combining the two, the size dependence of the similarity coefficient is significantly reduced.

Following the creation of a distance matrix for a given set of ligands using either the Tanimoto coefficient or the modified Tanimoto coefficient, the ligands were clustered with the popular hierarchical agglomerative minimum variance method of Ward.[35] The stopping criterion for the clustering was a Mojena probability of 0.05.[36] This probability represents the significance below which all new cluster fusions must differ from the set of all previous fusions.

Validation Set. The various implementations of the above interaction fingerprint method were evaluated on ligand-protein complexes from a series of docking experiments and molecules generated with an automated ligand generation method. For the docking method we made use of the validation data for the EasyDock algorithm, which implements a quantum stochastic tunneling optimization method.[37] The data set used was the "clean" subset of the CCDC/Astex data set.[38] This data set represents the largest populated structural classes of proteins, with each of the complexes having manually assigned protonation and tautomeric states. The clean subset comprises 220 ligand-protein complexes and is the result of removing those structures from the complete data set that contain either errors or inconsistencies in their PDB files, ligands with unlikely conformations, and complexes with severe clashes between protein and ligand atoms. For each member of the clean set, the cocrystallized ligand was repeatedly docked, and a large number of different binding modes (poses) were generated. Full details of this docking validation study have been published elsewhere.[37] The generated poses were subsequently analyzed and clustered using the different implementations of the interaction fingerprint method, and their ability to increase the likelihood of correctly identifying the crystallographic binding mode was calculated. Such analysis should be applicable to the results of any efficient docking method and/or scoring function.

Computer-aided de novo ligand design was carried out using the Skelgen algorithm[7] to generate 1000 ligands for the target protein. Skelgen can incrementally construct and/or modify a ligand in the binding site of a target protein using a Monte Carlo simulated annealing optimization algorithm. The program uses a set of common ring and acyclic fragments that are assembled together into a ligand structure following chemical rules. Ligand structures are modified through fragment additions, removals, and mutations as well as by molecular translations and rotations and conformational changes in torsional space. These modifications allow for previously incorporated fragments to be removed or replaced with different fragments, enabling the ligand to gradually satisfy the protein binding site constraints. This process is carried out in a stochastic manner to gradually optimize the interaction properties and chemical features of the generated ligand during the annealing optimization. The assembled ligands must satisfy user-defined geometric constraints, such as those defining hydrogen bond distances and angles for preselected donor and acceptor sitepoint targets and the steric constraints imposed by the structure of the binding site. Full details of this algorithm can be found elsewhere.[7,39,40]

The target protein for the Skelgen algorithm was the oedema factor (EF) protein of the anthrax virus. EF, an adenylate cyclase, is one of three component proteins of the toxin secreted by *Bacillus anthracis*. EF impairs a host's defense system through a number of mechanisms including the inhibition of phagocytosis.[41,42] The other two components of the toxin are the lethal factor, which causes lysis of macrophages through cleavage of the mitogen-activated protein kinase kinase,[43−45] and the protective antigen, which acts as a delivery system into the cytosol for the lethal and oedema factor.[46]

The extent of the binding site of EF defined for the Skelgen algorithm was determined from the predicted binding mode of a ligand known to bind with high affinity in the active site of the protein.[47] The binding mode was determined using the EasyDock algorithm to dock the ligand into the crystallographic binding site of the noncyclizable substrate analogue 3′dATP (PDB code: 1K90).[48] In the predicted binding mode (Figure 2), the quinazoline group of the ligand is superimposed onto the adenosine ring of 3′dATP, with amine substituents overlaid. This alignment results in the remaining groups of the ligand extending in the opposite direction to the triphosphate groups of 3′dATP, forming additional hydrogen bonds with the binding site. Based on the crystallographic binding mode of 3′dATP and the docked binding mode of the experimentally active ligand, key sitepoints were identified within the active site of EF. These formed the core target sitepoints for the Skelgen algorithm (Figure 2). The core sitepoints were divided into three groups for the purposes of the Skelgen algorithm. The first group (red spheres in Figure 2) contained the side chain OH and backbone NH and C═O of Thr 548, along with the backbone C═O of Thr 579 and Gly 578. The second group (blue spheres in Figure 2) contained an oxygen of the carboxylate groups of Asp 493, Glu 580, and Glu 588, the side chain guanidine NH of Arg 329, and the backbone C═O and NH of Leu 494. The third and final group (cyan spheres in Figure 2) contained the backbone C═O of Asp 582, the backbone NH of Glu 584, Glu 585, and Phe 586, along with the side chain amide C═O and NH of Asp 583. Each ligand generated by Skelgen was required to satisfy hydrogen-

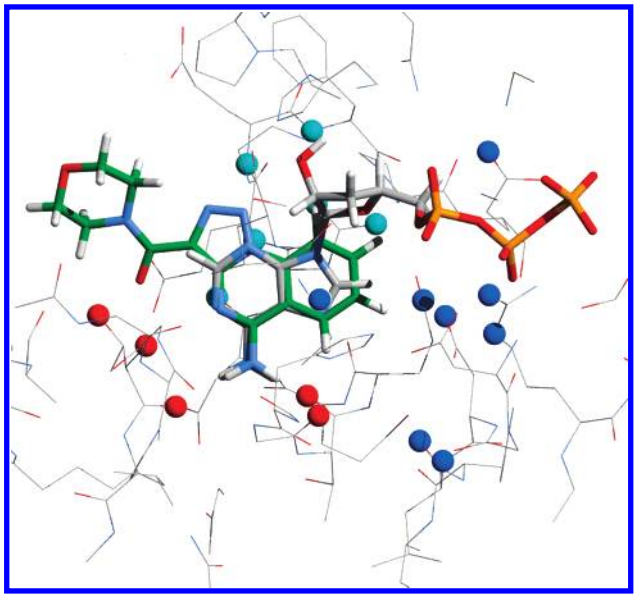

Figure 2. The ATP-binding site of EF, displaying the crystallographic binding mode of 3′dATP (carbon atoms colored in gray) and the docked binding mode of the experimentally active ligand (carbon atoms colored in green) (N.B. compound 3 in ref 47). The core sitepoints for the Skelgen algorithm are represented as spheres, divided by color into distinct subregions.

bonding interactions with a minimum number of members from each group. As the first group contained key sitepoints used by 3′dATP and the docked experimental ligand, each generated ligand had to establish hydrogen bonds with at least three of the six members. For the remaining two groups, one member of each had to establish hydrogen bonding interactions with the ligand.

The set of ligands generated by Skelgen were analyzed and clustered using the different implementations of the interaction-based clustering method. The effects of the different implementations on the clusters allocated were assessed manually in order to determine the relative merits of the different approaches for sorting the ligands into appropriate clusters according to their binding modes with the protein.

3. RESULTS AND DISCUSSION

Docking Output. The multiple binding modes generated for each of the proteins in the CCDC test set were clustered using the various implementations of the expanded interaction fingerprint method described above. For each method, the ability to increase the likelihood of identifying the experimental binding mode of the ligand was assessed. Each docking run generated a large number of potential binding modes, which were ranked by their calculated ScreenScore[9] value. For the control calculations, we calculated the percentage of proteins whose top-ranked docked solution had an RMSD of less than 2.0 Å from the crystallographic solution. Equivalent percentages were calculated if the correct crystallographic solution (within an RMSD of 2.0 Å) could be found when taking more of the top-ranked solutions up to a maximum of ten. The resultant values are shown in Figure 3.

In an attempt to improve on the values of the control method, we investigated the effects of postfiltering the docking output using the different implementations of the

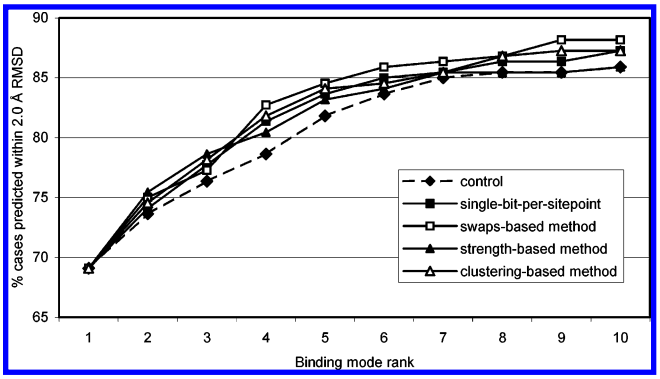

Figure 3. A comparison of the effects of using the various implementations of the interaction bitstring method to postprocess the docking-generated poses on the likelihood of identifying the crystallographic binding mode within the top 10 ranked solutions.

interaction fingerprint method. For each protein, all members of the ranked full list of docking poses were assigned to one of a set of clusters as calculated by each of the implementations of the interaction fingerprint method. For the control method, the percentages were calculated using the top 10 solutions as determined purely by their ScreenScore values, as stated above. However, for the fingerprint methods, the top 10 solutions were determined based upon both their ScreenScore and the clusters to which they belonged, as determined by each implementation of the interaction fingerprint method. The top ranking solution is again the highest as ranked by ScreenScore. The next best solution is the next highest ranked solution by ScreenScore that belongs to a different cluster to that of the top ranking solution. This process ensures that the second solution presents a significantly different binding interaction with the protein to that of the first solution. The third solution is calculated as the next highest ranked solution by ScreenScore that belongs to a different cluster to both of the previously determined higher-ranking solutions. This process is repeated successively to generate a total of 10 new top solutions.

By implementing this modified approach, one would expect to survey a greater range of binding modes than when just using the top 10 as ranked by ScreenScore as, for example, the two highest ranked solutions could occupy near-identical orientations in the pocket. By increasing the number of distinct poses sampled, one would expect a greater chance of identifying the crystallographic binding mode when considering a limited number of ligand poses. The consideration of multiple binding modes has recently been shown to improve the search for the correct binding mode in ligand-protein docking.[11] The results generated for different implementations of the interaction fingerprint method are shown in Figure 3. The similarity coefficient used to generate the data shown was the modified Tanimoto coefficient. While displaying only minor improvements over the standard Tanimoto index in the examples studied, it was selected due to its resilience to the size-dependent effects suffered by the standard Tanimoto coefficient.

As expected, the four different implementations of the fingerprint method all show an improvement over the control approach. For example, when postprocessing the docking output with the SWAPS-based interaction fingerprint method, the percentage of proteins having a docking solution with an RMSD of less than 2.0 Å away from the crystallographic

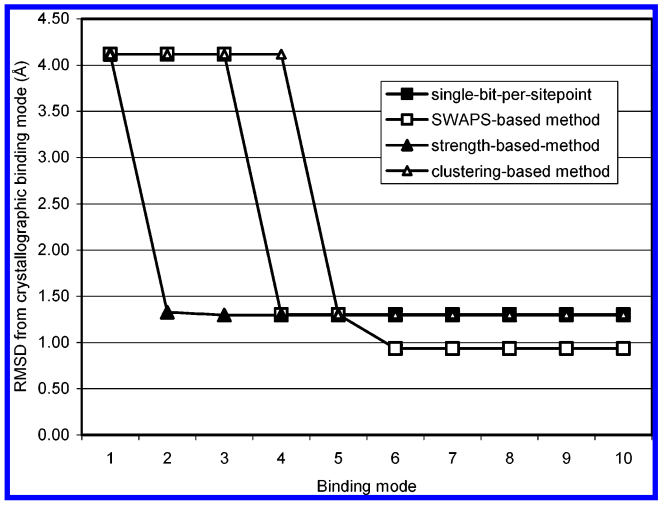

Figure 4. The individual analysis for HIV protease (PDB code: 4phv) as an example of the differing abilities of the methods in different cases. In this case, the hydrogen-bond strength-modified method is able to identify the crystallographic solution (within 2.0 Å) with the fewest number of distinct binding modes.

solution is greater by over 4% points (when considering the top four ligand poses). The level of improvement in the success rate of finding the correct binding mode is consistent with that reported recently on the basis of RMSD differences between docking poses.[11]

HIV protease in PDB entry 4phv is identified in fewer binding modes with the hydrogen-bond strength-modified implementation when compared with the other implementations, as is illustrated in Figure 4. This implementation is able to identify the correct solution with only two distinct binding modes, whereas other methods require either four or five binding modes. The superior performance of the hydrogen-bond strength-modified implementation in this example is due to the presence of a geometrically optimal hydrogen bond between the inhibitor and the backbone N of Asp 29 along with a near-optimal interaction to the backbone O of Gly 120. These optimal interactions carry a greater weight in the interaction bitstring compared to the other interactions and therefore provide a dominant influence in their clustering. This would ensure that those poses not sharing these interactions are clustered separately from those that do. With the other implementations (those not considering the geometry of a hydrogen bond), a successful ligand pose, within a 1.3 Å RMSD of the crystallographically determined solution, is in fact clustered together with poses worse than a 4.0 Å RMSD with respect to the crystallographic solution, due to additional hydrogen-bonds that they have in common. As this more distant pose is ranked higher by the scoring function (ScreenScore), the successful solution is effectively skipped. However, the hydrogen-bond strength-modified implementation ensures such poses are clustered separately, thereby improving the likelihood of identifying the crystallographically determined binding mode in fewer binding modes.

The suitability of the hydrogen-bond strength-modified method to this example could not have been predicted, since the existence of geometrically optimal and near optimal hydrogen bonds would not be known in advance. In other cases, different implementations may prove more suitable. For example, if the crystallographically determined binding mode of a ligand formed a hydrogen bond with a highly ranked sitepoint (high SWAPS value), the SWAPS-based method is likely to prove more efficient in identifying this binding mode, due to the higher weight assigned to such sitepoints in the clustering procedure. Prior knowledge of this interaction would be dependent upon such key sitepoints being targeted in the binding mode of known active ligands.

Table 1. Differences between the Mean Percentage of Cluster Members Using High Scoring Sitepoints (SWAPS ≥ 2.0) and Low Scoring Sitepoints (SWAPS < 2.0) for the Basic One-Bit-Per-Sitepoint Implementation of the Interaction Fingerprint Method, Compared to the Modified SWAPS-Dependent Fingerprint Method

	mean percentage (%) usage of sitepoints	
	SWAPS ≥ 2.0	SWAPS < 2.0
Basic method	32.60	50.34
SWAPS-based method	70.15	40.67

De Novo Design Output. To assess the specific performance benefits of the different implementations of the interaction fingerprint method, 1000 potential ligands were generated for the oedema factor protein of the anthrax virus. The 1000 ligands generated by Skelgen were clustered using each of the implementations of the interaction fingerprint method described above. The set of 19 clusters generated with the basic "one-bit-per-sitepoint" method were taken as the control set, and any modifications of these clusters with the extended implementations were analyzed.

The first modification of the basic method was the inclusion of the SWAPS value for each sitepoint used. The reason for this modification was to assign a higher similarity to those ligands sharing a sitepoint of greater importance (i.e. high SWAPS) when compared to ligands sharing sitepoints of lower importance. To test this effect, the variation in the sitepoints used by the different ligands of each cluster was analyzed. If a greater similarity is to be assigned to those ligands sharing a sitepoint with high SWAPS, one would expect to observe that a high scoring sitepoint used by one member of a cluster is more likely to be used by all other members of that cluster when compared to a low scoring sitepoint. This is due to high scoring sitepoints lending a greater weight to the determination of similarity (or dissimilarity). To quantify this, the mean percentage of members of each cluster sharing a high scoring sitepoint (SWAPS ≥ 2.0) was compared to the percentage of members sharing a low scoring sitepoint (SWAPS < 2.0). This was calculated for both the control set of clusters and the set of clusters generated by the modified method to account for SWAPS (see Table 1). As the data shows, the clusters produced by the modified method clearly demonstrate a grouping of clusters determined to a larger extent by high scoring sitepoints, with the clusters generated by the SWAPS-modified method demonstrating a significantly higher percentage usage of high scoring sitepoints when compared to low scoring sitepoints ($P = 0.0022$ as determined by a one-tailed, unpaired t-test). Interestingly, for the control method the clusters actually show a higher percentage usage of low scoring sitepoints in comparison to high scoring sitepoints. This could be due to the greater number of low scoring sitepoints when compared to high scoring sitepoints.

An example comparing the clusters produced by the control and SWAPS-modified fingerprint methods can be

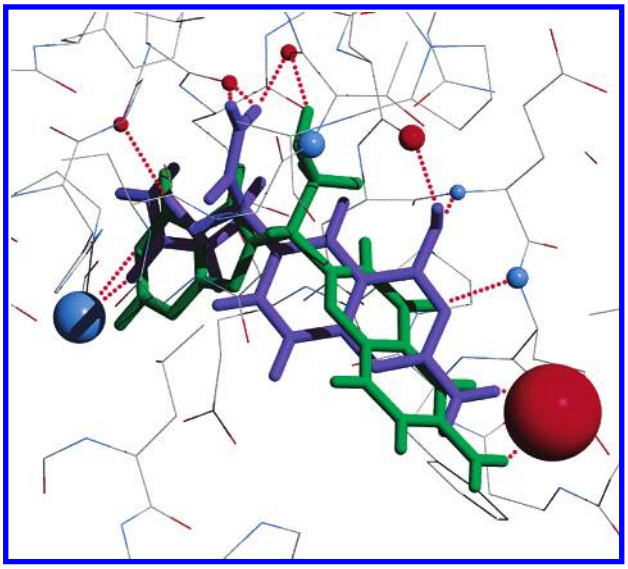

Figure 5. An example of the effect of including SWAPS in the fingerprint method on the clustering of ligands. With the basic fingerprint method, the purple and green ligands are clustered separately. However with the SWAPS-modified method, because they share interactions with the two high scoring sitepoints (largest red (oxygen) and blue (nitrogen) spheres, representing an oxygen of a carboxylic acid group and an amine of an arginine side chain, respectively) they are clustered together, despite not having all interactions with the low scoring (i.e. less important) sitepoints in common.

seen in Figure 5. Although the two ligands shown in this figure do not have all their hydrogen-bonding interactions with the protein in common and are clustered separately by the control method, they do share interactions with the two most important sitepoints as determined by their SWAPS (an almost fully accessible oxygen of a carboxylic acid group and a fully accessible amine group of an arginine side chain). This occurrence is captured by the SWAPS-modified method where the ligands are clustered together, despite the differing interactions with the lower scoring sitepoints. This modified grouping of the ligands is the preferred result since the two high scoring sitepoints are likely to represent a greater contribution to the overall binding energy of the ligand.

Another modification of the basic interaction fingerprint method was the inclusion of the strength of hydrogen bonds formed between the ligands and the protein. The reason for this modification was to assign a greater similarity to those ligands sharing a strong hydrogen-bonding interaction when compared to those ligands sharing weak hydrogen-bonding interactions. This modification was tested using a similar approach as for the SWAPS-modified fingerprint method described above. In this case, for each sitepoint used by the members of a particular cluster, the mean strength of interaction was calculated, along with the percentage of members of the cluster that interact with that sitepoint. If a greater similarity is to be assigned to those ligands sharing a strong hydrogen-bonding interaction, one would expect to observe that sitepoints with a high mean strength of interaction are more likely to be used by all members of the cluster than those sitepoints with a low mean interaction strength. To quantify this, the average usage of sitepoints with a mean interaction strength greater than or equal to 50% of the maximum was calculated and compared to the mean usage of sitepoints with a mean interaction strength less than 50% of the maximum. This was calculated for both the clusters generated with the control fingerprint method and those generated by the strength-modified method (see Table 2). As the data reveals, the clusters produced by the modified method clearly show a grouping of clusters determined more by interactions of greater strength, with the strength-modified method demonstrating significantly greater usage of those sitepoints that, on average, form strong interactions, when compared to those forming generally weaker interactions ($P = 7.9 \times 10^{-10}$ as determined by a one-tailed, unpaired t-test). With the control method, a greater usage is also observed for those sitepoints generally forming strong interactions, although the difference is less significant ($P = 0.022$).

Table 2. Differences between the Mean Usage of Sitepoints with a Mean Interaction Strength Greater than or Equal to 50% of the Maximum When Compared to Those with a Mean Interaction Strength Less than 50% of the Maximum, for Both the Basic One-Bit-Per-Sitepoint Implementation of the Interaction Fingerprint Method and the Modified Strength-Dependent Fingerprint Method

	mean percentage (%) usage of sitepoints	
	strength ≥ 50%	strength < 50%
Basic method	59.7	44.2
Strength-based method	69.8	28.6

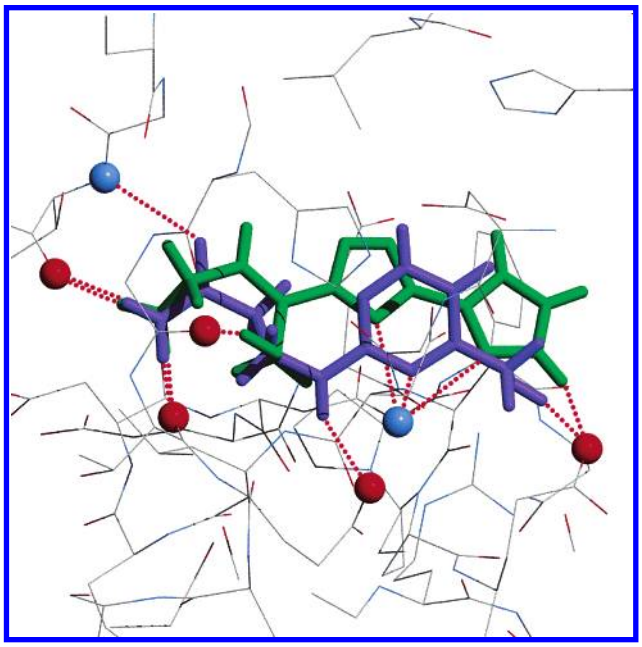

Figure 6. An example of the effect of including the strength of the hydrogen bonds in the fingerprint method on the clustering of ligands. With the basic fingerprint method, the purple and green ligands are clustered separately. However with the strength-modified method, because they share three "ideal" hydrogen bonds (with the two oxygens (red spheres) furthest left and the oxygen furthest right) they are clustered together, despite not having all weaker hydrogen bonds in common.

An example comparing the clusters produced by the control and strength-modified fingerprint methods can be seen in Figure 6. As with the SWAPS example shown in Figure 5, the two ligands shown in this figure do not have all their hydrogen-bonding interactions with the protein in common and are clustered separately by the control method. However, they do have three "ideal" hydrogen bonding-interactions with the protein in common. Based upon these

Table 3. Differences between the Mean Percentage Usage of a Sitepoint Cluster within a Cluster of Ligands, for the Basic One-Bit-Per-Sitepoint Implementation of the Fingerprint Interaction Method and the Sitepoint-Clustering (SC) Modified Fingerprint Method with a Range of Penalty Weighting Factors

	penalty weighting factor	mean percentage usage (%)
Basic method	N/A	80.60
SC-modified method	0.75	84.05
SC-modified method	0.5	85.46
SC-modified method	0.25	89.10
SC-modified method	0.1	90.33

interactions, and the fact that the interactions by which they differ consist of weaker interactions, the ligands are clustered together by the strength-modified fingerprint method. This modified grouping of the ligands is the preferred result since it is these three "ideal" interactions that are shared by the two ligands that would be expected to provide the greatest contribution to the overall binding energy of the ligand. The sitepoints with which these "ideal" interactions are formed (Thr548 C═O, Thr579 C═O, and Glu588 OE2) are all used by 3′dATP in its crystallographic binding mode and are members of the core sitepoints used in the Skelgen run.

The final modification of the basic method was the inclusion of a preclustering of the sitepoints and then assigning a greater distance value to those ligands showing an intercluster difference in their interaction pattern when compared to an intracluster difference. The purpose of this modification was to minimize the effect of small changes in the binding mode of a ligand when "switching" between closely neighboring alternate sitepoints and deem them more similar when compared to more dramatic intercluster changes in the binding mode. This effect was tested as follows: for each cluster of ligands, the percentage of members using each sitepoint cluster was calculated. The greater this value, the fewer the intercluster differences between the ligands of that cluster, as the sitepoints are used more consistently by these ligands. The mean of all percentages computed for each cluster of sitepoints for each ligand cluster was then calculated. This mean was obtained for the control set of clusters generated by the basic fingerprint method in addition to the modified method implementing different penalty weights for intercluster differences (see Table 3). As expected, fewer intercluster differences are observed within ligand clusters generated with the modified fingerprint method, when compared to the control method ($P = 0.044$ with a penalty weighting factor of 0.1). This effect becomes more pronounced as the weighting factor moves closer to zero, thereby penalizing intercluster differences more.

An example comparing the clusters produced by the control and sitepoint clustering-modified fingerprint methods can be seen in Figure 7. The two ligands in this figure share all but one interaction with the protein and are clustered separately by the control method. However, since it is only an intracluster difference that separates them, the ligands are clustered together by the modified method. Given that it would seem reasonable that such an intracluster difference represents only a small change in the binding mode of a ligand, this modified clustering is the preferred result. Such an effect cannot be captured by the control method, since it does not account for the geometric arrangement of the sitepoints. Incidentally, this effect would also be missed by the SIFt method as all interaction differences between residues are treated equally.

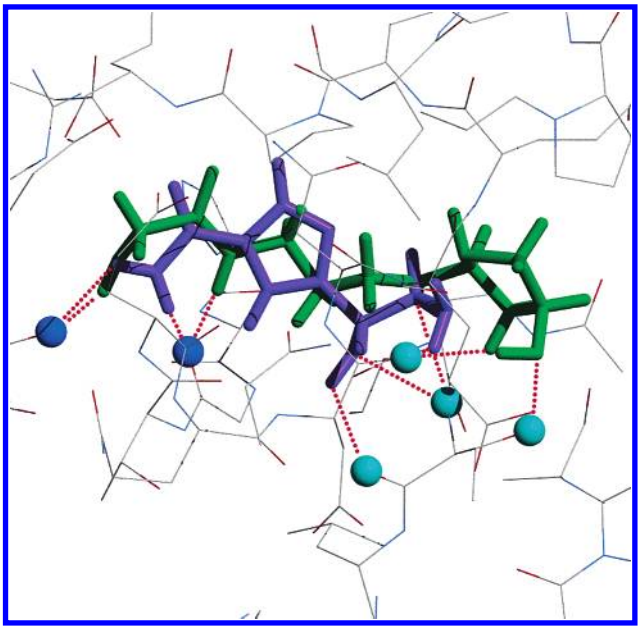

Figure 7. An example of the effect of including sitepoint clustering in the fingerprint method on the clustering of ligands. The sitepoints are colored to indicate the clusters to which they belong. With the basic fingerprint method, the purple and green ligands are clustered separately as not all interactions are shared. However with the sitepoint-clustering-modified method the ligands are clustered together as it is only an intracluster difference that separates them (i.e. the purple ligand interacts with the lowermost sitepoint, whereas the green ligand interacts with the next lowest sitepoint, both members of the same "cyan" cluster).

It is clear that the interaction fingerprint methods that we have developed are able to capture the various hydrogen-bonding interaction or geometric features that characterize ligand-protein interactions. Consequently, the choice of method would depend on the structural characteristics of the binding site and the weight that one wishes to assign to the relative importance of hydrogen bonding sitepoints and/or the strength of the hydrogen bonds made with putative ligands.

The encoding of hydrogen-bonding interactions described above can be readily expanded to incorporate additional interactions via nonpolar and/or steric sitepoints by extending the interaction fingerprint binary string. However, it is the explicit and detailed assessment of each ligand-protein interaction that confers a clear advantage over the original SIFt method. For instance, our method can account for multiple interactions of the same type to the same side chain (such as an acidic side chain accepting two hydrogen bonds). We have seen that it can also take into account variations in the hydrogen-bonding strength through the consideration of either sitepoint chemical type or hydrogen-bond geometry (this may be expanded to hydrophobic interactions). Furthermore, our method is able to distinguish between subtle switching among neighboring sitepoints and more conspicuous changes in the binding mode of a ligand.

One useful application of our method would be the screening of data sets of either predocked ligands or de novo designed molecules in order to identify and classify those sharing the same or similar binding mode. For example, ligands displaying similar interactions to those of a known

active compound could be readily and unambiguously retrieved.

4. CONCLUSIONS

An expanded interaction fingerprint method has been developed for analyzing the binding modes of ligands in docking and structure-based design methods. We have expanded the concept of representing a ligand in terms of a binary string that denotes its interaction with a target protein by including interaction-specific information. To provide a better representation of a ligand-protein interaction, we have considered the hydrogen-bonding strength and/or the accessibility of the hydrogen-bonding groups within a binding site as well as their geometric arrangement.

These various expanded methods have been applied to the postprocessing of binding poses generated in a docking study for 220 different proteins and to the analysis of ligands generated by an automated ligand-generation algorithm for the anthrax oedema factor. The application of even a basic implementation of the fingerprint method to the analysis of docking output resulted in an increased success rate for identifying the crystallographic binding mode.

In the analysis of the ligands generated for the anthrax oedema factor, the incorporation of additional interaction-specific information, such as hydrogen bond strength and the geometric arrangement of the hydrogen bonding groups, resulted in a more intuitive and comprehensive analysis of automated ligand-generation output.

While our initial methods have been limited to the consideration of hydrogen-bonding interactions alone, they have nonetheless shown a good performance in the analysis of docking and automated ligand-generation output. This performance will most likely be improved with the consideration of additional types of ligand-protein interactions, particularly nonpolar interactions, providing a more comprehensive representation of the ligand-protein interaction.

ACKNOWLEDGMENT

M.D.K. gratefully acknowledges De Novo Pharmaceuticals for the award of a post-graduate studentship. R.L.M. is also a Research Fellow of Hughes Hall, Cambridge.

REFERENCES AND NOTES

(1) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.

(2) Glen, R. C.; Allen, S. C. Ligand-Protein Docking: Cancer Research at the Interface between Biology and Chemistry. *Curr. Med. Chem.* **2003**, *10*, 767−782.

(3) 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.

(4) Blaney, J. M.; Dixon, J. S. A Good Ligand is Hard to Find: Automated Docking Methods. *Perp. Drug Discuss. Des.* **1993**, *1*, 301−319.

(5) Lyne, P. Structure-Based Virtual Screening: An Overview. *Drug Discuss. Today.* **2002**, *7*, 1047−1055.

(6) Stahl, M.; Schulz-Gasch, T. Practical Database Screening with Docking Tools. *Ernst Schering Res. Found. Workshop.* **2003**, *42*, 127−151.

(7) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H.-J.; Dean, P. M. A Validation Study on the Practical Use of Automated *de novo* Design. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 459−478.

(8) Lewis, R. A.; Leach, A. R. Current Methods for Site-Directed Structure Generation. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 467−475.

(9) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(10) 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.

(11) Källblad, P.; Mancera, R. L.; Todorov, N. P. Assessment of Multiple Binding Modes in Ligand-Protein Docking. *J. Med. Chem.* **2004**, *47*, 3334−3337.

(12) Carhart, R. E.; 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.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.

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

(15) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.

(16) Fisanick, W.; Cross, K. P.; 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.

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

(18) Kearsley, S. K.; Sallamack, 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.; Wild, D. J.; Willet, 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) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete. Appl. Math.* **1988**, *19*, 17−44.

(21) 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.

(22) 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.

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

(24) Holliday, J. D.; Salim, N.; Whittle, M.; Willet, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.

(25) Whittle, M.; Willet, P.; Klaffke, W.; van-Noort, P. Evaluation of Similarity Measures For Searching the Dictionary of Natural Products Database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449−457.

(26) Eldridge, M.; Murray, C. W.; Auton, T. A.; Paolini, G. V.; Lee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(27) Muegge, I.; Martin, Y. C. A General Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(28) Kelly, M. D.; Mancera, R. L. A New Method For Estimating the Importance of Hydrogen-Bonding Groups in the Binding Site of a Protein. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 401−414.

(29) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: A Library of Information about Nonbonded Interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525−537.

(30) Bohm, H.-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant For a Protein−Ligand Complex of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(31) Sokal, R. R.; Michener. C. D. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kans. Sci. Bull.* **1958**, *38*, 1409−1438

(32) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L.; Vajda, S. Algorithms for Computational Solvent Mapping of Proteins. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 340−351.

(33) Tanimoto, T. T. An Elementary Mathematical Theory of Classification and Prediction. *IBM Internal Report;* 1958.

(34) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *0*, 1−10.

(35) Ward, J. H. Hierarchical Grouping to Optimise an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(36) Mojena, R. Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *Comput. J.* **1977**, *20*, 359−363.
(37) Mancera, R. L.; Källblad, P.; Todorov, N. P. Ligand-Protein Docking Using A Quantum Stochastic Tunneling Optimization Method. *J. Comput. Chem.* **2004**, *25*, 858−864.
(38) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein−Ligand Interaction. *Proteins* **2002**, *49*, 457−471.
(39) Todorov, N. P.; Dean, P. M. A Branch-And-Bound Method for Optimal Atom-Type Assignment in *de novo* Ligand Design. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 335−349.
(40) Todorov, N. P.; Dean, P. M. Evaluation of a Method for Controlling Molecular Scaffold Diversity in *de novo* Ligand Design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 175−192.
(41) Leppla, S. H. Anthrax Toxin Edema Factor: A Bacterial Adenylate Cycalse that Increases Cyclic AMP Concentrations Eukaryotic Cells. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 3162−3166.
(42) O'Brien, J.; Friedlander, A.; Dreier, T.; Ezzell, J.; Leppla, S. Effects of Anthrax Toxin Components on Human Neutrophils. *Infect. Immun.* **1985**, *47*, 306−310.
(43) Duesbery, N. S.; Webb, C. P.; Leppla, S. H.; Gordon, V. M.; Klimpel, K. R.; Copeland, T. D., Ahn, N. G., Oskarsson, M. K., Fukasawa, K., Paull, K. D., Woude, V. Proteolytic Inactivation of MAP−Kinase-Kinase by Anthrax Lethal Factor. *Infect. Immun.* **1985**, *47*, 306−310.
(44) Pellizzari, R.; Guidi-Rontani, C.; Vitale, G.; Mock, M.; Montecucco, C. Anthrax Lethal Factor Cleaves MKK3 in Macrophages and Inhibits the LPS/IFNgamma-induced Release of NO and TNFalpha. *FEBS Lett.* **1999**, *462*, 199−204.
(45) Friedlander, A. M. Macrophages are Sensitive to Anthrax Lethal Toxin Through an Acid-Dependent Process. *J. Biol. Chem. 261*, 7123−7126.
(46) Petosa, C.; Collier, R. J.; Klimpel, K. R.; Leppla, S. H.; Liddington, R. C. Crystal Structure of the Anthrax Toxin Protective Antigen. *Nature* **1997**, *385*, 833−838.
(47) Soelaiman, S.; Wei, B. Q.; Bergson, P.; Lee, Y.-S.; Shen, Y.; Mrksich, M.; Schoichet, B. K.; Tang, W.-J. Structure-Based Inhibitor Discovery Against Adenylyl Cyclase Toxins from Pathogenic Bacteria That Cause Anthrax and Whooping Cough. *J. Biol. Chem.* **2003**, *278*, 25990−25997.
(48) Drum, C. L.; Yan, S.-Z.; Bard, J.; Shen, Y.-Q.; Lu, D.; Soelaiman, S.; Grabarek, Z.; Bohm, A.; Tang, W.-J. Structural Basis for the Activation of Anthrax Adenylyl Cyclase Exotoxin by Calmodulin. *Nature* **2002**, *415*, 396−402.

CI049870G