Multiple-Docking and Affinity Fingerprint Methods for Protein Classification and Inhibitors Selection

Bo Li,[‡] Zhenming Liu,*[,†] Liangren Zhang,[†] and Lihe Zhang[†]

State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, China, and State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Received February 04, 2009

The function-based protein classification holds tremendous promise for molecular recognition and the structure-based design process. We describe here a new strategy combined with multiple-docking tools and "affinity fingerprint" analysis technology to detect functional relationships among proteins based on the substrate binding features and protein−ligand interaction matrix and apply it successfully for the family of phospholipase A2 to investigate protein−ligand binding, function-based protein classification, and inhibitor selection, evaluation. Binding data and matrix were generated by multiple versus multiple-docking among 12 PLA2s and 84 PLA2 inhibitors. Three kinds of statistic techniques, principal component analysis, multidimensional scaling, and cluster algorithms, were chosen to distinguish the groups with similar binding characteristics. The 12 PLA2s were automatically categorized into reasonable subfamilies on the basis of the protein−ligand binding matrix, and the classifying problem of cPLA2 (PDB ID: 1CJY) with relatively low homology was successfully dealt with. This approach was also used to identify and group the selective inhibitors against human nonpancreatic sPLA2. A sound pharmacophore has been defined from these selective inhibitors. It shows that the method is quite robust against individual data deviation, especially false positive, which makes it possible to be used in virtual screening with large enzyme families to generate selective inhibitors of targets on the basis of limited structural/function information.

INTRODUCTION

After the completion of the human genome roadmap sequencing, studies with the structural genomics techniques have rapidly increased the number of known structures of human gene products.[1] The important and true challenge is to assign biological functions to those raw sequences and gain a better understanding of the roles of specific proteins in both normal and disease processes.[2,3]

The classification of proteins on the basis of their functions, both biologically and structurally, has long been an active field of research life and chemical sciences.[4,5] The classification of proteins is important in many areas of research, including drug target identification, protein family characterization, and structure-based drug designs. Functional proteomics techniques have been used to address this challenge with sequence-alignment algorithms as a way to correlate biological functions to specific proteins. Indeed, this process has led to the categorization of a substantial number of enzymes and protein families. However, classifying proteins into functionally distinct families only on the basis of primary sequence information remains a difficult task. The limit action of classification using sequence similarity is mainly due to the presence of the so-called twilight zone, where similarity becomes indistinguishable from random matches. The domain comparison method is also used to investigate protein function classification on the basis of three-dimensional structure information, especially for those proteins that have low sequence similarities.[6]

Prediction and classification of protein function on the basis of binding-site character is another important strategy, which uses chemical probe or affinity fingerprint to annotate protein function and classify them into functional family. This method was first introduced in 1995 by Kauvae et al.,[7] but their emphasis mainly laid on the identification property of fingerprint. They selected out 10 proteins as the reference panel to make codes for more than 120 structurally diverse compounds. Application of multivariate regression techniques then took part in the data-processing step to draw useful information, and at last more than 75% of the small compounds met their expectation to get the identity code. Greenbaum et al. also stepped into this area in 2002, and their work[8] mainly focused on the classification of closely related proteins. They found that the fingerprint matrix was much more sensitive to the structural difference among proteins. Following, they did experiments and collected a large data set of affinity fingerprints for a pool of small molecules against the papain family of cysteine proteases, and then a dendrogram was generated by applying the cluster analysis. As compared to the dendrogram based on the sequence alignment, this classification method had the ability to find subtle differences among members of a protein family with high degree of sequence homology. In addition, they also classified those small molecules into three large groups, specific, nonspecific, and poor inhibitors, although the

* Corresponding author phone: 86-1082805514; fax: 86-1082802724; e-mail: zmliu@bjmu.edu.cn. Permanent address: School of Pharmaceutical Science, Peking University Health Center, Beijing 38 Xue Yuan Road, Beijing, China.
[†] School of Pharmaceutical Sciences.
[‡] College of Chemistry and Molecular Engineering.

10.1021/ci900044j CCC: $40.75
Published on Web 06/05/2009

boundaries among the inhibitors are not as clear as those among the proteins.

The method of affinity fingerprint is able to functionally characterize proteins and ligands and yield more information than those being categorized from the sequential or evolutional perspective. The previous work referred to above, however, was all based on the experimental data. It cannot image how many experiments need to be done to accomplish the final fingerprint matrix when the number of concerned proteins or ligands increases. Nonetheless, the reliability of affinity fingerprint is mainly based on the number of variables. To deal with the dilemma between reliability and feasibility, we ask for help from the DOCK strategy.[9] Starting from the crystal structure of proteins and corresponsive ligands, we can simulate the real binding interaction in silico. Although there are several unsatisfactory steps in such simulation procedure, especially the default score function, we can accept the accuracy of the DOCK program for sure due to the self-robustness of the method.

Besides the docking process, which generates the original data for the fingerprint matrix, a variety of statistical and artificial intelligence techniques are also applied in our research. These techniques have been introduced to the biorelated research for years, and some of their previous applications list as follows: Nendza and Seydel applied multiple linear regression (MLR) to analyze ecotoxicity data obtained in 11 biological test systems for more than 50 phenols, anilines, and hydrocarbons.[10] MLR showed that the relative toxicities of these compounds were well-determined in all testing systems and could be described as a function of lipophilicity. Ebert et al. used principal component analysis (PCA) to investigate the relative information contents of 15 routine antitumor tests on 13 para-substituted aryldimethyltriazenes.[11] PCA[12] is a method for re-expressing multivariate data. It follows the researcher to reorient the data so that the first few dimensions account for as much of the available information as possible. If there is substantial redundancy present in the data set, it may be possible to account for most of the information in the original data set with a relatively small number of dimensions. This dimension reduction makes visualization of the data more straightforward and subsequent data analysis more manageable. Other than PCA, multidimensional scaling (MDS) was also used in this field. Shi et al.[13] investigated the anticancer activity of more than 7000 compounds across 60 human cancer cell lines with MDS. Such statistical technique refers to a set of methods used to obtain spatial representations of similarities or proximities between the row elements of a data matrix. Its goal is to create a map of manageable dimensionality such that the distances between objects on the map correspond closely with the observed proximities. In some sense, PCA is also one sort of scaling method. The difference is that for MDS the coordinate locations of the objects are usually interval scaled variables that can be used in subsequent analysis, in addition to visualizing the configuration of objects like PCA.

Both PCA and MDS give a brief visualization of large amounts of data, but our final purpose involves categorization, which means dividing groups of observations into smaller groups so that the observations within each group possess largely the same characteristics. It has to say that between the results yielded by PCA (or MDS) and our final purpose there surely exists distance that requires further work for researchers, which will inevitably import subjective influences. The best solution for such dilemma is cluster analysis, which is undertaken with the objective of addressing data heterogeneity. Rather than dealing with one group of widely divergent observations, cluster analysis can divide the massive data group into more homogeneous subsets. Many different approaches of cluster analysis have been developed nowadays, and some of them exactly fit our requirement. The most frequently used one, agglomerative hierarchical cluster method,[14] is nicely one member. Its functional principle is rather simple: start with n clusters, each consisting of a single point; then repeatedly merge the two clusters with the highest affinity into a single super cluster, until only one cluster of size n remains. Greenbaum has established that such a clustering algorithm gave powerful help for the function-based classification of proteins and ligands,[8] so our research work mostly settles on this foundation.

We describe here a new strategy combined with multiple-docking tools and "affinity fingerprint" analysis technology to detect functional relationships among proteins on the basis of the substrate binding features and protein−ligand interaction matrix, which represents the multiple-docking procedures and a series of data processing. Information encoded by the output result of such method can help to gain essential insight into the binding activities among different proteins against different small molecules. For proteins, we can find the latent relationship of similar binding activities among distinct proteins from the perspective of structure or source; for ligands, we can also identify the subset of inhibitors that exhibit similar functional characteristics. Besides the docking process that generates the original data for the fingerprint matrix, a variety of statistical and artificial intelligence techniques are also applied in our research.

METHODS AND MATERIALS

Selection of Protein Family, Crystal Structure, and Ligands. Phospholipase A2 (PLA2), especially human nonpancreatic secretary phospholipase A2 and human cytosolic phospholipase A2, have long been located at the research focus because they are believed to play an important role in the pathogenesis of arthritis.[15−19] PLA2s exist widely in venoms of snakes and bees, and in the pancreas of mammals and leucocytes, and they have the ability to hydrolyze the acyl group from the *sn*-2 position of glycerophospholipids. Normally, PLA2s are characterized by their disulfide bonds, size, dependence on Ca^{2+}, and their relative preferences for arachidonic acid at the *sn*-2 position. The majority of PLA2s extracted from venom and body fluid in animals belong to type II PLA2 enzymes. The members of this subfamily share the same characteristics such as 14 kDa molecular weight, Ca^{2+}-dependent active site, and no preference for arachidonic acid. Type IV PLA2 enzyme, which includes human cPLA2, has a larger molecular weight of 85 kDa and bears the binding tendency for arachidonic acid at the *sn*-2 position.

We selected 12 structures of PLA2 from the PDB database, most of which are sPLA2, to prepare for further studies. Some of them are sequence-identical, such as 1SV3, 1OXL, and 1SV9, but with different crystal resolution (Table 1).

Table 1. The 12 Selected PLA2s from the Protein Bank Database

ID	PDB ID	source	resolution (Å)	state	ligand
01	1MF4	andaman cobra	1.9	Apo	
02	1SV3	Russell's viper	1.35	Holo	4-methoxybenzoic acid
03	1CJY	human cytosolic PLA2	2.5	Apo	
04	1OXL	Russell's viper	1.8	Holo	(2-carbamoylmethyl-5-propyl-octahydro-indol-7-yl)-*ac*-acetic acid
05	1SV9	Russell's viper	2.71	Holo	2-[(2,6-dichlorophenyl)amino] benzeneacetic acid
06	1AE7	mainland tiger snake	2.00	Apo	
07	1POC	bee venom	2.00	Holo	1-*o*-octyl-2-heptylphosphonyl-*sn*-glycero-3-phosphoethanolamine
08	5P2P	porcine pancreas	2.40	Holo	phosphonic acid 2-dodecanoylamino-hexyl ester propyl ester
09	1KVO	human nonpancreatic sPLA2	2.00	Holo	4-(*s*)-[(1-oxo-7-phenylheptyl)anino]-5-[4-(phenylmethyl)phenyl thio]phentanoic acid
10	1DB4	human nonpancreatic sPLA2	2.2	Holo	[3-(1-benzyl-3-carbamoylmethyl-2-methyl-1*h*-indol-5-yloxy)-propyl]-phosphonic acid
11	1BP2	bovine pancreas	1.70	Holo	(4*s*)-2-methyl-2,4-pentanediol
12	1G4I	bovine pancreas	0.97	Holo	(4*r*)-2-methylpentane-2,4-diol

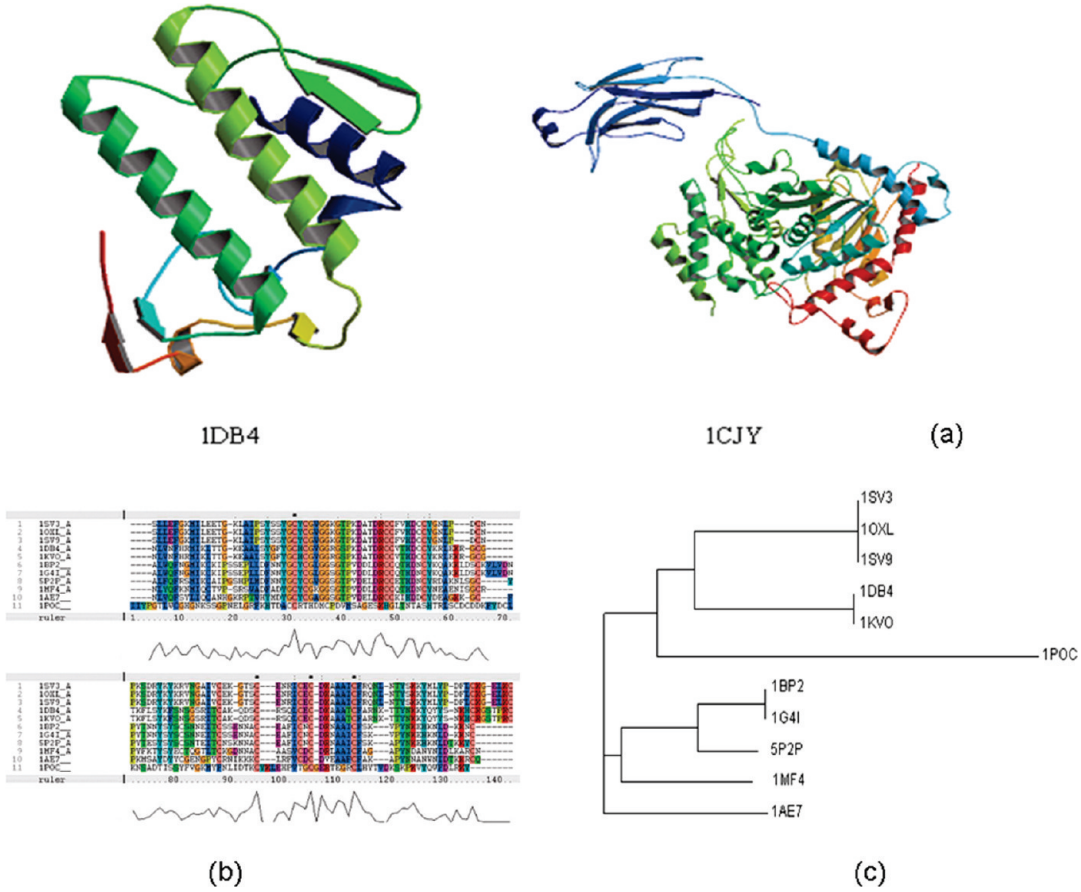

Figure 1. (a) Crystal structures of human nonpancreatic sPLA2 (PDB ID: 1DB4) and cPLA2 (PDB ID: 1CJY). The figures were generated by molscript and Raster3D. (b) The result of multiple sequence alignment given by ClusterW (1CJY not included). (c) The N−S dendrogram drawn by Tree Explorer according to the result of ClusterW.

We intend to check how much influence it will make for the variable crystal resolution to the generated fingerprint.

Because of biological evolution, the sequences of PLA2 from diverse species are dissimilar to each other. Figure 1b shows the result of multiple sequence alignment given by ClusterW.[20] We can see clearly that there are bundles of differences among these 11 sequences, the most obvious one of which is that the PLA2s from human body fluid and Russell's viper have comparatively longer sequences, about six amino acids longer than PLA2s from other sources. One thing has to be mentioned here that we can not examine the ascription of cPLA2 (PDB ID: 1CJY) because it has a much longer sequence than others. If we performed the multiple sequence alignment with it anyway, no information we will obtained from it because 1CJY would certainly be allocated to one single branch of N−S tree generated by ClusterW.

According to the result of clusterW, human nonpancreatic sPLA2 shares lots of similarities with those PLA2s extracted from the venom of Russell's viper, and it has been a demonstrated fact that these two kinds of sPLA2s have a close evolutionary relationship with each other.[21] The sPLA2s from other mammals exhibit more differences from them, but are relatively closer than those extracted from Andaman cobra or mainland tiger snake.

In the next step, we intend to deal with the functional classification of PLA2s, applying our method of molecular fingerprint. The 12 proteins have been selected as objectives to be sorted, and then we need to define the correspondent variables for the constitution of "fingerprint". Here are the principles of selection: those small molecules as variables cannot take on poor binding activities with all PLA2s, and at least one PLA2 in the 12 selected ones has the binding experimental data with them to make the reference panel for comparison with the computational data. We found that there is one such group of chemicals in the MDDR (MDL Drug Data report) named as "PLA inhibitors" that are capable of meeting our requirement (MDDR Index: 78348), in that they all exhibit some extent of inhibitory effect against a certain kind of PLA2s, respectively. A total of 65 chemicals are selected from such group. Besides that, we also add another 19 chemicals to our variable set, which experimentally present activities against human nonpancreatic sPLA2 according to the research work of Lilly Comp.[22−24] Such kind of sPLA2 correlates more with us human beings, so naturally we cast more attention on its inhibitors.

RESULTS AND DISCUSSION

Affinity Fingerprint Data Analysis. Processing of the 12 proteins and 84 small molecules is manually performed. There are some structural errors such as wrong atomic type or chemical bond, which cannot be autofixed. Each of the 84 small molecules was docked into the 12 proteins, respectively, resulting in a total of more than 1000 docking processes. To save time, the DOCK4.0 program was used. Every round it generated 500 conformations, and they were ranked by binding scores. For any docked solution, the binding score was calculated normally as the sum of van der Waals and electrostatic interactions. From the 500 generated conformations and their corresponsive binding scores, it is quite a difficult work to extract the most representative one as the "correct" combination. Because of the defect of score functions, docking is always harassed by a false positive, which means assigning an abnormally high score to a false binding model. There are lots of methods having been developed for reducing the false positive,[25] and one of them argued that a protein used as "background" should be docked with the original receptor together. The hypothesis is that those small molecules prone to generating the phenomenon of false positive usually have the same attitude toward all of the receptors. If the docking scores gained by the receptor and "background" protein with the same small molecule are both high, we cast deep doubt on their reliability. Such a method has been demonstrated effectively. In addition, we can call it a simplified edition of molecular fingerprint, with the number of objective (docked proteins) fixed to 2 only. Therefore, we can infer that our fingerprint method of multiple-docking has the natural resistance to false positive. To demonstrate that, we must first find the true binding model of each docking complex and compare its categorizing result with the DOCK-selected ones. The classification method is believed to be robust only when the two processes yield acceptable similar outcomes. The true binding model, however, is hard to determine doubtlessly except for those complexes already resolved by X-ray experiment. To step out of the predicament, we make one hypothesis: the true binding model is most closed to the first cluster of conformations ranked by score function. Those conformations with the highest docking score are all in the shade of false positive, but the "true" false positive should be an individual phenomenon. It rarely occurs on large amounts of conformations, especially those congregating together with unnoticeable rmsd value. In another word, they have a minor possibility to be false positive. If we then acknowledge that the true binding model merges in the top ranked conformations, the first cluster, especially its first member, should be the best representative of the true binding model in the nature of things. To facilitate our work, we arbitrarily define that such kind of "conformation cluster" should have more than five (included) members, and 0.05 Å was set as the maximum difference of rmsd value among the adjacent cluster members. To do comparisons, we obtained two result matrices recording the "best" scores of docking and "cluster" scores of docking, respectively, and these two matrices are converted into gray maps (show in Figure 2a) for illustration.

Enzymes Classification Based on Fingerprint Clustering. In the next step, we perform the cluster analysis for all 12 proteins. For the agglomerative hierarchical cluster, the structure of dendrogram is determined by both the distance matrix and the clustering method. According to Leming et al.,[15] we settled on average linkage clustering and a distance matrix of $(1 - r)$. Average linkage clustering means the distance between two clusters is the average of the distances between the data points in one cluster and the other. For the distance matrix of $(1 - r)$, r is the Pearson correlation coefficient between the activity patterns of two proteins. Figure 2b gives the cluster results for the two scoring methods referred to above.

We can see that these two methods generated rather similar results. That is to say, the false positive gives tiny influence on the classification process of proteins, which primarily demonstrates that our fingerprint method is pretty robust. From the fingerprint maps generated by two scoring ways, we can also illustrate this phenomenon. The reason is that similar structures certainly go with the similar activity, and also the "similar" false positive. If one small molecule behaves false positive with one protein, those molecules with similar structures will have more possibility to exhibit the phenomenon of false positive with this protein. Because the classification is based on the relative distance of objectives, the distance between similar false positive and the distance between similar true binding models will make no difference to the results. That is the key point to understand what has happened. The only difference was located in the middle level of the dendrogram, where some position exchange happened among 1SV3, 1SV9, 1OXL, and 1CJY. The difference is meaningless, however, because such classification method is relatively rough that we cannot and need not speculate on the tiny branches of cluster tree. From the two dendrograms, we can only get the information that these four proteins, 1SV3, 1SV9, 1OXL, and 1CJY, are all adjacent with each other and share similar binding activities, but no pair of them can be chosen as the closer partners as compared to the rest.

On the basis of the cluster result generated by sequence alignment, 1SV3, 1SV9, and 1OXL have the same first structure, so they naturally present similar binding activities and yield the similar fingerprint. The same case also occurred

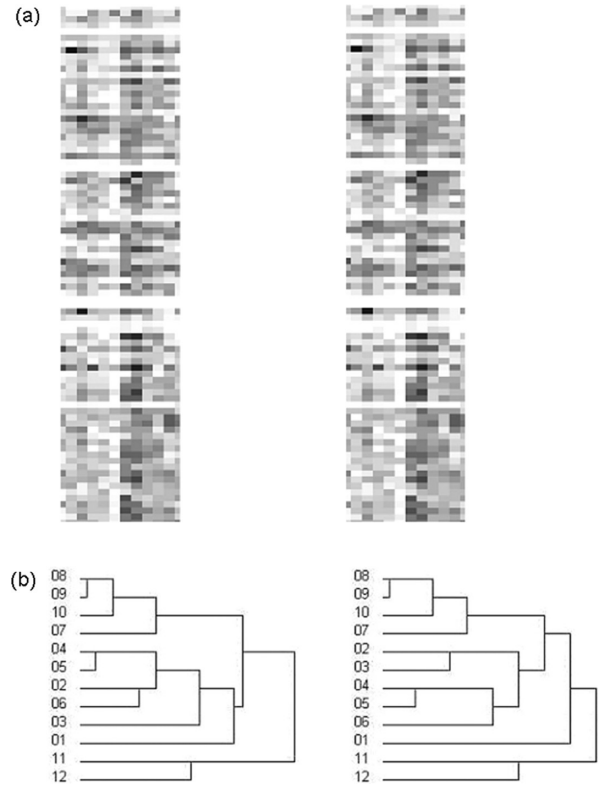

Figure 2. (a) Gray maps for the "best" score and the "cluster" score record. The *X* axis of the matrix represents the 12 PLA2s, and the *Y* axis represents the 84 PLA2 inhibitors. Every point in the left map represents the best score given by the DOCK4.0 program for one pair of complex, and every point in the right map represents the best score in the first cluster of conformation ranked by the DOCK4.0 program. The gray maps are produced by Origin 7.5. It gives a simple relationship between the value and gray scale, which means that the "deepness" of black color has a linear relationship with the binding intensity. (b) The cluster dendrogram for the 12 PLA2s according to their docking affinities. The left one was produced on the basis of the "best" score record and the right one on the basis of the "cluster" score record. The hierarchical cluster analysis was performed by SPSS 12.0.[26]

in the pairs of 1DB4 and 1KVO, 1BP2 and 1G4I, which are all neighborhoods in the dendrogram. We then could draw the conclusion that the resolution of X-ray structures makes little influence on the fingerprint analysis if the differences are acceptable. Although it is believed that fine resolution leads to finer docking result, the robustness of fingerprint will compensate the lost of rough resolution if given enough number of variables. Such a conclusion increases the feasibility of our fingerprint method because various resolutions can be found in the PDB database.

Despite the coherence, there are some differences deserving our attention. 5P2P, which is the crystal structure of porcine pancreatic sPLA2, acted differently with the methods of sequence alignment and fingerprint. From the perspective of evolution, the PLA2s from porcine pancreas have the closest relationship with those from bovine pancreas, 1BP2 and 1G4I, as demonstrated by sequence alignment. However, functional classification by fingerprint showed 5P2P is more similar to human PLA2s. Another example of the differences of sequence analysis and fingerprint analysis was 1MF4 and 1AE7 proteins, which are extracted from the venom of two uncommon snakes. Although ClusterW showed that these two PLA2s have different sequences, they still function similarly with PLA2s from other snakes, as demonstrated by their location in the middle branch of the dendrogram. These examples showed that the similarity does not always lead to activity similarity. Although primary sequence determines the tertiary structure and the function of proteins, some residues like those around the binding pocket likely make more contribution to the binding of substrates. We call them "key" residues. Most softwares for sequence alignment, like ClusterW, compare the whole sequences and rank them by similarity, but cannot distinguish those "key" residues from others. It is well-known that those with low sequence homology may show the same function because they share similar key residues and 3D structure. The function-based classification by fingerprint distinguishes those key residues from others and therefore has more important application in structure-oriented research, such as structure-based drug design. Moreover, fingerprint also has the further ability to categorize those proteins that are unsuitable to be clustered by sequence alignment. For example, 1CJY is the only cPLA2 in the 12 selected PLA2s, and ClusterW refuses to attribute it to any branches. This kind of classification is totally meaningless to 1CJY because of its abnormal molecular weight as compared to other sPLA2s. The latent similarity of those key residues is submerged in the huge difference of the whole sequences. From the perspective of fingerprint, we can see that the obstacle does not exist. Although making little differences in the two scoring methods, it is quite obvious that 1CJY behaves more similarly in the binding process with the sPLA2s extracted from venom, which gives some illustration for the toxic mechanism of snake's saliva.

Investigation of Inhibitors Selectivity Based on Fingerprint Clustering. Now we return to inspect the functional relationship of 84 small molecules, concerning their cross-reactivity and binding specificity to the 12 selected

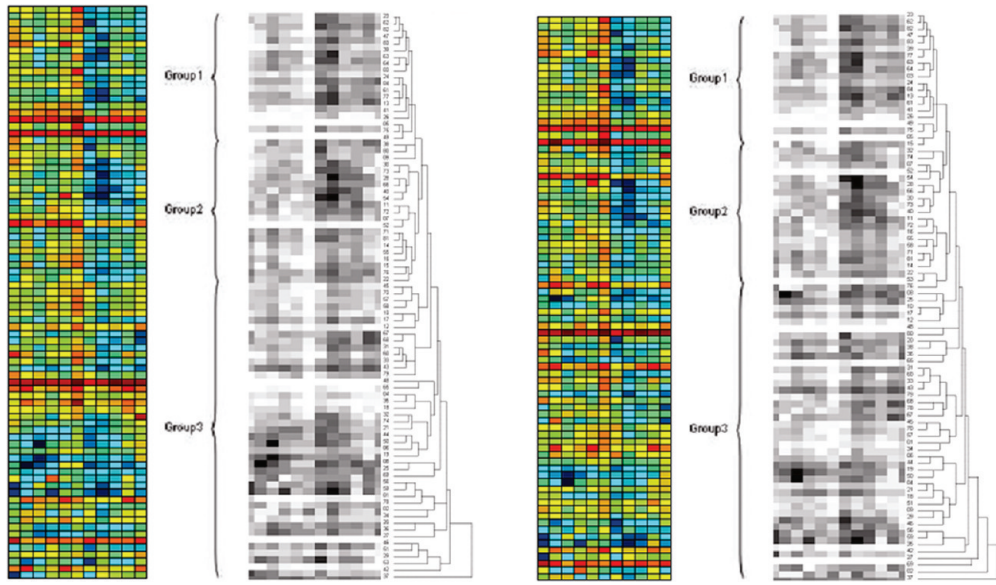

Figure 3. (a) The dendrogram and color maps based on "best" score record for the 84 small molecules. The dendrogram is generated by SPSS12.0 with agglomerative hierarchical cluster analysis, and the score matrix is reorganized according to order of cluster tree. (b) The dendrogram and color maps based on "cluster" score record for the 84 small molecules. The dendrogram is generated by SPSS12.0 with agglomerative hierarchical cluster analysis, and the score matrix is reorganized according to order of cluster tree.

PLA2s. The specificity to the binding target is extremely important during the evaluation process of inhibitors. The human body should be viewed as a whole biological system; therefore, higher binding specificity means fewer disturbances to its normal function in view that there are lots of latent "targets" for the inhibitor in human body. For our fingerprint method, it is expected that the clustering of the PLA2 inhibitors will reveal the underlying patterns of inhibition by categorizing molecules into groups of overall poor binding, promiscuous binding, and specific binding. The most concerned drug target in the 12 PLA2s is the human nonpancreatic sPLA2, which has been demonstrated to correlate with arthritis. Consequently, the most arrestive cluster of molecules are the selective inhibitors against 1DB4, 1KVO, or members in their clustered branch, which includes 5P2P and 1POC.

Figure 3 presents the dendrogram of cluster results for the 84 small molecules. On the left of each cluster tree, there are two kinds of illustration maps giving a clear and easy-perceived outline of the whole fingerprint matrix: one is the colorful scale map, and the other is the gray scale map. Both kinds of maps have been reorganized according to the order of dendrogram. First, we can find some interesting properties in the color maps. Almost all of the 84 small molecules have strong affinity with 5P2P and 1POC, and poor affinity with 1AE7. Selective inhibitors against human nonpancreatic sPLA2s do exist and are clustered together, but strong inhibitors against PLA2s from venom usually have the same binding intensity with others. One possible reason is that the binding site of those toxic PLA2s is the conservative structure for all of the PLA2s, and these molecules are all synthesized aiming to be inhibitors of human nonpancreatic sPLA2s, not in consideration of their inhibitory specificities.

Next, we compare the two groups of color maps, from which we can roughly argue that the results given by two scoring methods are rather similar: the top 20 molecules (labeled with "group 1") exhibit more affinity to those nontoxic PLA2s, especially those from bovine and porcine pancreas. Although group 1 shows quite a preference toward human nonpancreatic sPLA2s, its binding specificity is not enough. The next bundle of compounds labeled by "group 2", however, seems to be more satisfactory because they bear even more binding affinity with 1DB4, 1KVO and less with 1BP2, 1G4I. If not strictly required, compounds of both groups 1 and 2 can be viewed as PLA2 inhibitors with selectivity, which should receive much more attention during the process of drug design as PLA2 antagonists. The rest of the molecules constitute group 3, which include both poor inhibitors and promiscuous inhibitors. These two categories twist with each other in group 3, and our method cannot separate them. In the two fingerprints of different scoring methods, there seem to be many more differences for group 3, but these detailed differences are not worth paying attention to. As referred to before, our fingerprint method is relatively "rough". We apply hieratical cluster analysis to "group" the small molecules according to their binding characteristics, but the minor branches of cluster tree should not be taken seriously because they are too sensitive. The accuracy of molecular docking is far from satisfactory, and some errors do exist in our data matrix even after our deliberate revision to reduce the false positive. For a hierarchical cluster tree, its root is most robust, because all of the branches meet together at the root no matter how the data would change. The robustness will become less and less if we go along from the root to the end of branch. Fortunately, we need not put much emphasis on those details. To investigate the inhibitory specificities, we need only to find out the boundaries among groups in the dendrogram, and this task can be done by inspecting the reorganized matrix and color map. It is how we mark off the compound pool of groups 1, 2, and 3.

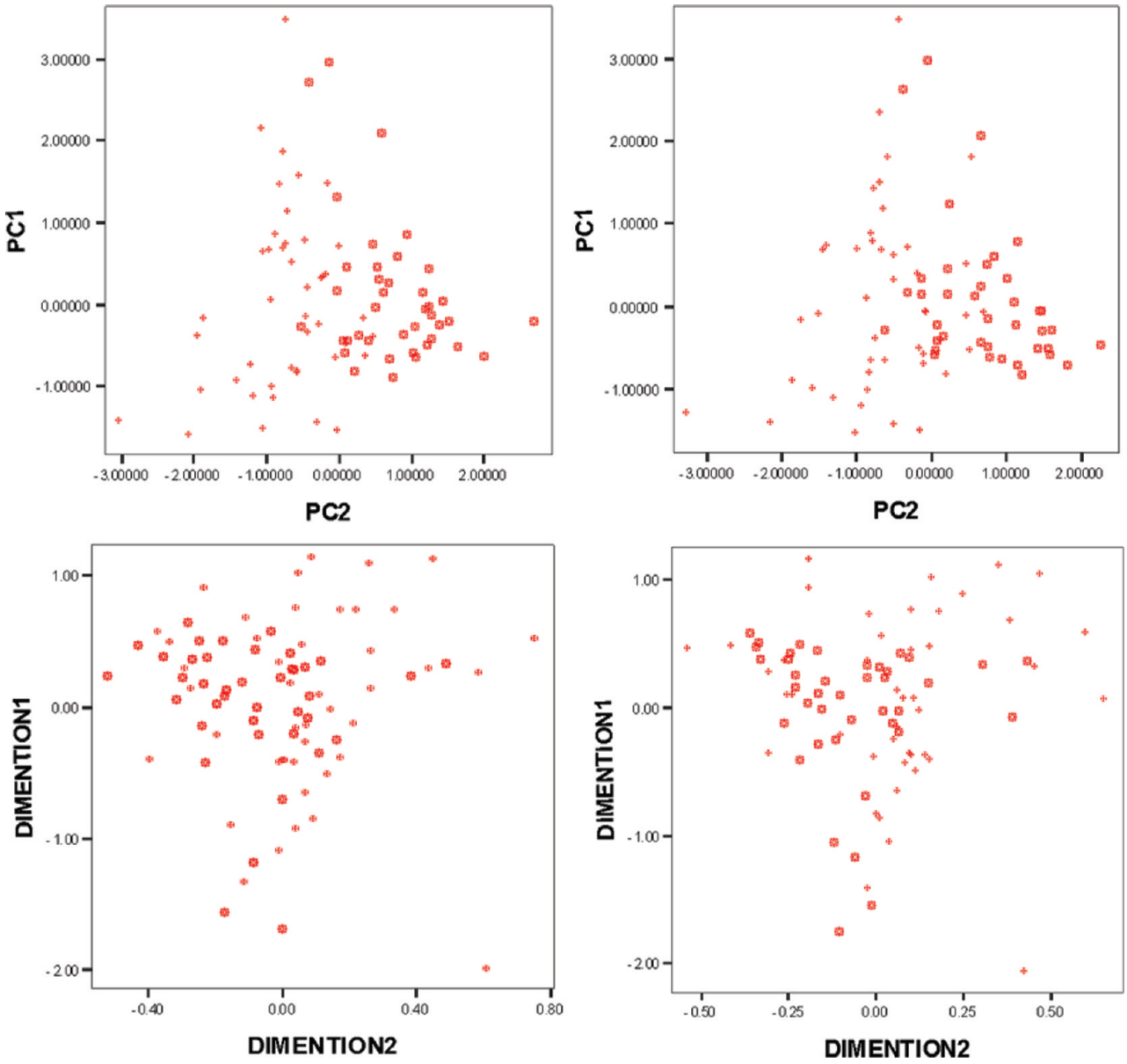

Figure 4. (a) The result of principal component analysis by SPSS 12.0. The left graph represents the "best" score, and the right accounts for the "cluster" one. We labeled all of the selective inhibitors with "□" symbol and others with "+". (b) The result of multidimensional scaling analysis by SPSS 12.0. The left graph represents the "best" score result, and the right one is for the "cluster". The selective inhibitors were labeled with "□" and others with "+".

PCA and MDS Analysis of Affinity Fingerprint. Except for cluster analysis, we also applied PCA and MDS on the fingerprint matrix as shown in Figure 4. For PCA (see Figure 4a), the outcomes of two scoring methods are as similar as cluster analysis. Take the "cluster" score record, for instance, the first PC accounts for 62.48% of variance in activity, and the sequential five PCs are 7.28%, 6.59%, 4.68%, 4.03%, and 3.16%, respectively. An immediate conclusion that can be drawn is that the binding characteristics of these PLA2s are quite similar in that they share lots of common information with each other. Despite that, we can also find that their limited dissimilarities are able to successfully distinguish the selective inhibitors from others by PCA. The "□" symbols, which represent selective inhibitors, conglomerate mostly at the botton-right corner of PCA score plot. Switching to the "best" score, which has more false positive data, the PCA result is almost the same. The first six PCs account for 62.71%, 7.40%, 6.30%, 4.36%, 4.12%, and 3.12% of variance, respectively, and the symbol dispersion in the score plot shows few differences from that from cluster analysis. Generally, PCA has the same strong error-tolerating ability as cluster analysis.

As to MDS (see Figure 4b), it mainly concerns inspecting the patterns of proximities of the data matrix, and we found that such a data-processing method is rather sensitive to the outliers. Because the false positive data are caused by the outliers, we obtain two different space plots for the two scoring methods. The sensitivity to outliers would impair the robustness of our fingerprint method. Another major drawback of MDS is that it generates worse categorization. For either MDS plot, those selective inhibitors are scattered along with nonselective ones everywhere. Although there is a trend to cluster inclination at the upper-left corner of plot, selective inhibitors are all surrounded and permeated so that no clear boundary can be outlined readily as PCA or cluster analysis. Therefore, MDS is unsuitable for our fingerprint method.

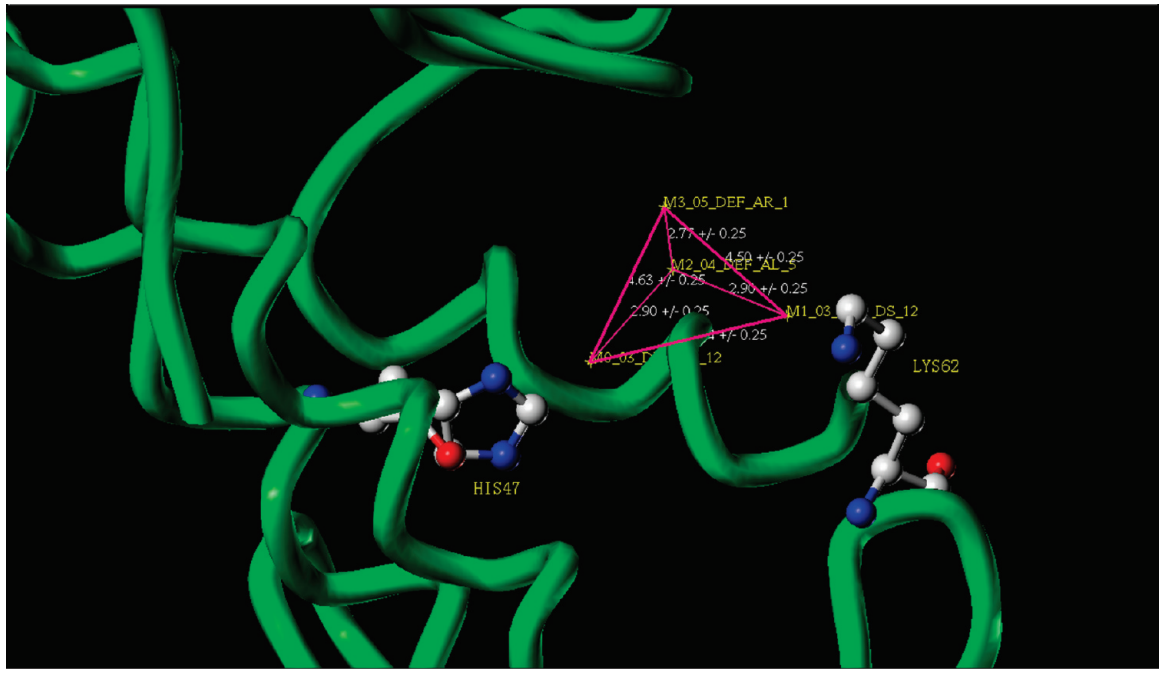

Figure 5. The pharmacophore showed above is defined by the program DISCO (distance comparison) in the software package of SYBYL 6.9. Single conformation is generated for each molecule, and the distance tolerance is set to 0.5 Å. Other parameters remain as default. Symbol "M" refers to the average center of key fragments recognized by DISCO.

Generate Pharmacophore Model from Affinity Fingerprint Analysis. In medical research today, one significant issue is the pharmacophore identification for drug-use molecules. Definitely, a pharmacophore is the three-dimensional arrangement of chemical features that causes activation or inhibition of the receptor. Well-identified pharmacophore extremely facilitates the design and modification of drug candidates and enhances the hit-rate of virtual screening. Nonetheless, the members of inhibitors are hard to decide in the pharmacophore-identification process, because we do not know which are selective inhibitors and which are promiscuous ones. The cross-reactivity of concerned inhibitors is rarely inspected by experimentalists in consideration of the time and money cost. Take PLA2 inhibitor for example, all of the 84 molecules we selected from MDDR database have more or less binding activity against human nonpancreatic sPLA2, but few have their IC_{50} values with other PLA2s. Without other information, we cannot distinguish those true selective inhibitors. If performing pharmacophore mapping for the whole group, surely we would obtain the common chemical features among all PLA2s, and consequently those inhibitors based on this pharmacophore would have strong inclination for promiscuous binding. With the help of the fingerprint method, however, the trouble can be handled with ease. We have identified that group 2 is the set of inhibitors with the best selectivity, so the pharmacophore extracted from compounds of group 2 is theoretically the best one for selective inhibitors against human nonpancreatic PLA2.

M0 and M1 are both electron-donor sites, which means that they interact with some side chains carrying positive charges. For M0, we found that His47 is in its 5 Å radius around, and Volwerk[27] has established that this residue plays an indispensable role in the catalytic process. Therefore, we may conclude that M0 is a critical site. M1 is also important because it is only about 2 Å away from Lys62, another key residue responsible for the constitution of an active passway with Leu2, Phe5, Cys44, Phe98, etc. The long side chain of Lys62 stretches out so it can be readily "held" by the negative group of inhibitor. M2 and M3 are both hydrophobic parts of inhibitor, which take the representation of aromatic rings and their accessories. They mainly interact with the hydrophobic pocket of the protein receptor. So, the pharmacophore defined by group 2 is rather reasonable and precise.

CONCLUSION

Instead of the traditional way of docking and reverse docking, we have developed a strategy combined with multiple-docking method in silico and "affinity fingerprint" analysis technology, which has never been introduced in the experimental field. Several multivariate statistical and artificial intelligence techniques are applied to the generated data matrix. Cluster analysis is determined as the principal tool due to its clear view of categorization. Principal component analysis is also feasible due to its similar result and robustness as cluster analysis. Multidimensional scaling, however, is unsuitable for our analysis because of its outlier-sensitivity and unsatisfactory classification.

As compared to the cluster result given by sequence alignment, fingerprint generates the protein classification dendrogram from the functional perspective. These two kinds of methods have lots of differences from each other, and the latter is more significant for structure-based drug design, because the proteins with low homology but similar binding sites can be categorized together. If applied to ligands, the cluster results make much sense too. We successfully distinguish the selective inhibitors against human nonpancreatic PLA2 from others. A well-defined pharmacophore has also been identified on the basis of this group of compounds.

Our method is quite robust against data deviation, which has been demonstrated by the same classification of PLA2s,

respectively, based on two data matrix, one with false positive and another without. Therefore, it has a promising application in the field of virtual screening, to help pick out or validate the false positive data.

Such a fingerprint method still deserves further improvement. We have already demonstrated that it has the promising ability of antifalse positive due to the robustness of cluster analysis. Yet due to the limitation of computational speed, large quantities of data will consume so much computer resources that we cannot apply this fingerprint method to a huge data set. We need a systemic and provable method to help us determine the most appropriate quantity of variables, which can give the acceptable result with minimum requirement of computer utilities.

Ultimately, the fingerprint method can dig out more binding information among proteins and small molecules. Its function-based classification gives a strong arm in the process of target selection, prioritization, and drug design. It should become more and more popular with the development of molecular docking software and computer speed.

ACKNOWLEDGMENT

We want to thank Prof. Lai Luhua and Prof. Qian Minping for related conversations and suggestions.

Supporting Information Available: Table containing the structures of all 84 small molecules used for multiple-docking and "fingerprint" analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **2001**, *294*, 93–96.
(2) Butcher, E. C.; Berg, E. L.; Kunkel, E. J. Systems Biology in Drug Discovery. *Nat. Biotechnol.* **2004**, *22*, 1253–1259.
(3) Cascante, M.; Boros, L. G.; Comin-Anduix, B.; de Atauri, P.; Centelles, J. J.; Lee, P. W. Metabolic Control Analysis in Drug Discovery and Disease. *Nat. Biotechnol.* **2002**, *20*, 243–249.
(4) Alam, I.; Dress, A.; Rehmsmeier, M.; Fuellen, G. Comparative Homology Agreement Search: An Effective Combination of Homology-Search Methods. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13814–13819.
(5) Lau, A. Y.; Chasman, D. I. Functional Classification of Proteins and Protein Variants. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6576–6581.
(6) Campbell-Valois, F. X.; Tarassov, K.; Michnick, S. W. Massive Sequence Perturbation of A Small Protein. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 14988–14993.
(7) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.
(8) Greenbaum, D. C.; Arnold, W. D.; Lu, F.; Hayrapetian, L.; Baruch, A.; Krumrine, J.; Toba, S.; Chehade, K.; Brömme, D.; Kuntz, I. D.; Bogyo, M. Small Molecule Affinity Fingerprinting: a Tool for Enzyme Family Subclassification, Target Identification, and Inhibitor Design. *Chem. Biol.* **2002**, *9*, 1085–1094.
(9) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(10) Nendza, M.; Seydel, J. K. Multivariate Data Analysis of Various Biological Test Systems Used for Quantification of Ecotoxic Compounds. *Quant. Struct.-Act. Relat.* **1988**, *7*, 165–174.
(11) Ebert, C.; Lassiani, L.; Linda, P.; Nisi, C.; Alunni, S.; Clementi, S. Chemometric Investigation of Antitumor Tests. *Quant. Struct.-Act. Relat.* **1984**, *3*, 143–147.
(12) Lattin, J. M.; Carroll, J. D.; Green, P. E. PCA for histogram-valued data proceedings. *Analyzing Multivariate Data*, 2nd ed.; Brooks/Cole, an imprint of Thomson Learning: New York, 2002; Vol. 1, p 126.
(13) Leming, M. S.; Yi, F.; Jae, K. L.; Mark, W.; Andrews, D. T.; Uwe, S.; Paull, K. D.; Weinstein, J. N. Mining and Visualizing Large Anticancer Drug Discovery Databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 367–379.
(14) Johnson, S. C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32*, 241–254.
(15) Komatsubara, T.; Tojo, H.; Zhao, Y.; Tomita, T.; Ochi, T.; Okamoto, M. Serum Phospholipase A2 Activity and Immunoreactive Group II Phospholipase A2 in Rheumatoid Arthritis. *Clin. Chim. Acta* **1995**, *236*, 109–112.
(16) Perrier, H.; Prasit, P.; Street, I.; Wang, Z. Bis(Aryloxy)Alkanes as Inhibitors of Phospholipase A2 Enzymes. *Biotechnol. Adv.* **1996**, *14*, 536–536.
(17) Lin, M. K.; Farewell, V.; Vadas, P.; Bookman, A. A.; Keystone, E. C.; Pruzanski, W. Secretory Phospholipase A2 as an Index of Disease Activity in Rheumatoid Arthritis: Prospective Double Blind Study of 212 Patients. *J. Rheumatol.* **1996**, *23*, 1162–11666.
(18) Michaels, R. M.; Reading, J. C.; Beezhold, D. H.; Ward, J. R. Serum Phospholipase A2 Activity in Patients With Rheumatoid Arthritis before and after Treatment with Methotrexate, Auranofin, or Combination of the Two. *J. Rheumatol.* **1996**, *23*, 226–229.
(19) Lin, M. K.; Katz, A.; van den Bosch, H.; Kennedy, B.; Stefanski, E.; Vadas, P.; Pruzanski, W. Induction of Secretory Phospholipase A2 Confirmst the Systemic Inflammatory Nature of Adjuvant Arthritis. *Inflammation* **1998**, *22*, 161–173.
(20) Jeanmougin, F.; Thompson, J. D.; Gouy, M.; Higgins, D. G.; Gibson, T. J. Multiple Sequence Alignment With Clustal X. *Trends Biochem. Sci.* **1998**, *23*, 403–405.
(21) Davidson, F. F.; Dennis, E. A. Evolutionary Relationships and Implications for The Regulation of Phospholipase A2 from Snake Venom to Human Secreted Forms. *J. Mol. Evol.* **1990**, *31*, 228–238.
(22) Dillard, R. D.; Bach, N. J.; Draheim, S. E.; Berry, D. R.; Carlson, D. G.; Chirgadze, N. Y.; Clawson, D. K.; Hartley, L. W.; Johnson, L. M.; Jones, N. D.; McKinney, E. R.; Mihelich, E. D.; Olkowski, J. L.; Schevitz, R. W.; Smith, A. C.; Snyder, D. W.; Sommers, C. D.; Wery, J.-P. Indole Inhibitors of Human Nonpancreatic Secretory Phospholipase A2. 1. Indole-3-acetamides. *J. Med. Chem.* **1996**, *39*, 5119–5136.
(23) Dillard, R. D.; Bach, N. J.; Draheim, S. E.; Berry, D. R.; Carlson, D. G.; Chirgadze, N. Y.; Clawson, D. K.; Hartley, L. W.; Johnson, L. M.; Jones, N. D.; McKinney, E. R.; Mihelich, E. D.; Olkowski, J. L.; Schevitz, R. W.; Smith, A. C.; Snyder, D. W.; Sommers, C. D.; Wery, J.-P. Indole Inhibitors of Human Nonpancreatic Secretory Phospholipase A2. 2. Indole-3-acetamides with Additional Functionality. *J. Med. Chem.* **1996**, *39*, 5137–5158.
(24) Draheim, S. E.; Bach, N. J.; Dillard, R. D.; Berry, D. R.; Carlson, D. G.; Chirgadze, N. Y.; Clawson, D. K.; Hartley, L. W.; Johnson, L. M.; Jones, N. D.; McKinney, E. R.; Mihelich, E. D.; Olkowski, J. L.; Schevitz, R. W.; Smith, A. C.; Snyder, D. W.; Sommers, C. D.; Wery, J.-P. Indole Inhibitors of Human Nonpancreatic Secretory Phospholipase A2. 3. Indole-3-glyoxamides. *J. Med. Chem.* **1996**, *39*, 5159–5175.
(25) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview Of Search Algorithms And A Guide To Scoring Functions. *Proteins* **2002**, *47*, 409–443.
(26) *SPSS, version 12*; SPSS Base 12.0 for Windows User's Guide; SPSS Inc.: Chicago, IL, 2005.
(27) Volwerk, J. J.; Pieterson, W. A.; de Haas, G. H. Histidine at the Active Site of Phospholipase A2. *Biochemistry* **1974**, *13*, 1446–1454.

CI900044J