SitePrint: Three-Dimensional Pharmacophore Descriptors Derived from Protein Binding Sites for Family Based Active Site Analysis, Classification, and Drug Design

James R. Arnold,[†] Keith W. Burdick,[#] Scott C.-H. Pegg,[§] Samuel Toba,[||] Michelle L. Lamb,[⊥] and Irwin D. Kuntz*[,‡]

Department of Pharmaceutical Chemistry, School of Pharmacy, University of California San Francisco, Box 2240, N474-A Genentech Hall, San Francisco, California 94143-2240

Received June 5, 2004

Integrating biological and chemical information is one key task in drug discovery, and one approach to attaining this goal is via three-dimensional pharmacophore descriptors derived from protein binding sites. The SitePrint program generates, aligns, scores, and classifies three-dimensional pharmacophore descriptors, active site grids, and ligand surfaces. The descriptors are formed from molecular fragments that have been docked, minimized, filtered, and clustered in protein active sites. The descriptors have geometric coordinates derived from the fragment positions, and they capture the shape, electrostatics, locations, and angles of entry into pockets of the recognition sites: they also provide a direct link to databases of organic molecules. The descriptors have been shown to be robust with respect to small changes in protein structure observed when multiple compounds are cocrystallized in a protein. Five aligned thrombin cocrystals with an average core α-carbon RMSD of 0.7 Å gave three-dimensional pharmacophore descriptors with an average RMSD of 1.1 Å. On a larger test set, alignment and scoring of the descriptors using clique-based alignment, and a best first search strategy with an adapted forward-looking Ullmann heuristic was able to select the global minimum three-dimensional alignment in twenty-nine out of thirty cases in less than one CPU second on a workstation. A protein family based analysis was then performed to demonstrate the usefulness of the method in producing a correlation of active site pharmacophore descriptors to protein function. Each protein in a test set of thirty was assigned membership to a family based on computed active site similarity to the following families: kinases, nuclear receptors, the aspartyl, cysteine, serine, and metallo proteases. This method of classifying proteins is complementary to approaches based on sequence or fold homology. The values within protein families for correctly assigning membership of a protein to a family ranged from 25% to 80%.

INTRODUCTION

One challenge in drug discovery projects is integrating information from screening, biology, and perhaps biological structure with classes of organic compounds.[1−3] Often different classes or series of compounds must be categorized and prioritized for further optimization within a set or family of targets. Given the advent of high throughput protein crystallography and computational model building techniques, information from protein structures and active sites can be used as one way of linking biological data to compound classes and prioritizing compounds. A computer program (SitePrint) is described that bridges information extracted from protein binding sites with small organic molecules such as those in screening collections, focused combinatorial libraries, and lead series. The primary objectives in the approach described here are to derive 3D pharmacophore descriptors from protein active sites and to show an initial study in which families of proteins are classified by comparing descriptors formed from their active sites. This report is intended to describe the methodology and algorithms used in the SitePrint approach.

Several methods have been described in the literature to characterize protein active sites, while several other programs have been described to create 3D pharmacophore descriptors from aligned ligands. The program GRID characterizes active sites by embedding the protein in a three-dimensional grid and mapping physiochemical properties of the atoms onto the grid points.[4] The volume of the box used to designate the boundaries of the site is often manually defined based on known function or ligand coordinates. LIGSITE also detects binding sites using a grid,[5] while the method in CavBase[6] combines the LIGSITE grid with potential hydrogen bonding, hydrophobic, and aromatic interaction locations[7] to create a descriptor that includes shape and binding properties. SPHGEN generates a set of spheres in cavities on a protein surface that map the points in space that ligand atoms can occupy.[8] The α-shapes method automatically

* Corresponding author phone: (415)476-1937; fax: (415)502-1411; e-mail: kuntz@cgl.ucsf.edu.

[†] Current address: AstraZeneca CNS Discovery, 1800 Concord Pike, Wilmington, DE 19850.

[‡] Current address: Department of Pharmaceutical Chemistry, University of California San Francisco, Genentech Hall, 600 16th Street, San Francisco, CA 94143-2240.

[§] Current address: Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, CA 94143-2240.

[||] Current address: Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121.

[⊥] Current address: AstraZeneca R&D Boston, 35 Gatehouse Drive, Waltham, MA 02451.

[#] Current address: Rigaku/MSC, 4045 Sorrento Valley Boulevard, San Diego, CA 92121.

10.1021/ci049814f CCC: $27.50
Published on Web 10/02/2004

detects cavities in a given protein structure using Delaunay complexes derived from Voronoi diagrams, giving a readily visualized surface.[9] Analysis of protein surface properties and curvature has been used to characterize binding and interaction sites[10] and to identify conserved patches with functional importance.[11] Functionally related binding sites have been classified using solvent-accessible surface patches that have been assigned physicochemical properties using a self-organizing neural net.[12] Protein similarity searches have been conducted using critical points assigned to areas of local curvature on a Connolly surface of the protein.[13,14] The methods ASSAM and TESS use clique detection and geometric hashing, respectively, to retrieve matches in proteins to predefined amino acid templates.[15,16] These descriptions allow protein-to-protein comparisons based on important residue locations and types.

Three-dimensional pharmacophore similarity searching methods recover useful and potent molecules.[17−22] Additional methods have been developed with the goal of augmenting the accuracy of three-dimensional similarity searches by reducing potential sources of error, and the method reported here builds on a substantial body of work in attempting to formulate a set of descriptors with complementary attributes. Potential sources of inaccuracy in 3D pharmacophore searches include those in generating bioactive conformations, in alignment, absence of explicit representation of molecular shape, absence of explicitly represented excluded volume, the lack of precision in the scoring function used to calculate similarity, and an inherent error because similarity to an active molecule is used to estimate free energy of binding to a biological target. Additional approaches that influenced the work reported here include COMFA[23] and the receptor surface models in Catalyst[24] which form three-dimensional inverse representations of putative binding sites from ensembles of aligned ligands. Both programs create a visualizable hypothesis of the volume and properties required for activity. Excluded volume can also be used as a constraint in search queries to improve the orientations of conformations to pharmacophores.[22] Other methods have been examined that use the overall surface,[12,25] or skin,[26] of active molecules as a similarity search query, which several other representations of molecular shape[27] or scaffold vectors[28] have also been used to perform similarity searches. All of the 3D representations discussed capture information in a different way than 2D methods such as Daylight fingerprints,[29] and one of the strengths of the 3D methods is that they facilitate structure jumping. The 2D and 3D methods are complementary, and they have been combined to enhance search coverage and accuracy.[30] Higher-order descriptors[31,32] and pharmacophore descriptor fingerprints[33] have also been used to improve the precision and amount of information present in pharmacophore searches. The structures of known drugs have been shown to reduce to a subset of small organic fragments,[34] and binding sites have been characterized by clustering small molecular fragments that have been placed in the sites.[35−37] 3D descriptors derived from protein active sites have been used in pharmacophore searching[31] and docking.[38,39]

The SitePrint method was developed to create a direct link between protein structures and organic molecules. In building on the work described above, it was desired that this method should enable comparisons between protein active sites, classification based on active site features, visualization of surfaces and properties, be used directly in ligand-based and structure-based searches, and be amenable to high throughput computation. The six point pharmacophore descriptors discussed in this paper also provide information about pockets and "scaffold locations" in active sites, where a "scaffold location" is a sitepoint from which R groups on an organic scaffold can access pockets in the site. Sitepoints in this work that are connected to only one other sitepoint are generally in pockets, while those connected to more than one other sitepoint are generally in scaffold locations. Information about scaffold locations and relative orientations of pockets in protein subfamilies has been shown to be useful in lead generation.[40−43]

METHODS

The algorithm described here produces 3D descriptors from protein binding sites where the proteins are not prealigned. One module of the program creates sitepoints, grids, and molecular surfaces from molecular fragments that have been docked, minimized, and clustered in protein binding sites.[4,18−27] A second module aligns and computes similarities between the descriptors. A third module assigns set membership of a given pharmacophore descriptor based on similarity to a group of descriptors where the function or activity is known. Set membership is used to predict the function of protein structures as a positive control for the accuracy of the descriptors and similarity scoring methods. The six-point descriptors used in these studies capture the rough shape of the binding sites, the scaffold-to-pocket relationships between regions of a site, and they allow for rapid alignment and scoring of descriptors. The grids allow for more detailed calculations of the van der Waals and electrostatic contributions to binding in the given site, and the surfaces facilitate visualization of the sites and how ligands may bind within them. Figure 1 shows a descriptor formed for thrombin in the context of the entire protein and the active site.[44] The terminal node in the S1 pocket of thrombin is shown at the bottom of Figure 1c: Asp 189 and Ser 195 are shown for reference with the Asp being at the bottom of the figure. The method for producing the descriptors is described below.

Each protein is automatically prepared for docking using a Perl script developed in the Kuntz laboratory at UCSF called AutoMolPrep. The script writes the protein, ligand, and water molecules to separate files. It adds hydrogens to the protein at pH 7 and assigns AMBER95 charges using the BioPolymer module of Sybyl.[45] SPHGEN is run over the entire protein and spheres within 10 Å of a manually specified central residue in the active site are saved for docking.[8] The final step in the script is to call grid to generate a 0.3 Å grid around the protein active site. The SitePrint method uses a set of eighteen fragments that are roughly based on those used in MCSS,[35] the drug frameworks derived by Bemis and Murcko,[34] and a reduced alphabet of the amino acids, as shown in Figure 2. Karplus and co-workers have shown the use of probing binding sites by performing dynamics on fragments in the sites.[35] The method reported here reuses the C++ docking code libraries of Makino and Ewing[46] to rapidly probe a site with fragments before generating descriptors. This has been done to allow the

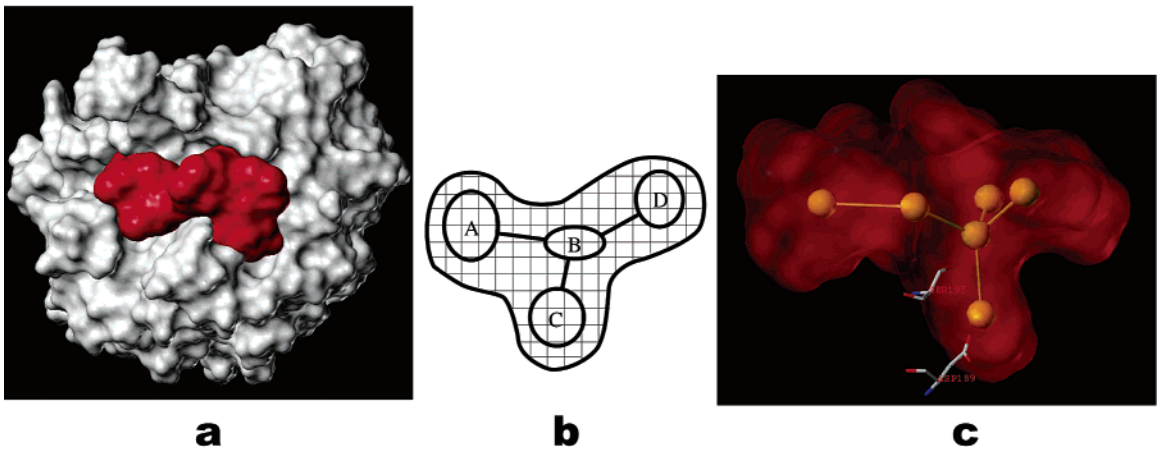

Figure 1. (a) The molecular surface of thrombin from the PDB structure 1dwb[44] is shown in white with the surface of the ensemble of fragments that map the binding site shown in red. (b) A schematic of the graph, grid, and surface that is generated for each binding site by the SitePrint algorithm. (c) The surface and a six-point pharmacophore descriptor generated for the binding site of thrombin is shown in a perspective shifted ninety degrees from that of **a**, with the S1 pocket of the site shown at the bottom.

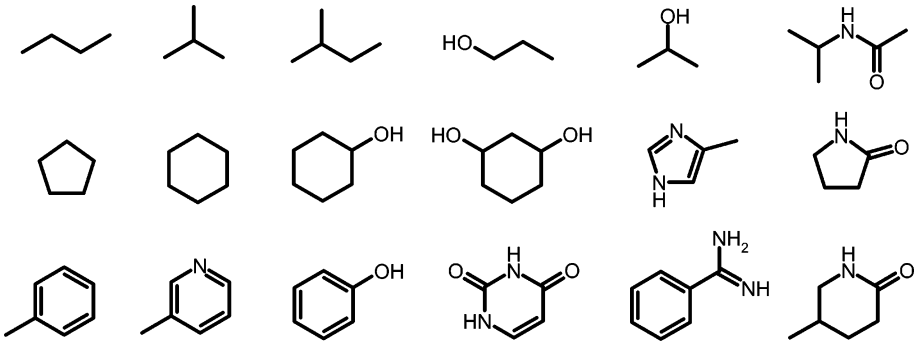

Figure 2. Fragments used to probe protein active sites.

method to be scaled up to operate on large numbers of biological structures. Each docked configuration of a given fragment is minimized, using the simplex implementation by Makino, to a gradient of 1.0 DOCK force field units. The DOCK force field scores are the sum of grid-based intermolecular van der Waals and electrostatic terms between the protein and the fragment configuration.[47,48] Configurations with DOCK force field scores more favorable than zero are retained. The resulting ensemble is smoothed using a greedy algorithm (sometimes called best-first clustering) that selects the best scoring fragment and removes all within 0.5 Å RMSD of it. The algorithm then continues to select the next best scoring fragment from those remaining and remove those close to it, until it has examined all of the fragments. This procedure removes unusually dense collections of fragments that are due to phenomena such as ionic interactions at the base of well defined pockets. This permits less dense collections of fragments, such as those bound in poorly defined hydrophobic pockets, to be counted on the same scale as the ionic collections, and it causes the resulting pharmacophore descriptors to be based more on shape and pocket location than on electrostatics. The electrostatic contributions are accounted for in grids after the descriptors have been formed.

The entire configuration of the smoothed fragments is clustered 10 different times using an implementation of the k-medoid algorithm[49] to produce 10 pharmacophore descriptors of between five and fifteen sitepoints. Each of the 10 separate descriptors contains between five and fifteen nodes where each node is a centroid of a cluster of fragment positions: six-point descriptors have been used in the studies reported here. The k-medoid algorithm was selected to perform the clustering because it is known to be tolerant of small variations in data. An example of a variation is the difference in the active site of a protein cocrystallized with two different ligands where one or more side chains are in different conformations due to small differences between the bound ligands. Each set of clustered nodes is then connected into both a tree and a graph. The trees are generated by an implementation of Kruskal's minimal spanning tree algorithm[50] which initially sorts all possible edges by distances between their nodes. The graphs are composed by starting with all possible edges and removing those that penetrate a wall of the protein or that are longer than 9.0 Å. Grids of the binding sites are created that consist of all grid points (in a grid of 0.3 Å resolution) that are within 1.5 Å of a heavy atom in the probe ensemble. The van der Waals and electrostatic information calculated from the protein is stored along with the grid point coordinates. Connolly surfaces are generated over the probe ensemble by saving the grid points as atoms and passing that molecule to the MOLCAD module of Sybyl.[21] The clusters, graphs, trees, grids, and surfaces for an individual active site are generated in less than a CPU minute on a workstation.

PROGRAM DESIGN

A modular program was built from the system level architecture design shown in Figure 3. The implementation is based on the C++ libraries that had been encoded by Makino.[46] Assessing the algorithmic choices, the simplex

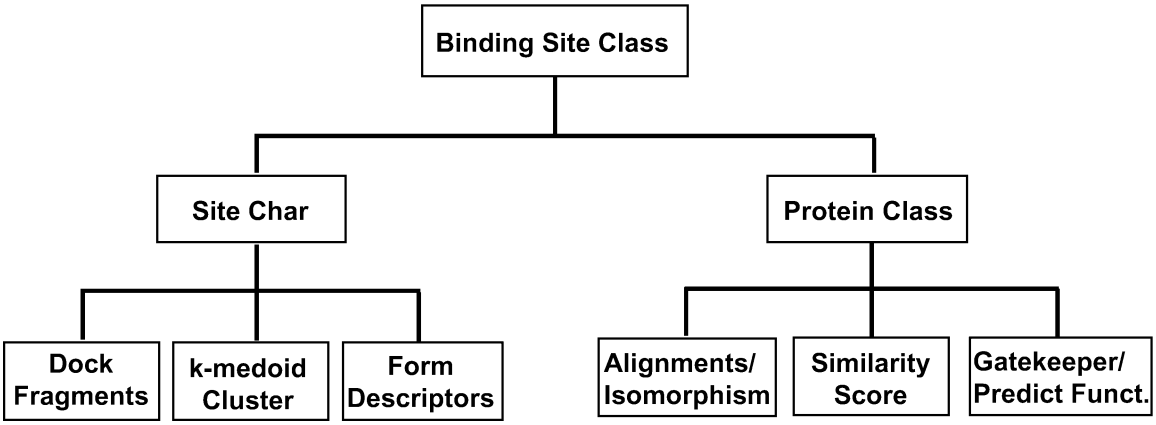

Figure 3. The architecture level design of two C++ programs and modules is shown. On the left side of the hierarchy is the program that generates the binding site descriptors. On the right is the program that compares the descriptors.

method is used to minimize fragments before clustering, and it is known to converge effectively on problems containing this number of degrees of freedom. The k-medoid algorithm[49] is of order $O(n^3)$, but it is known to be tolerant of outliers, and in our hands it has proven to be robust against small conformational changes in protein structure and also given changes in the fragment set. The docked and minimized ensembles of fragments number on average less than 10 000, so the accuracy of the clustering algorithm is more important than its algorithmic efficiency. The clique-based alignment implementation of Ewing and Makino[46] has been previously used to efficiently and accurately dock a given molecule to a set of spheres formed by SPHGEN.[8] The methods have been reused here to create multiple alignments of one pharmacophore onto another. Finally, Willett had suggested using the Ullmann heuristic[51,52] in conjunction with known tree searching algorithms[53] to accomplish graph isomorphism. The method has been extended to 3D pharmacophores to rapidly obtain one-to-one mappings of sitepoints in aligned pharmacophores prior to scoring the alignments.

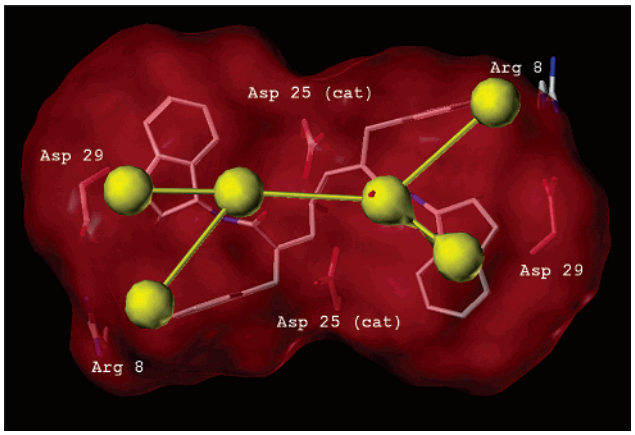

Figure 4. The top view of the active site of HIV protease with a six point pharmacophore descriptor shown in yellow, the surface mapped out by the probe ensemble in transparent red, and also the Merck inhibitor L700417 and side chains of residues in the site.[54] The terminal nodes of the graph are located in pockets, while the nonterminal nodes represent positions that might be occupied by scaffolds.

ANALYSIS OF FEATURES OF DESCRIPTORS

The SitePrint program was applied to HIV protease and thrombin in early validation studies before being used to characterize proteins in the kinase, nuclear receptor, aspartyl, serine, cysteine, and metallo protease families. A binding site descriptor formed from HIV protease complexed with the Merck inhibitor L700417[54] is shown in Figure 4 to illustrate the attributes of these descriptors. Aromatic groups of the ligand occupy the four pockets of the site, and those pockets are also described by sitepoints in the pharmacophore descriptor, where each sitepoint represents the location of a cluster of fragments that has been positioned in the site. The four terminal nodes in the graph are positioned in the pockets of the site. Each connection from one of the central, nonterminal nodes to one of the terminal nodes defines a possible approach into that pocket that might be engineered into a combinatorial library designed from a central scaffold. Connecting the pharmacophore points with a minimal spanning tree structure causes the topology of the graphs to contain information about pocket locations because bases of the pockets are at terminal positions in the tree. The overall position and connectivity of the sitepoints shows the rough shape of the active site. This aspartyl protease descriptor has a roughly linear shape with pockets staggered on either side of the site, which is characteristic of the β-sheet the protein has evolved to recognize. One deficiency found in the descriptor shown in Figure 4 is that the node in the upper left corner is not deep enough into the hydrophobic pocket because many of the fragments in the ensemble are positioned close to the mouth of the active site near the familiar salt bridge between Arg 8 and Asp 29. The DOCK force field is used as a scoring function in positioning the fragments, and it is known to overemphasize electrostatic interactions within some active sites. It is expected that a more accurate representation of solvation[55] effects would diminish the value of the ionic interaction and enhance the relative population of fragments binding in the hydrophobic pocket.

In contrast to the HIV protease example in Figure 4, the thrombin descriptor shown in Figure 1c demonstrates an example of a pocket of a protein occupied by a terminal node that indicates a possible warhead position. A *warhead* is a molecular fragment that binds to a recognition feature in a protein where often the pocket is deep, strong binding contacts are made, and a limited number of R-group substitution patterns from the fragment are possible because it is deeply bound within the pocket. An example of a warhead bound to a protein is benzamidine complexed in

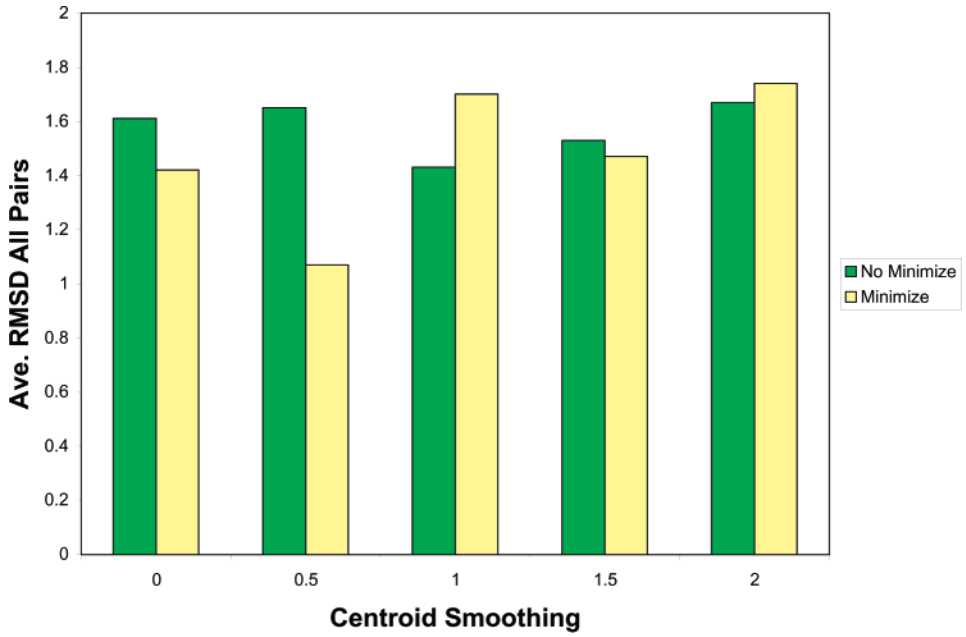

Figure 5. Consistently similar pharmacophore descriptors have been generated, aligned, and scored for RMSD similarity in this data set of five thrombin cocrystals. The yellow bars indicate that each configuration in the ensemble was minimized before smoothing was performed, while the green bars indicate no minimization was used. The *x* axis shows the different centroid smoothing parameters used on the fragment ensembles before the ensembles were clustered. The *y* axis reports the value of the averaged similarity for each five by five matrix of thrombin binding site descriptors.

the S1 pocket of thrombin.[40,44] Alternately, multivalent nodes such as those shown in Figure 4 correspond to scaffold positions of a combinatorial library from which multiple R-group substitution patterns are possible. The sitepoint descriptors may correspond to molecular design strategies that might be pursued for a given active site. The pharmacophore sitepoints can be used to search databases by programs such as the pharmacophore search module in MOE[22] or in work by Mason[31] or by de novo design programs[56] such as SPROUT[57] or INVENTON.[58] Protein pharmacophores have also been shown to be useful in the docking process by Joseph-McCarthy and Alvarez.[38,39]

TESTING ROBUSTNESS TO SMALL CHANGES IN PROTEIN STRUCTURE

It is important that the descriptors are consistently similar when they are produced from active sites that are different by very small amounts; i.e., the methods of generating and comparing site descriptors should be robust with respect to small changes in the positions of side chains in an active site. This was shown using a training set of five thrombin structures in which four different ligands were bound in protein cocrystals and the resolution of the structures ranged from 1.9 Å to 3.3 Å.[44] The average RMSD difference of the α carbons of the aligned thrombin structures was 1.5 Å. The loop regions of the proteins were then excluded from the computation because they can be disordered or can take different conformations due to changes in the packing forces imparted by different space groups in the crystals. The RMSD of only the α carbons of the protein cores differed by an average of 0.7 Å.

The study shown in Figure 5 examines the effect of fragment minimization and the RMSD smoothing parameter on the robustness of the resulting descriptors to induced fit changes in the protein. Pharmacophore descriptors were generated for each of the five thrombin structures where the docked fragment ensemble was either minimized or not, and the centroid smoothing parameter was systematically varied in five steps from 0.0 to 2.0 Å. This produced 10 descriptors for each active site where five fragment ensembles had been minimized before smoothing and five had not. The descriptors generated for each of the five proteins under a given parameter setting were then compared to each other in a five by five matrix, leaving the diagonal out. For example, five descriptors that were generated with minimization and the smoothing parameter set to 0.5 Å were aligned to each other, and RMSD of their mapped sitepoints was calculated for the resulting matrix. Each bar in the graph in Figure 5 represents the average value from twenty comparisons as only the scores above the diagonal in the matrix were used. An average similarity of 1.1 Å was achieved when the probe ensemble was minimized, and a smoothing filter of 0.5 Å was used. This is the parameter setting used to generate all descriptors discussed in later sections. The SitePrint method is robust in the thrombin system with respect to resolution of the structures and small conformational changes in active site side chains induced by different ligands bound in the cocrystal.

METHODS APPLIED TO A DATA SET OF PROTEIN FAMILIES

A larger data set for binding site classification was formed. It contains six families of proteins that have been actively pursued as drug design targets. In each family there are many structures available that are of high resolution, and each structure contains a bound ligand. Apo structures were not examined in this study given the increasing number of cocrystals that are being generated due to high throughput crystallography. The protein families and individual structures are listed in Table 1.

Table 1. Protein Structures in the Data Set Used To Test Family Based Analysis Using Active Site Pharmacophore Descriptors

family	protein	PDB	resolution (Å)
kinase	cyclic dependent (CDK2)	1atp[59]	2.2
kinase	transforming growth factor β	1b6c[60]	2.6
kinase	cyclin dependent	1hck[61]	1.9
kinase	insulin receptor	1ir3[62]	1.9
kinase	NAD dependent	2nad[63]	2.0
NR	estrogen α	3erd[64]	2.0
NR	PPAR γ	2prg[65]	2.3
NR	thyroid β	1bsx[66]	3.7
asp	renin	1rne[67]	2.4
asp	pepsin	1pso[68]	2.0
asp	cathepsin D	1lyb[69]	2.5
asp	HIV protease	4phv[55]	2.1
asp	endothiapepsin	4er2[70]	2.0
cys	cruzain	1aim[71]	2.0
cys	cathepsin B	1csb[72]	2.1
cys	papain	1pip[73]	1.7
cys	interleukin converting enzyme	1bmq[74]	2.5
ser	trypsin	1tps[75]	1.9
ser	trypsin	1trn[76]	2.2
ser	elastase	1fle[77]	1.9
ser	thrombin	1dwb[44]	3.3
ser	thrombin	1dwc[44]	3.3
ser	thrombin	1dwd[44]	3.3
ser	thrombin	1ppb[78]	1.9
met	aminopeptidase	1igb[79]	2.0
met	stromelysin	1sln[80]	2.3
met	thermolysin	4tmn[81]	1.7
met	collagenase-3	830c[82]	1.6
met	carboxypeptidase	8cpa[83]	2.0

The primary objective in examining many protein structures and multiple protein families is to confirm that the methods of comparing binding sites are accurate and operate quickly enough to be used on a large scale. The keys to comparing descriptors are rapidly generating a group of alignments and finding an isomorphic relationship between nodes in the aligned graphs so that nodes can be compared for scoring. The clique-based alignment technique used in this module has previously been validated by Makino and Ewing. Isomorphisms between graphs oriented in three-dimensional space are discovered using tree search algorithms that are guided by an adapted Ullmann heuristic.[51] The implementation of the tree search algorithms permits depth-first, best-first, and exhaustive methods to be easily compared.[53]

The efficiencies of the three tree searching methods were tested in the data set given in Table 1, and they are shown in Figure 6. In this study, all pairs of sitepoint descriptors for the thirty proteins in the training set have been aligned, and one to one isomorphisms have been generated for each alignment using the Ullmann matrix, and the RMSD score for the mapping is calculated. This procedure was repeated four times with the clique matching tolerance for the alignments varied systematically from 1 to 4 Å for each of the three tree searching algorithms. The average time required to align, map, and score all pairs of sitepoint descriptors in the training set is plotted. Increasing the clique matching tolerance increases the number of alignments examined. An average of five hundred alignments is produced for each protein pair when using a matching tolerance of 4 Å. This tolerance was required to generate the global minimum alignment in all cases.

In this study, the exhaustive search always generated the global optimum solution. However, it is also the least

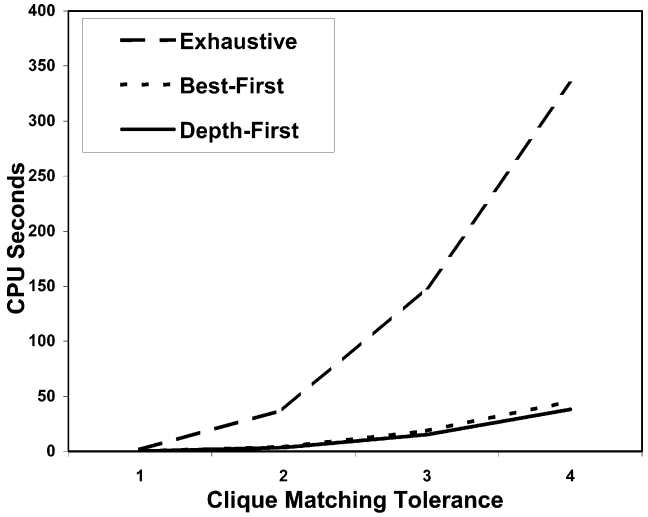

Figure 6. A comparison of the efficiency of exhaustive, depth-first, and best-first tree searching methods for performing graph isomorphism on SitePoint descriptors.

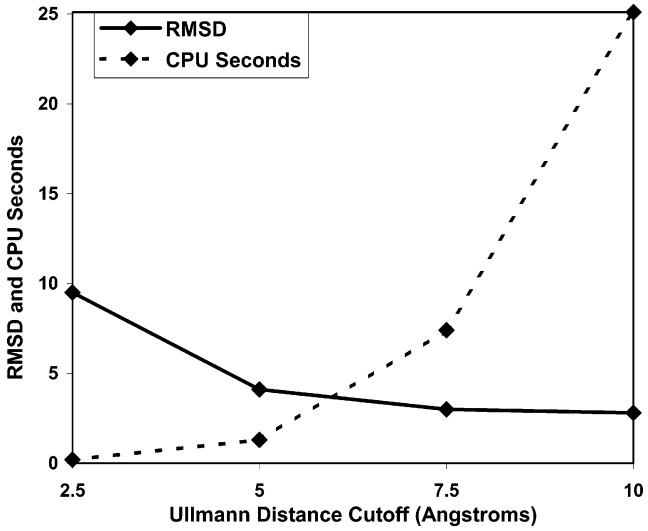

Figure 7. Evaluation of the effect of changing the distance cutoff for the Ullmann heuristic[51,52] on the speed and accuracy of using best-first search to perform graph isomorphism on aligned, three-dimensional, SitePoint descriptors is shown. The solid line shows the change in the RMSD of the best alignment as the distance cutoff is extended to include more sitepoints. The dashed line shows the cost in CPU time for including more potential alignments. At 7.5 Å the Ullmann directed best-first search achieves almost perfect accuracy: reproducing the global minimum in twenty-nine of thirty cases.

efficient. The best-first search, guided by the Ullmann heuristic adapted to 3D, is approximately 10 times more efficient. It also generates the global optimum alignment and mapping for 29 out of the 30 binding site descriptors. The Ullmann heuristic uses a look-ahead in the search, and it is critical to making this method fast and accurate enough to be used on a large scale. The Ullmann method is used to identify isomorphisms for the 3D alignments in the study shown in Figure 7.

The Ullmann heuristic speeds and directs the search process by pruning the breadth of the search tree. We have adapted it to operate on three-dimensional graphs aligned in space. It functions by examining two aligned graphs and creating a distance matrix for each pair of nodes in the graphs. Only nodes within a certain distance (8.0 Å) of a

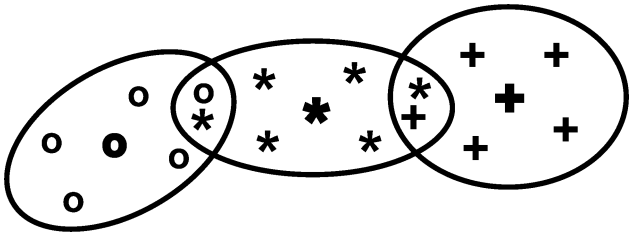

Figure 8. A schematic of three hypothetical protein families organized by the similarity of pharmacophore descriptors generated from their binding sites. Gatekeepers, which are the centroid of each cluster, are drawn in bold.

given node receive a value in the matrix that indicates they might map to the given node. A row or column of 0's indicates that one of the graph nodes cannot be mapped. Those alignments cannot succeed and are removed without attempting a search—which is a form of look ahead. The tree searching methods use the matrix to reduce the branching factor of the trees by only creating members of a new breadth when there is a 1 between two nodes being considered. Comparing all possible orientations between two pharmacophore descriptors (typically greater than 300) is accomplished in less than a CPU second on an SGI workstation. Using the Ullmann matrix as a heuristic speeds execution of a best-first search with very little loss of accuracy.

CORRELATING ACTIVE SITE PHARMACOPHORE DESCRIPTORS WITH ACTIVITY OR FUNCTION

A leave-one-out validation study was performed for each protein in the test set in order to evaluate the accuracy of the methodology in correlating the descriptors with function or activity. In this case protein function is defined based on activity, and the most representative pharmacophore descriptor for each family is called the *gatekeeper*. Gatekeepers were defined as being closest to the centroid of descriptors for that family. Hypothetical protein families and gatekeepers are shown in Figure 8. In the study, the descriptor that was left out was not included in defining the gatekeeper descriptor for the family. Similarity of the descriptor that was left out to all of the family gatekeepers was calculated, and family membership was assigned as that of the most similar gatekeeper. In the absence of a large amount of screening data this positive control study shows how accurately assigned set membership for 3D pharmacophore descriptors (based on a cluster centroids) relates to known function. This measurement of protein family membership is made based on active site descriptors, and it is complementary to methods that are based on sequence similarity or fold homology.

Metrics were defined based on aspects of cluster membership. The *tightness* of the protein families is the averaged sum of the RMSD distances between the descriptors derived from the family members. Family *interpenetration* is the fraction of members that are closer to gatekeepers from other families than to their own. *Function* was assigned to the protein that has been left out by measuring its similarity to that of the gatekeepers and assigning its function as that of the most similar gatekeeper. The protein that was most often predicted to be the gatekeeper for a given family in the leave one out study is reported in Table 2.

This initial study is intended to test the algorithms and performance of the SitePrint approach in terms of characterizing active sites and correlating the descriptors with biological activity. Large-scale studies of protein active sites that investigate the biology of the systems will also be undertaken. The results in Table 2 and Figure 9 show that the centroid-based family prediction gives better performance for tightly grouped families.

Table 2. Values for the Metrics in the Analysis of the Protein Binding Sites

protein family	tightness	interpenetration	% correct classification	gatekeeper
kinase	1.9 +− 1.3	0.40	60	c-AMP
NR	1.2 +− 0.9	0.33	67	TRβ
Asp	1.5 +− 0.8	0.20	80	pepsin
Ser	2.1 +− 0.9	0.29	71	thrombin (1dwd)
Cys	2.6 +− 1.6	0.75	25	cruzain
Met	2.2 +− 1.3	0.60	40	collagenase 3

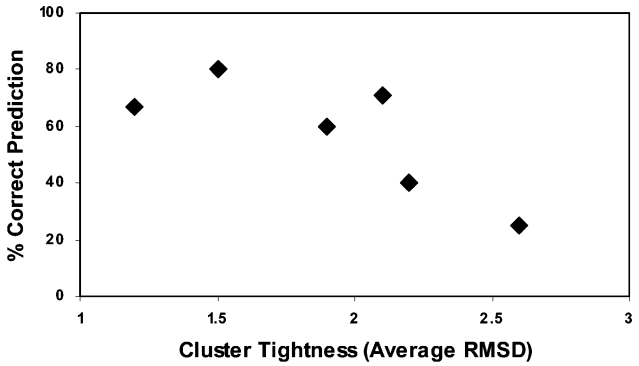

Figure 9. The covariance of cluster tightness and correct assignment of function of the protein families used in this study. Superior classification of function is seen for tighter clusters.

The pharmacophore descriptors for the nuclear receptors give a tight, highly correlated group. The estrogen and peroxisome proliferator-activated receptor bind steroids and are cocrystallized with estradiol and rosiglitazone, respectively, while the thyroid receptor is cocrystallized with triiodothyronine. There is similarity between the overall shapes of the three binding sites. Conversely, interactions with metals in the metalloprotease active sites are known to be a distinguishing feature for ligands binding them. The overall shape of the metalloprotease binding sites, in this study, is a less useful feature in classifying the members of that family. The primary result from the study shown in Table 2 is that the methods of forming, aligning, and calculating similarity of the SitePrint descriptors have been used to predict protein function in four out of the six families in the study.

CONCLUSIONS

The SitePrint program creates 3D descriptors from ensembles of molecular fragments docked into protein active sites. The descriptors have been formed in the volume occupied by ligands with the goal of later using them in database searches to augment screening collections and focus combinatorial libraries with compounds biased either toward, or away from, families of proteins. The method is robust against small conformational changes such as those that can be induced by different ligands binding to the same protein or by different crystallization conditions. In a positive control study using thrombin, consistently similar descriptors were generated where the RMSD of aligned pharmacophores is

0.4 Å greater than the RMSD of α-carbons in the cores of the aligned proteins. The proteins were not prealigned before the descriptors were created, so the pharmacophore descriptors were aligned, mapped, and scored as similar.

Both speed and accuracy of execution are issues in using these methods on a large scale, such as studies where all known proteins structures are evaluated. It has been demonstrated that an efficient strategy for orienting and performing graph isomorphism on SitePoint descriptors overlaid in three dimensions has been developed. This approach uses clique-based alignment and a version of the Ullmann heuristic adapted to 3D space in combination with a best-first search. The speed of execution of the method is an order of magnitude faster than an exhaustive search: comparing all possible orientations between two descriptors is accomplished in less than a CPU second on an SGI Octane. The method reproduces the globally optimal solution in twenty-nine out of thirty cases and differs by 0.2 Å in the other case. The adapted Ullmann heuristic is an effective and accurate way of performing a three-dimensional similarity search.

The methodology has been used to select representative members from protein families based on structural features of their active sites. Selecting gatekeepers imparts a hierarchy to proteins based on both known functions and calculated similarity of binding sites rather than on sequence similarity or overall fold homology. Potential uses of the gatekeepers include predicting function from structure and organizing proteins so that ligand binding information for a family may be obtained by screening against a limited number of biological targets. This is a practical consideration for screening groups working in protein families such as the kinases that number greater than five hundred members where primary and secondary screening can only be performed on a subset of the family. The SitePrint descriptors have been used to compare proteins based on their active site features and predict the functions of therapeutically significant classes of proteins. This demonstrates a correlation between the binding site descriptors and biological classes in a way that may have utility in family based drug discovery. Further studies will be carried out to show enrichment rates obtained from 3D pharmacophore searches with these descriptors.

ACKNOWLEDGMENT

Financial support was provided by the UCSF/NIGMS fellowship program and NIH Grant GM-56531 (P. Ortiz de Montellano, P.I.).

REFERENCES AND NOTES

(1) Greer, J.; Erickson, J. W.; Baldwin, J. J.; Varney, M. D. *J. Med. Chem.* **1994**, *37*, 1035−1054.
(2) Babine, R. E.; Bender, S. L. *Chem. Rev.* **1997**, *97*, 1359−1472.
(3) Bohm, H. J.; Klebe, G. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588−2614.
(4) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(5) Hendlich, M.; Rippman, F.; Barnickel, G. *J. Mol. Graph. Model.* **1997**, *15*, 359−363.
(6) Schmitt, S.; Kuhn, D.; Klebe, G. *J. Mol. Biol.* **2002**, *323*, 387−406.
(7) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.
(8) Shoichet, B.; Bodian, D.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 380−397.
(9) Liang, J.; Edelsbrunner, H.; Woodward, C. *Protein Sci.* **1998**, *7*, 1884−1897.
(10) Goldman, B. B.; Wipke, W. T. *Proteins: Struct., Funct., Genet.* **2000**, *38*, 79−94.
(11) Rosen, M.; Liang, S. L.; Wolfson, H.; Nussinov, R. *J. Mol. Biol.* **1998**, *11*, 263−277.
(12) Stahl, M.; Taroni, C.; Schneider, G. *Protein Eng.* **2000**, *13*, 83−88.
(13) Connolly, M. L. *J. Appl. Cryst.* **1983**, *16*, 548−558.
(14) Lin, S. L.; Nussinov, R.; Fischer, D.; Wolfson, H. J. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 94−101.
(15) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. *J. Mol. Biol.* **1994**, *243*, 327−344.
(16) Wallace, A. C.; Borkakoti, N.; Thornton, J. M. *Protein Sci.* **1997**, *6*, 2308−2323.
(17) Gund, P.; Wipke, W. T.; Langridge, R. In *Computers in Chemical Research and Education*; Elsevier: Amsterdam, 1973; Vol. II, p 33.
(18) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(19) ROCS; OpenEye Scientific Software: Santa Fe, NM.
(20) Catalyst; Accelrys: San Diego, CA.
(21) Sybyl; Tripos: St. Louis, MO.
(22) Molecular Operating Environment; Chemical Computing Group: Montreal, Quebec.
(23) Cramer, R. D., III; Patterson, J. D.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(24) Hahn, M.; Rogers, D. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(25) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. *J. Comput.-Aided. Mol. Des.* **1994**, *8*, 635−652.
(26) Van Drie, J. H. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 38−42.
(27) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17*(14), 1653−1666.
(28) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. *Molecular Recognit.* **1989**, 182−196.
(29) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(30) Aronov, A.; Goldman, B. B. *Bioorg. Med. Chem.* **2004**, *12*, 2307−2315.
(31) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(32) Martin, E. J.; Hoeffel, T. J. *J. Mol. Graphics Modell.* **2000**, *18*, 383−403.
(33) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D.; Spellmeyer, D. C.; Miller, J. L. *J. Med. Chem.* **2000**, *43*(14), 2770−2774.
(34) Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(35) Caflisch, A.; Miranker, A.; Karplus, M. *J. Med. Chem.* **1993**, *36*, 2142−2167.
(36) Mattos, C.; Ringe, D. *Nat. Biotech.* **1996**, *14*, 595−599.
(37) Kortvelyesi, T.; Silberstein, M.; Sheldon, D.; Vajda, S. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 173−186.
(38) Joseph-McCarthy, D.; Alvarez, J. C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 189−202.
(39) Joseph-McCarthy D.; Thomas, B. E. IVth; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 172−188.
(40) Bohm, H. J.; Banner, D. W.; Weber, L. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 51−56.
(41) Fejzo, J.; Lepre, C. A.; Peng, J. W.; Bemis, G. W.; Ajay; Murcko, M. A.; Moore, J. M. *Chem. Biol.* **1999**, *6*, 755−769.
(42) Blundell, T. L.; Jhoti, H.; Abell, C. *Nature Rev. Drug Discov.* **2002**, *1*, 45−54.
(43) Carr, R.; Hann, M. *Modern Drug Discov.* **2002**, April, 45−48.
(44) Banner, D. W.; Hadvary, P. *J. Biol. Chem.* **1991**, *266*, 30, 20085−20093.
(45) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(46) Makino, S.; Ewing, T. J. A.; Kuntz, I. D. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(47) Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 266−278.
(48) Shoichet, B. K.; Kuntz, I. D. *Protein Eng.* **1993**, *6*, 223−232.
(49) Kaufman, L.; Rousseeuv, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis;* 1990.
(50) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 1990.
(51) Ullmann, J. R. *J. ACM* **1976**, *16*, 31−42.
(52) Brint, A. T.; Willett, P. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.
(53) Knight, E.; Rich, K. *Artificial Intelligence*, 2nd ed.; 1992.
(54) Bone, R.; Vacca, J. P.; Anderson, P. S.; Holloway, M. K. *J. Am. Chem. Soc.* **1991**, *113*, 9382−9384.
(55) Zou, X.; Sun, Y.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033−8043
(56) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593−606.

(57) Gillet, V. J.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127−153.
(58) Pitman, M. C. Fragment Assembly in the Automated Molecular Invention System: INVENTON; Ph.D. Dissertation, University of California, Santa Cruz, CA, 1995.
(59) Zheng, J. H.; Trafny, E. A.; Knighton, D. R.; Xuong, N. H.; Taylor, S. S.; Teneyck, L. F.; Sowadski, J. M. *Acta Crystallogr. D Biol. Crystallogr.* **1993**, *49*, 362−365.
(60) Huse, M.; Chen, Y. G.; Massague, J.; Kuriyan, J. *Cell* **1999**, *96*, 425−436.
(61) Schulze-Gahmen, U.; De Bondt, H. L.; Kim, S. H. *J. Med. Chem.* **1996**, *39*, 4540−4546.
(62) Hubbard, S. R. *EMBO J.* **1997**, *16*, 5572−5581.
(63) Lamzin, V. S.; Dauter, Z.; Popov, V. O.; Harutyunyan, E. H.; Wilson, K. S. *J. Mol. Biol.* **1994**, *236*, 759−764.
(64) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. *Cell* **1998**, *95*, 927−937.
(65) Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. *Nature* **1998**, *395*, 137−143.
(66) Wagner, R. L.; Darimont, B. D.; Apriletti, J. W.; Stallcup, M. R.; Kushner, P. J.; Baxter, J. D.; Fletterick, R. J.; Yamamoto, K. R. *Genes Dev.* **1998**, *12*, 3343−3356.
(67) Rahuel, J.; Priestle, J. P.; Grutter, M. G. *J. Struct. Biol.* **1991**, *107*, 227−236.
(68) Fujinaga, M.; Chernaia, M. M.; Tarasova, N. I.; Mosimann, S. C.; James, M. N. *Protein Sci.* **1995**, *4*, 960−972.
(69) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Hosur, M. V., Sowder II, R. C.; Cachau, R. E., Collins, J.; Silva, A. M.; Erickson, J. W. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6796−6800.
(70) Pearl, L.; Blundell, T. *FEBS Lett.* **1984**, *174*, 96−101.
(71) Gillmor, S. A.; Craik, C. S.; Fletterick, R. J. *Protein Sci.* **1997**, *6*, 1603−1611.
(72) Turk, D.; Podobnik, M.; Popovic, T.; Katunuma, N.; Bode, W.; Huber, R.; Turk, V. *Biochemistry* **1995**, *34*, 4791−4797.
(73) Yamamoto, A.; Tomoo, K.; Doi, M.; Ohishi, H.; Inoue, M.; Ishida, T.; Yamamoto, D.; Tsuboi, S.; Okamoto, H.; Okada, Y. *Biochemistry* **1992**, *31*, 11305−11309.
(74) Okamoto, Y.; Anan, H.; Nakai, E.; Morihira, K.; Yonetoku, Y.; Kurihara, H.; Sakashita, H.; Terai, Y.; Takeuchi, M.; Shibanuma, T.; Isomura, Y. *Chem. Pharm. Bull.* **1999**, *47*, 11−21.
(75) Lee, A. Y.; Smitka, T. A.; Bonjouklian, R.; Clardy, *J. Chem. Biol.* **1994**, *1*, 113−117.
(76) Gaboriaud, C.; Serre, L.; Guy-Crotte, O.; Forest, E.; Fontecilla-Camps, J. C. *J. Mol. Biol.* **1996**, *259*, 995−1010.
(77) Tsunemi, M.; Matsuura, Y.; Sakakibara, S.; Katsube, Y. *Biochemistry* **1996**, *35*, 11570−11576.
(78) Bode, W.; Mayr, I.; Baumann, U.; Huber, R.; Stone, S. R.; Hofsteenge, J. *EMBO J.* **1989**, *8*, 3467−3475.
(79) Chevrier, B.; D'Orchymont, H.; Schalk, C.; Tarnus, C.; Moras, D. *Eur. J. Biochem.* **1996**, *237*, 393−398.
(80) Becker, J. W.; Marcy, A. I.; Rokosz, L. L.; Axel, M. G.; Burbaum, J. J.; Fitzgerald, P. M.; Cameron, P. M.; Esser, C. K.; Hagmann, W. K.; Hermes, J. D. *Protein Sci.* **1995**, *4*, 1966−1976.
(81) Holden, H. M.; Tronrud, D. E.; Monzingo, A. F.; Weaver, L. H.; Matthews, B. W. *Biochemistry* **1987**, *26*, 8542−8553.
(82) Lovejoy, B.; Welch, A. R.; Carr, S.; Luong, C.; Broka, C.; Hendricks, R. T.; Campbell, J. A.; Walker, K. A.; Martin, R.; Van Wart, H.; Browner, M. F. *Nat. Struct. Biol.* **1999**, *6*, 217−221.
(83) Kim, H.; Lipscomb, W. N. *Biochemistry* **1991**, *30*, 8171−8180.

CI049814F