Published in final edited form as:
Adv Protein Chem Struct Biol. 2008 ; 75: 107–141. doi:10.1016/S0065-3233(07)75004-0.

Predicting and characterizing protein functions through matching geometric and evolutionary patterns of protein binding surfaces

Jie Liang[a], **Yan-Yuan Tseng**[c], **Joseph Dundas**[a], **Andrew Binkowski**[b], **Andrzej Joachimiak**[b], **Zheng Ouyang**[a], and **Larisa Adamian**[a]

[a] Program of Bioinformatics, Depart of Bioengineering, University of Illinois at Chicago

[b] Structural Biology Center and Midwest Center for Structural Genomics, Biosciences Division, Argonne National Laboratory, Illinois 60439, USA

[c] Department of Ecology and Evolution, University of Chicago, Illinois 60637, USA

Abstract

Predicting protein functions from structures is an important and challenging task. Although proteins are often thought to be packed as tightly as solids, closer examination based on geometric computation reveals that they contain numerous voids and pockets. Most of them are of random nature, but some are binding sites providing surfaces to interact with other molecules. A promising approach for function inference is to infer functions through discovery of similarity in local binding pockets, as proteins binding to similar substrates/ligands and carrying out similar functions have similar physical constraints for binding and reactions. In this chapter, we describe computational methods to distinguish those surface pockets that are likely to be involved in important biological functions, and methods to identify key residues in these pockets. We further describe how to predict protein functions at large scale (millions) from structures by detecting binding surfaces similar in residue make-ups, shape and orientation. We also describe a Bayesian Monte Carlo method that can seperate selection pressure due to biological function from pressure due to protein folding. We show how this method can be used to reconstruct the evolutionary history of binding surfaces for detecting similar binding surfaces. In addition, we briefly discuss how the negative image of a binding pocket can be casted, and how such information can be used to facilitate drug discovery.

Keywords

Local binding surface; protein function; pocket; void; Bayesian Monte Carlo; CASTp; pvSOAR; alpha shape

1. Introduction

The structural genomics projects have made significant contributions to our current body of knowledge of protein structures [1]. They have further faciliated the establishment of a comprehensive view of the global universe of protein structures, and have provided a foundation with a wealth of information for developing model and computational tools that can be used to understand the molecular mechanism how individual proteins carry out their biological roles and how protein functions evolve.

Functional characterization of proteins with unassigned functions is an important task. By design, a large number of newly determined protein structures from structural genomics are not related to other known proteins, and bioinformatics tools based on sequence alignment often cannot provide accurate information about the functional roles of these proteins. Several early studies showed that reliable functional assignment will require sequence identity of 60–70% between the protein of unknown function and a well-studied protein [2,3].

Recently, the approach of inferring protein functions by detecting local spatial regions on protein structures with similar patterns has been shown to be very effective [4,5,6,7,8,9,10,11,12]. The rationale behind this approach is intuitive and appealing. For proteins binding to similar substrates or ligands and carrying out similar functions, they are constrained by the requirement of providing the necessary microenvironment for similar binding and biochemical reactions to occur. These physical constraints are reflected by similarity in the shape of local binding surfaces and in the physico-chemical texture of the binding surfaces. In order for similar functions to occur, the evolution of residues involved in binding and reaction will be constrained and this results in similarly allowed and forbidden residue substitution on binding surfaces [11].

In this chapter, we discuss our approach to predict and characterize protein functions from protein structures by comparing local surfaces. We first discuss the existence of voids and pockets, and their distribution in proteins [13]. We then describe how to identify those that are likely to be functionally important, as well as the key residues on them [14]. This is followed by a discussion on how to match local surfaces and how to assess their similarity in both sequence order dependent and independent fashion [5]. Next we discuss how to extract evolution patterns of small local regions directly related to protein function and unaffected by folding requirement using a Bayesian Monte Carlo method, and how this approach improves protein function prediction [11]. We then describe three examples of protein function prediction and characterizations using proteins generated from the Midwest Center for Structural Genomics [15]. This is followed by a brief discussion on how further information from computed protein local binding pockets can be extracted in the form of negative image to guide for selecting inhibitors from a collection of candidate compounds [16].

2. Voids and Pockets in Protein Structures and their Origins

Protein structure is known to be packed tightly. The packing density of protein interior is comparable to that of solid, with low compressibility [17]. Protein packing has been described as a jig-saw puzzle [18]. However, detailed study using the technique of alpha shape [19,20,21,22] revealed that there are numerous voids and pockets in protein structures (Fig 1) [13].

Here voids are enclosed empty space that is inaccessible to a water molecule modeled as a probe of 1.4 Å radius, and pocket is an empty space in the protein that has a constricted opening to the bulk exterior and is accessible to a water molecule (Fig 1). The size of the void or pocket in this study is required to be large enough to contain at least one water molecule. In fact, there is a scaling relationship between the number of voids and pocket and the chain length of the protein (Fig 2a). On average, there is an increase of 15 voids or pockets for every 100 amino acid residues [13]. For example, the binding sites of HIV-1 protease and phosphatidylinositol transfer protein (PITP) both corresponds to well-defined surface pockets (Fig 3).

Various scaling relationships suggest that protein packing is of random nature [13]. For example, if we use a simple solid ball packing as a model of protein, we would expect that the volume $V = 4\pi r^3/3$ and the area $A = 4\pi r^2$ should have a scaling relationship of $V \propto A^{3/2}$. In reality, this scaling relationship is linear (Fig 2b). This linear relationship is reminiscent of the scaling relationship of clustered random spheres in off-lattice and on-lattice models [23,24].

To further investigate the nature of protein packing and the origin of voids and pockets, we have studied the packing behavior of random chain polymer in off-lattice three-dimensional space [25]. Other than the requirement that these polymer chains are compact and self-avoiding, there is no relationship between these studied chains and real protein. The task of assessing the ensemble properties of packing of these chain polymers in a statistically accurate manner is technically very challenging, as one needs to generate adequate samples that are independent and properly weighted. This relates to the well-known attrition problem: the success rate of generating self-avoiding chain polymers is rapidly diminishing with the increase of chain length, as it becomes exponentially difficult to maintain the self-avoiding requirement. For example, even for a short chain of length 48, the success rate of using simple growth method would be only 0.79% [26].

Using the sequential Monte Carlo method [27,28], we have overcome this technical difficulty, and succeeded in generating properly weighted ensemble of thousands of self-avoiding chains up to length 2,000 [25]. We have carried out the same geometric analysis on these chain polymer structures, just as we did with protein structures. The results indicate that both the scaling relationship of the coordination number, and the packing density with the chain length show characteristically the same scaling relationship as that of proteins [25]. Altogether, these findings provide strong evidence that proteins are not optimized by evolution to eliminate voids and pockets. Rather, the majority of the voids and pockets simply emerge from the requirement of packing self avoiding chains in a compact space.

3. Identifying Functional Surfaces of Proteins

The existence of numerous voids and pockets poses two challenging problems. First, how do we identify the void(s) and pocket(s) that are biologically important, *e.g.*, how to distinguish those involved in binding and biochemical reactions from those formed by random chance. Second, for a given pocket or voids found on a protein structure, how do we know if it is important for some biological functions known or yet to be discovered?

We have developed a method to address these problems for enzymes. In this method, we do not directly compare the structure or function of a well-characterized protein with the protein in question. Rather, we seek to recognize pocket or void that might be involved in enzyme function based on general characteristics. We discuss in later sections the comparative approach when the unknown query protein is compared with a database of protein structures.

Typically, about 10%–30% of all residues in an enzyme participate in the formation of the binding pocket [14]. Compared to the full length primary sequences, the usage of residues in forming pocket is biased. Often His, Asp, Glu, Ser, and Cys account for the most important active site residues [29,5,8,14]. These are residues known to be important for catalytic functions. On the other hand, nonpolar residues such as Val, Leu, Pro are far less frequent in enzyme binding pocket [14]. Although these hydrophobic residues are frequently conserved for maintaining protein structures and for protein folding, they are often not directly involved in molecular functions of enzymes. In fact, the composition of residue on binding surfaces of enzyme is very different from that of the overall sequences (Fig 4).

In our method for identifying functional region from enzyme structures [14], we examine the occurrence of the *atomic pattern* of a residue with exposed surface in the binding pocket. That is, we record the residue type and all of the exposed atoms from this residue, along with the secondary structure environment this residue belongs to. A probability function for each atom pattern, residue type, and secondary structure is then constructed based on statistical analysis of a database of annotated key residues of enzymes. After evaluating this probability function for each residue in a candidate pocket, we can sum up the probability values for all residues in the identified pocket, and if its above a threshold value, a functional binding pocket is predicted, and the few residues with the highest probability values are further predicted to be functionally important key residues.

This method has been shown to work well in a 10-fold cross-validation test of 3,503 protein surfaces from 70 proteins, with a sensitivity of 92.9% and specificity of 99.88% [14]. We have also shown that for four enzyme families (2,3-dihydroxybiphenyl dioxygenase, E.C. 1.13.11.39; adenosine deaminase, E.C. 3.5.4.4; 2-haloacid dehalogenase, E.C. 3.8.1.2, and phosphopyruvate hydratase, E.C. 4.2.1.11), the key residues predicted are also consistent with annotated information contained in the Structure-Function Linkage Database (SFLD) [30]. Fig 5 illustrates the example of predicted binding surface and key residue on a structure of alpha amylase.

4. Matching Local Binding Surfaces

A different approach that can potentially yield rich information is to compare the local surface of a binding pocket to a database of local surfaces, some of which have known biological characterization. Fig 6 illustrates an example. The cAMP dependent protein kinase (1cdk) and Tyr protein kinase c-src (pdb 2src) share only 13% sequence identity. However, the ATP binding pocket have similar shape and chemical texture. Once these ATP binding pockets are identified and computed from their structures, we can select the residues located on the wall of the binding pocket, and remove residues on the loops connecting these wall residues. It is clear that the remaining sequence fragments have much higher sequence identity (51%). In both cases, the residues forming the pocket wall come from diverse regions in the primary sequences.

The simple example shown in Fig 6 suggests an effective strategy that can rapidly decide if two pocket surfaces are similar. We can derive surface patterns from the residues forming the walls of pockets (called pvSOAR patterns for pocket and void surface patterns of amino acid residues), and rapidly compare these patterns. Once a pair of protein surfaces are found to be similar, we can further examine their shape and chemical texture in detail, and determine the statistical significance of their overall similarity. This approach is generally applicable to any two surface patterns of pockets and voids [5].

There are several technical problems to be solved for this approach to be generally useful. We need to identify and generate local surfaces automatically and accurately. This can be achieved by applying void and pocket algorithm for exhaustive identification and measurement of voids and pockets from protein structures [20,21,22]. We also need to rapidly and accurately assess surface similarity. Once a pair of similar local surfaces are found, we need to evaluate whether the similarity is statistical significant.

4.1. Comparison of Sequence Patterns of Surface Pockets and Voids

Sequence order dependent method—By concatenating wall residues of a pocket or void on a peptide chain, we have compiled a database of pvSOAR sequence patterns for all protein structures in the protein data bank. This database is part of the CASTp database [31,32]. It currently (August, 2008) contains 46,071 protein structures, with 1,582,472 voids and 1,555,994 pockets. We can rapidly query a protein surface pocket against CASTp database through alignment of sequence fragments using standard dynamic programming technique, allowing gap insertion [5]. In this approch, we assume that the residues in the sequence pattern are positioned following their order of the primary sequence.

Sequence order independent comparison—The alignment of pvSOAR sequence fragments through dynamic programming can discover many similar binding pockets. However, there are many cases where two proteins with similar placement of amino acids in their tertiary structures have different relative positioning of these amino acids in their primary structures (see Fig 7 for stromelysin). When comparing two local surface pockets, we also need to detect similar residue patterns while ignoring their strict positioning in the primary structures. This is the problem of finding which amino acid on the query protein surface pocket is equivalent to which amino acid on the target protein surface pocket.

Sequence order independent matching of pockets can be formulated as a maximum weight bipartite matching problem, where graph nodes represent amino acids (*e.g.*, using C_α atoms) from the two protein pockets. Directed edges are used to connect nodes from the query protein to nodes of the target protein, if the two nodes share some similarity (*e.g.*, by a scoring function based on shape and chemistry). Each edge is given a weight that is based on the similarity measure. The problem is to find a set of edges connecting nodes in query pocket to nodes in target pocket, with maximized total edge weight, while insisting only at most one edge is selected for each residue [33].

One way to solve this problem is by using the Hungarian algorithm [34] as described in [35] with modifications. This is an iterative method that uses the Bellman-Ford algorithm [36]. First, we add a fictitious source node s that connects to every query node with 0-weight. We then add a fictitious destination node d that connects to every target node with 0-weight. The Bellman-Ford algorithm computes the distance $F(i)$ of the shortest path(s) from the source node to each of the remaining node i. The weight for each edge that does not contain the source node is then updated. The new weight $w'(i, j)$ for edge $e(i, j)$ starting from node i to node j is:

$$w'(i,j)=w(i,j)+[F(i)-F(j)].$$

An overall score F_{all}, initialized to 0, is now updated as $F'_{\text{all}}=F_{\text{all}}-F(d)$. Next, we flip the directions of all edges in the shortest path from the source s to the destination d.

We then apply the Bellman-Ford algorithm on this new graph, and this is repeated until either there is no directed path from s to d as edges have been flipped, or the shortest distance $F(d)$ to the destination is greater than the current overall score F_{all}. The output of the Hungarian algorithm includes a set of directed edges starting from target nodes to query nodes, and these provide the equivalence relationship, namely, which residue in the target pocket should be aligned to which residue in the query pocket. Based on this equivalence relationship, we can then compute the shape similarity between these two surface pockets at atomic details, as described below. When we use atoms as nodes instead of residues, the results will be atomic alignment of pocket surfaces.

4.2. Comparison of Shapes of Surface Pockets and Voids

Once two voids or pockets are found to have significant sequence similarity, we then follow up with more detailed shape analysis using two methods. First, we compute the coordinate root mean square distance (cRMSD) between the subset of equivalent residues or atoms. This equivalence relationship is established by the local alignment of pocket sequence fragments. The cRMSD distance is measured when the subset of residues are optimally aligned with rigid motion and has the least RMSD value. This alignment and the cRMSD value can be computed from the singular value decomposition of the correlation matrix of the coordinates of the point sets [37].

cRMSD is not a perfect measure of shape similarity. It works well when two structures are similar, but is sensitive to outliers. If a protein experiences conformational change, its binding pocket may expand or shrink and its residues may retain the relative orientational relationship, but with significantly altered Euclidean distances. To address this deficiency, we can use the orientational RMSD (oRMSD) measure [5]. We first place a unit sphere at the geometric center of the pocket. The location of each residue is then projected onto the unit sphere along the direction of the vector from the geometric center. The projected pocket is therefore represented by a set of unit vectors on the unit sphere, which preserves the original orientational relationship. The RMSD of the two sets of unit vectors for the two pockets in comparison can then be measured, which gives the oRMSD value [5].

For sequence-order-independent comparison of two surface pockets, we start from a crude initial equivalence relationship that represents the initial correspondence between residues from query and target pockets. We then apply the optimal rotation matrix and translation vector computed using [37] to this initial alignment. The Euclidean distances between residues (or atoms) in the query pocket and target pocket are then computed after the optimal superposition. Those that are below a threshold are updated with new weights computed using a similarity scoring function. The Bellman-Ford algorithm and the SVD based optimal alignment and update of Euclidean distances are then repeated iteratively. One can stop this iterative process if the improvement is less then a threshold. As the overall alignment shape score may deteriorate temporarily when a new equivalence relationship is found and new superposition applied, simulated annealing allowing a probability that structural alignment may temporarily deteriorate can also be applied here [35].

As an illustration, the sequence order independent alignment of surface pockets in two structures of stromelysin shown in Fig 7. It has an overall cRMSD of 0.76 Å for 29 atoms from 10 residues. The C_α atoms from these 10 residues align with an cRMSD of 1.05 Å. The alignment obtained in a sequence-order-dependent fashion contains 16 residues. If we select the subset of 10 residues from these 16 residues that overlap most with that of the sequence order independent alignment, the alignment of their C_α atoms has a cRMSD value of 3.71 Å. This example illustrates that this method of sequence-order-independent comparison of two surface pockets works well, and often can identify excellent surface matches that are challenging for other methods (Dundas and Liang, unpublished).

4.3. Statistical Significance

After the similarity of two surface pockets is calculated, we need to assess its statistical significance to aid in biological interpretation. pvSOAR sequence patterns are typically short, and are of different composition from the full chain sequences. In addition, frequently the two pocket sequence patterns in comparison have different number of residues. Although the theoretical model of extreme value distribution (EVD) provides accurate description of gapless local alignment of random sequences [38], no exact theoretical models are known in general for local sequence alignment of very short sequences with gaps.

We have developed a heuristic approach to assess the statistical significance of two pocket pvSOAR sequences aligned in sequence order. By removing the largest peak in the low-score region of the distribution of alignment scores of random short sequences which often contain just one or two matched residues, we found that the remaining distribution can be described by an extreme value distribution well [5]. Specifically, the Smith-Waterman scores of the search results of a query sequence pvSOAR pattern to a database of randomly shuffled pocket sequences are collected. They are then fitted to an EVD distribution, and the goodness-of-fit is then evaluated using the Kolmogorov-Smirnov test [39]. If the observed Kolmogorov-Smirnov statistic doe not indicate that the random scores are inconsistent with an EVD distribution, we further estimate the statistical significance p-value using the calculated *z*-score $z = (S - \mu)/\sigma$, where *S* is the similarity score, μ the mean of random scores, and σ the standard deviation. The *p*-value can be estimated from the *z*-score as [5]:

$$p(Z>z)=1-\exp(-e^{-1.282z-0.5772}).$$

The expected number *E* of random pocket sequences with the same or better score can be calculated as:

$$E=p\times N_r,$$

where N_r is the number of randomly shuffled sequence fragments. The *p*-value or *E*-value can be used to exclude matched pairs of pocket pvSOAR sequences that are unlikely to be biologically relevant.

Once the cRMSD or oRMSD value is calculated for two surface pockets, we also need to evaluate the statistical significance of shape comparison. As illustrated above, a common practice for determining statistical significance is to assume the similarity score are drawn randomly from a specific underlying distribution. The parameters of the assumed distribution are then estimated by curve-fitting the distribution of scores from the random comparison of protein pockets. The derived parameters can then be used to find the *Z*-score or *p*-value of a given similarity score [40,41,42,43]. We found that the distribution of both cRMSD and oRMSD for random surfaces on protein structures do not follow known parametric model such as the extreme value distribution [5]. We empirically estimate the probability *p* of obtaining a specific cRMSD or oRMSD value for *n* number of matched positions from a set of randomly generated surface pockets and voids. By collecting cRMSD and oRMSD values of millions of randomly matched pockets with different number of selected matched residues, we can estimate the *p*-value of a specific cRMSD or oRMSD of with a specific number of matched residues. This can be found by finding the closest value of the rank order statistic in the randomly collected cRMSD or oRMSD data of the same number of residues [4,5].

5. Uncovering Evolutionary Patterns of Local Binding Surfaces

Fast comparison of pvSOAR sequence fragments is a key step when querying a specific surface pocket/void against a database of precomputed pocket/voids, as the database can contain hundreds of thousands or millions of entries. This is possible by applying fast dynamic programming method to align the sequence fragments representing the two pockets/voids. This step is carried out before promising hits are identified and further detailed shape comparison is carried out.

The specific scoring matrix used to assess the similarity of two aligned pocket/void sequence fragments is critical for detecting functionally related binding pockets/voids. A convenient choice is to adopt widely used PAM matrice or BLOSUM matrice [44,45]. A disadvantage of this approach is that these are precomputed matrice and have implicit parameters with values predetermined from the analysis of large quantities of sequences, which contain little information of the protein of interest. Another approach is to use position specific scoring matrix (PSSM) such as those generated by the PSI-BLAST program [46]. The drawback of this latter approach is that it often leads to serious bias as the PSSM is derived from all sequences aligned to the query sequence satisfying certain statistical significance requirement. Bias comes from the fact that all aligned sequences contribute equally to the derivation of PSSM, regardless how closely or distantly they are related. This is particularly problematic if the query results from the database is dominated by closely related proteins.

5.1. Evolution Model

To resolve these issues, we have adopted an approach that models the evolutionary process using a continuous time Markov process and an explicit phylogenetic tree [11]. Markovian evolutionary models are parametric models and do not have pre-specified parameter values. These values are instead estimated from specific sequence data relevant to the protein of interests [47]. This approach has been shown to be more effective in deriving informative rate matrice with significant advantage over matrices obtained from other methods [47].

We assume that a reasonably accurate phylogenetic tree $\boldsymbol{T}$, the branch lengths of the tree representing divergence time, and an accurate multiple sequence alignment are known. These can be computed using maximum likelihood method or Bayesian method [48,49,50]. The subset of columns in the multiple sequence alignment corresponding to the residues in the binding pocket are then identified based on pocket calculation [5,51,11]. Our model assumes that the evolution of the residues in the binding pocket can be modeled by a Markovian process characterized by a 20×20 matrix $\boldsymbol{Q} = \{q_{ij}\}$ of instantaneous substitution rates. The divergence time t is measured in the unit of the expected number of residue changes per 100 sites between the sequences.

Once the instantaneous substitution rate matrix $\boldsymbol{Q} = \{q_{ij}\}$ is known, the matrix of probabilities of substitution of residue i by residue j in the time interval t can be computed as:

$$\boldsymbol{P}(t)=\{p_{ij}(t)\}=\exp(\boldsymbol{Q}\cdot t).$$

For symmetric $\boldsymbol{Q}$, the matrix exponential can be conveniently computed as:

$$\exp(\boldsymbol{Q}\cdot t)=\boldsymbol{U}\exp(\Lambda t)\boldsymbol{U}^{-1},$$

where $\boldsymbol{U}$ is the matrix of right eigenvectors of $\boldsymbol{Q}$, and $\boldsymbol{U}^{-1}$ is that of the left eigenvectors. A technique to construct a more general non-symmetric instantaneous rate matrix $\boldsymbol{Q}$ that can be symmetrized can be found in [52,11].

For a column in the multiple sequence, we follow the phylogenetic tree $\boldsymbol{T}$ and compute the transition probability $p_{x_i x_j}(t_{ij})$ for each of the edge in the tree, whose length denotes the time interval $t_{i,j}$. Here x_i and x_j are the residues at the positions corresponding to the nodes connected by the edge. If we knew all the ancestral sequences (corresponding to the internal nodes in the phylogenetic tree) of the extant sequences (corresponding to the leaf nodes), the likelihood given the tree $\boldsymbol{T}$ and the instantaneous rates $\boldsymbol{Q}$ for this column h can be obtained by combining probabilities along all edges:

$$p(\boldsymbol{x}_h|\boldsymbol{T},\boldsymbol{Q})=\pi_{x_k}\prod p_{x_i x_j}(t_{ij}).$$

Here the $\boldsymbol{\pi}_{xk}$ is the prior probability of an arbitrarily chosen node k as the starting node taking its residue as type x_k at column h. $\boldsymbol{\pi}_{xk}$ typically can be computed as the composition of the aligned sequences. The product sign Π is over all edges in the phylogenetic tree. Since in reality we do not know the identities of the residues in ancestral sequences, we sum over all possible values the ancestral sequence might take in this column, and the probability $p(\boldsymbol{x}_h|\boldsymbol{T}, \boldsymbol{Q})$ of observing this particular column h in the multiple sequence alignment is:

$$p(\boldsymbol{x}_h|\boldsymbol{T}, \boldsymbol{Q}) = \pi_{x_k} \sum \prod p_{x_i x_j}(t_{ij}).$$

Here the summation sign Σ is overall all possible residues in this column for each of the ancestral sequences.

Treating each column independently, the probability $P(\mathcal{S}|\boldsymbol{T}, \boldsymbol{Q})$ of observing all residues in the selected columns for the functional region $\mathcal{S}$ is:

$$P(\mathscr{S}|\boldsymbol{T}, \boldsymbol{Q}) = P(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s|\boldsymbol{T}, \boldsymbol{Q}) = \prod p(\boldsymbol{x}_h|\boldsymbol{T}, \boldsymbol{Q}).$$

Here the product Π sign is over all columns.

5.2. Estimating Model Parameters *Q* and Bayesian Monte Carlo

We adopt a Bayesian framework, and each model parameter is described with a distribution instead of a single value. The *posterior probability* $\boldsymbol{\pi}(\boldsymbol{Q}|\mathcal{S}, \boldsymbol{T})$ of the rate matrix for a given aligned pocket region $\mathcal{S}$ and the phylogenetic tree $\boldsymbol{T}$ integrates our prior information (represented by the prior distribution $\boldsymbol{\pi}(\boldsymbol{Q})$) on the model parameters, and the likelihood function-related probability $P(\mathcal{S}|\boldsymbol{T}, \boldsymbol{Q})$ derived from the observed data:

$$\pi(\boldsymbol{Q}|\mathscr{S}, \boldsymbol{T}) \propto \int P(\mathscr{S}|\boldsymbol{T}, \boldsymbol{Q}) \cdot \pi(\boldsymbol{Q}) d\boldsymbol{Q}.$$

Once this posterior distribution is known, we can calculate the posterior mean of the parameters:

$$\mathbb{E}_{\pi}(\boldsymbol{Q}) = \int \boldsymbol{Q} \cdot \pi(\boldsymbol{Q}|\mathscr{S}, \boldsymbol{T}) d\boldsymbol{Q}.$$

In practice, we generate correlated samples from the posterior distribution, and the posterior means of the model parameters are estimated from these samples:

$$\mathbb{E}_{\pi}(\boldsymbol{Q}) \approx \sum \boldsymbol{Q}_i \cdot \pi(\boldsymbol{Q}_i|\mathscr{S}, \boldsymbol{T}).$$

Samples drawn from the desired posterior distribution $\boldsymbol{\pi}(\boldsymbol{Q}|\mathcal{S}, \boldsymbol{T})$ are generated by running a Markov chain. Briefly, we start with an initial set of parameter values for $\boldsymbol{Q}$. The new parameter set $\boldsymbol{Q}_{t+1}$ at time $t + 1$ is generated from a proposal transition function $T(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1})$. It will be either accepted or rejected by following the acceptance rule denoted as $A(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1})$. The criterion in designing the acceptance rule is to ensure that the detailed balance

$$\pi(\boldsymbol{Q}_t|\mathscr{S}, \boldsymbol{T}) \cdot A(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1}) = \pi(\boldsymbol{Q}_{t+1}|\mathscr{S}, \boldsymbol{T}) \cdot A(\boldsymbol{Q}_{t+1}, \boldsymbol{Q}_t),$$

is observed. This is necessary for the samples generated by the Markov chain to follow the desired posterior probability distribution $\boldsymbol{\pi}(\boldsymbol{Q}|\mathcal{S}, \boldsymbol{T})$. The move set behind the proposal transition function that generates new trial parameter set is very important for efficient computation. Its design is discussed in [11].

The Metropolis-Hastings acceptance rule

$$r(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1}) = \min\{1, \frac{\pi(\boldsymbol{Q}_{t+1}|\mathscr{S}, \boldsymbol{T}) \cdot T(\boldsymbol{Q}_{t+1}, \boldsymbol{Q}_t)}{\pi(\boldsymbol{Q}_t|\mathscr{S}, \boldsymbol{T}) \cdot T(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1})}\},$$

is a rule that ensures detailed balance. It either accepts or rejects the proposed new parameter set $\boldsymbol{Q}_{t+1}$ by evaluating whether a random number u generated from the uniform distribution between 0 and 1 is no greater than $r(\boldsymbol{Q}_t, \boldsymbol{Q}_{t+1})$.

5.3. Deriving Scoring Matrice from Rate Matrix

Once the expected values for the rate matrix $\boldsymbol{Q}$ is obtained, we follow the framework by Karlin and Altschul and derived scoring matrix used for assessing the similarity between residues at different time interval [46]. For residue i and residue j at time interval t, the similarity score $b_{ij}(t)$ can be computed as:

$$b_{ij}(t) = \frac{1}{\lambda} \log \frac{p_{ij}(t)}{\pi_j} = \frac{1}{\lambda} \log \frac{m_{ij}(t)}{\pi_i \pi_j},$$

where $m_{ij}(t)$ is the joint probability of observing both residue type i and j at the two nodes separated by time t, and λ is a scalar [46].

5.4. Validity of the Evolutionary Model

The validity of this approach is confirmed by extensive simulation test. In [11], an explicit phylogenetic tree and 16 artificially evolved sequences of carboxypeptidase A2 are used to test if the underlying model of substitution rate parameters of Jones, Taylor and Thornton (JTT) [53] used to generate the artificial sequences can be recovered. In 50 independent simulations, the recovered rates and the true JTT parameters all have the weighted mean error (as defined in [54]) less than 0.0045. In addition, the parameters can be recovered with acceptable accuracy when only about 20 residues in total size is used [11].

5.5. Evolutionary Rates of Binding Surfaces and Other Surfaces Are Different

We have calculated the substitutionrate matrix for both the binding surface region and the remaining surface region of alpha amylase. The distinct selection pressure for functional surface is also clearly evident in the different patterns of the inferred substitution rates for binding region and for the rest of the protein surface region (Fig 8)[11]. In addition, both substitution patterns are also very different from the precomputed JTT model [53]. This example illustrates the need of extracting evolution pattern specific to the functional surfaces of a particular protein for constructing sensitive and specific scoring matrix for detecting functionally related protein surfaces. It also indicates that selection pressure specific for protein function can be extracted without being altered by selection pressure due to folding.

6. Predicting protein function by detecting similar biochemical binding surfaces

Amylase and other enzymes

Alpha amylase (Enzyme Classification number 3.3.1.1) is an enzyme that breaks down starch, glycogen, and other related polysaccharides and oligosaccharides. An objective test for protein function prediction is to take a known amylase structure and ask if it is used as a template, whether we can find all other amylase structures in the protein data bank (PDB) and nothing else. This is a challenging task, as amylase exist in diverse species, and some of them have very low sequence identity (< 25%), which is challenging for function inference.

Using the template structure `1bag` from *B. subtilis*, we are able to identify one of the computed pocket containing 18 residues as the binding pocket, (Fig 9). With multiple sequence alignment of 14 sequences homologous to the template `1bag`, all with < 90% seqeunce identity to the template or to each other, we have constructed a phylogenetic tree using the Molphy package (Fig 9a) [48]. The rate matrix $\boldsymbol{Q}$ for the binding region (which correspond to the positions of the 18 residues) are then estimated using the Bayesian Monte Carlo method we developed [11]. Scoring matrice of different divergence time are then generated from this rate matrix $\boldsymbol{Q}$. These scoring matrice are then used to evaluate the similarity for each of the >2 million precomputed pocket/void sequence fragment contained in the pvSOAR database [55] with the query sequence fragment. This comparison is carried out using the Smith-Waterman method as implemented in the FastA package [39]. Promising hits with *E*-value < 0.1 are then selected for further shape analysis. Those with cRMSD or oRMSD values with the template surface pocket at a statistical significance of $p < 0.01$ [5] are then chosen as predicted hits, namely, proteins that are predicted as alpha amylase.

Using this template, we are able to predict 58 other PDB structures as alpha amylase. Indeed, all of them are found to have the same EC number as that of `1bag`. When following the same procedure but using a different PDB template `1bg9` from the plant barley, we can predict 48 other PDB structures to be alpha amylase, again in this case all are of the same E.C. number as that of `1bg9` and `1bag` [11]. Combining the hits using these two templates together, we are able to identify 69 PDB structures of alpha amylase among the 75 known alpha amylase structures. This method using specific matrix estimated by Bayesian Monte Carlo compares more favorably than using the general JTT matrix, and than using the iterative dynamic programming sequence alignment method Psi-blast. Details can be found in [11].

This method has been tested for other enzymes. The results for 2,3-dihydroxybiphenly dioxygenase (E.C. 1.13.11.39), adnosine deaminase (E.C. 3.5.4.4), 2-haloacid dehalogenase (E.C. 3.8.1.2) and phosphopyrovate hydratase (E.C. 4.2.1.11) are described in [11], where all other protein structures of the same E.C. numbers are correctly predicted. In a recent study, we have selected a set of 100 enzyme families with about 6,000 structures and 770,000 precomputed binding surface pockets/voids for testing. By taking the structure with the best resolution and R-factor as template, we test if our method can identify other members of the same protein family and nothing else. After calculating the overall sensitivity and specificity of predictions of all 100 protein families, the accuracy of predictions for the functions of all 6,000+ structures from the 100 protein family is 92%, and the best Mathews coefficient is 86.6% (Tseng and Liang, unpublished).

Identifying metal cofactor of YecM from E. coli

The problem of predicting ion-specificity of YecM protein structure is studied in [15]. YecM protein (pdb 1k4n) from *E. coli* was chosen as a structural genomics target, as it does not have recognizable similarity to other proteins of known structures. Structural analysis indicates that YecM shares some similarity to an isomerase and several oxidorectases [56]. As these proteins all contain a divalent metal cation, it was predicted that YecM is a metal binding protein, but the preferred metal ions were not known.

In order to predict the metal cofactor more accurately, the putative metal binding pocket on the YecM structure was compared against all known metal binding surfaces in the PDB database using pvSOAR [15,55]. The results of surface alignment indicates that several zinc binding surfaces from diverse species (*Rattus norvevgicus, Bacillus thermoproteolyticus*, and *Bacillus anthracis*) share strong similarity to that of YecM, all with significant *p*-values [15]. In fact, the top 30% of a rank ordered list of all significant hits are zinc binding surfaces. In contrast, binding surfaces for other metal ions (*i.e.* Co, Mn, Fe, and Mg) have less significant similarity to that of YecM. This result suggest that YecM is likely to have zinc as its preferred metal cofactor.

Locating the active site of ribose 5-phosphate isomerase

pvSOAR analysis helped to identify the active site of another protein from structural genomics project [15]. RpiB protein from *E. coli* (pdb 1nn4) is known to have ribose-5-phosphate isomerase activity. However, the active site on this protein is unknown [57]. Although RpiA and RpiB have similar function, these two proteins belong to two different structural folds [15]. The active site of RpiA as identified by mutagenesis and co-crystal structure with inhibitor is absent on RpiB structure [57]. A ligand docking study suggested that the active site of RpiB from *M. tuberculosis* is located at the dimer interface [15].

Pairwise comparisons of the active sites using pVSOAR show that the active sites of RpiA and RpiB from *E. coli* and *M. tuberculosis* have similar area and volume, and the active sites on RpiB from *E. coli* and *M. tuberculosis* have almost identical geometry measured in both cRMSD and oRMSD, with strongly conserved phosphate binding residues. Detailed analysis further reveals that the most notable difference between RpiA and RpiB is in the composition of basic residues, where His/Arg in RpiB are replaced by Lys in RpiA. The surface patches of positively charged residues, and the orientation of acidic and basic residues important for catalysis are all conserved for these proteins to carrying out similar functions.

Although biochemical assays clearly indicates that all three proteins have the same substrate, and they are likely to have very similar binding surfaces, the location and identities of the binding surfaces cannot be detected without surface comparison, as RpiA and RpiB have no detectable similarity in overall sequence and structural fold. This study indicates that pvSOAR analysis can help to understand how two seemingly different binding surfaces performed the same function.

Putative adenine nucleotide binding site on CBS domain

CBS-domains are present in many species and have unknown specific functions, but are thought to be part of an energy status sensor complex [58]. They appear in AMP-activated protein kinase, IMP dehydrogenase-2, and chloride channel CLC2 binding adenosyl moieties (such as AMP, ATP, or S-adenosyl methionine), and are often found in tandem pairs [59,58]. Their biochemical roles and the locations of the active sites are uncharacterized.

In the study of [15], three structures of different proteins from different species of archaea and bacteria containing CBS domains are analyzed (Fig 10). These domains have about 20% sequence identities, which is insufficient for functional inference. Surface patches from the structures of these domains are identified and searched against a library of AMP and ATP binding surfaces for potential matches. Among these, well-defined interface pockets are identified by CastP computation, and strong hits of diverse AMP and ATP binding surfaces are found that are similar to these interface surfaces [15]. The results suggest that both tandem CBS domains from protein mt1622 (pdb 1pbj from *M. thermoautotrophicum*) and inosine-5′-monophate dehydrogenase (IMPDH from *S. pyogenes*, pdb 1zfj) can bind to AMP and ATP, consistent with experimental studies [58].

An unexpected finding for hypothetical protein Ta549 CBS from *T. acidophilum* is that an alternative binding surface is found to have formed by a C-terminal additional insert of the singleton CBS domain, and a CBS domain tandem pair on a different chain. This binding surface has only weak similarity to the above-mentioned binding surface of the tandem CBS pairs, but showed strong similarity to ATP binding surface on saicar-synthase from *S. cerevisiae*. This finding suggests the existence of multiple binding sites in a CBS binding domain, stabilized by a third CBS domain.

7. Adaptive patterns of spectral tuning of proteorhodopsin from metagenomics projects

Our method can also be applied to protein sequences with only limited structural information to gain biological insight [60]. Proteorhodopsins (PR) are a class of newly discovered retinal-containing rhodopsins with structural and functional similarities to archaeal bacteriorhodopsins [61,62]. They are found in numerous marine bacteria and archaea through metagenomics studies of the communities of marine organisms. A number of homologous proteorhodopsins were functionally expressed in *E. coli* and found to form active, light-driven proton pumps in the presence of retinal [61,63,64,65].

The absorption maxima of light wavelength of several subfamilies of protorhodopsins span the spectral range from blue (490 nm) to green (525 nm) [66]. The absorption maxima correlate with the depth at which the samples were collected, *e.g.*, green absorbing pigments (GPR) are found at the surface, and blue absorbing pigments (BPR) are found at the deeper waters [62]. Spectroscopic and mutagenesis analyses indicate that a single residue difference at the position 105 (Leu in GPR and Gln in BPR) functions as a spectral tuning switch and accounts for most of the spectral differences [66]. Residues A, E, M, and V also appear at the position 105 in the family of green absorbing pigments, each with a specific absorption maximum [66,67].

Based on sequence similarity to the archaeal bacteriorhodopsin with known structures, we have mapped out 13 non-redundant putative retinal-binding pocket sequence fragments from 99 sequences of proteorhodopsins [60]. The substitution rates for the amino acid residues forming the putative retinal-binding pocket are then calcualted using the Bayesian Markov Chain Monte Carlo method [11]. Fig 11 shows the putative proteorhodopsin retinal binding pocket sequences, along with the phylogenetic tree and the bubble-plot of amino acid substitution rates. The amino acid substitution rates indicate very fast exchange rate between the pairs of amino acid residues at position 105 (Fig 11c)), such as A/E, A/L, A/V, E/Q, L/Q, E/L, and E/V, indicating that this position of the retinal-binding pocket is the important location of the functional adaptation of the proteorhodopsin. Results from this analysis support the model that proteorhodopsins experience fast adaptation to the environmental conditions (ocean depth) of their habitat by mutating at position 105, rather than acquiring a new function (such as signal transduction). As light is at a premium at ocean depth, spectral tuning is very important, as a well-tuned pigment would be more effective at capturing light [62,66,68].

8. Generating binding site negative images for drug discovery

We can also construct the negative image of a binding pocket, and use it as a shape template for understanding substrate/ligand and protein binding. With additional chemical texture mapped on the template, negative images of binding pockets can be used for rapid screening of compounds in order to identify those that might bind to the proteins [16].

The negative image of a binding pocket can be constructed using a set of circumscribing spheres for the discrete set of Delaunay tetrahedra and triangles that defines the binding pocket [22,16]. First, the orthogonal centers of each Delaunay tetrahedron contained in the binding pocket is calculated. Circumscribed spheres are then generated with the orthogonal centers taken as their spherical centers. The radii of the circumscribed spheres are then further optimized so the resulting collection of spheres most faithfully represent the negative shape of the binding pocket [16]. Fig 12 gives an example of the negative image computed for the isoflurane binding pocket in apoferritin, which provides the only soluble protein model known to contain the structural motif thought to be important for strong anesthetic binding [69].

When combined with pharmacophore information, the negative images of protein binding pockets are found to be very effective in enriching inhibitors when examining and ranking a long list of chemical compounds for potential binding activities [16]. Results for HIV-1 protease, phosphodieterase 4B, estrogen receptor alpha, HIV-1 reverse transcriptase, and thymidine kinase show that the enriched compounds are of generally diverse chemical nature [16]. This offers an advantage for further development of drug-like compounds based on these leads.

9. Summary and Conclusion

Structural genomics projects have significantly advanced our understanding of the structural basis of the protein universe. It provides a wealth of information for tackling the challenging problem of understanding protein functions. By providing a large amount and standardized data, the success of structural genomics enables development of new and well-tailored computational methodology to interrogate a variety problems in functional understanding of the biological roles of protein molecules.

In this Chapter, we have discussed our approach of studying protein local surfaces for function inference and function characterization. The approach described in this chapter combines computational geometric characterization of protein structure, sequence and shape matching, and uncovers evolutionary signal of protein function. Our results suggests that this approach is effective in detecting enzyme functional surfaces, in inferring and characterizing protein functions, and in gaining biological insight of the relevant cellular processes. An important advantage of this integrated approach is that it gives clear location information about the region of protein surfaces where biological function occurs. Another important advantage is that by generating well-defined surface pockets and interior voids, by identifying those surfaces related to binding, and by applying the Bayesian Monte Carlo method as developed in [11], we are now able to achieve the important task of separating selection pressure due to protein function from that due to protein stability and folding. This is evidence by the improved ability in predicting protein functions when using customized scoring matrice computed using our approach vs. using precomputed scoring matrice.

It is envisioned that this approach of local surface analysis and comparison can be generalized to study the challenging problem of physical protein-protein interactions. Additional development in surface partition, shape matching, and evolutionary signal detection will likely to yield new insight.

Acknowledgments

We thank Drs. Rong Chen, Ken Dill, Rod Eckenhoff, Herbert Edelsbrunner, Jinfeng Zhang, Weifan Zheng, and Clare Woodward for fruitful collaborations. This work is supported by grants from National Science Foundation (CAREER DBI0133856, DBI0078270, DBI0646035, and DMS-0800257), National Institute of Health (GM68958, GM079804, and GM081682), O3ce of Naval Research (N000140310329), and Whitaker Foundation (TF-04-0023).

References

1. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. Science. 2006; 311(5759):347–51. [PubMed: 16424331]
2. Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318:595–608. [PubMed: 12051862]
3. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333:863–882. [PubMed: 14568541]
4. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol. 1998; 279:1211–1227. [PubMed: 9642096]
5. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol. 2003; 332:505–526. [PubMed: 12948498]
6. Glaser F, Pupko T, Paz I, Bell RE, Shental D, Martz E, Tal N. Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics. 2003; 19:163–164. [PubMed: 12499312]
7. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A. 2004; 101(41):14754–9. [PubMed: 15456910]
8. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. J Mol Biol. 2005; 351:614–26. [PubMed: 16019027]
9. Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol. 2005; 347:565–81. [PubMed: 15755451]
10. Najmanovich RJ, Torrance JW, Thornton JM. Prediction of protein function from structure: Insights from methods for the detection of local structural similarities. BioTechniques. 2005; 38(6):847–851. [PubMed: 16018542]
11. Tseng YY, Liang J. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: A Bayesian Monte Carlo approach. Mol Biol Evol. Feb; 2006 23(2):421–436. [PubMed: 16251508]
12. Gold ND, Jackson RM. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol. 2006; 355:1112–24. [PubMed: 16359705]
13. Liang J, Dill KA. Are proteins well-packed? Biophys J. 2001; 81(2):751–766. [PubMed: 11463623]
14. Tseng YY, Liang J. Predicting enzyme functional surfaces and locating key residues automatically from structures. Annals of Biomedical Engineering. 2007; 35(6):1037–42. [PubMed: 17294116]
15. Binkowski TA, Joachimia A, Liang J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. Protein Science. 2005; 14:2972–2981. [PubMed: 16322579]
16. Ebalunode JO, Ouyang Z, Liang J, Zheng W. Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. J Chem Inf Model. 2008; 48(4):889–901. [PubMed: 18396858]
17. Gavish B, Gratton E, Hardy CJ. Adiabatic compressibility of globular proteins. Proc Natl Acad Sci. 1983; 80:750–754. [PubMed: 6572366]

18. Richards FM, Lim WA. An analysis of packing in the protein folding problem. Q Rev Biophys. 1994; 26:423–498. [PubMed: 8058892]

19. Edelsbrunner H, Mücke EP. Three-dimensional alpha shapes. ACM Trans Graphics. 1994; 13:43–72.

20. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules I: Molecular area and volume through alpha-shape. Proteins. 1998; 33:1–17. [PubMed: 9741840]

21. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules II: Identification and computation of inaccessible cavities inside proteins. Proteins. 1998; 33:18–29. [PubMed: 9741841]

22. Edelsbrunner H, Facello M, Liang J. On the definition and the construction of pockets in macromolecules. Disc Appl Math. 1998; 88(83–102)

23. Stauffer, D. Introduction to percolation theory. Taylor & Francis; London: 1985.

24. Lorenz B, Orgzall I, Heuer HO. Universality and cluster structures in continuum models of percolation with two different radius distributions. J Phys A: Math Gen. 1993; 26:4711–4722.

25. Zhang J, Chen R, Tang C, Liang J. Origin of scaling behavior of protein packing density: A sequential monte carlo study of compact long chain polymers. J Chem Phys. 2003; 118:6102–6109.

26. Liu, JS. Monte Carlo strategies in scientific computing. Springer-Verlag; New York: 2001.

27. Liu JS, Chen R. Sequential monte carlo methods for dynamic systems. Journal of the American Statistical Association. 1998; 93:1032–1044.

28. Doucet, A.; De Freitas, N.; Gordon, N.; Smith, A. Sequential Monte Carlo Methods in Practice. Springer Verlag; 2001.

29. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol. 2002; 324:105–21. [PubMed: 12421562]

30. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry. 2006; 45:2545–55. [PubMed: 16489747]

31. Binkowski TA, Naghibzadeh S, Liang J. CASTp: Computed atlas of surface topography of proteins. Nucleic Acids Res. 2003; 31:3352–3355. [PubMed: 12824325]

32. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. 2006; 34(Web Server issue):W116–8. [PubMed: 16844972]

33. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. Introduction to Algorithms. 2.

34. Kuhn, Harold W. The hungarian method for the assignment problem. Naval Research Logistics Quarterly. 1955; 2:83–97.

35. Chen L, Wu LY, Wang R, Wang Y, Zhang S, Zhang XS. Comparison of protein structures by multi-objective optimization. Genome Informatics. 2005; 16(2):114–124. [PubMed: 16901095]

36. Bellman, Richard. On a routing problem. Quarterly of Applied Mathematics. 1958; 16(1):87–90.

37. Umeyama S. Least-square estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Annalysis and Machine Intelligence. 1991; 13:376–380.

38. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA. 1990; 87:2264–2268. [PubMed: 2315319]

39. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991; 11:635–650. [PubMed: 1774068]

40. Ye Y, Godzik A. Database searching by flexible protein structure alignment. Protein Science. 2004; 13:1841–1850. [PubMed: 15215527]

41. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. PNAS. 1998; 95:5913–5920. [PubMed: 9600892]

42. Jia Y, Dewey G, Shindyalov IN, Bourne PE. A new scoring function and associated statistical significance for structure alignment by ce. Journal of Computational Biology. 2004; 11(5):787–799. [PubMed: 15700402]

43. Zhu J, Weng Z. A novel protein structure alignment algorithm. Proteins: Structure, Function and Bioinformatics. 2005; 14:417–423.

44. Dayhoff, MO.; Schwartz, RM.; Orcutt, BC. Atlas of Protein Sequence and Structure. National Biomedical Research Fundation; Washington, D.C: 1978.

45. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992; 89:10915–10919. [PubMed: 1438297]

46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

47. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in Genet. 2001; 17:262–272. [PubMed: 11335036]

48. Adachi J, Hasegawa M. MOLPHY, version 2.3. Programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr Inst Stat Math Tokyo. 1996; 28:1–150.

49. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997; 13:555–556. [PubMed: 9367129]

50. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 2001; 294:2310–4. [PubMed: 11743192]

51. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Science. 1998; 7:1884–1897. [PubMed: 9761470]

52. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001; 18:691–699. [PubMed: 11319253]

53. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS. 1992; 8:275–282. [PubMed: 1633570]

54. Mayrose I, Graur D, Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. Mol Biol Evol. 2004; 21:1781–1791. [PubMed: 15201400]

55. Binkowski TA, Freeman P, Liang J. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res. 2004; 32:W555–558. [PubMed: 15215448]

56. Zhang RG, Duke N, Laskowski R, Evdokimova E, Skarina T, Edwards A, Joachimiak A, Savchenko A. Conserved protein YecM from Escherichia coli shows structural homology to metal-binding isomerases and oxygenases. Proteins. 2003; 51(2):311–4. [PubMed: 12660999]

57. Zhang RG, Andersson CE, Skarina T, Evdokimova E, Edwards AM, Joachimiak A, Savchenko A, Mowbray SL. The 2.2 A resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction. J Mol Biol. 2003; 332(5):1083–94. [PubMed: 14499611]

58. Scott JW, Hawley SA, Green KA, Anis M, Stewart G, Scullion GA, Norman DG, Hardie DG. CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations. J Clin Invest. 2004; 113(2):274–84. [PubMed: 14722619]

59. Bateman A. The structure of a domain common to archaebacteria and the homocystinuria disease protein. Trends Biochem Sci. 1997; 22(1):12–3. [PubMed: 9020585]

60. Adamian L, Ouyang Z, Tseng YY, Liang J. Evolutionary patterns of retinal-binding pockets of type i rhodopsins and their functions. Photochem Photobiol. 2006; 82(6):1426–1435. [PubMed: 16922602]

61. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF. Bacterial rhodopsin:evidence for a new type of phototrophy in the sea. Science. 2000; 289:1902–1906. [PubMed: 10988064]

62. Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. Proteorhodopsin phototrophy in the ocean. Nature. 2001; 411:786–789. [PubMed: 11459054]

63. Friedrich T, Geibel S, Kalmbach R, Chizhov I, Ataka K, Heberle J, Engelhard M, Bamberg E. Proteorhodopsin is a light-driven proton pump with variable vectoriality. J Mol Biol. 2002; 321(5): 821–838. [PubMed: 12206764]

64. Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, Isaacson T, Hirschberg J, Wagner M, Beja O. New insights into metabolic properties of marine bacteria encoding proteorhodopsins. PLoS Biol. 2005; 3(8):e273. [PubMed: 16008504]

65. Kim SY, Waschuk SA, Brown LS, Jung KH. Screening and characterization of proteorhodopsin color-tuning mutations in escherichia coli with endogenous retinal synthesis.

66. Man DL, Wang WW, Sabehi G, Aravind L, Post AF, Massana R, Spudich EN, Spudich JL, Beja O. Diversification and spectral tuning in marine proteorhodopsins. EMBO J. 2003; 22(3):1725–1731. [PubMed: 12682005]

67. Gomez-Consarnau L, Gonzalez JM, Coll-Llado M, Gourdon P, Pascher T, Neutze R, Pedros-Alio C, Pinhassi J. Light stimulates growth of proteorhodopsin-containing marine flavobacteria. Nature. 2007; 445(7124):210–213. [PubMed: 17215843]

68. Sabehi G, Massana R, Bielawski JP, Rosenberg M, Delong EF, Bj O. Novel proteorhodopsin variants from the mediterranean and red seas. Environ Microbiol. 2003; 5(10):842–849. [PubMed: 14510837]

69. Liu R, Loll PJ, Eckenhoff RG. Structural basis for high-affinity volatile anesthetic binding in a natural 4-helix bundle protein. FASEB J. 2005; 19(6):567–76. [PubMed: 15791007]

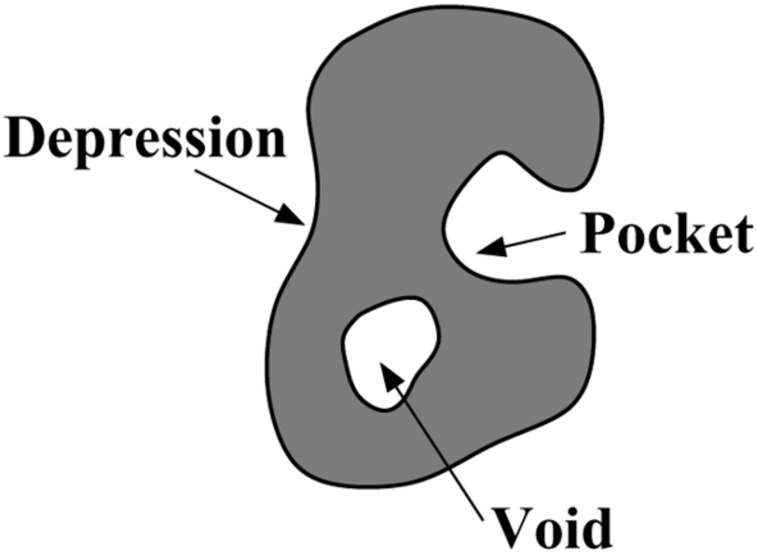


Figure 1.
Pockets and voids in proteins. There are three types of unfilled space on protein surfaces. *Voids* are fully enclosed and have no outlet, *pockets* are accessible from the outside but with constriction at mouths, and shallow *depressions* have wide openings. We use the general term *surface pockets* to include both pockets and voids (Adapted from [13]).

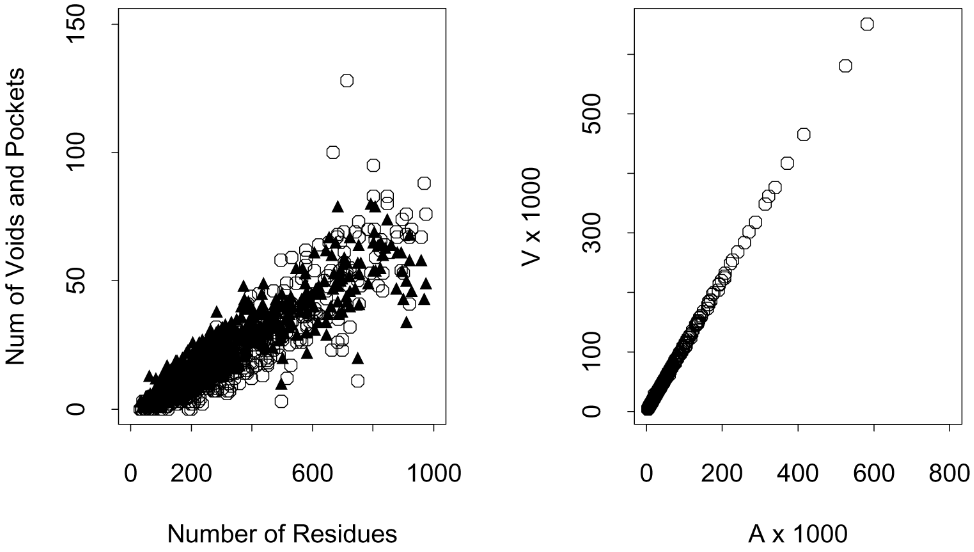


Figure 2.

Voids and pockets in protein structures. (a) Number of voids and pockets scale roughly linearly with protein length for a representative set of 636 proteins. Here circles and solid triangles represent the numbers of voids and pockets, respectively. (b) The volume of protein as calculated using van der Waals model scales linearly with the van der Waals area of protein (Adapted from [13]).

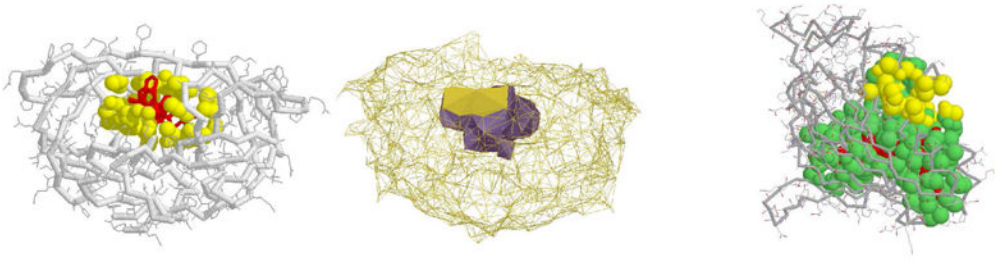


Figure 3.

The binding pockets on HIV-1 protease and phosphatidylinositol transfer protein (PITP). (Left): Binding pocket (yellow) on HIV-1 shown in van der Waals space filling model. Ligand is colored red. (Middle): The alpha shape of the HIV-1 binding site. Its mouth opening is colored gold. (Right): Binding pocket (green) on PITP for phoshpolipid (red) and a regulatory site on a different region (yellow) of the same protein.

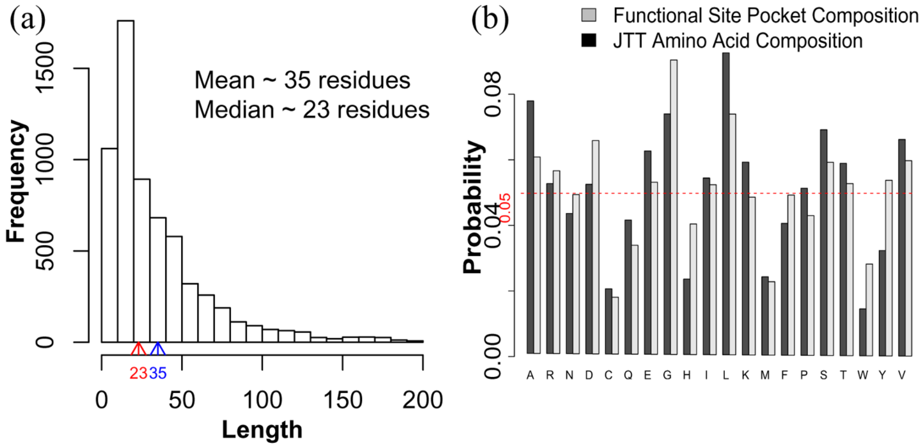


Figure 4.

The length distribution and residue composition of functional surfaces for 3,275 enzyme proteins containing known functional key residues. (a) Functional surfaces usually consist of 8–200 residues, with the mean at 35 residues. (b) The amino acid residue composition of functional surfaces is different from the composition of sequences used to construct the Jones-Taylor-Thornton (JTT) model (Adapted from [14]).

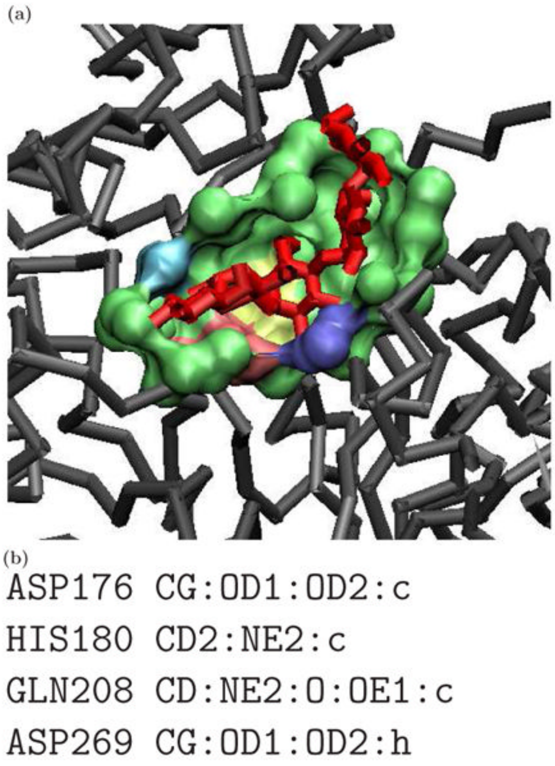
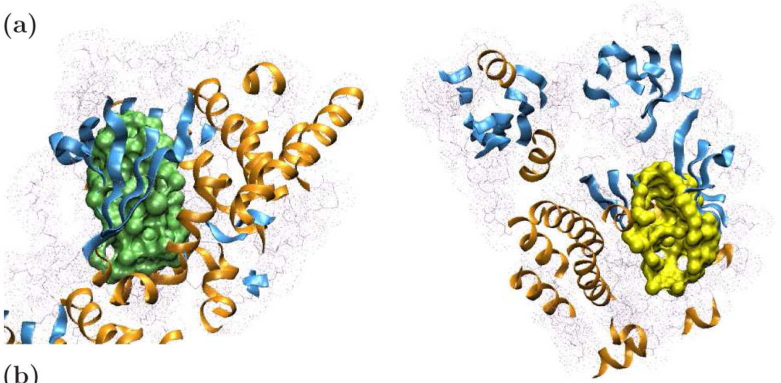


Figure 5.

The binding surface (green) and key residues predicted from a structure of alpha amylase. Here the predicted four key residues are colored yellow (D176), cyan (H180), pink (N208) and blue (D269). They contain several high propensity atomic patterns from our library of 1,031 functional atomic patterns. Their classes of secondary structural environment (sheet s, helix h, and coil c) are also listed. The substrate molecule is colored red (adapted from [14]).



(b)

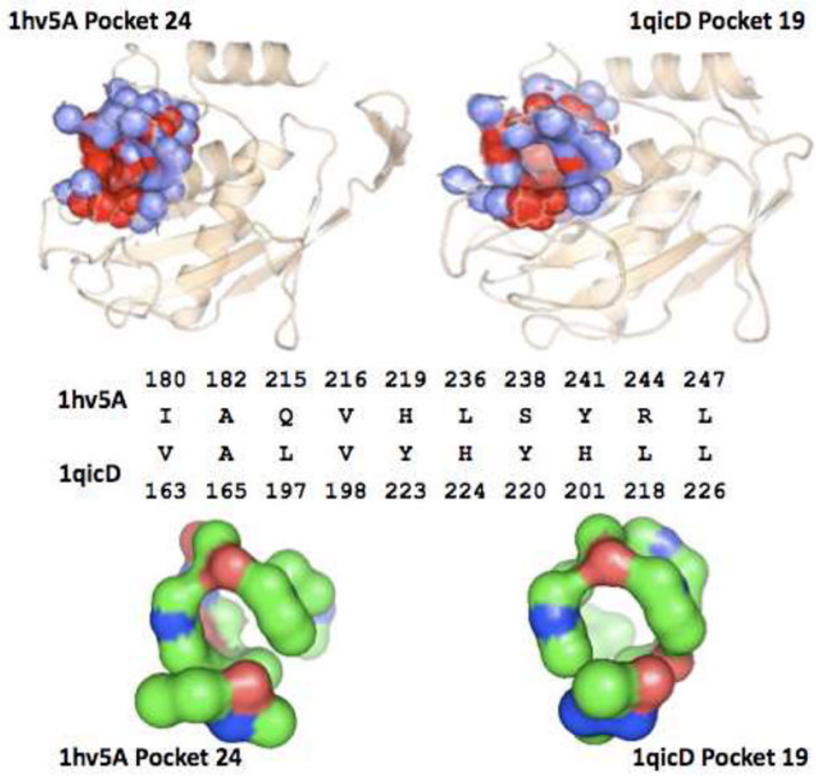
```
>1cdk_A
KGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGN
HFAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGE
MFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDF
>2src_
SLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVMKKLRHEKLVQL
YAVVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHR
DLRAANILVGENLVCKVADF
```

(c)

```
1cdk_A CASTP104
LGTGSFGRVAKLKVLQHTELVMMEYV---EDKENLTDF
2src_ CASTP51
LGQGCFGEVA-IKLMFAMVLVITEYMGSLDDRANLADF
```

Figure 6.

Functional surfaces on the catalytic domains of cAMP-dependent protein kinase (1cdk) and tyrosine protein kinase (2src). (a) In both cases, the active sites are computed as surface pockets. (b) Residues defining the pockets are well dispersed throughout the primary sequences (full sequence identity = 16%), (c) The identity of their surface sequence patterns is much higher (51%).



1hv5A	180	182	215	216	219	236	238	241	244	247
	I	A	Q	V	H	L	S	Y	R	L
1qicD	V	A	L	V	Y	H	Y	H	L	L
	163	165	197	198	223	224	220	201	218	226

Figure 7.

The binding pockets from two different stromelysin catalytic domains (pocket 29 from pdb 1hv5.A and pocket 19 from 1qic.D). They are aligned in a sequence order independent fashion with an cRMSD of 0.76 Å for 29 atoms from 10 residues. (Top) The binding pockets on the two protein structures, with pocket atoms shown in space filling form. The aligned atoms are colored in red. (Middle) The alignment of residues of these two surface pockets. Atomic details of the alignment are not shown. Sequence numbers are listed above and below the residue names for 1hv5 and 1qic, respectively. Residues in 1hv5 are arranged in order, but it is clear that the aligned residues in 1qic are not in sequence order. This residue alignment is derived from detailed alignment of atoms from surface pockets. (Bottom) Aligned atoms from these two surface pockets, with N atoms in blue, O in red, and C in green.

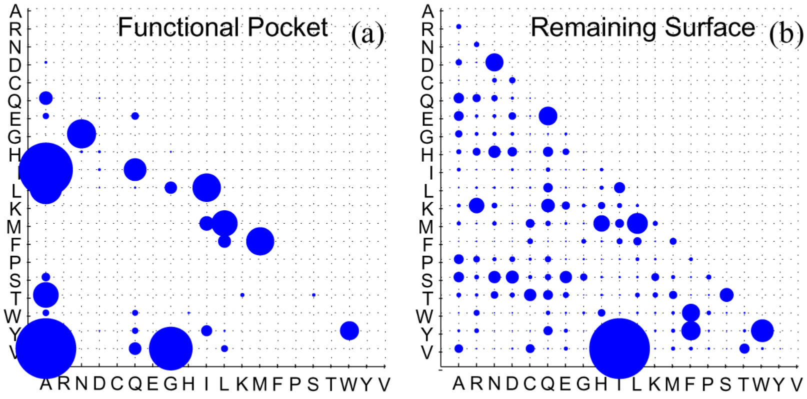


Figure 8.

Substitution rates of residues in the functional binding surface and the remaining surface of alpha-amylase (pdb `1bag`). (a) Substitution rates of residues on functional binding surface (values represented by bubble sizes). (b) Substitution rates of residues on the remaining surface on `1bag`. The values and overall pattern of substitutions that appear in both surface regions are very different (adapted from [11]).

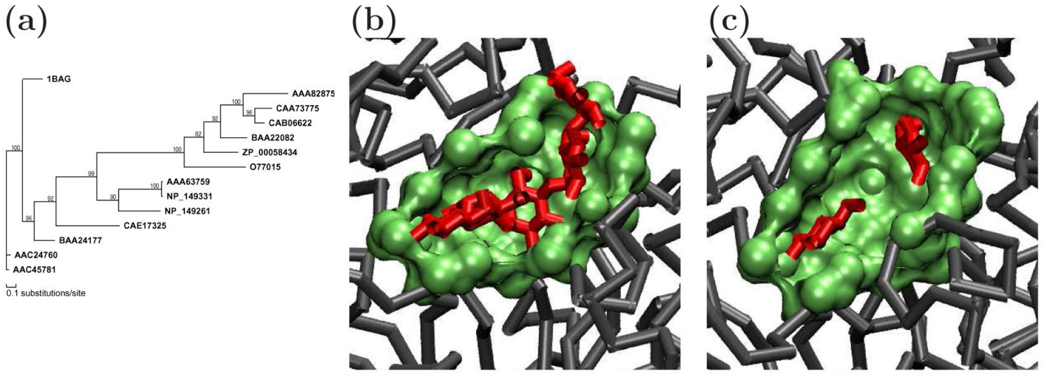


Figure 9.

Function prediction of alpha amylases. (a) The phylogenetic tree for Pdb structure `1bag` from *B. subtilis*. (b) The functional binding pocket of alpha amylase on `1bag`. (c) A matched binding surface on a different protein structure (`1b2y` from human, full sequence identity 22%) obtained by querying with the binding surface of `1bag` (adapted from [11]).

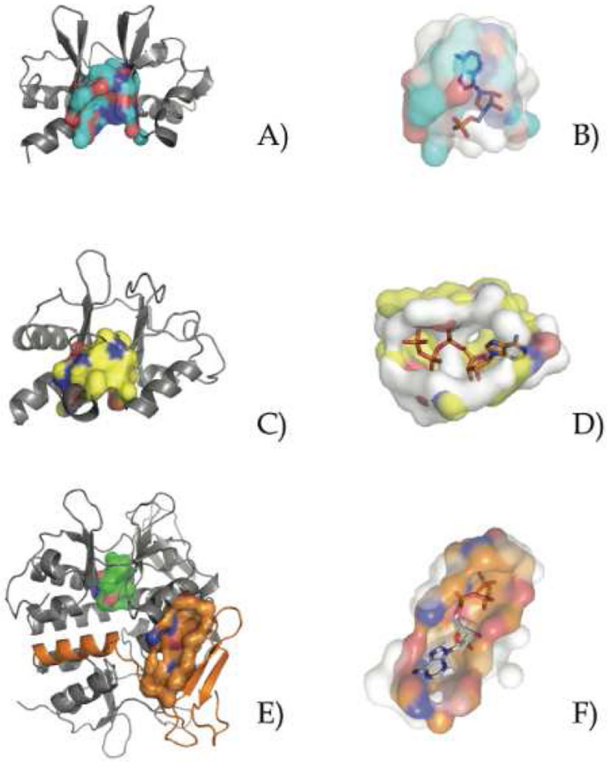


Figure 10.

Structures containing the CBS domain: (A) CBS domain protein mt1622 from *M. thermoautotrophicum* (PDB ID=1pbj), (C) inosine-5′-monosphate dehydrogenase (IMPDH) from *S. pyogenes* (PDB ID=1zfj), and (E) conserved hypothetical protein Ta549 from *T. acidophilum* (PDB ID=1pvm). The proposed nucleotide bindings surface of mt1622 (CASTp ID=9, cyan, A) is shown superposited to a flavoprotein (PDB ID=1efp, white) with bound AMP molecule (B). The IMPDH binding surface (CASTp ID=31, yellow) is show superpositioned with ATP bound cyclin-dependent kinase 2 (PDB ID=1b38, white) (D). Ta549 contains an additional C terminus CBS domain (C, orange) opposite the tandem domain interface surface (CASTp ID=27, C, green). The domain insert creates a novel surface (CASTp ID=30, orange) that shares similarity to an ATP binding surface from saicar-synthase (PDB ID=1obd, white) (F).

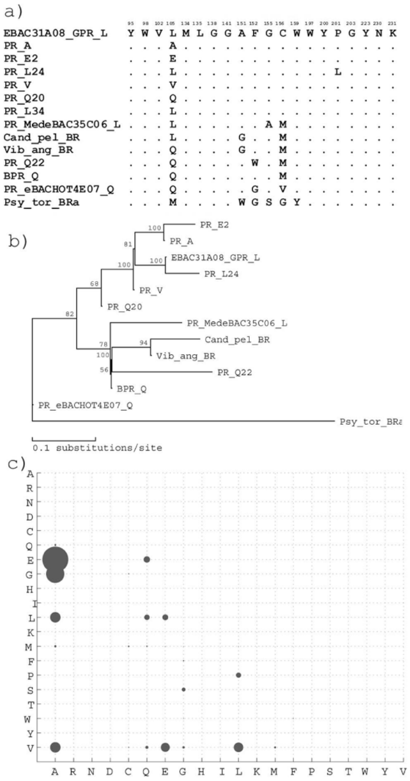


Figure 11.

Amino acid substitution rates in the putative retinal-binding pockets of proteorhodopsins. a) Alignment of putative pocket sequences. The 20 pocket residue positions are mapped from retinal-binding pocket in bacteriorhodpsin structure 1KGB. Residues that are identical with the residues in the first sequence are substituted with ".". b) Phylogenetic tree of the full-length protorhodopsin sequences. c) The plot of amino acid substitution rates for residues in the putative retinal binding pocket. The area of the circles is proportional to the substitution rate. The exchange pairs with the fastest rates are found at positions 93 and 137 in PR (following BR numbering). These are: A/L, A/V, A/E, E/Q, E/L, L/Q, L/V, and M/T (Adapted from [60].

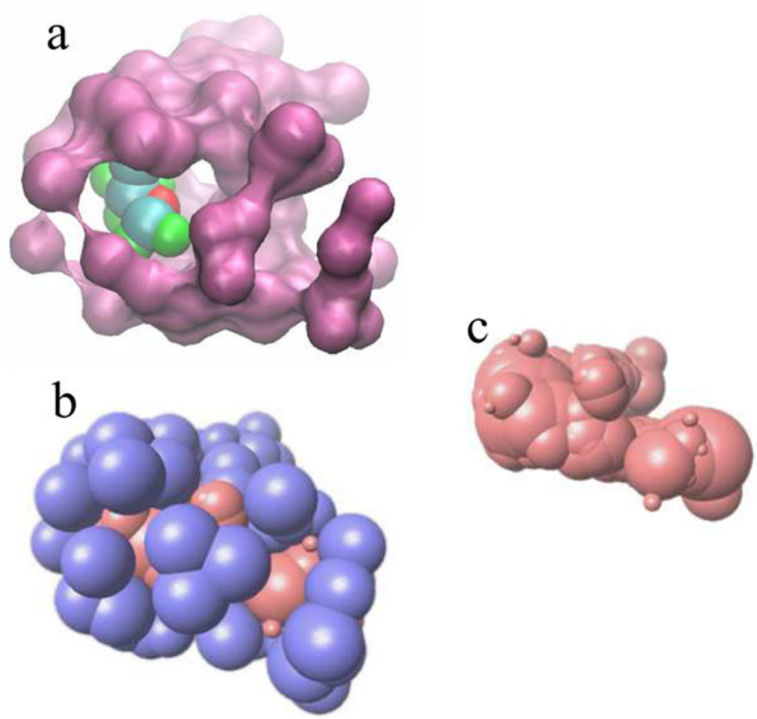


Figure 12.

The generation of a negative image of a binding pocket. (a) The surface pocket in apoferritin that binds isoflurane, (b) the atoms forming the binding pocket and its computed negative image, and (c) negative image of the binding pocket.