doi:10.1016/S0022-2836(03)00882-9

J. Mol. Biol. (2003) **332**, 505–526

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns

T. Andrew Binkowski, Larisa Adamian and Jie Liang*

Department of Bioengineering
SEO, MC-063, University of
Illinois at Chicago, 851 S.
Morgan Street, Room 218
Chicago, IL 60607-7052, USA

*Corresponding author

We describe a novel approach for inferring functional relationship of proteins by detecting sequence and spatial patterns of protein surfaces. Well-formed concave surface regions in the form of pockets and voids are examined to identify similarity relationship that might be directly related to protein function. We first exhaustively identify and measure analytically all 910,379 surface pockets and interior voids on 12,177 protein structures from the Protein Data Bank. The similarity of patterns of residues forming pockets and voids are then assessed in sequence, in spatial arrangement, and in orientational arrangement. Statistical significance in the form of E and p-values is then estimated for each of the three types of similarity measurements. Our method is fully automated without human intervention and can be used without input of query patterns. It does not assume any prior knowledge of functional residues of a protein, and can detect similarity based on surface patterns small and large. It also tolerates, to some extent, conformational flexibility of functional sites. We show with examples that this method can detect functional relationship with specificity for members of the same protein family and superfamily, as well as remotely related functional surfaces from proteins of different fold structures. We envision that this method can be used for discovering novel functional relationship of protein surfaces, for functional annotation of protein structures with unknown biological roles, and for further inquiries on evolutionary origins of structural elements important for protein function.

Keywords: protein surface; surface pattern; protein function; pocket sequence; pocket shape

Introduction

With rapid progress in the determination of protein structures,[1,2] protein structural analysis has become an important source of information for understanding functional roles of proteins.[3–7] Conservation of protein structures often reveals very distant evolutionary relationships, which are otherwise difficult to detect by sequence analysis alone.[8] Analysis of protein structure can provide insightful ideas about the biochemical functions and mechanisms of proteins (e.g. active sites, catalytic residues, and substrate interactions).[9–11]

An important approach of studying protein structures is fold analysis.[3–7] Identifying the correct tertiary fold of protein is often helpful for inferring protein function. In many cases, fold assignment alone can provide valuable functional inference.[12] Nevertheless, the relationship between protein fold and protein function in general is complex.[8] A protein fold can adopt many different functions,[13] while a biological function can have many different structural supports.[14] This complex relationship is lucidly illustrated for a subset of proteins, whose functional roles can be explicitly described by Enzyme Classification (E.C.) labels.[10,15–17] It was found that functional inference between a pair of enzymes becomes difficult when sequence identity drops below 40%.[16] Jaroszewski and Godzik[18] further demonstrated that if descriptions other than the secondary structures are used, unexpected structural similarities can be found between proteins of different structural classes.

Abbreviations used: CASTp, Computed Atlas of Surface Topography of Proteins.

E-mail address of the corresponding author: jliang@uic.edu

They showed the example of tenascin (1ten, all β) and phosphotransferase (1poh, α + β), which are of different folds but have strong similarity in the geometry of their backbone traces. These results imply that different classification systems of protein structures other than current fold classification would also be possible.

Proteins fulfill their cellular roles by interacting with other molecules. A fundamental challenge in identifying protein function from sequence is that the functional surface of a protein often involves only a small number of key residues. These interacting residues are dispersed in diverse regions of the primary sequences and are difficult to detect if the only information available is the primary sequence. Discovery of local spatial motifs from structures that are functionally relevant is therefore an important task.

Several methods have been developed for analyzing local spatial patterns in proteins. Artymiuk *et al.* developed an algorithm based on subgraph isomorphism detection.[19] By representing residue side-chains as simplified pseudo-atoms, a molecular graph is constructed to represent the patterns of side-chain pseudo-atoms and their interatomic distances. A user defined query pattern can then be searched rapidly against the Protein Data Bank for similarity relationship. Another widely used approach is the method of geometric hashing. By examining spatial patterns of atoms, Fischer *et al.* developed an algorithm that can detect surface similarity of proteins.[20,21] This method has also been applied by Wallace *et al.* for the derivation and matching of spatial templates.[22] Russell developed a different algorithm that detects side-chain geometric patterns common to two protein structures.[23] With the evaluation of statistical significance of measured root mean square distance (RMSD), several new examples of convergent evolution were discovered, where common patterns of side-chains were found to reside on different tertiary folds. Further development of an elegant parametric model for assessing significance of matched side-chain patterns can be found in Ref. 24. Schmitt *et al.* recently described a method that detects similar surface cavities using descriptors generated from pre-computed cavities and a clique detection algorithm.[25]

Several studies combine protein structural context information with conserved sequence patterns to identify distantly related members of a protein family, or to infer protein functions. Yu *et al.* developed a protein surface similarity measure for WD protein family by combining structural information of beta propeller encoded in a hidden Markov model with sequence profiles of two maximally conserved sequence regions.[26,27] The idea is that protein surfaces are more functionally diagnostic than the full protein sequence, because strong hydrophobicity and size constraints for protein interior, similarity due to internal buried residues between proteins can be misleading in inferring functional similarity. This similarity measure was successfully applied for subfamily clustering of WD proteins for function prediction.[26] Zvelebil and Sternberg developed a method to predict catalytic residues that combines sequence and spatial local averages of residue conservation derived from multiple sequence alignment.[28] This idea is further expanded by Ota *et al.*, where additional geometric information and destabilizing mutation data are incorporated for predicting functionally important catalytic residues of enzymes.[29]

In this study, we describe a novel approach for detecting similar patterns of local motifs of protein structures. Because protein functional surfaces are frequently associated with surface regions of prominent concavity,[30,31] we focus on surfaces of pockets and voids on a protein structure. We do not assume prior knowledge of functional site residues, and do not require any similarity in either primary sequence or backbone fold structures. In addition, our method has no limitation in the size of the spatially derived motif and can successfully detect patterns small and large. Our method is also different from previous efforts to embed functional patterns into protein structural context. Instead of using general fold or architectural information, we obtain direct structural information of protein surfaces in the form of geometrically computed pockets and voids. Currently, our method cannot detect similar surface patterns whose underlying primary sequences have different order, such as those seen in the catalytic triad found in some examples of serine protease.

We first compute the alpha shapes of 12,177 protein structures in the PDB databank,[32–35] and exhaustively identify all surface pockets and interior voids for each of the protein structures.[33,34] For each pocket and void, the residues forming the wall are then concatenated to form a short sequence fragment of amino acid residues, while ignoring all intervening residues that do not participate in the formation of the wall. Two sequence fragments derived from pocket surface residues are then compared using dynamic programming. The similarity score for any observed match is assessed for statistical significance using an empirical randomization model constructed for short sequence patterns. Results from database search indicate that such short surface patterns of pocket and void residues are informative and often discriminating. In addition, we further assess the shape similarity of two pocket or void surfaces in Euclidean space, as well as in relative orientation using a new approach inspired by the work of Kedem *et al.*[36] We show examples of detection of similar functional surfaces among proteins of the same fold but low sequence identities, and among proteins of different fold. An all-against-all database search of all PDB structures reveals global pictures of surface similarity of currently known protein structures. We discuss how our method can be used in exploring protein structure–function relationship.

Matching Spatial Surface Patterns

Surface pockets and interior voids in proteins

Proteins are tightly packed. Their packing densities are comparable to that of crystalline solids.[37–39] Yet there are numerous packing defects in the form of pockets and voids in protein structures with broad size distributions.[40] For example, the volume v and area a of proteins do not scale as $v \sim a^{3/2}$, which would be expected for models of tight packing. Rather, v and a scale linearly with each other.[40] This and other scaling studies of geometric parameters of real proteins and off-lattice near-compact chain polymers as generated by sequential Monte Carlo indicate that proteins pack like random polymers[41] under loose compactness criterion.[42,43] Furthermore, the interior of proteins is more like Swiss cheese with many holes than tightly packed jigsaw puzzles.[40]

In this study, we follow[31,33,40] and define a pocket as an empty concavity on a protein surface into which solvent can gain access, i.e. these concavities have mouth openings connecting their interior with the outside bulk solution. A void is an interior unoccupied space that is not accessible to the solvent probe. It has no mouth openings to the outside bulk solution. Protein pockets and interior voids are computed using the weighted Delauney triangulation and alpha shape method developed by Edelsbrunner and colleagues, as described in Refs. 31–33,44,45. Detailed descriptions of the computational techniques can be found in Refs. 32–34. Precomputed pockets and voids of each protein structure in the PDB databank are conveniently organized as the database of Computed Atlas of Surface Topography of Proteins (CASTp)†.[35]

With the criterion that a void or pocket needs to be large enough to contain at least one water molecule, we find that there are 910,379 voids and pockets on 12,177 structures from the Protein Data Bank. On average, there are 15 voids or pockets for every 100 residues.[40] Figure 1(a) shows the size distribution of voids and pockets for the 12,177 protein structures and Figure 1(b) for 1641 proteins from the PDBSELECT database,[46] where no two protein structures have more than 25% sequence identity. The majority of pockets and voids are formed by 4–20 amino acid residues.

Compared to the full length primary sequences of proteins, the amino acid residues forming pockets and voids are compositionally different (Figure 2(a)). Figure 3(a) shows the ratio of composition for each of the 20 amino acid residues in pockets and voids and in the full primary sequences. We find that aromatic residues (F, W, and Y) are favored to be located in pockets and voids. We also compare the composition of surface residues and buried residues (Figures 2(b) and 3(b)). As expected, ionizable residues and polar residues are favored on protein surfaces, hydrophobic residues are favored in the interior. Residues located in pockets and voids show similar patterns when compared with interior buried residues (Figures 2(c) and 3(c)). For surface residues with >0.0 solvent accessibility that are not located in pockets or voids, similar patterns are found (Figures 2(d) and 3(d)). The bias in amino acid residue usage for pockets and voids is further demonstrated in Figures 2(e) and 3(e), where we compare surface residues that are located in pockets or voids with the rest of surface residues. We find that aromatic residues (F, W, and Y), residues often known to be functionally important (R and H), as well as branched hydrophobic residues (L, I, and V) have higher propensity to be in pocket or voids. This is consistent with the observation that H and W residues have high catalytic propensity for enzyme reactions.[47] Similar bias persists when we compare the composition of residues in pockets and voids containing functionally annotated residues by SwissProt with the composition of the full sequence of the proteins (Figure 2(a)).

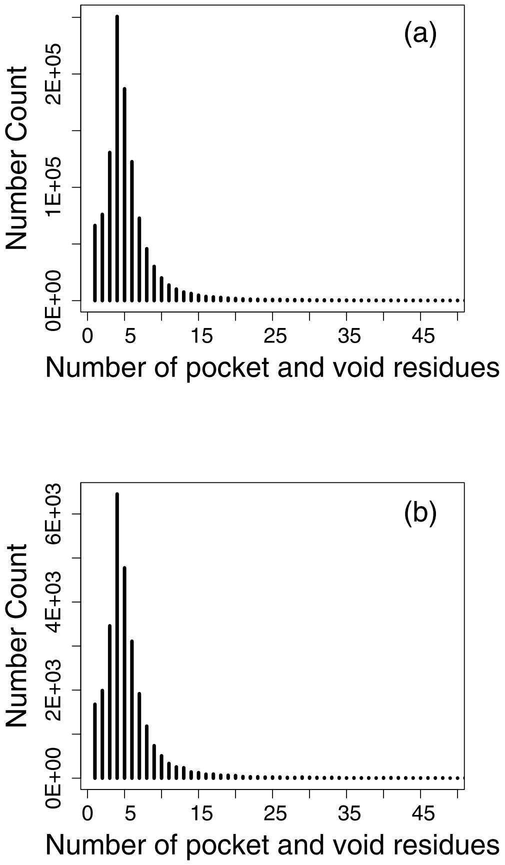


Figure 1. Size distribution of pockets and voids from (a) all 12,177 PDB structures studied and (b) 1641 structures from the PDBSELECT database (<25% sequence identity).

† http://cast.engr.uic.edu

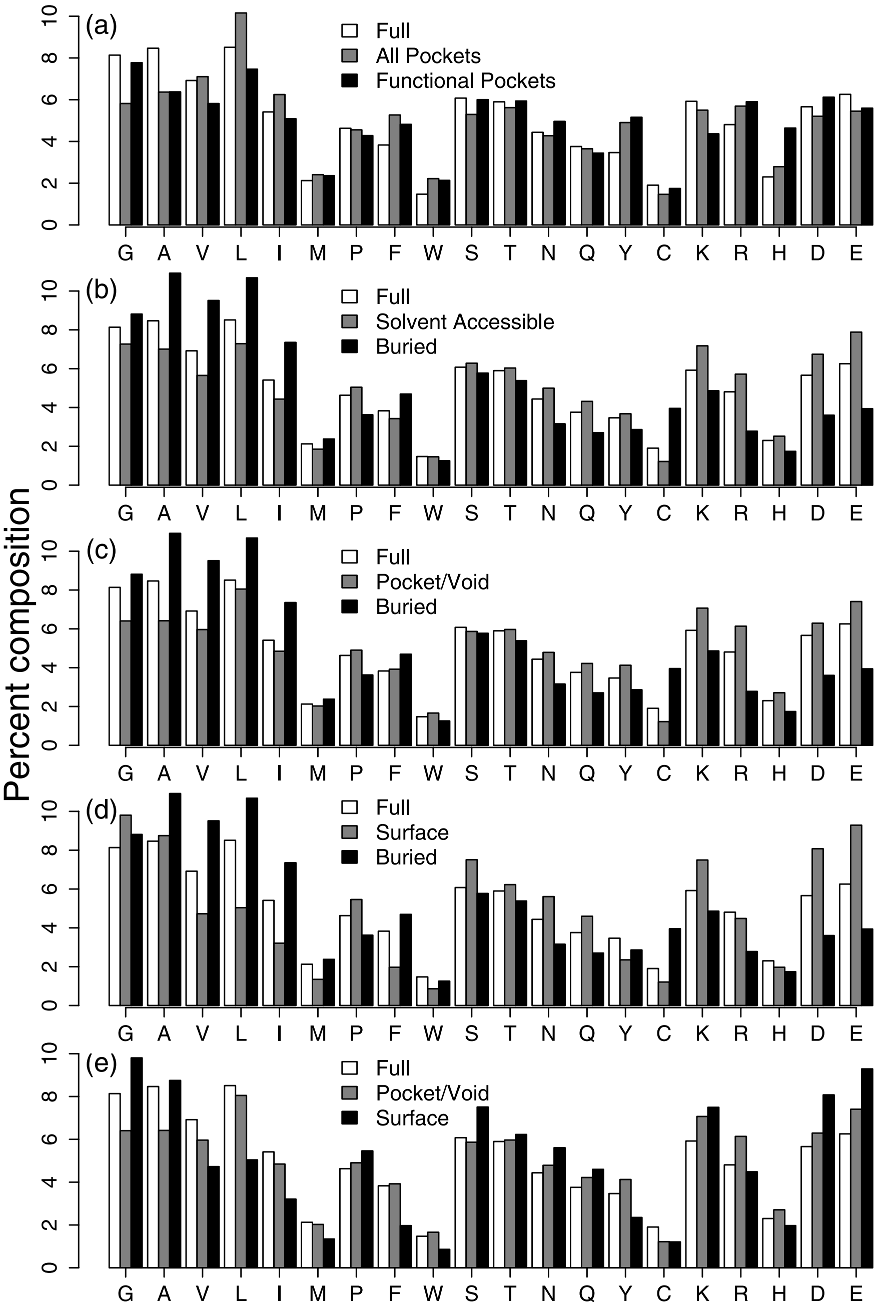


Figure 2. The composition of amino acid residues for PDB structures used in this study. (a) Fractions of different types of amino acid residues for the full sequence of the protein, for all pockets and voids, and for pockets and voids with functional annotation by SwissProt. (b) Fractions of different types of amino acid residues for the full sequence of the protein, for residues with >0.0 Å^2 solvent accessibility, and for residues that are buried with 0.0 solvent accessibility. (c) Fractions of different types of amino acid residues for the full sequence of the protein, for residues located in pockets or voids with >0.0 solvent accessibility, and for residues that are buried with 0.0 solvent accessibility. (d) Fractions of different types of amino acid residues for the full sequence of the protein, for surface residues not located in pockets or voids but with >0.0 solvent accessibility, and for residues that are buried with 0.0 solvent accessibility. (e) Fractions of different types of amino acid residues for the full sequence of the protein, for residues located in pockets or voids with >0.0 solvent accessibility, and for other amino acid residues with >0.0 solvent accessibility but are not located in a pocket or a void.

Comparison of sequence patterns of surface pockets and voids

We derived a set of protein surface patterns from the residues forming the walls of both pockets and voids as shown in Figure 4. We call these "pocket and void surface patterns of amino acid residues"† (pvSOAR patterns). The sequences of these

† A web server of pvSOAR can be accessed at http://pvsoar.engr.uic.edu

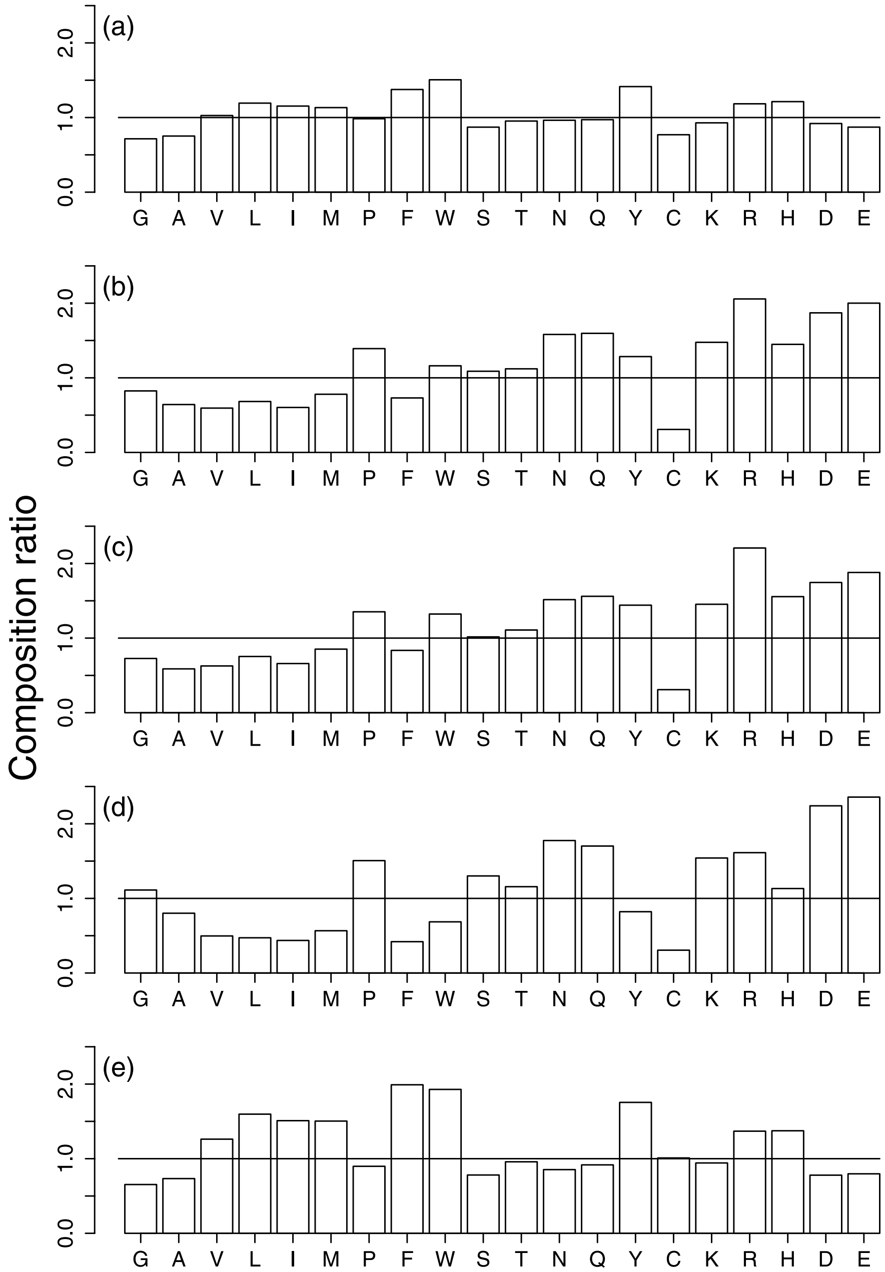
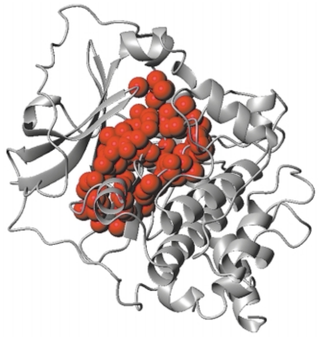


Figure 3. Ratio of amino acid residue compositions in all 12,177 PDB structures. Similar patterns are seen for proteins in PDBSELECT. (a) Ratio of composition of amino acid residues located in a pocket or a void *versus* composition of residues from the full sequence of the protein. (b) Ratio of composition of amino acid residues with >0.0 solvent accessibility *versus* composition of all amino acid residues that are buried with 0.0 solvent accessibility. (c) Ratio of composition of amino acid residues located in pockets or voids *versus* composition of all amino acid residues that are buried with 0.0 solvent accessibility. (d) Ratio of composition of surface amino acid residues not located in pockets or voids but with >0.0 solvent accessibility *versus* composition of all amino acid residues that are buried with 0.0 solvent accessibility. (e) Ratio of composition of amino acid residues located in pockets or voids with >0.0 solvent accessibility *versus* composition of other surface amino acid residues with >0.0 solvent accessibility but are not located in a pocket or a void.

patterns can be used to assess the similarity relationship among protein surfaces. Figure 5 provides an illustrative example. The catalytic subunits of cAMP-dependent protein kinase (1cdk) and tyrosine protein kinase c-src (2src) both bind to AMP or AMP analogs. The overall sequence identity between their primary sequence is low (16%). However, the AMP binding sites have similar shape and chemical texture as identified geometrically (Figure 5(a),(b)). In both cases, the residues participating in the formation of pocket walls come from diverse regions in the primary sequences (Figure 5(c)). When these residues are concatenated, the shorter sequences of binding site residues have much higher sequence identity (51%, Figure 5(d)).



```
(b) 49LGTGSFGRVMLVKHKETGNHFAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNL
    YMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFG
    FAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVR
    FPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSN
    F327

(c) 49LGTGSFGRV------------A-K-L---KV--L-Q--HT--E---L--------V------------
    -M-MEYV---E--------------------------------------D-K-EN-L---------TDFG
    F

(d) LGTGSFGRVAKLKVLQHTELVMMEYVEDKENLTDFGF
```

Figure 4. Protein surface sequence pattern is created by concatenating residues forming the wall of a pocket. For cAMP-dependent protein kinase (1cdk, chain A), the pocket residues of the protein structure and visualization as accessed through CASTp[35] is shown in (a). These residues are highlighted in the primary sequence (b), and are extracted (c), and concatenated (d) to from a sequence of pocket and void surface pattern.

This approach is applicable to the sequences of any two surface patterns of pockets or voids. By concatenating wall residues of a pocket or void on the same polypeptide chain, a sequence pattern of surface residues is compiled for each protein pocket and void in CASTp database. Collecting the pocket sequences of all 910,379 pockets and voids on 12,177 PDB structures, we constructed a database called pvSOAR database. We then used the Smith–Waterman algorithm as implemented in SSEARCH by Pearson[48] to compare the similarity of two pocket sequence patterns. In this study, we use BLOSUM50 as default scoring matrix,[49] and concatenate only wall residues that are on the same polypeptide chain.

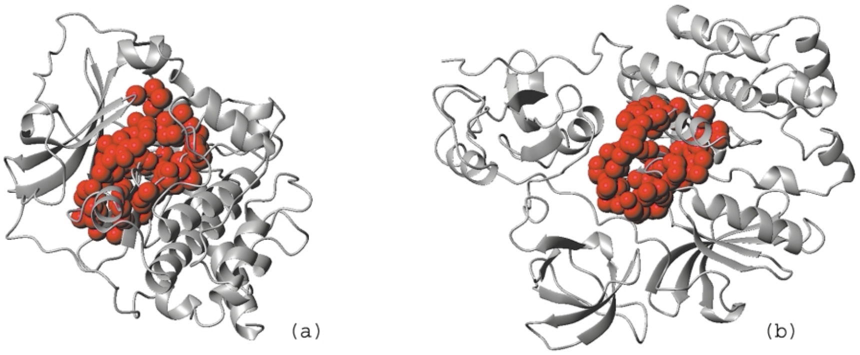
Statistical significance

When two sequences of pocket surface patterns are found to be similar, it is essential to assess the significance of detected similarity to aid in biological interpretation. For gapless local sequence alignment, the theoretical model of extreme value distribution (EVD) provides accurate description of alignment scores of random sequences.[50] This allows rapid assessment of statistical significance in the form of p and E-values.[48,51]

Assessment of statistical significance of matched sequences of pocket surface patterns is more challenging. First, pocket sequences are usually short. Unlike alignment of protein sequences where a peptide chain frequently has hundreds of residues, the majority of pocket patterns have between 5–20 amino acid residues. Second, the composition of pockets is biased and is different from that of the full chain sequences (Figure 2). Third, two pocket sequence patterns frequently have different number of residues, therefore the introduction of gaps in alignments is necessary. Although recent work has obtained analytical results for local alignment with gaps using selected scoring systems†,[52] no exact theoretical models are known in general for local sequence alignment of very short sequences with gaps. As an example, Figure 6(a) shows that the distribution of Smith–Waterman scores for querying randomly shuffled pocket sequences with a sequence pattern of the ANP binding pocket in tyrosine protein kinase c-src (2src) is very different from that of an EVD model.

We found that once the largest peak in the low-score region of the distribution of alignment scores of random short sequences is removed, the remaining distribution frequently follows an EVD. We have developed a heuristic approach for assessing statistical significance of matched similarity by exploiting this observation. Specifically, a query sequence of a surface pocket is first searched against all pocket sequences in the pvSOAR database, which contains $N_{\text{all}} = 910,379$ pocket sequences. Pocket sequences with Smith–Waterman scores below 20 are then removed, with N_t pocket sequences remaining. Pocket sequences

† citeseer.nj.nec.com/bundschuh99analytic.html



(c)

```
>1cdk_A
GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHFAMKILD
KQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQI
VLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALG
VLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKDGVNDIKNHKWFATT
DWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF

>2src_
MVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRR
ESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYS
KHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEA
FLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN
YVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELT
TKGRVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQXQPGE
NL
```

(d)

```
1cdk_104 LGTGSFGRVAKLKVLQHTELVMMEYV---EDKENLTDF
2src_51  LGQGCFGEVA-IKLMFAMVLVITEYMGSLDDRANLADF
```

Figure 5. Comparison of sequence patterns of the catalytic site of two kinases. (a) Active site of the catalytic subunit of cAMP-dependent protein kinase (1cdk, CASTp id = 104, chain A), and (b) of tyrosine protein kinase c-src (2src, CASTp id = 51). Both kinases bind to AMP or AMP analogs and their binding sites are similar. (c) The residues forming the geometrically defined pockets (colored in red) are well dispersed throughout the primary sequence. The overall identity between the primary sequences of these two kinases is low (at 16%), but (d) the identity of their surface sequence patterns have much higher sequence identity (51%).

removed typically contained only one or two aligned residues (as are alignments generating the peak in Figure 6(a)). Next, we randomly select 200,000 pocket sequences from the set of N_t pocket sequences, or all of them if $N_t < 200,000$. These sequences are randomly shuffled, and the query pattern is searched against this shuffled database (Figure 6(b)). The Smith–Waterman scores of the search are then collected. The goodness-of fit of theses scores to an EVD distribution is then evaluated using the non-parametric Kolmogorov–Smirnov test, which is provided as part of the SSEARCH tools.[53]

If the observed Kolmogorov–Smirnov statistic indicates that the random scores are not inconsistent with an EVD distribution, we further estimate the significance level p of the detected similarity. p value represents the probability of obtaining the same or better score $Z > z$ by chance, where z is the observed score when searching the query pattern against pvSOAR database. It is calculated as $z = (S - \mu)/\sigma$, where S is the similarity score, μ the mean of random scores, and σ the standard deviation. For EVD, p-value can be estimated from the z score match:[48]

$$p(Z > z) = 1 - \exp(-e^{z \cdot \pi/\sqrt{6} - \Gamma'(1)}) \qquad (1)$$

$$= 1 - \exp(-e^{-1.282z - 0.5772}) \qquad (2)$$

E-value represents the number of random pocket sequences with the same or better score that would be matched by random chance. It is calculated as:

$$E = p \times N_t$$

We use the estimated E-value to exclude matched pairs of pocket sequences that are unlikely to have biological significance.

Comparison of shapes of surface pockets and voids

Alignment of sequence patterns from pvSOAR database identifies residues that are conserved between two geometrically well-defined pockets

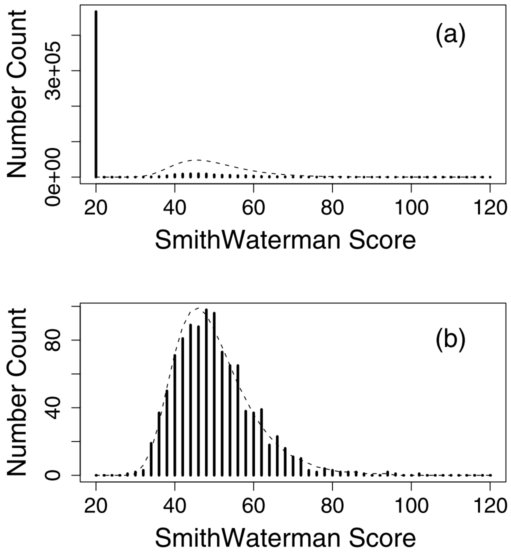


Figure 6. Distribution of Smith–Waterman scores for the ANP binding pocket (CASTp id = 51) of tyrosine protein kinase (2src). (a) The distribution of random scores as calculated using FASTA is very different form an extreme value distribution model (dotted line). (b) Distribution of Smith–Waterman scores of random pocket sequences after removing those with Smith–Waterman scores ≤20. Kolmogorov–Smirnov test statistic for goodness-of-fit is 0.0195, indicating that the observed data is not inconsistent with an EVD distribution.

or voids, and provides an equivalence relationship between pocket residues of the two proteins. It is often informative to further compare the shapes of the subset of equivalent pocket residues to further assess their geometric similarity.

cRMSD

A simple method for measuring geometric dissimilarity is to calculate the coordinate RMSD (cRMSD) between the subset of equivalent residues. We only need to find the optimal superposition of a subset of pocket residues from one protein to a subset of pocket residues from another protein. Such optimal structural alignment can be found following the method of Umeyama,[54] which is based on singular value decomposition (SVD) of the correlation matrix of the coordinates of the sets of points.[55] This method provides the least square rotational matrix and translational vector, as well as the cRMSD values. It is similar to that of McLachlin[56] and Kabsch.[57] Here, the structural alignment is based on optimization for least RMSD at the residue level. We use one point to represent a residue. When multiple atoms from the same residue contribute to the wall of a pocket or void, the geometric center of these atoms is used.

oRMSD

cRMSD is useful for assessing structures that are very similar, but this measure is very sensitive to outliers, namely, a few outlier residues in the two sets of pocket residues that would otherwise be very similar will dominate the cRMSD value.[58] In addition, cRMSD between two structures increases as the number of aligned residues increases.

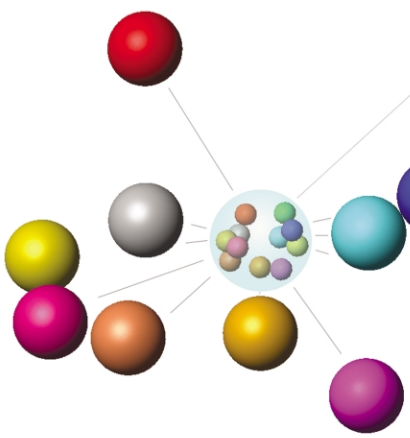


Figure 7. Unit sphere transformation of the geldanamycin binding pocket (CASTp id = 33) of human heat shock protein 90 (HSP90) (PDB id = 1yes). The location of each residue is projected onto the unit sphere $\mathbb{S}^2$ where the center is the geometric center of the pocket. The resulting structure is a collection of unit vectors on $\mathbb{S}^2$.

An elegant alternative measure of dissimilarity is the unit vector RMSD (uRMSD), originally developed in Ref. 36 where the unit vectors connecting consecutive C^α atoms are mapped to a unit sphere $\mathbb{S}^2$, and the dissimilarity of the backbones of two proteins are measured by calculating the RMSD between the two series of unit vectors from the two proteins on $\mathbb{S}^2$.

We modify this method for measuring geometric dissimilarity of two protein pockets and voids. First, we place a unit sphere $\mathbb{S}^2$ at the geometric center of the pocket $\mathbf{x_0} \in \mathbb{R}^3$. Second, the location of each residue $\mathbf{x} = (x, y, z)^T$ is then projected onto the unit sphere along the direction of the vector from the geometric center: $\mathbf{u} = (\mathbf{x} - \mathbf{x_0})/\|\mathbf{x} - \mathbf{x_0}\|$. The projected pocket is represented by a collection of unit vectors located on $\mathbb{S}^2$ and the original orientation of residues in the pocket is preserved (for example see Figure 7). Third, we measure the RMSD of the two sets of unit vectors derived from the two pockets, which we call oRMSD for "orientation RMSD" to distinguish from the original uRMSD. uRMSD was derived naturally from consecutive C^α atoms along the backbone of a protein, and is not affected by possible bias introduced by placing the centers of the unit sphere to the geometric center of pocket residues. The orientational relationship between the two sets of residues on $\mathbb{S}^2$ is used for further discrimination.

Statistical significance of matched shapes

To evaluate the significance of shape similarity detected by either cRMSD or oRMSD between two surface spatial patterns, we estimate the probability p of obtaining a specific cRMSD or oRMSD value for n_{res} matched positions from a set of randomly generated surface pockets and voids. This is similar to that of Reference 23. We choose two pockets at random from all 910,379 pockets with the criterion that each has at least n_{res} residues. For each residue contained in a pocket, we calculate the coordinates of the geometric center of those atoms appearing on the wall of the pocket or void, and use the geometric center to represent this residue. For each pocket, we choose N_{res} residues randomly. The cRMSD values are then calculated for the n_{res} residues from the selected two random pockets. This process is repeated with differently chosen random pockets and differently chosen random n_{res} residues for oRMSD measurement. We collect about 38 million cRMSD values and separately oRMSD values for $n_{\text{res}} = 3$, and about one million cRMSD and oRMSD values for $n_{\text{res}} = 100$. The p-value for a specific cRMSD or oRMSD value can then be assessed by finding the closest value of the rank order statistic in the randomly collected cRMSD or oRMSD data for n_{res} residues, respectively. When n_{res} is small (e.g. 3–5 residues), we can have estimated p-value down to 10^{-8}. When n_{res} is large (e.g. 50 residues), we can have estimated p-value down to 10^{-6}.

Because exact p-values in the tail region of $p < 10^{-8}$ cannot be assessed, we do not calculate E-value of observed cRMSD and oRMSD for matched spatial surface patterns.

Data selection

To evaluate results from all-against-all searching of large databases with nearly a million entries, we use heuristics to prune the data for identifying biologically interesting similarity relationships. We noted that many surface pockets with more than 100 residues are typically protein–protein interfaces. Although important in their own right, in this study we focus on smaller functional surfaces that are more likely to be involved in ligand and substrate binding. Therefore we exclude large surface pockets with more than 100 residues.

Because of the uneven distribution of structures in the Protein Data Bank, a large number of matched pocket patterns with significant E and p-values come from proteins with identical or strongly homologous sequences. We exclude these relatively easy cases. First, we mark proteins in the pvSOAR database with hierarchical structural classification label as extracted from the CATH[4] and the SCOP[3] databases. Every residue on a pocket or void is labeled by either a SCOP or a CATH label. We exclude pockets without simultaneously both classification labels. Second, we exclude matched pair of pocket sequences if the full primary sequence identity of the two proteins exceeds 30% as measured by SSEARCH, because these similarity relationships can be easily detected by other methods such as PSI-BLAST.[59]

In the random model used for estimating E-values for pocket sequence alignment, we assume that each residue appearing in a pocket is drawn from a random position of the sequence. We therefore further select only matched pocket sequences from residues that are not all contiguous sequence neighbors. We use the following sequence separation measure d_{s} :

$$d_{\text{s}} = \frac{\sum_{i \in P} n_i - n_{i-1}}{|P| - 1}$$

where P is the set of matched pocket residues, which has a total of $|P|$ residues, and i is the ith matched pocket residue after ordering them by sequence number n_i. The number n_{i-1} is the sequence number of the preceding residue. If $d_{\text{s}} < 2$ for the set of aligned residues in a matched pair of pocket sequences, this pair of matched sequence fragments is excluded from analysis. To further ensure that similar surface patterns are statistically significant and to allow cRMSD and oRMSD to be calculated, we require that a matched surface pattern contain at least four residues.

Results

We begin discussion of results with three types of examples. First, we describe the detection of similar functional surfaces from members of the same protein family. Examples are given for acetylcholinesterase, where matching of functional pocket surface pattern is shown to be specific, namely, all proteins containing significantly matched surface patterns are members of the acetylcholinesterase family. Second, we describe the detection of functionally related binding surfaces among proteins of the same tertiary fold but from different protein families with overall low sequence identity. For this, we discuss alpha-amylases in detail. Third, we describe detection of related functional surface between proteins not only of overall low sequence identity but also of different tertiary fold or class. We discuss the example of HIV-1 protease and heat shock protein-90 in some detail. We further describe an intriguing similarity between functional surface of aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase, which are again of different class and fold. We conclude with a preliminary statistical summary of results from a global all-against-all search of all surface pockets and voids of protein structures in the pvSOAR database, which represent most structures in the Protein Data Bank.

Functional surfaces from the same protein family: acetylcholinesterase

Acetylcholinesterase is a serine hydrolase that belongs to the esterase family. Its function is to catalyze the hydrolysis of the neurotransmitter acetylcholine by transferring the acetyl group to water, forming choline and acetate.[60] It acts to stop neurotransmission at cholinergic synapses frequently found in the brain. It is an α/β protein (CATH code 3.40.50.950, SCOP code c.69.1.1). The active site contains a catalytic triad (S200, H440, and E327), which is located in the "aromatic gorge" heavily lined with aromatic residues. Two of the catalytic residues on the structure of 2ack, S200 and H440, are located in a prominent surface pocket identified by CASTp (pocket id = 68, solvent accessible surface area 352 $Å^2$, volume 180 $Å^3$). In addition, this pocket contains six G residues (residue number 117–119, 123, 335), five Y (70, 121, 130, 334, 442), four F (282, 288, 290, 330, 331), four S (81, 122, 200, 286), three W (84, 233, 279), two L (127, 282), two I (287, 444), and one for each of R, D, E, H, N, Q, and P residues. The third residue E327 of the catalytic triad is not directly located in this pocket, but is located in another pocket that opens up in an opposite direction (id = 66, area 44 $Å^2$, volume 11 $Å^3$) and is immediately behind S200 and H440 in the structure of 2ack.

Results of searching the pvSOAR database with the sequence pattern of the pocket containing S200 and H440 on 2ack are shown in Table 1. For this highly conserved functional surface, all significant hits at the level of $E < 0.1$ are surface patterns from members of the same acetylcholinesterase-like family. Many proteins in this family have strong overall sequence identity with the query protein. The lack of matches with proteins from any other families indicates that acetylcholinesterase proteins exhibit significant similarity in the surface pattern of the active site, and this pattern is unique to the acetylcholinesterase protein family. This example demonstrates that in some cases functionally related surfaces can be identified with specificity.

Similar functional surfaces from different protein families

We discuss detecting functionally related binding surfaces from proteins of different families but of the same superfamily with varying overall sequence identities. Alpha amylase is an enzyme that catalyzes the breakdown of amylase and amylopectin through hydrolysis at 1–4 glycosidic bonds (E.C. 3.2.1.1). Alpha-amylase from *Bacillus subtilis* (1bag) contains two domains: an α/β TIM barrel domain (CATH code 3.20.20.80, SCOP code c.1.8.1) and a β sandwich domain (CATH code 2.60.40.1180, SCOP code b.71.1.1). Its substrates are starch, glycogen and polysaccharide, and the product of the enzyme reaction is oligosaccharide. The substrate binding site (CASTp id = 60) for 1bag is located on the TIM barrel domain, and is formed by four L residues (141, 142,144, 210), three H (102, 180,268), two Y (59, 62), two D (176, 269), two Q (63, 208), and one each of R (174), K (179), N (273), W (58) and A (177) residues. It is the largest pocket on the protein, with solvent accessible area of 181 $Å^2$ and volume of 137 $Å^3$. This enzyme belongs to the glycosidase homologous superfamily within the TIM barrel topology (CATH code 3.20.20.80.25).

Results of searching the pvSOAR database with the sequence of the substrate binding site are partly shown in Table 2. There are 46 hits with significance value of $E < 0.01$, several of which have overall sequence identity with 1bag below 25%, as measured by sequence alignment using SSEARCH. These include structures of orthologous alpha amylase proteins from other species, as well as other members of the amylase family with related function. For example, the alpha amylase from *Bacillus stearothermophilus* (1qho, CATH label 3.20.20.80.14) takes glucan as substrate and produces alpha-maltose, a smaller molecule than oligosaccharide produced by alpha-amylase from *B. subtilis*. The matched pocket (CASTp id = 96 on chain A) contains many residues that are in the substrate-binding site. If we only had access to primary sequence information of these two proteins, a Smith–Waterman alignment will not provide convincing evidence with an E-value 3.4×10^2 (from SSEARCH[48]) that these two proteins are functionally related, since their overall sequence identity is about 23%, well below the 30–40% threshold when functional inference becomes difficult.[11] The alignment of the sequences of the two pocket surface patterns (Figure 8) shows a 60% sequence identity, corresponding to a significant E-value of 0.00042. Structural comparison between the pockets (Figure 8(e)) shows that the 11 conserved residues superimpose nearly perfectly with an cRMSD of 1.44 Å and a p-value of 9.3×10^{-8}. The only positional difference in the structural alignment is between N273 from 1bag and N371 from 1qho.

In addition to alpha amylases, several structures (e.g. 1cgw, 1cgv, 2dij) of cyclodextrin/cyclomaltodextrin glycosyltranferase (E.C. 2.4.1.19) are also found to have similar functional surfaces. These proteins degrade starch to cyclodextrins by formation of a 1,4-alpha-D-glucosidic bond. They are members of the glycosyltransferase sequence family, a different branch of the glycosidases superfamily as annotated by CATH.[4] Although their overall sequence identity to alpha amylase (1bag) are low (22% for 1cgw and 1cgv, 25% for 2dij), both are detected by matching sequence of surface patterns with significant E-values ($E = 1.1 \times 10^{-4}$ and $E = 1.8 \times 10^{-3}$ for 1cgw and 1cgv, respectively). The shapes of these two pockets are also conserved (cRMSD p-value $= 3.1 \times 10^{-4}$, 5.7×10^{-4} for 1cgw and 1cgv, respectively).

Similar functional surfaces from different protein fold and class

Proteins of different overall fold can also have similar biological function, but such cases are considerably more difficult to detect. We describe here examples of inferring remotely related

Table 1. Search results with the sequence of the surface pattern of the functional pocket (CASTp id = 68) forming the catalytic triad from acetylcholinesterase (2ack)

PDB code	CASTp id	Chain id	*E*-value	CATH id	SCOP id	Structure	Backbone seq. id	Aligned residues	cRMSD	*p*-value
1efj	64	A	8.10×10^{-20}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.61	1.6×10^{-11}
2ace	62	0	8.10×10^{-20}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.66	3.2×10^{-11}
1ax9	71	0	8.10×10^{-20}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.36	3.0×10^{-13}
1qie	79	A	1.50×10^{-18}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.51	3.5×10^{-12}
1qii	83	A	1.50×10^{-18}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.60	1.3×10^{-11}
1ea5	79	A	1.80×10^{-18}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.53	3.0×10^{-12}
1acl	50	0	1.80×10^{-18}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.66	2.2×10^{-11}
1som	70	A	2.50×10^{-18}	n/a	c.69.1.1	Acetylcholinesterase	1.000	33	0.60	1.4×10^{-11}
1qih	80	A	4.90×10^{-18}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	32	0.51	5.6×10^{-12}
1qim	84	A	2.80×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.69	5.2×10^{-11}
1qif	74	A	2.80×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.56	7.4×10^{-12}
1qig	78	A	2.80×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.58	1.0×10^{-11}
1qij	84	A	2.80×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.65	2.8×10^{-12}
1maa	286	D	5.30×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.590	32	0.82	4.7×10^{-10}
1maa	285	A	5.30×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.590	32	0.80	3.3×10^{-10}
1oce	63	0	8.90×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	32	0.63	3.5×10^{-11}
1qik	83	A	9.20×10^{-17}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	32	0.66	4.9×10^{-11}
1vxr	76	A	5.10×10^{-16}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	28	0.67	7.2×10^{-13}
1eve	73	0	6.10×10^{-16}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.61	9.3×10^{-12}
1qid	73	A	6.30×10^{-16}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.54	3.6×10^{-12}
1vxo	85	A	7.80×10^{-16}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.59	7.6×10^{-12}
1amn	78	0	8.70×10^{-16}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.991	33	0.60	1.3×10^{-11}
1dx6	75	A	1.00×10^{-15}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	0.69	3.3×10^{-11}
1vot	57	0	2.80×10^{-15}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	32	0.65	4.1×10^{-11}
1fss	73	A	4.60×10^{-15}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	32	1.10	1.3×10^{-08}
1qti	69	A	9.50×10^{-15}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	28	0.81	8.6×10^{-12}
1c2o	318	C	2.00×10^{-13}	n/a	c.69.1.1	Acetylcholinesterase	0.590	31	0.88	4.6×10^{-10}
1c2b	87	A	2.00×10^{-13}	n/a	c.69.1.1	Acetylcholinesterase	0.590	31	0.88	4.6×10^{-10}
1c2o	319	A	2.00×10^{-13}	n/a	c.69.1.1	Acetylcholinesterase	0.590	31	0.88	4.6×10^{-10}
1acj	79	0	8.30×10^{-13}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	33	0.89	6.6×10^{-10}
1f8u	97	A	1.80×10^{-12}	n/a	c.69.1.1	Acetylcholinesterase	0.574	25	7.39	3.2×10^{-01}
1b4l	104	A	2.80×10^{-12}	n/a	c.69.1.1	Acetylcholinesterase	0.578	27	7.52	3.4×10^{-01}
1e3q	83	A	5.20×10^{-12}	3.40.50.950	c.69.1.1	Acetylcholinesterase	1.000	31	2.10	4.1×10^{-05}
1mah	107	A	5.70×10^{-12}	n/a	c.69.1.1	Acetylcholinesterase	0.590	30	0.87	2.3×10^{-11}
1maa	288	C	2.30×10^{-11}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.590	30	3.16	6.2×10^{-03}
2dfp	78	A	1.20×10^{-10}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.998	28	1.50	8.2×10^{-08}
1maa	287	B	2.00×10^{-10}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.590	27	2.80	4.8×10^{-03}
1qo9	86	A	2.10×10^{-06}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.370	21	4.24	8.5×10^{-02}
1c2o	326	D	5.70×10^{-06}	n/a	c.69.1.1	Acetylcholinesterase	0.590	24	4.66	1.5×10^{-01}
1qon	86	A	1.30×10^{-03}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.370	18	1.77	1.7×10^{-04}
1c2o	326	B	1.30×10^{-03}	n/a	c.69.1.1	Acetylcholinesterase	0.590	24	4.63	1.4×10^{-01}
1dx4	76	A	5.30×10^{-02}	3.40.50.950	c.69.1.1	Acetylcholinesterase	0.370	18	2.89	1.9×10^{-02}

This Table gives the PDB code, the pocket identification number generated by CASTp, the chain identification, pvSOAR sequence alignment *E*-value, structural classification labels from both CATH and SCOP, name of the protein, sequence identity of primary sequence as obtained by SSEARCH alignment, length of alignment of pocket sequences, cRMSD value and associated *p*-value. The symbol "n/a" indicates that no information was available at the time of calculation.

biological functions by detecting similar binding surfaces on the structures of HIV-1 protease and HSP-90, and on aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase.

HIV-1 protease and heat shock protein-90

Human immunodeficiency virus type-1 protease is a member of the retroviral protease family (SCOP label b.50.1.1). It is an all-β dimer protein with identical single domain-chains, each containing a (6,10) barrel. The active site of HIV-1 protease is located at the dimer interface. On the structure of HIV-1 protease complexed with substrate-based inhibitor acetylpepstatin (5hvp),[61] the active site is the largest pocket on the protein (CASTp id = 21, solvent accessible area = 529.9 Å^2, volume = 415.0 Å^3). It has two mouths and its wall is formed by a series of loops and gaps in the flexible region of the protein. The inhibitor acetyl-pepstatin (isovaleral-Val-Val-Sta-Ala-Sta) interacts with the protein through both hydrogen bonding and hydrophobic interactions. Of the ten residues that participate in hydrogen bonding with the inhibitor, nine are located within the pocket.

An unexpected similar pocket surface pattern was identified on the structure of human heat shock protein 90 (HSP90, SCOP classification d.122.1.1, PDB id = 1yes) complexed with geldanamycin. HSP90 is of different fold from that of HIV-1 protease, and is a molecular chaperone participating in the conformational maturation of nuclear hormone receptors and protein kinases, as well as functioning in cellular stress response.[62]

Table 2. Results of searching against the pvSOAR database with the sequence pattern of the substrate binding pocket (CASTp id = 60) of alpha-amylase from *B. subtilis* (1bag)

PDB code	CASTp id	Chain id	*E*-value	Name	Backbone seq. id.	Aligned residues	cRMSD	*p*-value
1jae	57	0	8×10^{-09}	Alpha-amylase	0.244	14	0.48	1.686×10^{-07}
1b2y	80	A	9.8×10^{-09}	Alpha-amylase	0.237	14	0.49	1.686×10^{-07}
1kgu	77	A	1.1×10^{-07}	Alpha-amylase, pancreatic	0.244	15	1.88	3.621×10^{-03}
1jfh	81	0	1.8×10^{-07}	Alpha-amylase	0.233	14	0.43	1.686×10^{-07}
1kgw	63	A	5.6×10^{-07}	Alpha-amylase, pancreatic	0.244	14	0.52	1.686×10^{-07}
2cpu	70	A	1.5×10^{-06}	Alpha-amylase	0.235	11	0.41	9.257×10^{-08}
1pif	75	0	5.2×10^{-06}	Alpha-amylase	0.239	14	1.98	5.211×10^{-03}
1pig	85	0	5.4×10^{-06}	Alpha-amylase	0.239	14	0.44	1.686×10^{-07}
1ose	82	0	5.4×10^{-06}	Porcine alpha-amylase	0.233	14	0.50	1.686×10^{-07}
1cpu	84	A	9.6×10^{-06}	Alpha-amylase	0.237	14	1.17	1.383×10^{-05}
3cpu	68	A	1.1×10^{-05}	Alpha-amylase	0.235	12	0.66	1.135×10^{-07}
1hx0	82	A	1.4×10^{-05}	Alpha-amylase (Ppa)	0.236	13	0.42	1.385×10^{-07}
1ppi	81	0	1.4×10^{-05}	Alpha-amylase (Ppa)(E.C.3.2.1.1)	0.236	13	0.45	1.385×10^{-07}
1c8q	70	A	1.6×10^{-05}	Alpha-amylase	0.239	14	2.24	1.919×10^{-02}
1hny	82	0	2.3×10^{-05}	Human pancreatic alpha-amylase	0.237	10	3.59	6.797×10^{-01}
1jxk	77	A	3.8×10^{-05}	Alpha-amylase, salivary	0.242	10	3.65	7.644×10^{-01}
1smd	86	0	4.0×10^{-05}	Amylase	0.239	10	3.58	6.662×10^{-01}
1b0i	73	A	4.4×10^{-05}	Alpha-amylase	0.254	14	0.53	1.686×10^{-07}
1e3z	70	A	4.6×10^{-05}	Alpha-amylase	0.228	11	2.26	1.154×10^{-02}
1bli	67	0	4.6×10^{-05}	Alpha-amylase	0.235	11	2.44	2.536×10^{-02}
1e3z	70	A	4.6×10^{-05}	Alpha-amylase	0.228	11	2.26	1.154×10^{-02}
1hvx	69	A	4.6×10^{-05}	Alpha-amylase	0.235	11	2.37	1.881×10^{-02}
2dij	90	0	4.8×10^{-05}	Cyclodextrin glycosyltransferase	0.221	11	1.41	7.109×10^{-05}
1g94	66	A	5.3×10^{-05}	Alpha-amylase	0.254	11	0.33	9.257×10^{-08}
1bsi	69	0	5.6×10^{-05}	Alpha-amylase	0.237	13	1.83	1.854×10^{-03}
2cxg	85	0	7.8×10^{-05}	Cyclodextrin glycosyltransferase	0.221	11	1.43	8.238×10^{-05}
1kck	91	A	7.8×10^{-05}	Cyclodextrin glycosyltransferase	0.223	11	1.53	1.774×10^{-04}
1aqh	71	0	8.5×10^{-05}	Alpha-amylase	0.254	13	0.51	1.385×10^{-07}
1aqm	69	0	8.5×10^{-05}	Alpha-amylase	0.254	13	0.43	1.385×10^{-07}
1cgw	93	0	0.00011	Cyclomaltodextrin glucanotransferase	0.223	11	1.61	3.070×10^{-04}
7taa	82	0	0.00018	Taka amylase	0.249	11	1.13	4.628×10^{-06}
1qho	96	A	0.00042	Alpha-amylase	0.220	11	1.44	8.896×10^{-05}
1qhp	101	A	0.00045	Alpha-amylase	0.220	11	1.40	6.563×10^{-05}
1e40	67	A	0.00093	Alpha-amylase	0.228	10	1.72	5.222×10^{-04}
1e43	61	A	0.00093	Alpha-amylase	0.228	10	1.79	7.970×10^{-04}
1e3x	68	A	0.00093	Alpha-amylase	0.228	10	1.92	1.669×10^{-03}
1jxj	77	A	0.001	Alpha-amylase, salivary	0.237	9	2.58	2.924×10^{-02}
1cgv	77	0	0.0018	Cyclomaltodextrin glucanotransferase2	0.221	11	1.70	5.651×10^{-04}
5cgt	88	0	0.002	Cyclodextrin glycosyltransferase	0.232	10	2.39	1.668×10^{-02}
1cxh	84	0	0.0024	Cyclodextrin glycosyltransferase	0.221	11	1.61	3.070×10^{-04}
1kcl	88	A	0.0024	Cyclodextrin glycosyltransferase	0.221	11	1.67	4.627×10^{-04}
1i75	149	A	0.0027	Cyclodextrin glycosyltransferase	0.240	11	1.65	4.037×10^{-04}
1dtu	84	A	0.0049	Cyclodextrin glycosyltransferase	0.225	10	2.30	1.125×10^{-02}
1cgt	91	0	0.007	Cyclodextrin glycosyltransferase	0.234	11	5.19	3.961×10^{-01}
1cgy	76	0	0.0096	Cyclomaltodextrin glucanotransferase	0.221	9	1.79	6.476×10^{-06}

The list contains many significant hits from proteins with sequence identity below 25%. These include orthologous alpha-amylase proteins as well as other members of the amylase family.

HSP90 consists of nine helices and an anti-parallel β sheet of eight strands that fold into an α/β sandwich. A deep binding pocket is formed by three helices, a loop, as well as β-sheets forming the bottom. This binding pocket is also the largest pocket on the protein (CASTp id = 33, solvent accessible surface area = 322.0 Å^2, volume 252.5 Å^3).

Experimental data suggested that this pocket is a substrate-binding site that shares extensive similarities to a typical enzyme active site. The substrate is a segment of protein whose maturation and refolding is regulated by HSP90.[62] Substrate geldanamycin in the structure of lyes is highly compact and differs from the conformation of free geldanamycin. Its ansa ring is very similar to a penta-peptide in a turn conformation, and the carbamate group may serve as a mimic of a Trp residue in the biological substrate peptide.[62]

The alignment of sequences of the pocket surface pattern is shown in Figure 9. The sequences of surface pattern from HIV-1 protease and HSP90 are aligned with an *E*-value of 8.3×10^{-3}. They superimpose with an cRMSD of 7.21 Å, with an insignificant *p*-value of 7.9×10^{-1} (Figure 9(e)). However, the oRMSD between the two surface patterns on a unit sphere is 0.73 Å with a significance level of

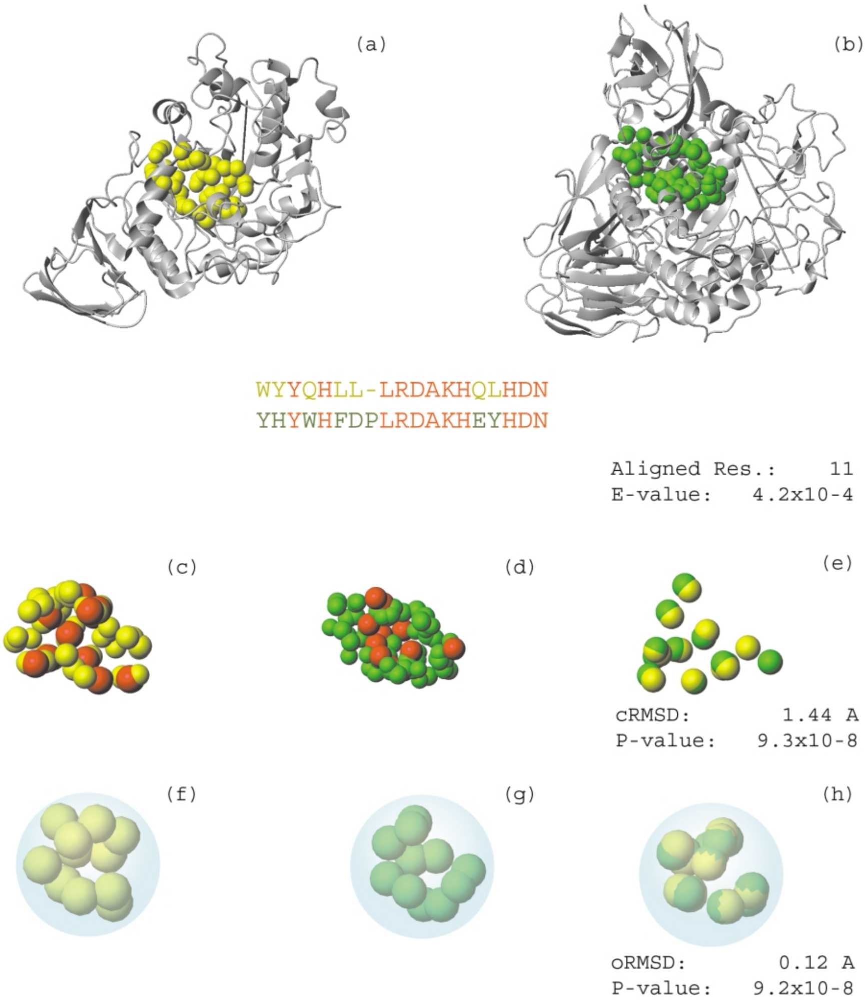


Figure 8. The pvSOAR alignment of the substrate binding site (CASTp id = 60) of alpha-amylase from *B. subtilis* (1bag) (a) to the substrate binding site of alpha-amylase from *B. stearothermophilus* (1qho, CASTp id = 96 on chain A) (b). The conserved residues in the pocket of (c) 1bag and (d) of 1qho are shown in red. Their superposition is shown in (e), where conserved residues are colored in yellow for 1bag and green for 1qho. The alignment of unit vectors on the unit sphere for oRMSD calculation for 1bag (e) and 1qho (f) is shown in (h).

p-value $= 6.3 \times 10^{-6}$ (Figure 9(h)). This suggests that the relative positions of conserved residues in these two active sites are similar.

Although the cRMSD between the binding-site surface patterns is large, it is not inconsistent with the suggestion that HIV-1 protease and HSP90 protein have similar functional surfaces. There are other examples where related functional surfaces of proteins have large cRMSD. For example, querying against the pvSOAR database using the sequence pattern of the functional site of acetylcholinesterase 2ack found several hits with significant E-values (e.g. 1b41 and 1f8u, Table 1). The p-values associated with cRMSD for these hits are all in the order of 10^{-1}. This indicates that the active site pocket of acetylcholinesterase may experience some conformational change, and searching by spatial similarity alone would have missed these hits.

The large cRMSD between the two pockets of HIV-1 protease and HSP90 is not surprising. HIV-1 protease undergoes substantial conformational change upon ligand binding. For HSP90, the size and accessibility of the active site pocket are also altered with conformational change.[62] This example again shows that shape similarity measured by cRMSD is not so informative for proteins experiencing functionally important conformational change, and in some cases oRMSD provides better assessment.

There are ten matched residues out of the 15 aligned residues between the two pocket sequences: K58, I91, D93, G97, D102, G132, G135, V136, G137, F138 from HSP90 (1yes) and R207, L223, D225, G227, D229, G248, G249, I250, G252, F253 from HIV-1 protease (5hvp). The key pocket residues from HIV-1 protease involved in substrate binding are all conserved in HSP90. Among these, D229, G227, D225 in the body of HIV-1 protease and residue G248 in the gap region form hydrogen bonds with the peptide inhibitor. Corresponding residues from HSP90 (D93, G97, D102, G132) are

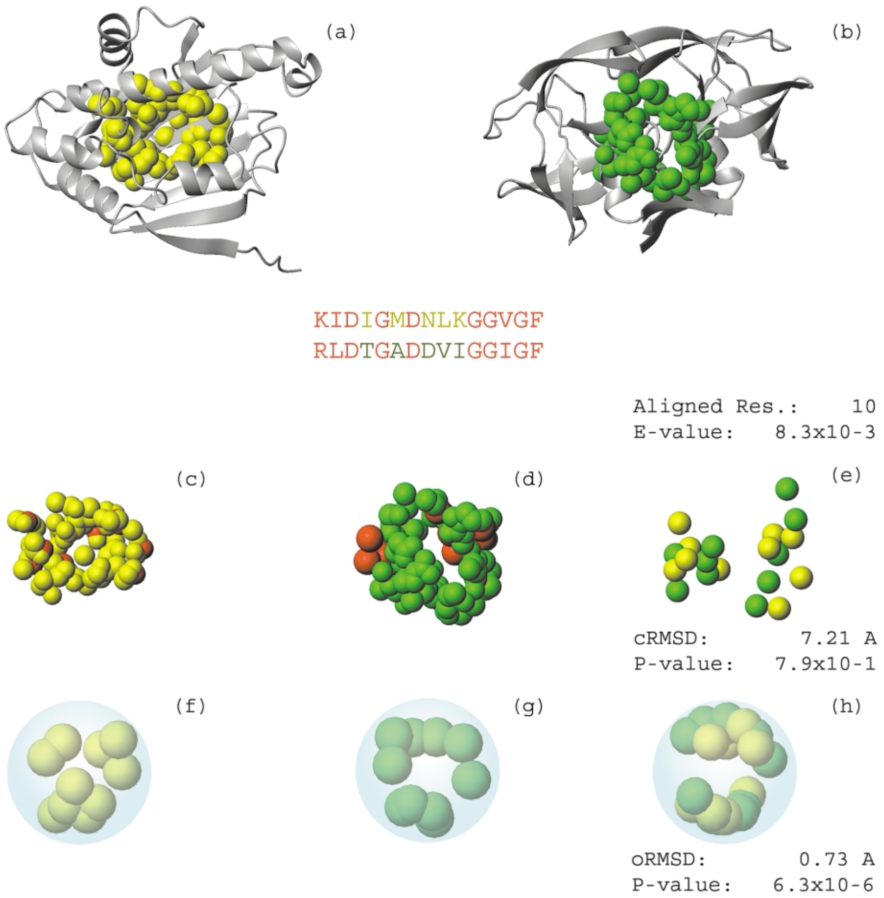


Figure 9. The pvSOAR alignment of the substrate binding site (CASTp id = 33) of human heat shock protein 90 (HSP90) (1yes) (a) to the substrate binding site (CASTp id = 21) of Human immunodeficiency virus type-1 protease (HIV-1) (5hvp) (b). The structural alignment of the conserved residues (shown in red) in the pockets for 1yes (c) and 5hvp (d) is shown in (e). The alignment for the pocket unit sphere for 1yes (e) and 5hvp (f) is shown in (h).

also known to be involved in substrate binding. A key residue in substrate binding of HSP90 is D93, which provides a strong hydrogen bond network with geldanamycin. G97 participates in hydrogen bonding, and residue D102 is involved in van der Waals interactions with geldanamycin. The critical role of these aligned residues in substrate binding offers some explanation for the detected similarity in pocket sequence and in oRMSD measurement. These two active sites both bind polypeptide substrates, despite the fact that they have structural supports of different protein folds and belong to different protein class.

Aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase

Aromatic amino acid transferase (AroAT) from *P. denitrificans* (2ay5) is a pyridoxal 5′-phosphate (PLP) cofactor dependent enzyme that catalyzes the transamination reaction. It can take both acidic and aromatic amino acid residues as substrates.[63] A series of aliphatic monocarboxylates attached to the bulky hydrophobic groups can bind to the active sites.[64] These compounds contain three moieties: the carboxylic group, an aliphatic chain of 2–4 carbon atoms, and a functional hydrophobic probing group. The substrate binding site is found to be the most prominent pocket on 2ay5 (CASTp id = 110, solvent accessible area = 797 Å^2 and volume = 514 Å^3). It is formed at the dimer interface, but the majority (45 residues) of the 51 wall residues come from chain A.

Results of searching pvSOAR database with the sequence of pocket surface pattern from chain A are listed in Table 3. As expected, the highest scoring match is the query pattern itself, as well as sequence patterns from other PDB structures of aromatic amino transferase (not listed). Additional high scoring matches include many surface patterns from structures of aspartic amino transferase.

A surprising match is 17-β-hydroxysteroid dehydrogenase (17-β-HD (1fdw)) at significant *E*-value of 2.1×10^{-4}. 17-β-HD belongs to the NADP-binding Rossman fold, which is different by SCOP from that of aromatic amino transferase (which is PLP-dependent transferase fold). A key enzyme in the estrone metabolic pathway, it catalyzes the conversion of estradiol-17-β to estrone. This is a different chemical reaction than that catalyzed by aromatic amino transferase. The substrate binding site of 17-β-HD is located at the most

Table 3. PDB structures containing pocket surface patterns that are similar to the functional site of aromatic aminotransferase (2ay5)

PDB code	Pocket id	Chain id	*E*-value	Name	Sequence identity
2ay5	110	A	5.1×10^{-26}	Aromatic amino acid aminotransferase	1.00
1aam	63	0	1.3×10^{-11}	Aspartate aminotransferase	0.46
1asl	125	A	1.1×10^{-05}	Aspartate aminotransferase	0.46
2aat	83	0	1.6×10^{-05}	Aspartate aminotransferase	0.46
1asn	140	A	1.6×10^{-05}	Aspartate aminotransferase	0.46
8aat	127	B	2.2×10^{-05}	Aspartate aminotransferase	0.36
1arg	138	A	2.9×10^{-05}	Aspartate aminotransferase	0.46
1asm	150	A	2.9×10^{-05}	Aspartate aminotransferase	0.46
1ahe	132	B	3.8×10^{-05}	Aspartate aminotransferase	0.45
1ajs	112	A	4.3×10^{-05}	Aspartate aminotransferase	0.37
1asm	149	B	7.8×10^{-05}	Aspartate aminotransferase	0.46
1tas	119	A	1.4×10^{-04}	Aspartate aminotransferase	0.36
1art	66	0	1.7×10^{-04}	Aspartate aminotransferase	0.46
1arg	139	B	2.1×10^{-04}	Aspartate aminotransferase	0.46
1fdw	39	0	2.1×10^{-04}	17-Beta-hydroxysteroid dehydrogenase	0.28
1ari	146	A	2.7×10^{-04}	Aspartate aminotransferase	0.46
1aka	130	A	3.9×10^{-04}	Aspartate aminotransferase	0.36
1ajs	113	B	5.1×10^{-04}	Aspartate aminotransferase	0.36
1qir	67	A	6.4×10^{-04}	Aspartate aminotransferase	0.46
1ama	52	0	6.5×10^{-04}	Aspartate aminotransferase	0.36
1maq	64	0	7.5×10^{-04}	Aspartate aminotransferase (Maspat)	0.36
1ahg	150	B	8.6×10^{-04}	Aspartate aminotransferase	0.45
1ajr	99	B	8.9×10^{-04}	Aspartate aminotransferase	0.36
1ivr	56	A	1.5×10^{-03}	Aspartate aminotransferase	0.36
1ari	147	B	1.5×10^{-03}	Aspartate aminotransferase	0.46
1tat	136	A	2.7×10^{-03}	Aspartate aminotransferase (Maspat)	0.36
1ajr	98	A	3.7×10^{-03}	Aspartate aminotransferase	0.36
1oxp	56	0	4.2×10^{-03}	Aspartate aminotransferase	0.36
1ams	62	0	4.8×10^{-03}	Aspartate aminotransferase	0.46
1yaa	229	D	6.0×10^{-03}	Aspartate aminotransferase	0.34
1tat	137	B	7.4×10^{-03}	Aspartate aminotransferase	0.36
1yaa	231	B	8.0×10^{-03}	Aspartate aminotransferase	0.34
1bhs	30	0	8.6×10^{-03}	17-Beta-hydroxysteroid dehydrogenase	0.28

Sequence identity of primary sequence as obtained by SSEARCH alignment are also listed. The hits listed are obtained by querying pvSOAR database with the pattern obtained from the active-site pocket (CASTp id = 110) on chain A of 2ay5. All hits have significant *E* values ≤0.01. The most significant hit is the query pattern itself. There are 87 hits from structures of aromatic aminotransferase and aspartic aminotransferase with *E*-values between 5.1×10^{-26} and 1.1×10^{-5}. Only one (1aam) is listed for brevity. All hits with *E* values between 1.0×10^{-5} and 0.01 are listed. Two 17-β-hydroxysteroid dehydrogenase structures are identified with significant *E* values of 0.00021 and 0.0086.

prominent pocket on 1fdw (CASTp id = 39, solvent accessible area = 818 Å^2 and volume = 844 Å^3). This binding site pocket contains 59 residues. When searching pvSOAR database with the sequence pattern from 1fdw, the strongest matches are surface patterns from other structures of 17-β-hydroxysteroid dehydrogenase, as expected. But structures of aromatic amino transferase are also detected at significant levels (*E*-value = 5.3×10^{-4} for the structure with the highest matching scores of AroAT, Table 4). The success of bi-directional search using both surface patterns of substrate-binding sites from AroAT and 17-β-hydroxysteroid dehydrogenase as query in identifying the other indicates that the similarity relationship between the functional surfaces of these two proteins can be detected robustly.

The functional roles of the matched residues in these two patterns provide some rationalization of the detected surface similarity. The alignment of the pocket residues for these two proteins is shown in Figure 10. Among these, 17 residue pairs are identical or are physico-chemically homologous. G36 and F360 from AroAT interact with the carboxyl group and the aliphatic group of the substrate. N142 and T109 recognize the aromatic groups through van der Waals interactions with the substrate. K258, G108, T109, S257, and Y225 bind to PLP. All these residues are conserved in 17-β-HD. Conversely, six conserved residues in the binding site of 17-β-HD interact with the hydrophobic group of the substrate, S142, P187, Y218, S222, F226, F259, and E282. The corresponding conserved residues on AroAT are T109, P195, Y225, S257, F360, Y380, and D384. Altogether, ten of the 17 conserved residue pairs have clear functional role in binding substrate in either AroAT or in 17-β-HD, as assessed from the structures of 2ay5 and 1fdw. These results suggest that similar sequence patterns of the binding surfaces of aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase may be related to their shared similar functional role of binding a bulky and hydrophobic group.

Unlike the example of HIV-1 protease and HSP-90, the orientational similarity as measured by oRMSD is not significant (oRMSD = 1.02 Å, *p*-value = 9.2×10^{-2}). The overall cRMSD measure

Table 4. Several structures of aromatic aminotransferase are among the list of hits of proteins with surface patterns similar to the functional site of 17-β-hydroxysteroid dehydrogenase on 1fdw

PDB code	Pocket id	Chain id	*E*-value	Name	Sequence identity
1fdw	39	0	9.2×10^{-30}	17-Beta-hydroxysteroid deydrogenase	1.00
1bhs	30	0	9.3×10^{-26}	17Beta-hydroxysteroid deydrogenase	0.99
1fdv	158	D	1.4×10^{-22}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdu	156	A	3.2×10^{-22}	17-Beta-hydroxysteroid deydrogenase	0.99
1equ	82	0	2.2×10^{-21}	Estradiol 17-beta-dehydrogenase	0.99
1fdv	161	C	4.9×10^{-20}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdu	154	D	5.9×10^{-20}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdv	159	A	2.8×10^{-19}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdv	160	B	1.1×10^{-18}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdu	155	B	4.2×10^{-18}	17-Beta-hydroxysteroid deydrogenase	0.99
1a27	31	0	5.2×10^{-17}	17-Beta-hydroxysteroid deydrogenase	0.99
1fdt	32	0	4.3×10^{-16}	17-Beta-hydroxysteroid deydrogenase	0.99
1iol	42	0	5.6×10^{-15}	Estrogenic 17-beta hydroxysteroid dehydrogenase	0.99
1equ	81	0	3.8×10^{-13}	Estradiol 17-beta-dehydrogenase	0.99
1dht	35	A	5.6×10^{-12}	Estrogenic 17-beta-hydroxysteroid dehydrogenase	0.99
1fdu	156	C	1.5×10^{-11}	17-Beta-hydroxysteroid deydrogenase	0.99
1fds	31	0	1.6×10^{-11}	17-Beta-hydroxysteroid deydrogenase	0.99
3dhe	43	A	4.3×10^{-11}	Estrogenic 17-beta-hydroxysteroid dehydrogenase	0.99
2ay5	110	A	0.00053	Aromatic amino acid aminotransferase	0.27
2ay4	124	A	0.0032	Aromatic amino acid aminotransferase	0.27
2ay8	120	A	0.0084	Aromatic amino acid aminotransferase	0.27

The listed hits all have *E* value ≤0.01 and are obtained by querying pvSOAR database with the pattern of functional site obtained from pocket 39 of 1fdw.

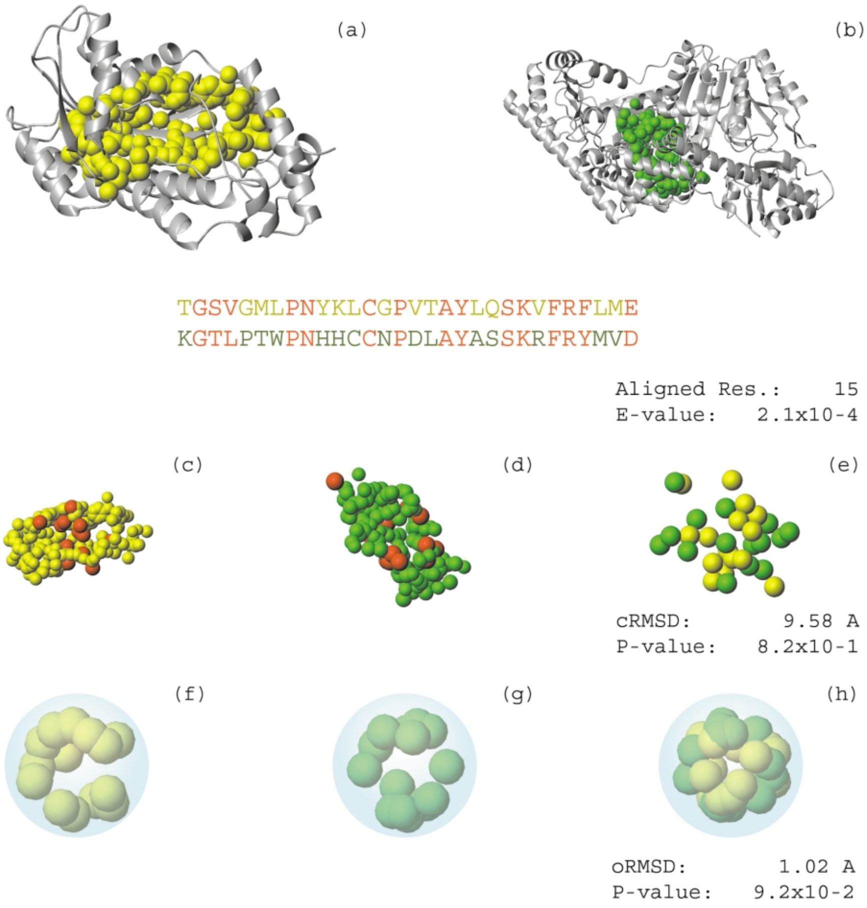


Figure 10. The substrate binding sites of (a) aromatic aminotransferase from *P. dentrificans* (2ay5, CASTp id = 110) and (b) 17-β-hydroxysteroid dehydrogenase (17-β-HD) (1fdw, CASTp id = 39). The alignment of the sequences of these two active sites is also shown, where identical or conserved residues are colored in red. The conserved residues in the pocket of (c) 1fdw and (d) of 2ay5 are shown in red. Their superposition is shown in (e), where conserved residues are colored in yellow for 1fdw and green for 2ay5. The alignment of unit vectors on the unit sphere for oRMSD calculation for 1fdw (e) and 2ay5 (f) is shown in (h).

(9.58 Å) between the spatial locations of the two pockets is also not significant (p-value of 8.2×10^{-1}). As seen earlier in the example of acetylcholine esterase (Table 1), the active site of a protein may be flexible and can adopt a number of different conformations in X-ray crystallographic structures. In many cases the shape similarity as measured by cRMSD or even oRMSD can no longer be detected with statistical significance. For example, 1f8u and 1b41 in Table 1 are both structures of acetylcholinesterase, as in the query structure 2ack. The p-values of the measured cRMSDs to the active site of 2ack are insignificant (3.2×10^{-1} and 3.4×10^{-1}, respectively for 1f8u and 1b41). However, in both cases, similarity in sequence patterns reveals close functional relationship of these protein structures. It is to this type of similar but flexible binding surface that we believe aromatic amino acid transferase and 17-β-hydroxysteroid dehydrogenase belong.

All-against-all comparison of structures in pvSOAR database

To assess globally the relationship of functional surfaces on all known protein structures, we carried out a preliminary all-against-all search of similar surface sequence and shape patterns for each pocket and void on each protein structure contained in the pvSOAR database. We found numerous examples where pocket surface patterns alone can detect functional relationship of proteins from the same family, similar to recently reported results.[25,65] We also examine unusual functional relationship of protein surfaces, that is, functional relationship between proteins of different superfamily, different fold, and sometimes different class, such as the example of HIV-1 protease and heat shock protein-90, and the example of aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase. Because of the preliminary nature of our study, here we have not removed redundant structures of the same protein from the data set, nor the structures of highly homologous proteins.

We restrict ourselves to proteins with known fold classification. We use the SCOP and CATH hierarchical classification systems to identify matched pairs of pocket surface patterns from two proteins, each of which must have both SCOP or CATH fold classifications. Only 10,429 protein structures in a total of 12,177 structures in the pvSOAR database have both SCOP and CATH classifications.

Summary results of similar pocket surface patterns identified from proteins of different SCOP class, fold, superfamily and family as obtained from an all-against-all comparison are listed in Table 5 at various statistical significance E-values. Similar results organized by CATH class, architecture, topology, homologous superfamily, and family classification are also listed in Table 5. These include many redundant and highly homologous entries in the PDB database, similar to the example of aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase in Tables 3 and 4. The full details of the results of all-against-all searches will be available on the web.

The all-against-all comparison identifies a total of 18,470 and 13,018 surface patterns with $10^{-9} < E < 10^{-3}$ that belong to different SCOP and CATH class, respectively. As an example, a matched surface pattern is found between 1cla and 1a28 (Table 6). These two proteins have only 19% overall full primary sequence identity. For similar surface patterns from proteins of different SCOP fold classification, we found a total of 29,085 matches at significance level of $10^{-9} < E < 10^{-3}$. Table 6 shows two examples from this search. The matched surfaces between 1xla and 1esn and between 1qsl and 1djy have 23% and 24% backbone sequence identity, respectively. Similarly, we have identified matched surface patterns using CATH classification. For example, the all-against-all comparison identifies a total of 24,249 surface patterns with $10^{-9} < E < 10^{-3}$ that belong to different superfamilies by CATH classification. As an example, the matched surfaces between 4rub and 1de6 are shown in Table 6. These two proteins have 27% overall backbone sequence identity. The all-against-all comparison, in addition, identifies a total of 30,425 surface patterns with $10^{-8} < E < 10^{-3}$ that belong to different SCOP families (no significant hits are found at $10^{-9} < E < 10^{-8}$). Table 6 shows a list of matches extracted from the all-against-all results.

These examples indicate that there exist similar protein surfaces from different family and sometimes different superfamily, fold, and class, either by SCOP or by CATH classification. It is likely that there is sometimes remote relationship in the biological functional roles of these matched surfaces, as elaborated in the examples of HSP90 and HIV-1, aromatic aminotransferase and 17β-hydroxysteroid dehydrogenase. More detailed analysis on these examples will provide additional information on the global relationship of biological functions and protein structures.

Discussion

In this study, we describe a new method for detecting similar patterns of protein structures that suggests related biological function. This method is fully automated without human intervention and does not require human input of query patterns. Our method is similar to other methods for detecting common local structure patterns and aims to uncover similarity of a small spatial region on protein structures.[19,22–24] These methods complement fold recognition methods and provide additional information for understanding protein structure and protein function relationship. Unlike side-chain based pattern discovery methods,[19,22,23] our method examines well

Table 5. A summary of similar pocket surfaces identified from proteins at different SCOP and CATH classification levels with various statistical significance *E*-values

E-value	Class	Fold	Superfamily	Family	Same	N/A	Total
SCOP							
10×10^{-9}	0	0	0	0	36363	11619	47982
10×10^{-8}	0	0	0	3	42406	13147	55556
10×10^{-7}	2	4	4	15	47683	15884	63582
10×10^{-6}	17	31	32	90	58353	21798	80241
10×10^{-5}	122	201	202	301	61624	24773	86698
10×10^{-4}	1175	1702	1718	1895	67887	31569	101351
10×10^{-3}	17163	27147	27529	28121	86040	67191	181352
10×10^{-2}	282157	439774	444712	448501	124269	468495	1041265
10×10^{-1}	3780513	5802962	5865460	5911283	248929	5439821	11600033

E-value	Class	Architecture	Topology	Homologous superfamily	Total	Same	N/A
CATH							
10×10^{-9}	32	32	35	35	32543	15404	47982
10×10^{-8}	54	54	58	58	38140	17358	55556
10×10^{-7}	84	85	89	89	43177	20316	63582
10×10^{-6}	116	129	146	147	52790	27304	80241
10×10^{-5}	253	291	324	328	55828	30542	86698
10×10^{-4}	997	1354	1509	1573	60800	38978	101351
10×10^{-3}	11482	17449	21060	22019	77137	82196	181352
10×10^{-2}	185133	279348	335486	352359	108415	580491	1041265
10×10^{-1}	2503177	3732045	4472269	4688335	229803	6681895	1160033

Matches are collected into different groups by the criteria whether the highest level of difference is at class, fold, superfamily, or family level for SCOP classification, and at class, architecture, topology, or superfamily, for CATH classification. Matches at lower level of classification automatically includes all matches at a higher level. For example, the 31 pairs of similar pocket surfaces from proteins of different SCOP fold at 10×10^{-6} will include all 17 cases of similar pairs of pocket surfaces from proteins of different SCOP class.

The total number (Total column) of significant matches at a specific significance level and the number of matches with identical CATH or SCOP classification (the Same column) are also listed. Since both SCOP and CATH classifications were used to identify a difference at the class level, the Total column represents the combined (from SCOP and CATH) number of matches. N/A indicates the number of matches where at least one protein did not have SCOP or CATH classification.

formed surface regions of pronounced concavity in the form of pockets and voids, which are frequently associated with protein function.[30,31] Because proteins play cellular roles through binding interactions with other molecules, this method helps to identify similarity relationship of protein surfaces that might be directly related to protein function. In addition, because these patterns are derived from short sequence fragments of pockets and voids, our method is not overly sensitive to small conformational changes in functional sites. Although our method detects order-dependent sequence pattern, it does not require residues to be adjacent in primary sequence.

A major challenge in analyzing local spatial patterns is to assess the significance of detected similarity. This is of fundamental importance for unambiguously establishing similarity relationship that is biologically interesting. We approach this problem using three different methods. First, we

Table 6. Partial list of significant matches between protein pockets belonging to different SCOP class, SCOP fold, CATH superfamily, and SCOP family from all-against-all search of similar surface sequence and shape patterns

Query			Match			PvSOAR alignment		ORMSD		CRMSD	
PDB code	Pocket id	Chain id	PDB code	Pocket id	Chain id	*E*-value	Aligned residues	ORMSD	*p*-value	cRMSD	*p*-value
Class											
1cla	28	0	1a28	70	B	2.6×10^{-03}	7	0.68	3.885×10^{-02}	3.98	3.623×10^{-01}
Fold											
1xla	116	B	1esn	145	A	5.3×10^{-03}	7	0.35	6.703×10^{-05}	2.30	5.105×10^{-03}
1qsl	76	A	1diy	125	B	1.0×10^{-02}	7	0.35	6.703×10^{-05}	3.68	2.484×10^{-01}
Superfamily											
4rub	298	A	1de6	227	A	2.6×10^{-03}	5	0.75	3.796×10^{-01}	2.87	2.397×10^{-01}
Family											
1qfy	65	A	1ddi	46	A	2.0×10^{-04}	12	0.09	1.130×10^{-07}	0.83	2.270×10^{-07}
1b38	36	A	1qcf	56	A	5.2×10^{-03}	21	0.28	5.190×10^{-07}	2.45	5.210×10^{-07}
1arm	38	0	1obr	40	0	1.9×10^{-03}	15	0.11	2.025×10^{-07}	0.92	1.016×10^{-06}

examine the significance of matched short sequences of the surface patterns. Since these patterns are derived from well-formed concave pockets and voids on protein structures, they are characteristic of the protein local surfaces. Because they are very short and have biased composition, we develop a heuristic method based on randomization test and non-parametric Kolmogorov–Smirnov test. Second, the relationship between the spatial arrangement of residues of pockets in Euclidean space provides further information about similarity relationship. We use standard cRMSD measurement to assess this relationship. Statistical significance of cRMSD value has been the subject of several important studies.[23,24,66–68] In this study, we followed the approach developed in Reference 23 for side-chain matching and constructed a random model by sampling unrelated surface pockets computationally for the matched number of pocket residues $n_{\rm res}$ between 3–100 residues, and derive empirically the p-value of observed cRMSD values by rank order statistic.

Third, a challenging problem in comparing spatial patterns of pockets is to identify related surface patterns on proteins that are subject to conformational changes. For these cases, we develop a method based on a novel measure oRMSD by assessing the relationship of two sets of equivalent orientational unit vectors on unit sphere representing residues from two surface pockets. Preliminary studies indicate that this method can suggest interesting similarity relationship between protein surfaces that do not superimpose well by rigid motion.

In most cases, the query surface patterns used for database search can be found automatically. For a given protein structure, we can first search against the pvSOAR database using all of its surface pockets and voids. We are then interested in surface pockets or voids on the query structure that matches with a pocket or void from another protein structure in the database with statistically significant low E-value. The examples discussed in Results are identified using this approach. Other strategies for identifying query patterns for database search are also possible. First, the largest or the second largest pocket or void is often of interest because it frequently is the binding site for enzymes.[30,31] Second, we can query with surface pockets selected by the criterion that they contain functionally annotated residues as described by SwissProt or reported in the literature. Third, we can query patterns that contain residues known to interact with ligand based on the PDB structure.

An important issue that cannot be addressed with our method is to detect similar protein surface patterns with different underlying primary sequence order. When convergent evolution occurs, nature discovers the same functional surfaces multiple times, as is the case of the catalytic triad in serine protease. It is likely that there may be many such examples where proteins with similar functional surfaces have different underlying protein core architecture, and specifically, the key residues important for function may have different order in primary sequences. Our method currently cannot detect such similarity. Further development of methods assessing similarity of order-independent surface patterns will be fruitful for studying this important issue.

There are additional possible further improvements. For detecting the similarity relationship between pocket sequences, we currently use BLOSUM50 amino acid substitution matrix for sequence alignment. We have shown that pockets and voids on proteins have different amino acid compositions from the composition of the entire primary sequences (Figure 2). Because BLOSUM and another widely used substitution matrix PAM are all constructed based on overall sequence similarity,[69,70] their use in our method is inconsistent with the goal of identifying surface pockets and voids that are functionally related. It is unclear how this would adversely affect our method. A substitution matrix tailored to the characteristics of binding surfaces may improve the sensitivity and specificity of our method. In principle, the heuristic approach of BLOSUM and PAM can be applied to develop a similarity matrix based on pocket sequence patterns, which would provide better measure for the task of identifying similar functional surfaces. An alternative approach that is based on a more satisfactory statistical model can be found in Ref. 71. In this study, the entries of the substitution matrix are derived from a model of continuous time reversible Markov process, whose parameters were estimated by approximate maximum likelihood method from multiple alignments of over 100 protein families.

In addition, it will be important in the future to cluster systematically all surface patterns found on protein structures by similarity in sequence and shape and to develop a classification system, such that protein surface pattern can be queried in an organized fashion, and understood in the context of other proteins with related functional surfaces.

Our method can be used to analyze details of protein functional surfaces and compare their patterns in sequence, spatial distance, and spatial orientation. Such analysis may be useful for elucidating the structural basis of specificity in binding. Another application of this method is to discover previously unknown relationship of protein surfaces. The all-against-all search conducted in this study already indicates that there are such examples in the existing Protein Data Bank. Further close examinations are required for understanding these examples. Obviously, when the query protein is a structure with unknown function, such as those obtained through structural genomics project, this method may be useful to identify the biological function of the unknown protein. Furthermore, when the requirement of statistical significance is relaxed, our method may help to detect potential targets of promiscuous binding of chemicals such as drugs to unintended proteins.

The functional relationship suggested by our method can be enhanced by independent corroborations, such as Gene Ontology classification of proteins,[72] their metabolic and/or regulatory pathway relationship, and operon organizational relationship in the case of bacterial proteins. In addition, experimental studies such as cross-binding assay of selected chemical compounds and peptide modulators will provide ultimate evidence of suggested functional relationship between protein surfaces.

One important further application of our method is to help to study evolution of protein structure and function relationship. The functional roles of proteins are results of evolution and natural selection. Biological functions are carried out through binding events occurring on the surface regions of proteins. These functionally important surfaces are architectured on main-chain folds of proteins. The relationship between functional binding surface and main-chain fold, and the way they influence each other during evolution is therefore fundamental for understanding the structure–function relationship of proteins. Extensive study of evolution of protein function and protein fold showed that a protein fold can have many functions and that similar functions can have many different structural solutions.[11,14] Our study provides additional examples of very remote similarity relationships between protein surfaces involving a large number of residues that is sometimes independent of main-chain fold, as in the case of the similar active sites of HSP90 and HIV-1 protease, and the similar binding surfaces of aromatic aminotransferase and 17-β-hydroxysteroid dehydrogenase. These examples, together with previously reported common local side-chain patterns,[19,23,24] may be results of convergent evolution. It is possible that preservation of the functional pockets may predate the emergence and maturation of protein fold. The requirement of these functional surfaces also puts some constraints on protein conformations. In this case, an intriguing possibility is to examine the structural biology of protein evolution not at protein fold level, but at the level of functional surfaces and related secondary structures. The existence of such evolutionary structural units that preserve key atoms and their spatial orientation to provide favorable locations for functional activity may be ancient, predating the formation of protein fold and domain. Computational method developed in this study may help to uncover these evolutionary structural units.

Acknowledgements

This work is supported by grants from National Science Foundation CAREER DB10133856, DB10078270, and American Chemical Society/Petroleum Research Fund 35616-G7. We thank Drs Bosco Ho, Roderic Eckenhoff, Eugene Koonin, Antony Kossiakoff, Charles Lawrence, Lee Makowski, Dagmar Ringe, and Dong Xu for helpful discussions. We thank Yaron Turpaz for assistance in data preparation and Patrick Freeman for aiding in setting up cluster computing. The results of this work will be made available at http://pvsoar.bioengr.uic.edu

References

1. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J. *et al.* (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
2. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H. *et al.* (2000). The protein data bank. *Nucl. Acids Res.* **28**, 235–242.
3. Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
4. Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. & Thornton, J. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
5. Shindyalov, I. & Bourne, P. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
6. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–602.
7. Gibrat, J., Madej, T. & Bryant, S. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385.
8. Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
9. Holm, L. & Sander, C. (1997). New structure: novel fold? *Structure*, **5**, 165–171.
10. Martin, A., Orengo, C., Hutchinson, E., Michie, A., Wallace, A., Jones, M. & Thornton, J. (1998). Protein folds and functions. *Structure*, **6**, 875–884.
11. Orengo, C., Todd, A. & Thornton, J. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
12. Sanchez, R. & Sali, A. (1998). Large scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
13. Orengo, C., Pearl, F., Bray, J., Todd, A., Martin, A., Lo, C. & Thornton, J. (1999). The CATH database provides insight into protein structure/function relationships. *Nucl. Acids Res.* **27**, 275–279.
14. Russell, R., Sasieni, P. & Sternberg, J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903–918.
15. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
16. Wilson, C., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
17. Devos, D. & Valencia, A. (2000). Practical limits of

function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.

18. Jaroszewski, L. & Godzik, A. (2000). Search for a new descriptor of protein topology and local structure. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (ISMB), pp. 211–217. AAAI Press La Jolla, CA.
19. Artymiuk, P., Poirrette, A., Grindley, H., Rice, D. & Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structure. *J. Mol. Biol.* **243**, 327–344.
20. Fischer, D., Norel, R., Wolfson, H. & Nussinov, R. (1993). Surface motifs by a computer vision technique: searches, detection, and implications for protein–ligand recognition. *Proteins: Struct. Funct. Genet.* **16**, 278–292.
21. Norel, R., Fischer, D., Wolfson, H. & Nussinov, R. (1994). Molecular surface recognition by computer vision-based technique. *Protein Eng.* **7**, 39–46.
22. Wallace, A., Borkakoti, N. & Thornton, J. (1997). TESS: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
23. Russell, R. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
24. Stark, A., Sunyaev, S. & Russell, R. (2003). A model for statistical significance of local similarities in structure. *J. Mol. Biol.* **326**, 1307–1316.
25. Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323**, 387–406.
26. Yu, L., Gaitatzes, C., Neer, E. & Smith, T. (2000). Thirty-plus functional families from a single motif. *Protein Sci.* **9**, 2470–2476.
27. Yu, L., White, J. & Smith, T. (1998). A homology identification method that combines protein sequence and structure information. *Protein Sci.* **7**, 2499–2510.
28. Zvelebil, M. & Sternberg, M. (1988). Analysis and prediction for the location of catalytic residues in enzymes. *Protein Eng.* **2**, 127–138.
29. Ota, M., Kinoshita, K. & Nishikawa, K. (2003). Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**, 1053–1064.
30. Laskowski, R., Luscombe, N., Swindells, M. & Thornton, J. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
31. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897.
32. Mücke, E. (1993). *Shapes and Implementations in Three-dimensional Geometry,* Department of Computer Science, University of Illinois at Urbana-Champaign.
33. Edelsbrunner, H., Facello, M. & Liang, J. (1998). On the definition and the construction of pockets in macromolecules. *Discrete Appl. Math.* **88**, 83–102.
34. Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. & Subramaniam, S. (1998). Analytic shape computation of macromolecules: II. Identification and computation of inaccessible cavities inside proteins. *Proteins: Struct. Funct. Genet.* **33**, 18–29.
35. Binkowski, T., Naghibzadeh, S. & Liang, J. (2003). CASTp: Computed atlas of surface topography of proteins. *Nucl. Acids Res.* **31**, 3352–3355.
36. Kedem, K., Chew, L. & Elber, R. (1999). Unit-vector rms (urms) as a tool to analyze molecular dynamics trajectories. *Proteins: Struct. Funct. Genet.* **37**, 554–564.
37. Richards, F. (1977). Areas, volumes, packing, and protein structures. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
38. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304–308.
39. Richards, F. & Lim, W. (1994). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423–498.
40. Liang, J. & Dill, K. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
41. Lorenz, B., Orgzall, I. & Heuer, H.-O. (1993). Universality and cluster structures in continuum models of percolation with two different radius distributions. *J. Phys. A: Math. Gen.* **26**, 4711–4722.
42. Liang, J., Zhang, J. & Chen, R. (2002). Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. *J. Chem. Phys.* **117**, 3511–3521.
43. Zhang, J., Chen, R., Tang, C. & Liang, J. (2003). Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. *J. Chem. Phys.* **118**, 6102–6109.
44. Edelsbrunner, H. (1995). The union of balls and its dual shape. *Discrete Comput. Geom. Des.* **13**, 415–440.
45. Facello, M. (1995). Implementation of a randomized algorithm for delaunay and regular triangulations in three dimensions. *Comput. Aided Geom. Des.* **12**, 349–370.
46. Hobohm, U. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* **1**, 409–417.
47. Batlett, G., Porter, C., Borkakoti, N. & Thornton, J. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.
48. Pearson, W. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.
49. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 915–919.
50. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
51. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.
52. Bundschuh, R. & Hwa, T. (1999). An analytic study of the phase transition line in local sequence alignment with gaps. In *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB)* (Istrail, S., Pevzner, P. & Waterman, M., eds), pp. 70–76, ACM Press, Lyon, France.
53. Pearson, W. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185–219.
54. Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 376–380.
55. Golub, G. & Van Loan, C. (1989). *Matrix Computations*, 2nd edit., Johns Hopkins University Press, Baltimore, MD.
56. McLachlan, A. (1979). Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.* **247**, 536–540.

57. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **32**, 922–923.
58. Chew, L. P., Huttenlocher, D. P., Kedem, K. & Kleinberg, J. M. (1999). Fast detection of common geometric substructure in proteins. *J. Comput. Biol.* **6**, 313–325.
59. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
60. Schumacher, M., Camp, S., Maulet, Y., Newton, M., MacPhee-Quigley, K., Taylor, S. *et al.* (1986). Primary structure of Torpedo califonica acetylcholinesterase deduced from its cDNA sequence. *Nature*, **319**, 407–409.
61. Fitzgerald, P., McKeever, B., Van Middlesworth, J. F., Springer, J., Heimbach, J. C., Leu, C. T. *et al.* (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0 Å resolution. *J. Biol. Chem.* **265**, 14209–14219.
62. Stebbins, C., Russo, A., Schneider, C., Rosen, N., Hartl, F. & Pavletich, N. (1998). Crystal structure of an hsp90–geldanamycin complex: targeting of a protein chaperone by an antitumor agent. *Cell*, **89**, 239–250.
63. Okamoto, A., Nakai, Y., Hayashi, K. & Kagamiyma, H. (1998). Crystal structures of *Paracoccus denitrificans* aromatic amino acid aminotransferase: a substrate recognition site constructed by rearrangement of hydrogen bond network. *J. Mol. Biol.* **280**, 1176–1999.
64. Okamoto, A., Ishii, S., Hirotsu, K. & Kagamiyama, H. (1999). The active site of *Paracoccus denitrificans* aromatic amino acid aminotransferase has contrary properties: flexibility and rigidity. *Biochemistry*, **38**, 1176–1184.
65. Di Gennaro, J., Siew, N., Hoffman, B., Zhang, L., Skolnick, J., Neilson, L. & Fetrow, J. (2001). Enhanced functional annotation of protein sequences *via* the use of structural descriptors. *J. Struct. Biol.* **134**, 232–245.
66. Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA*, **94**, 11911–11916.
67. Cohen, F. & Sternberg, M. (1980). On the prediction of protein structure: the significance of the root-mean square deviation. *J. Mol. Biol.* **138**, 321–333.
68. Reva, B., Finkelstein, A. & Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Fold. Des.* **3**, 141–147.
69. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
70. Altschul, S. (1991). Amino acid substitution matrices. *J. Mol. Biol.* **219**, 555–565.
71. Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699.
72. Consortium, T. G. O. (2000). Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29.

Edited by G. von Heijne

(Received 24 April 2003; received in revised form 2 July 2003; accepted 4 July 2003)