Pocket Similarity: Are α Carbons Enough?

Howard J Feldman* and Paul Labute

Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, Canada H3A 2R7

Received May 26, 2010

A novel method for measuring protein pocket similarity was devised, using only the α carbon positions of the pocket residues. Pockets were compared pairwise using an exhaustive three-dimensional Cα common subset search, grouping residues by physicochemical properties. At least five Cα matches were required for each hit, and distances between corresponding points were fit to an Extreme Value Distribution resulting in a probabilistic score or likelihood for any given superposition. A set of 85 structures from 13 diverse protein families was clustered based on binding sites alone, using this score. It was also successfully used to cluster 25 kinases into a number of subfamilies. Using a test kinase query to retrieve other kinase pockets, it was found that a specificity of 99.2% and sensitivity of 97.5% could be achieved using an appropriate cutoff score. The search itself took from 2 to 10 min on a single 3.4 GHz CPU to search the entire Protein Data Bank (133 800 pockets), depending on the number of hits returned.

INTRODUCTION

It is well understood that most enzymes work through a lock-and-key mechanism, with the binding site on the receptor fitting well to the intended ligand or ligands.[1] Thus it is useful to be able to measure similarity between protein structures in order to predict what ligands a protein of interest is likely to interact with. This can be important when trying to determine possible side effects of a new drug, for example.

In many cases sequence or structural similarity (rmsd) is sufficient to determine proteins of similar function. However, there are cases where proteins of seemingly different structure still have highly conserved binding sites and function.[2] This can occur as a result of divergent/convergent evolution[3] or domain swapping,[4−6] for example. In such cases, a more sophisticated approach is required to identify similar binding pockets without regard to the rest of the protein.

The problem can be further broken down into two distinct parts: pocket identification and similarity. Numerous methods have been used to identify likely 'druggable' pockets on the surface of a protein. Often a geometric- or energy-based calculation is involved, with the binding site being represented as a point set,[7] surface patches,[8] or by physicochemical properties.[9,10] Once a database of pockets has been generated, it can then be searched.

In this work we focus on the latter subproblem, a new method to compare and score protein−ligand binding pockets. Previous methods of measuring similarity have employed numerous techniques, such as graph theory,[11,12] geometric,[13,14] typed triangles,[15] spherical harmonics,[16] and physicochemical properties[17,18] of the site atoms, or even combinations of these.[19] The actual search may be done using a genetic algorithm,[20] geometric hashing,[21−23] or graph matching.[24,25] Here we set out to test the hypothesis that Cα positions and residue identities contain sufficient information to distinguish similar pockets. This allows a much simplified view of the pocket and implies that ligand recognition is for the most part encoded in small portions of the protein backbone.

METHODS

Pocket Database. A database of protein pockets was built and used to search against. For the purpose of developing a pocket similarity score, we make a similar simplifying assumption to other recent studies[14−16,22] taking the binding sites to be those residues within some short distance of a bound ligand. Pocket identification is thus treated as a separate problem. Preliminary tests of protein family clustering using a simple ligand-independent pocket definition produced comparable results, supporting this assumption (see Results and Discussion).

For each structure in the Protein Data Bank (PDB),[26] as of April 20, 2010, a pocket was defined as the set of residues within 4.5 Å of a ligand heavy atom, for each ligand in the structure. A ligand was defined as any nonprotein, nonnucleic acid molecule containing at least one torsion angle. This definition included covalently bound ligands, but only when they were represented as separate chains in the PDB file. A total of 133 800 ligand pockets were included in the pocket database, and for each, the coordinates of the Cα atoms and identities of the residues comprising the pocket were stored. In addition, pseudo-Cβ atom positions were stored for each residue to represent average side chain centroids.[27] These were positioned 2.4 Å from the Cα in the direction of the Cβ. For glycine and residues with missing side chains, the position where the Cβ would be was estimated from the local geometry.

Search. For a given query pocket, a similarity search against the pocket database was carried out by performing an exhaustive three-dimensional (3D) Cα common subset search. That is, for each candidate pocket in the database, all possible subsets of Cα from the query were optimally

* Corresponding author. E-mail: hfeldman@chemcomp.com. Telephone: +1-514-393-1055 ext. 144.

10.1021/ci100210c
Published on Web 08/06/2010

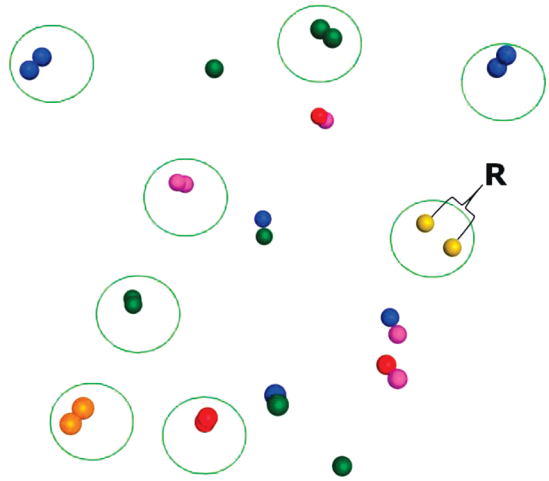

Figure 1. Graphical representation of the Cα's (colored spheres) used during the pocket similarity search process. Only spheres of the same color (i.e., residue class) count as a 'match' (circled) and count toward the score. The score also depends on the distance, R, between matched points. Points with $R > 1$ Å are not considered a match even if they are the same color.

Table 1. Amino Acid Equivalence Classes Used to Detect Pocket Residue Matches

class no.	class members
1	Phe, Ile, Leu, Met, Val, Cys
2	His, Asn
3	Ala, Cys, Xxx[a]
4	Ser, Thr
5	Asp, Glu
6	Gln, Glu
7	Arg, Gln
8	Lys, Arg
9	Phe, Trp, Tyr
10	Pro
11	Gly
12	Cys[b]

[a] Any nonstandard amino acid. [b] Half-cystine (disulfide-bonded Cys).

superposed on all possible subsets in the target. Once superposed, atom pairs that were within 1 Å of each other and were of the same residue class (see below) were considered to be a 'match'. A minimum of five matches were required for a target pocket to be considered a 'hit' to the query (see Figure 1) (thus all pockets were required to consist of at least five residues). In all cases, only the superposition with the largest number of matches was retained.

Since many residues are highly conserved in active sites, the search could be greatly sped up by allowing residues to match only when they were similar. Using a residue classification greatly reduces the number of possible superpositions that need to be tested but if not chosen carefully could also exclude valid hits. The set of allowable classes was derived from studies of amino acid substitutions in protein cores,[28,29] which should be comparable to most active site environments, and is given in Table 1. Thus any mutations which are not within the same class will not count toward the number of matches in the hit. Note that some residues, such as glutamate, appear in multiple classes, and cysteine is distinguished from half-cystine. If backbone interactions frequently play an important role in ligand binding, independent of the amino acid identity, then many

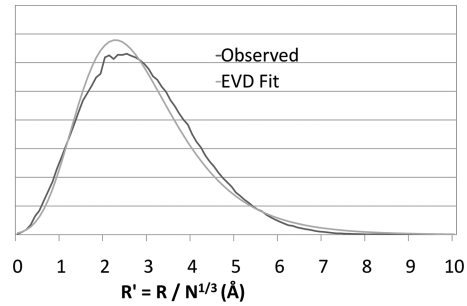

Figure 2. Distribution of normalized distance R' between pairs of randomly superposed Cα's for a given query pocket compared to each pocket in the database (dark curve). The EVD of best fit is shown by the light curve.

perfectly valid hits would be missed using this approach. Through the tests and examples outlined in the Results and Discussion Section, however, we feel that the number of potential hits missed due to residue classification is minimal.

Scoring. Although the number of matches between a query and target pocket gives a crude indication of how well the pockets match, a more sensitive scoring function was needed, especially for weak hits with only five or six matches. Investigations of the statistical distribution of randomly superposed proteins have shown that root mean square deviation (rmsd) fits well to an extreme value or Gumbel distribution (EVD).[30,31] Hence we chose to use the EVD to derive a statistics-based scoring function for pocket superposition with the goal of predicting the statistical significance of a given hit. The probability that a given random variable X which follows the EVD is less than or equal to some value x is given by

$$p(X \leq x) = e^{-e^{\langle \mu - x \rangle / \beta}} \quad (1)$$

where μ is the mode of the distribution, and β is proportional to the standard deviation.

Our goal was to estimate the probability of observing a particular pocket superposition by chance, from which a scoring function could easily be derived. To do so, a test set of 135 sites from 135 distinct, randomly chosen PDB files (available as Supporting Information), with a minimum of 5 Cα in each site was selected. The structures in this set comprised a diverse set of enzyme commission (EC) numbers in similar proportions to the distribution of EC numbers in the PDB. Each of these sites was superposed randomly to all 133 800 pockets in the pocket database—that is, random correspondences were chosen between the query and database site Cα's—and the distribution of distances R between corresponding Cα's was recorded. The number of correspondences was also randomized (but always at least five). This resulted in 135 distance distributions, each of which was fit to an EVD. An example of one such distribution is shown in Figure 2. The distance between corresponding points in a random superposition is expected to vary roughly with the cube root of the volume of the pocket, which is in turn proportional to the radius of gyration of the pocket (R_{gyr}). Hence R was normalized by the number of correspondences, N to give $R' = R/N^{1/3}$. This assumes that N is roughly proportional to the pocket volume.

As mentioned above, an EVD is defined by only two parameters: μ and β. Following the normalization step, the 135 EVD distributions were found to all have very similar

μ and β, however, they still varied slightly with the pocket size and shape. Two parameters of the query pocket were found to affect the EVD parameters: R_{gyr} and the shape D of the pocket Cα's. This latter term is defined as the variance along the longest axis (first principal component) through the set of points. It is computed as the largest eigenvalue of the covariance matrix of the points and gives a measure of the globularity or eccentricity of the points. Linear regression against these two parameters resulted in a fit for

$$\mu = 1.42 + 0.16R_{gyr} - 0.0048D \qquad (R^2 = 0.75) \tag{2}$$

$$\beta = 0.75 + 0.062R_{gyr} - 0.0014D \qquad (R^2 = 0.79) \tag{3}$$

These were used to derive the specific EVD distribution expected for an arbitrary given query pocket. It is important to note that the initial choice of pockets has little bearing on the parameter fitting, as only the general size (R_{gyr}) and shape (D) of the pocket have any bearing on the result. Constant μ and β values could have been used, but we found that allowing small variations with the query pocket size and shape produced better fits to the EVD.

Putting all the pieces together, a statistical scoring function was then designed to rank the superposition of arbitrary pockets. Starting with eq 1 and making the approximation that the normalized corresponding Cα distances after superposition, R'_i, are independent random variables for each pair i, we know for a given pocket of size N, the probability of observing a given superposition by chance is

$$p(\vec{r} \le \vec{R'}) = \prod_{i=1}^{N} e^{-e^{(\mu - R'_i)/\beta}} \tag{4}$$

with μ and β, as determined by eqs 2 and 3, constant for a given pocket. Taking the negative logarithm of the probability results in an energy-like score, S_a:

$$S_a = -\ln p = \sum_{i=1}^{N} e^{(\mu - R'_i)/\beta} \tag{5}$$

Eq 5 gives a simple, easily computed score as a sum of exponentials. Note that μ has a typical value of about 2 Å, while β was close to 1 Å, so a pair of corresponding points about 0.5 Å apart would contribute around 4.5 to the total score, and a perfectly superposed pair contributes 7.4.

We wished to test whether the orientation of the side chains had any effect on the scoring, so an additional score, S_b, was computed using the pseudo-Cβ positions described above. S_b was computed exactly the same as S_a, and superposition was still performed using Cα's, however R'_i is now measured between corresponding pseudo-Cβs. The S_b scoring function was found to be marginally better than S_a through the following procedure. The main pocket from PDB code 1FIN chain A (CDK2) was used as the query, which resulted in 7691 candidate hits in the pocket database (i.e., those that superposed with at least 5 residue matches). A total of 1636 pockets contained Interpro[32] domain IPR000719 (protein kinase core domain) according to the InterProScan[33] program and were considered the true positives for this query. From this data, a receiver operating characteristic (ROC) curve could be computed for various cutoffs of each scoring function, with the results shown in Figure 3. As shown, both scoring functions work remarkably well at distinguishing true kinase pockets based on the query with near ideal ROC curves. However looking closely, S_b has a slightly larger area under the curve than S_a. As a result, S_b was chosen as the scoring function to use and will be referred to as simply S from now on. Further, from the ROC curve, a cutoff score of 26 was chosen which yields true and false positive rates of 97.5 and 0.8%, respectively. Thus any hits to a query with $S < 26$ were rejected, and those with $S > 26$ were retained as 'true' hits.

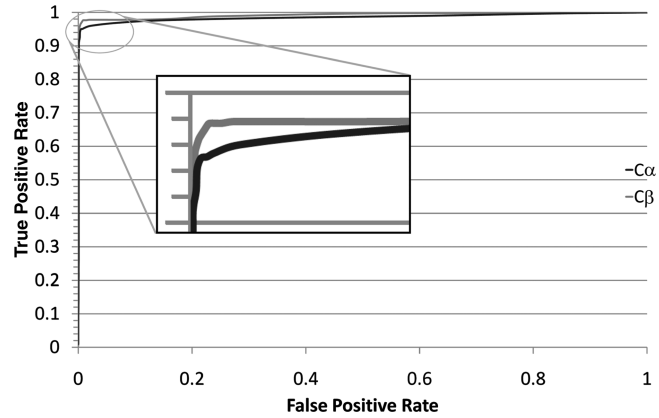

Figure 3. ROC curve for the S_a score (dark curve) which uses superposed Cα distances to compute the score, and the S_b score (light curve) which uses pseudo-Cβ distances. The inset shows finer detail where the curves differ most.

In summary then, for a given query pocket, first the μ and β for its EVD were computed with eqs 2 and 3. Next, it was superposed pairwise against every pocket in the database, maximizing the number of matched residues (matched Cα's within 1 Å of each other after superposition). If at least five residue matches were found, then the score S was computed as in eq 5. Recall that in computing S, $R'_i = R_i/N^{1/3}$, where R_i is the distance between pseudo-Cβs of matched pair i, and N is the number of residues matched. If S exceeds the cutoff of 26, the pocket is reported as a hit, otherwise it is assumed to be a false positive and ignored.

The algorithms described have been implemented in the Scientific Vector Language of the Molecular Operating Environment[34] version 2009.10. The pocket similarity search has also been integrated into the PSILO[35] database system version 2010.02.

RESULTS AND DISCUSSION

A novel scoring function was devised to search and superpose a given query pocket against a database of pockets. Pockets were compared pairwise using an exhaustive search of all possible Cα superpositions, and the highest scoring one retained. In order to test the performance of the new scoring function, a number of test cases of varying difficulty were examined. All tests were run against the entire pocket database of 133 800 pockets representing the entire PDB,[26] and timings are on a single 3.4 GHz Pentium D CPU.

Tyrosyl−tRNA Synthetase. TyrRS's role is to attach tyrosine to tyrosyl−tRNA for protein synthesis. This was considered an 'easy' query as this protein is quite distinctive and not closely related to any other non-tRNA synthetases,

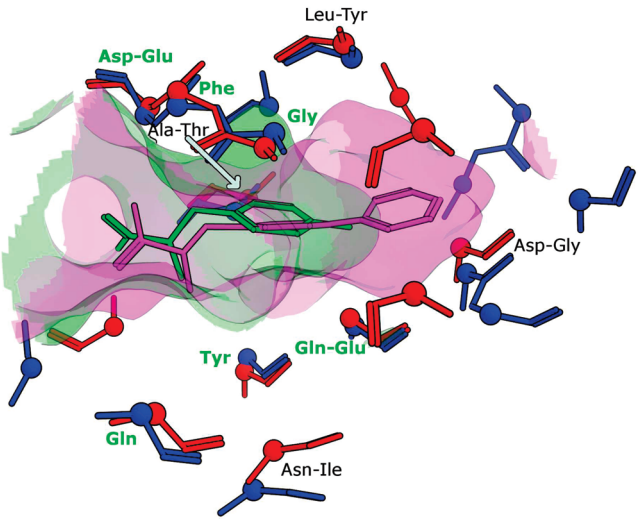

Figure 4. Optimal superposition of ligand pockets in the TyrRS query (1TYA chain E) and biptidylalanyl−tRNA synthetase (2PXH chain A), using only the matched pocket Cα's. 1TYA is shown in red with ligand and the pocket surface in green, and 2PXH is shown in blue with ligand and the pocket surface in purple. Only the backbones of pocket residues are shown for clarity, and Cα atoms are shown as spheres. Labels refer to the residue identities of superposed pairs of Cα's in the pocket. Green labels indicate the residues are identical or in the same equivalence class and count as a match. Red labels (if present) indicate that the residues would have counted as a match but are too far apart. Black labels indicate pairs which are not in the same equivalence class.

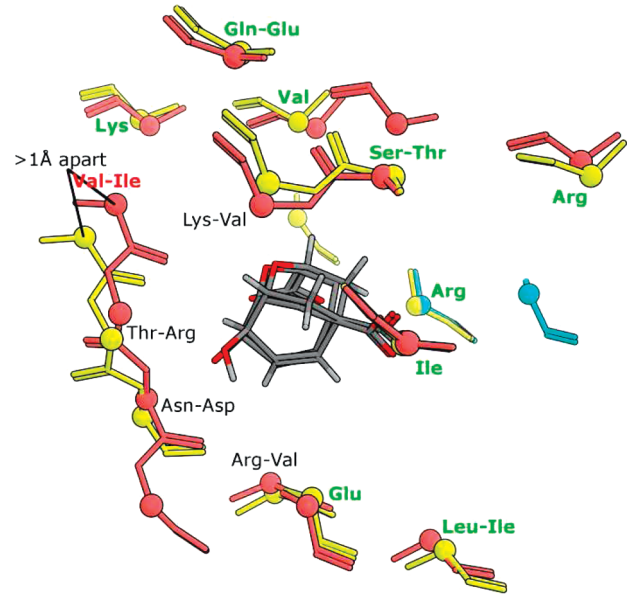

Figure 5. Optimal superposition of chorismate mutase ligand pockets: 4CSM chain B (yellow) and 1ECM (chain A cyan, chain B pink), using only the matched pocket Cα's. Only the backbones of pocket residues are shown, for clarity, and Cα atoms are shown as spheres. Labeling and display is as in Figure 4.

while all TyrRS have a high degree of sequence and structural similarity between species, being one of the fundamental building blocks of life. We searched for similar pockets to the tyrosine binding pocket of TyrRS (PDB 1TYA). The search took about 2 min to complete and produced 36 hits. The top hit, with a score of 111.6, was the query itself, 1TYA. This was followed by 25 more TyrRS's, all with scores above 40. Of the next seven hits, six were TyrRS mutants designed to encode unnatural amino acids, and the seventh (2J5B) was from mimivirus.[36] The 3 weakest hits, with scores below 30, were to TrpRS, interestingly from the same organism as the query, *Geobacillus Stearothermophilus*. This is not surprising since it is well-known that TyrRS and TrpRS are very closely related.[37]

To determine what structures, if any, were missed, we next performed a search for the gene ontology (GO)[38] term 'tyrosyl−tRNA aminoacylation' (GO:0006437) assigned by the InterProScan[33] program, which returned 41 hits. Since our current pocket database only includes pockets containing ligands, we must exclude nine of these hits which do not have ligands. This leaves 32 'true hits' that we expect to find (and do indeed) in the results of our pocket search. The four additional hits we find not classified by GO as TyrRS are the three TrpRS mentioned above, and the mimivirus TyrRS. Hence for this easy case, the search retrieved all known similar pockets as well as a few more distantly related ones which were correctly ranked near the bottom of the result set.

Figure 4 shows a superposition of the query pocket on the pocket from 2PXH, a low-scoring bipyridylalanyl−tRNA synthetase (score 33). This enzyme aminoacylates a much longer ligand than tyrosine and as a result requires a much deeper binding pocket. As shown, the outer half of the pockets and ligands superpose well, with six matched (i.e., substitutable) residues (just above our minimum requirement of five). In fact aside from the Ala−Thr pair, all residues where the pockets overlap are matched and thus highly conserved. But the pocket in 2PXH is much deeper than in 1TYA and is quite different from the query in the deepest region of the pocket, with no matching residues. Thus the lower score accurately reflects the fact that this pocket is only a partial match to the query.

Chorismate Mutase. Chorismate mutase is an enzyme involved in the biosynthesis of tyrosine and phenylalanine. This was a particularly interesting test case because *Escherichia coli* and *Saccharomyces cerevisiae* chorismate mutase are known to have very similar binding sites[2] yet less than 20% sequence identity. In fact, although they have similar secondary structure—a six-helix bundle—the *E.Coli* protein is an intertwined homodimer of two chains, while the *S. cerevisiae* protein has only one. It would be extremely difficult to find these two proteins through a standard sequence or structural similarity search.

The search was performed using *E.Coli* chorismate mutase (PDB 1ECM) as the query and took about 3 min to complete, resulting in 25 hits. Only 9 of these have a score above 30 however, which include 7 chorismate mutases and 2 isochorismate-pyruvate lyases, a low sequence identity structural homologue to chorismate mutase in the AroQα class.[39] *S. cerevisiae* chorismate mutase ranks fourth with a score of 54 resulting from 9 matched residues in the pocket, quite a strong hit. The superposition is shown in Figure 5 from which it is clear that there is a high degree of sequence similarity in the active site region, despite low overall sequence similarity. Again the ligands superpose almost perfectly since both proteins have the same binding mode, and the pocket Cα's also superpose well, though some deviation in the backbone is clearly visible. Specifically, the Val−Ile pair indicated in red in Figure 5, which normally would count as a match does not because the Cα's are more than 1 Å apart. There is still enough similarity in the remainder of the pocket

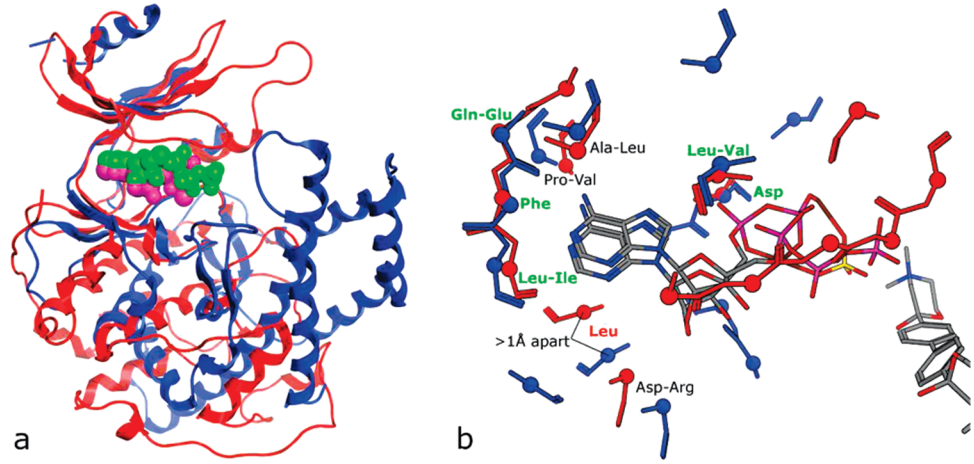

Figure 6. (a) Superposition of CDK2 (1FIN chain C) with choline kinase (3G15 chain B), a weak hit, using only the matched pocket Cα's for superposition. 1FIN is shown in red with ligand in green, and 3G15 is shown in blue with ligand in purple. Ligands are shown in spacefill for clarity. (b) Detailed view of ligand pockets of same chains as in a. Labeling and display is as in Figure 4.

in this case to clearly indicate that they are related. This example demonstrates that the method is able to detect similarities that a standard sequence search does not; neither BLAST nor PSI-BLAST find this hit, helping to confirm its utility as a bioinformatics tool. While the match does come up in a DALI[40] structure-based search, this is only as a weak hit ($Z = 6.6$, 18% identity), only slightly above several dozen false positives ($Z = 4.9$ and below).

Protein Kinase. Protein kinases represent one of the most diverse and well-studied family of proteins, with hundreds appearing in the human genome alone. Due to their key roles in cell signaling, kinases also make excellent drug targets for many diseases. Unlike the previous two examples, the active site can vary considerably from kinase to kinase, and they can have little or no structural similarity outside the core domain. Also with the sheer number of them in the PDB—well over 1000—it would be possible to get a precise estimate of the sensitivity of the method from the number of false positives and negatives resulting.

A cyclin-dependent kinase (CDK2) with ATP bound (1FIN chain C) was used as the query, and the search took under 9 min to complete, resulting in 1198 hits. A search for InterPro domain IPR000719 (protein kinase, core domain) assigned by InterProScan[33] retrieved 1481 hits, 1209 of which had ligands in or near the ATP-binding pocket. Of those, 1179 were found by the pocket search, and about 10 were not kinases at all (false positives from InterProScan). Thus our search missed 30 pockets labeled as kinase domains (false negatives) and found 19 pockets which were not protein kinases (false positives). It is instructive to look at an example from each of these cases.

Many of the 19 false positives (with a maximum score of 36) were in fact other types of kinases; choline kinase and 5-methylthioribose kinase were most common. Only four or five were not kinases at all (but still bound ADP/ATP). Figure 6a shows the superposition of the query (1FIN chain C) and one of the choline kinases with a score of 29 (3G15 chain B) using only the matched pocket Cα's to perform the superposition. Very good structural similarity in the hinge region, N lobe β-sheet, and catalytic region was evident, despite the complete lack of sequence similarity between the proteins and poor structural similarity elsewhere. 3G15 lacked density in the P loop region so it is unknown if this superposes well or not. The nucleotide portions of the ligands (Figure 6b) align quite well, and there are five conserved matches in the binding site: three in the hinge region, Asp from the DFG loop, and one residue from a strand in the N lobe. There was also a pair of Leu's in the catalytic region, but they were more than 1 Å apart and so were not counted as a match. These observations would suggest that the choline kinases are distantly related to the protein kinases which is supported by SCOP's[41] classification, placing them in the protein kinase-like superfamily. This is an example when the pocket search algorithm is able to find a remote homologue that is not detected using a basic sequence (BLAST) search. The hit does come up in a DALI[40] search also as a weak hit ($Z = 6.1$, 18% identity corresponding to similar fold).

An example of a weak hit that did not make it above the cutoff score of 26 (false negative) is a MAP kinase 14 (1WBS). Superposition with 1FIN (Figure 7a) shows that they do superpose reasonably well, however 1FIN is in the activated conformation, while 1WBS is in the inactivated (also called DFG out) form. Thus the location of the binding pocket is completely shifted between the two structures. Figure 7b shows the pockets in detail, with five matched Cα's: two from the hinge region, the Asp from the DFG loop and a conserved Val and conserved Ala in two strands of the N lobe. However, as seen in Figure 7b, both the Leu−Met pair and the conserved Asp have backbones pointing in very different directions due to the difference in protein conformation. This is a case where using the pseudo-Cβ positions to compute the score makes a difference. Looking only at the Cα's the five matched residues seem well superposed. However when the side chains are considered, the considerably different orientations of these two residue pairs down weight the overall pocket score sufficiently to give it only a 23, below our threshold of 26. The pockets themselves have very little in common in this example—less than half the residues comprising the pocket are shared between the two conformations—and so it is quite reasonable that this structure does not show up as a hit to our query. Eleven of the false negatives were in the inactivated conformation, while some of the other missed cases were simply due to there being too poor of a

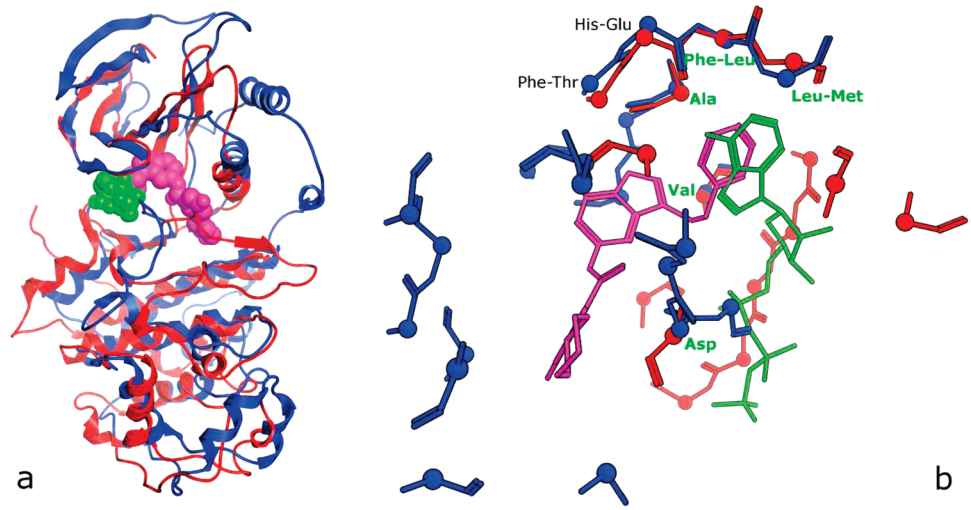

Figure 7. (a) Superposition of CDK2 (1FIN chain C) with mitogen-activated protein (MAP) kinase 14 (1WBS chain A), a kinase pocket missed by the search, using only the matched pocket Cα's for superposition. 1FIN is shown in red with ligand in green, and 1WBS is shown in blue with ligand in purple. Ligands are shown in spacefill for clarity. (b) Detailed view of ligand pockets of same chains as in a. Labeling and display is as in Figure 4.

superposition or the definition of the pocket not extending out quite far enough from the ligand (not shown).

Clustering Protein Families. It is clear from the previous examples that the pocket search algorithm and the scoring function are able to pick out similar pockets with a high degree of accuracy from proteins of related structure or function. Taking this one step further, we wanted to establish how well the scoring function worked at discriminating members of different protein families from each other. Kuhn et al.[42] have previously shown the ability of the Cavbase clustering procedure to group proteins from diverse families, so we have adopted their test set. However, since our pocket database only includes pockets with ligands bound, we had to exclude those structures from the test set that were without ligand, resulting in a total of 85 structures from 13 protein families.

The clustering procedure was as follows. All pockets from the 85 structures were compared pairwise, and the score S was computed using eq 5. For structures which contained more than one ligand-binding pocket in our database, we chose the one which maximized the overall pairwise similarity scores. This was accomplished using unary quadratic optimization (UQO)[43] which solves the following problem. Given an $n \times n$ symmetric matrix $\mathbf{G}$, an n-vector g and a scalar g_0, minimize the function:

$$F(x) = 0.5x^T\mathbf{G}x + g^T + g_0 \qquad (6)$$

subject to the constraint that x is in $\{0,1\}^n$, and the elements of x are divided into m partitions in each of which exactly one occurs, with the other elements zero. The x can be thought of as a state vector, composed of m independent variables each with multiple possible mutually exclusive states. In the present usage, the variables are the m structures being clustered, and the states for each are the various pockets we must choose from. The outcome of the UQO procedure was the set of states, or pockets, one per structure, which resulted in the largest sum of pairwise scores overall, and an 85 × 85 matrix of these optimal scores. From this **S** matrix, a new matrix **D** was computed where

$$\mathbf{D}_{ij} = \frac{\mathbf{S}_{ii} + \mathbf{S}_{jj}}{2} - \mathbf{S}_{ij}, \quad 1 \le i, \quad j \le 85 \qquad (7)$$

This converted the scores **S** into a distance, **D**, which could then be clustered by standard techniques. In this case, the hierarchical agglomerative clustering method described by Ward[44] was used, resulting in the tree diagram shown in

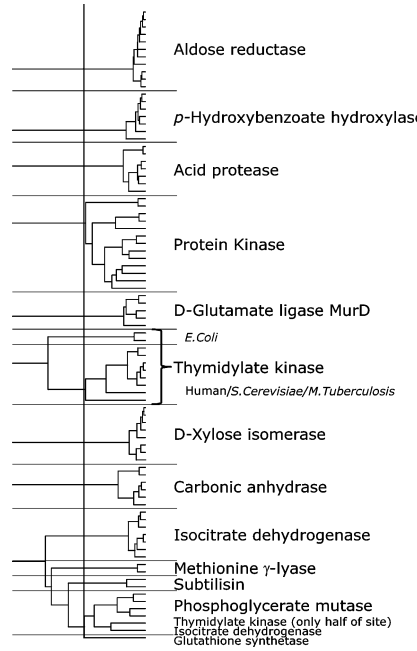

Figure 8. Hierarchical agglomerative clustering tree for 85 proteins from 13 diverse families. Pockets were clustered using Ward's method, and the vertical line represents an arbitrary cutoff to break the tree into clusters. The classification of all the structures in each cluster is shown to the right. Each family was put into its own cluster with the exception of thymidylate kinase, which was split into two subclusters as indicated, and the phosphoglycerate mutase, which had an isocitrate dehydrogenase and a thymidylate kinase pocket lumped into its cluster.

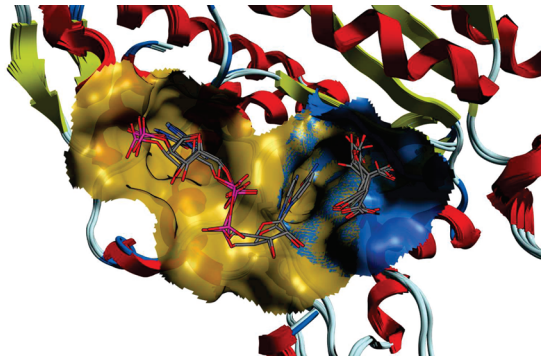

Figure 9. Superposition of the eight isocitrate dehydrogenases in the test set. The protein has two pockets, resulting in two clusters. In gold, on the left, is the NADP cofactor binding site, while on the right, in blue, is the isocitrate ligand binding pocket. Some structures have both pockets occupied, some only the NADP site, and some only the isocitrate one.

Figure 8. The proteins clustered quite clearly into their respective families, with two exceptions.

First, one of the isocitrate dehydrogenases (9ICD) was clustered with the phosphoglycerate mutase family. This is somewhat of an artifact of the data set, however, isocitrate dehydrogenase contains two binding sites, one for the isocitrate ligand and one for the nicotinamide adenine dinucleotide phosphate (NADP) cofactor. As shown in Figure 9, some of the structures (1CW1, 1GRO, 1GRP, and 8ICD) have only the isocitrate pocket filled (and thus in our pocket database), 9ICD contains only NADP, while 1AI2, 1AI3, and 1BL5 have both pockets filled. This results in two distinct clusters, one for each pocket depending on which pocket was chosen by the UQO in each case. The isocitrate pocket was chosen for all but 9ICD, which explains why this structure ended up separate from its family members. Once we are able to include apo-pockets in our database then we would expect this problem to go away, as all family members would have both pockets present.

Second, one of the thymidylate kinase structures (1E9A) has its binding pocket only partially occupied, and since our current definition of pocket residues includes only those atoms close to the ligand, the resulting pocket is much smaller than the other members of this family, and it ends up erroneously clustered with the phosphoglycerate mutases. This again is a result of our pocket definition, and changing our definition of pocket to be independent of the ligand would remedy this. The reason both 9ICD and 1E9A cluster with the phosphoglycerate mutase family, despite having no similarity whatsoever to them, is simply an artifact of our distance definition (eq 7). When $\mathbf{S}_{ij}$ is zero, $\mathbf{D}_{ij}$ will be smallest when $\mathbf{S}_{ii}$ and $\mathbf{S}_{jj}$ are minimized, which in turn occurs for the smallest sized pockets. This means small pockets with no similarity to anything else will tend to be clustered together, which is what is happening here.

Also worthy of note is that the thymidylate kinases from *E. coli* form a distinct cluster from the other thymidylate kinases. Lavie et al.[45] have reported that the former operate through a side-on interaction of Glu-12 with the 3′-hydroxyl of deoxythymidine monophosphate (dTMP), while the *S. cerevisiae* and the human versions of the protein have a different, bidentate mode of interaction between the Asp-14 of the P loop and the 3′-hydroxyl. The scoring function seems to be sensitive enough to distinguish this different mode of action at the active site.

Kuhn et al. are able to achieve perfect clustering but only for two out of a dozen combinations of clustering methods and scoring functions. Although we only tested hierarchical agglomerative clustering methods, they also found that agglomerative, and partitional, clustering work best. Choosing exactly how many clusters to create is a problem for both methods, however the approximate number of clusters can be determined fairly easily by manual inspection of the dendrogram. The few clustering errors we do observe are due to the limitations of our pocket definition rather than to the scoring scheme itself, and it is encouraging to see that the present, much simpler method seems to perform almost as well at distinguishing protein families. To help measure the dependence of the results on our pocket definition, a simple ligand-independent pocket definition (based on cavities generated by α spheres)[46] was tested as well, resulting in 80% of the structures being correctly clustered. No attempt was made to optimize this pocket definition however. It is difficult to compare this result to others as it is not clear from published accounts whether the pockets are manually selected from each structure or all pockets are considered as we have done here. Nevertheless we feel that the scoring method is not highly sensitive to the pocket definition.

A second test Kuhn et al. performed was to classify kinases into subfamilies. Again, using the technique described above and excluding three structures which had no ligands, we clustered the remaining set of 25 kinases into subfamilies. This is a considerably more difficult task than clustering protein families, as most of the proteins share similar or identical ligands. This is a real test of a method's ability to distinguish similarity in binding sites. In this case, we found average linkage hierarchical clustering to produce slightly better clusters than Ward, so it was used. Otherwise the procedure was carried out in the exact same way as the previous example, resulting in the tree shown in Figure 10.

Again there were some interesting exceptions. 1PME is an extracellular regulated protein kinase 2 (ERK2) bound to a p38 inhibitor and is found in a cluster by itself, in between the p38αs and the ERK2s. This is not all that surprising since it shares properties of both subfamilies. While the cAMP-dependent kinases do all cluster close together, there are also clearly three subclusters which are grouped roughly by ligand similarity and size: those with ATP/ANP bound (1ATP and 1CDK) and those with smaller ligands or inhibitors bound. The CDK2s also get split up into two subclusters, though this appears to be due to some genuine similarity between their binding pockets. For example 1FIN (CDK2) has 13 matches to 1YDS (cAMP) and only 12 to the members of the other CDK2 cluster (1B38 and 1HCK). Interestingly, our preliminary tests using the ligand-independent pocket definition (see above) result in all but one of the cAMPs forming one large cluster, and three of the four CDK2s forming a single cluster, so future work on tuning the pocket definition will likely improve subfamily discrimination further.

The last notable case is the tyrosine kinases (TK). These all cluster into their own clusters rather than with each other. Inspection shows that their pockets have several corresponding Cα greater than 1 Å apart and so are missed by our search. However, the three selected in this test set also have quite different phosphorylation targets: 1IR3 is an insulin

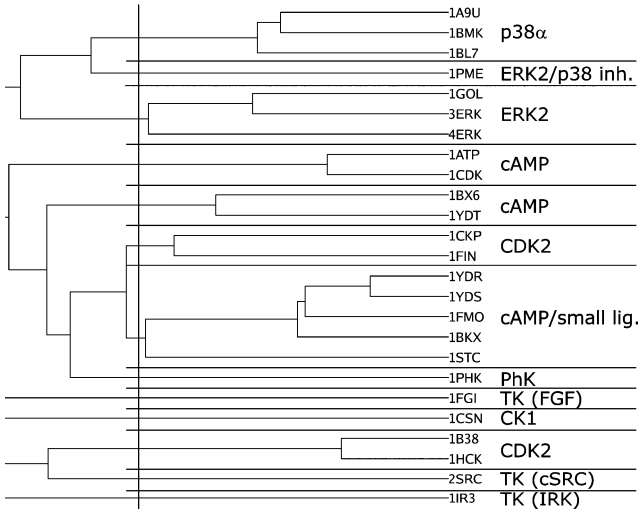

Figure 10. Hierarchical agglomerative clustering tree for 25 kinases belonging to 9 subfamilies. Pockets were clustered using the average linkage method, and the vertical line represents an arbitrary cutoff to break the tree into clusters. The classification of all the structures in each cluster is shown to the right. Each family was put into its own cluster by the procedure, and the cAMP, CDK2, and extracellular regulated protein kinase 2 (ERK2) families were broken up into two or more clusters as indicated. The lone structure between the p38α and ERK2 families was an ERK2 with a p38 inhibitor, and the largest cAMP cluster all had very small ligands in their binding pocket.

receptor kinase, 1FGI is from fibroblast growth factor, and 2SRC is a cSrc, so it seems reasonable to put these in separate clusters.

This is a more difficult test to judge because there is no single correct answer as with the previous example of clustering families. Often the differences between kinase subfamilies can be very subtle. Nevertheless the method seems to do a good job of grouping the structures according to the commonly used kinase classifications. Comparing again to Kuhn et al.'s findings, they produce a total of six clusters most of which are similar to ours. The main differences are in our splitting of the CDK2 and cAMP families; in the placement of casein (1CSN) and phosphorylase (1PHK) kinases which they cluster into a single cluster with 1STC (a cAMP dependent kinase); and the tyrosine kinases: Kuhn et al. group 1FGI with 1IR3 and 2SRC is lumped in with the CDK2s, while we put each in a cluster by itself. Thus the two approaches both make different types but similar numbers of misclassifications. Comparing to the SCOP 1.75[41] clustering at the protein level of the hierarchy, our clustering is 'perfect' aside from the case of 1PME and the subdividing of the CDK2s and cAMPs already mentioned. Hence despite the fact that very little information is being used to represent the binding pockets, it seems the Cα and residue types contain enough information to distinguish the subtle variations in the different kinase subfamilies with a good degree of accuracy.

What can be learned from the results that have been observed? The only information fed into the method are the Cα and pseudo-Cβ positions (which can be calculated from the Cα alone if necessary) and the amino acid identities. It is noteworthy that no information about potential hydrogen bonds, salt bridges, or even van der Waals interactions are explicitly included. Sufficient information required to identify a family of pockets, even with very weak sequence or structural similarity, is encoded in the protein backbone. Details of exact side chain conformation seem to be unimportant, for searching purposes at least. This allows for a much simplified representation of protein binding sites and for very fast searches, taking only minutes to search the entire PDB on a single CPU. Another advantage of the method over others is that it may be used when only backbone coordinates, or even only Cα coordinates, are available. Very accurate pocket superpositions, including the ligands, can be obtained using this minimal representation. Perhaps its most attractive feature is its simplicity. The method cannot withstand too many mutations in the active site of a protein, but the examples suggest that it is very unusual for proteins of the same family to have vastly different residues in the binding pocket. When a change in conformation is involved, such as the active and inactive forms of the protein kinases, the method will fail to detect similarity when there is significant conformational change in the pocket; however, if at least five or six residues can still be superposed well, then even these similarities can be detected.

Lastly, there are cases where detecting similarity cannot ever hope to succeed using a Cα-based method. For example, androgen receptor (2AM9) and 17β-hydroxysteroid dehydrogenase (1JTV) both bind testosterone, and the void volume of their pockets shows significant overlap. But their structures are completely different, and Cα's do not superpose at all. Only by using a surface- or volume-based approach could one hope to find similar pockets of this type. In this study, the free parameters of the method (1 Å superposition cutoff, minimum of 5 matches, and the significance cutoff score) were biased toward drug-sized molecules. With these settings, examples of convergent evolution[13] will not generally be detected, as the Cα's simply do not superpose well enough. They are detected with different parameters, but we did not retrain the statistical model for these settings. Relaxing the free parameters results in more distant relatives and binding motifs being detected but may also result in fewer statistically significant hits.

CONCLUSIONS

We have devised a simple energy-like scoring function for protein pocket superposition based on the extreme value distribution. By using an exhaustive search of all possible Cα superpositions to compare sites in a pairwise fashion in conjunction with this score, we are able to accurately pick out similar pockets from a large database with little or no error. We have also demonstrated the method's ability to reproduce standard protein classifications, through clustering, at both the family and subfamily levels. The search is fast enough to be run on a single CPU over the whole PDB, taking only a few minutes to complete, and has the advantage of being insensitive to missing or poorly resolved side chains. Its great speed can be attributed to the use of an efficient representation of the binding pockets and the use of residue equivalence classes, which permit a rapid reduction in the number of possible superpositions to be considered.

The pocket database used in this work was slightly biased due to the definition of the pocket residues depending on the ligand present, however our tests suggest that this bias is small. This would only be noticeable in cases of the same pocket binding ligands of vastly different size. If a more

robust pocket definition is used, independent of ligand atoms, then it is likely that this will not be a problem.

There are certainly cases of completely unrelated proteins binding the same ligand where the pocket Cα's simply do not superpose at all, and to find this sort of relation, a completely different kind of technique, such as a pocket volume-based approach, would be needed. Similarly, cases of unrelated proteins with similar motifs but where the Cα's do not superpose well will be missed. But it is encouraging to see that a very simple representation can go a long way for many cases even with very low sequence similarity. The pocket search algorithm is available as part of PSILO[35] version 2010.02.

ACKNOWLEDGMENT

We thank K. Kelly and J. Maier for assistance with the kinase searches and C. Williams and E. Sourial for helpful discussions and suggestions.

Supporting Information Available: The list of 135 PDB codes used in parametrizing the EVD for use in the pocket similarity scoring function. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Lehninger, A. L. Enzymes. In *Principles of Biochemistry*; Anderson, S., Fox, J., Eds.; Worth Publishers, Inc.: New York, 1982.
(2) Rosen, M.; Lin, S. L.; Wolfson, H.; Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.* **1998**, *11*, 263–277.
(3) Bruns, C. M.; Nowalk, A. J.; Arvai, A. S.; McTigue, M. A.; Vaughan, K. G.; Mietzner, T. A.; McRee, D. E. Structure of Haemophilus influenzae Fe(+3)-binding protein reveals convergent evolution within a superfamily. *Nat. Struct. Biol.* **1997**, *4*, 919–924.
(4) Hakansson, M.; Linse, S. Protein reconstitution and 3D domain swapping. *Curr. Protein Pept. Sci.* **2002**, *3*, 629–642.
(5) Bennett, M. J.; Schlunegger, M. P.; Eisenberg, D. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **1995**, *4*, 2455–2468.
(6) Zegers, I.; Deswarte, J.; Wyns, L. Trimeric domain-swapped barnase. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 818–822.
(7) Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
(8) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
(9) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.* **2007**, *47*, 2287–2292.
(10) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* **2007**, *47*, 400–406.
(11) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **1994**, *243*, 327–344.
(12) Konc, J.; Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160–1168.
(13) Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **1998**, *279*, 1211–1227.
(14) Das, S.; Kokardekar, A.; Breneman, C. M. Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.* **2009**, *49*, 2863–2872.
(15) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.
(16) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **2007**, *368*, 283–301.
(17) Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *11*, 99.
(18) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755–1778.
(19) Coleman, R. G.; Sharp, K. A. Protein pockets: inventory, shape, and comparison. *J. Chem. Inf. Model.* **2010**, *50*, 589–603.
(20) Lehtonen, J. V.; Denessiouk, K.; May, A. C.; Johnson, M. S. Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm. *Proteins* **1999**, *34*, 341–355.
(21) Weskamp, N.; Kuhn, D.; Hullermeier, E.; Klebe, G. Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics* **2004**, *20*, 1522–1526.
(22) Gold, N. D.; Jackson, R. M. A searchable database for comparing protein-ligand binding sites for the analysis of structure-function relationships. *J. Chem. Inf. Model.* **2006**, *46*, 736–742.
(23) Fischer, D.; Bachar, O.; Nussinov, R.; Wolfson, H. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn.* **1992**, *9*, 769–789.
(24) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–145.
(25) Kinoshita, K.; Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12*, 1589–1595.
(26) Kouranov, A.; Xie, L.; de la Cruz, J.; Chen, L.; Westbrook, J.; Bourne, P. E.; Berman, H. M. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **2006**, *34*, D302–305.
(27) Bryant, S. H.; Lawrence, C. E. An empirical energy function for threading protein sequence through the folding motif. *Proteins* **1993**, *16*, 92–112.
(28) Overington, J.; Donnelly, D.; Johnson, M. S.; Sali, A.; Blundell, T. L. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1992**, *1*, 216–226.
(29) Rodionov, M. A.; Blundell, T. L. Sequence and structure conservation in a protein core. *Proteins* **1998**, *33*, 358–366.
(30) Feldman, H. J.; Hogue, C. W. Probabilistic sampling of protein conformations: new hope for brute force. *Proteins* **2002**, *46*, 8–23.
(31) Jia, Y.; Dewey, T. G. A random polymer model of the statistical significance of structure alignment. *J. Comput. Biol.* **2005**, *12*, 298–313.
(32) Hunter, S.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Laugraud, A.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Mulder, N.; Natale, D.; Orengo, C.; Quinn, A. F.; Selengut, J. D.; Sigrist, C. J.; Thimma, M.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H.; Yeats, C. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **2009**, *37*, D211–215.
(33) Zdobnov, E. M.; Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* **2001**, *17*, 847–848.
(34) *Molecular Operating Environment*, version 2009.10; Chemical Computing Group: Montreal, Canada, 2009.
(35) *PSILO*, version 2010.02; Chemical Computing Group: Montreal, Canada, 2010.
(36) Abergel, C.; Rudinger-Thirion, J.; Giege, R.; Claverie, J. M. Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS. *J. Virol.* **2007**, *81*, 12406–12417.
(37) Doublie, S.; Bricogne, G.; Gilmore, C.; Carter, C. W., Jr. Tryptophanyl-tRNA synthetase crystal structure reveals an unexpected homology to tyrosyl-tRNA synthetase. *Structure* **1995**, *3*, 17–31.
(38) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29.
(39) Zaitseva, J.; Lu, J.; Olechoski, K. L.; Lamb, A. L. Two crystal structures of the isochorismate pyruvate lyase from Pseudomonas aeruginosa. *J. Biol. Chem.* **2006**, *281*, 33441–33449.
(40) Holm, L.; Kaariainen, S.; Wilton, C.; Plewczynski, D. Using Dali for structural comparison of proteins. *Curr. Protoc. Bioinformatics.* **2006**, Chapter 5, Unit 5.5.
(41) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.

(42) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023–1044.
(43) Labute, P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **2009**, *75*, 187–205.
(44) Ward, J. H. J. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
(45) Lavie, A.; Ostermann, N.; Brundiers, R.; Goody, R. S.; Reinstein, J.; Konrad, M.; Schlichting, I. Structural basis for efficient phosphorylation of 3′-azidothymidine monophosphate by Escherichia coli thymidylate kinase. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14045–14050.
(46) Edelsbrunner, H.; Facello, M.; Liang, J. On the definition and the construction of pockets in macromolecules. *Pac Symp Biocomput. 96* **1996**, 272–87.

CI100210C