# Geometric structures of proteins for understanding folding, discriminating natives and predicting biochemical functions

Jie Liang[1], Sema Kachalo[1], Xiang Li[1], Zheng Ouyang[1], Yan-Yuan Tseng[1], and Jinfeng Zhang[1]

Program in Bioinformatics, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, jliang@uic.edu

Proteins are the main working molecules of a cell. Typical protein molecules contain thousands or more atoms and have complex shapes. Understanding how atoms in proteins pack and how they form intricate shape in three dimensional space is an important task, as it helps us to gain insight on how proteins fold and how they carry out biological functions. There exists a large body of work using geometric constructs such as Voronoi diagram and Delaunay triangulation to study protein packing, folding, and its physical chemical properties. We refer readers to several excellent reviews that discuss these studies [1–4]. Among these, the review by Poupon provides a comprehensive and concise discussion of recent studies [4].

In this chapter, we focus on several important issues in which computation of the geometric structures of proteins improve our understanding of protein molecules. We first briefly describe the underlying geometric models for accurate description of the complex shapes of protein molecules. We then show how improved geometry based on weighted Delaunay triangulation and alpha shape can help to characterize protein folding speed. We further describe how such accurate geometric description can aid in the development of empirical statistical scoring function for protein structure prediction. We then discuss findings on packing defects in the form of voids and pockets in protein structures, their overall distributions, and their scaling properties with protein size. This is followed by a discussion on the origin of packing defects and the roles played by evolution. Finally, we discuss how to extract evolutionary patterns of protein binding pockets and how to predict enzyme binding activities and functions.

## 1 Geometric model for proteins.

The conformations of proteins and other biological macromolecules can be modeled geometrically as a set of fused balls, in which balls represent atoms and can have different radii [1, 2, 5, 6]. Our starting point is the weighted Voronoi diagram and the weighted Delaunay triangulation of this model of union of balls [6–11] (Fig 1).

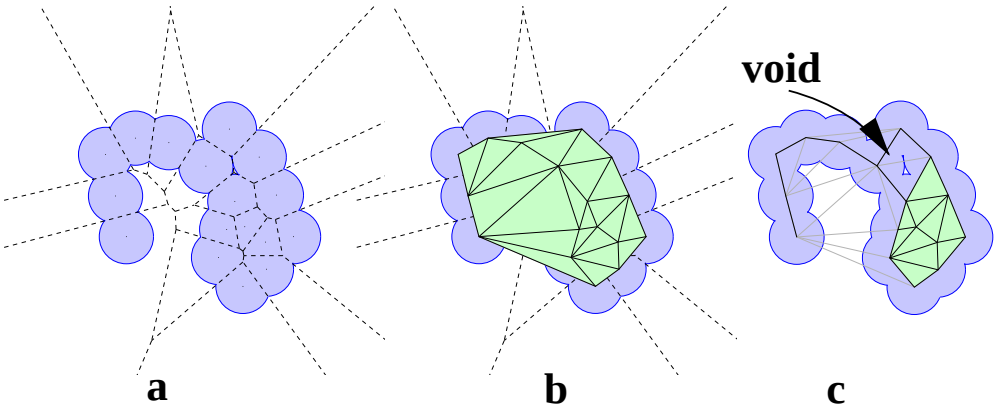



Fig. 1: Geometry of a simplified molecule in two-dimensional space for illustration. (a) The molecule formed by the union of atom disks. Voronoi diagram is in dashed lines. (b) The shape enclosed by the boundary polygon is tessellated by the *Delaunay triangulation*. (c) The alpha shape of the molecule is formed by removing those Delaunay edges and triangles whose corresponding Voronoi edges and Voronoi vertices do not intersect with the body of the molecule. A molecular void can be seen, and is represented in the alpha shape by two empty triangles (Adapted from [12]).

Briefly, the Voronoi region of an atom ball is the set of points closest to this ball by the power distance definition [6–10]. The power distance, denoted as $\pi_{\boldsymbol{x}}(\boldsymbol{y})$, of a point $\boldsymbol{y} \in \mathbb{R}^3$ from an atom ball $b(\boldsymbol{x}, r)$ centered at $\boldsymbol{x} \in \mathbb{R}^3$ with radius $r$ is defined as $\pi_{\boldsymbol{x}}(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{x}\|^2 - r^2$. The collection of Voronoi regions and their boundaries form the *weighted Voronoi diagram*, or the power diagram of the molecule [13]. For a set of balls $B$, the boundaries of their Voronoi regions decompose the space and the union of balls $\bigcup B$ into convex cells $V_B$. The well-studied weighted Delaunay triangulation is the dual structure of the Voronoi diagram. It is formed by a set of vertices representing atom centers, a set of edges connecting pairs of atoms whose Voronoi cells intersect, a set of triangles spanning three atoms whose bodies have a 3-overlap, and a set of tetrahedron whose vertices are centers of four atoms

with common intersection. These vertices, edges, triangles, and tetrahedra are called simplices and they form a simplicial complex [6, 7].

The alpha shape of a molecule is formed by a subset of the simplices in the weighted Delaunay triangulation [7]. It captures the connectivity of the convex Voronoi regions in the form of a *dual complex*, denoted as $\mathcal{K}_0$:

$$\mathcal{K}_0 = \{\sigma = \mathrm{conv}\boldsymbol{x}_B \mid \bigcap V_B \cap \bigcap B \neq \emptyset\},$$

where the intersection of the Voronoi cells of a set of balls ($\bigcap V_B$) overlap with the intersection of the balls themselves ($\bigcap B$). Here $\mathrm{conv}\boldsymbol{x}_B$ is the same as the simplex formed by the convex hull of the atom centers, denoted as $\boldsymbol{x}_B$. Details of the geometric model for protein structure can be found in [6, 7, 10, 12].

## 2 Improving understanding of protein structure and folding through geometry

Although describing protein structure using alpha shape is not as straightforward as commonly used heuristics, such as declaring neighboring relationship by a distance cut-off, or cubic grids for volume calculation, this approach offers important advantages, which often lead to fresh insights with improved understanding of protein molecules. Here we show how such geometric descriptions of protein contacts can lead to a better prediction of protein folding rates.

### 2.1 Native structure and folding rate

Proteins have complex three-dimensional native structures. A remarkable observation is that completely unfolded denatured proteins often could refold spontaneously to their native conformations. However, the folding speed of different proteins as measured by folding time ($\tau = 1/k_f$) can vary by 8-orders of magnitude, from about $10^{-6}$ to $10^2$ second (see [14] for a general discussion of protein folding rates). An important question therefore is: What determines the speed of protein folding?

One view postulates that the native structure of protein determines the folding speed. Using a set of about 20 proteins whose folding rates have been measured experimentally and whose three dimensional structures are known, Plaxco, Simons, and Baker discovered that folding rate correlates well with a parameter called contact order, which measures the localness of contact interactions in the folded structures of proteins [15]. However, as more experimental data become available, this correlation deteriorates [19].

One possible cause of the deterioration may lie in the inaccuracy of the contact description. In most computational studies, pairwise contacts between amino acid residues are declared if two residues are within a specific threshold of Euclidean distance. But this definition can potentially include many

implausible non-contacting neighbors, which have no significant physical interaction [16]. Whether or not a pair of residues can make physical contact depends not only on the distance between their center positions (such as $C_\alpha$ or $C_\beta$, or geometric centers of side chain), but also on the size and the orientations of side-chains [16]. Furthermore, two atoms close to each other may in fact be occluded from contacting each other. By using a distance threshold, fictitious contact interactions may be included. This is a well-recognized problem [11, 17, 18].

**Geometric Contact.** An alternative approach is to define geometric contacts between residues using criteria derived from the Voronoi diagram and the alpha shape of the protein structure [19]. In this case, one can identify interacting residue pairs following the edges in the alpha shape of the protein structure. A useful parameter is the *geometric contact number* $N_\alpha$, which is just the number count of residues connected by alpha edges. [1]

**Folding rate correlation.** Using $N_\alpha$ instead of contact order leads to improvement in correlation with folding rate. The correlation of $N_\alpha$ with $\ln k_f$ of experimentally determined folding rate $k_f$ for a set of 80 proteins with diverse structures is summarized in Figure 2. The folding rates of these proteins span a range over more than 8 orders of magnitude. For comparison, correlation with $\ln k_f$ for several other parameters reported in literature are also summarized in Figure 2. Among these, the relative contact order $RCO$ [15] is defined as $RCO = \sum \Delta S_{i,j}/(LN)$, where $N$ is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation between contacting residues $i$ and $j$, and $L$ is the total number of residues. It correlates poorly with folding rates of this set of 80 proteins. The absolute contact order $ACO$ is defined as $ACO = \sum \Delta S_{i,j}/N$ [20]. It has better correlations. In addition, protein chain length has a strong negative correlation for multi-state proteins ($R = -0.79$), but a weaker correlation for two-state proteins ($= -0.72$) [19]. In contrast, the quantity $N_\alpha$ computed from alpha shape correlates well with folding rates ($R = -0.83$ for all 80 proteins). These results indicate that accurate description of geometric contacts improves correlation of native protein structures with folding rates.

$N_\alpha$ is also a better predictor of folding rate than the simple measure of protein chain length. This is demonstrated by results from a random test, in which a subset of 30 proteins was selected out for the correlation analysis. The correlation coefficients between the folding rate $\ln k_f$ and the geometric contact number $N_\alpha$, between $\ln k_f$ and the chain length $L$ are recorded, respectively [19]. This is then repeated several times. Chain length $L$ is found not to be a consistently good predictor of protein folding rates: The correlation

[1] Two additional conditions are imposed, namely, contacts must be at least 4 residues apart in the primary sequence, and their spatial distance is no greater than 6.5 Å. Here the distance cutoff is used as an upper bound instead of a lower bound as cutoffs are used in other methods. As alpha shape of a protein is computed with all atomic radii incremented by the radius of a solvent (1.4 Å), this cutoff is necessary to exclude atoms not in physical contact but are within the diameter of a solvent.

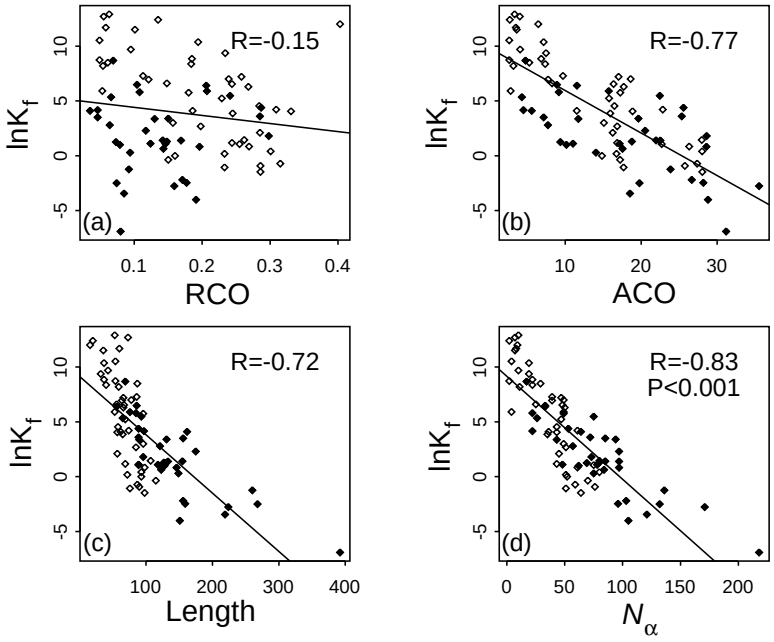



Fig. 2: Relationship between different structural parameters and folding rates of two-state (open squares) and multi-state (solid squares) proteins. Two state proteins are those whose folding behavior can be described by one exponent component, whereas multi-state proteins cannot be described by a single component. (a) Contact order RCO ($R$ = -0.15), (b) absolute contact order ACO ($R$ = -0.77), (c) c chain length ($R$ = -0.72), and (d) $N_\alpha$ ($R$ = -0.83) (Adapted from [19]).

$R$ is better than $-0.50$ only for two subsets, and can be as little as $-0.04$. In contrast, $N_\alpha$ gives consistently good correlations: all are better than $-0.58$, with the best value being $-0.79$. Overall, these results show that accurate description of geometric contacts can capture the nature of protein folding better than contacts defined by distance cut-off and chain length.

### 2.2 Understanding folding rates from studies of model proteins

Although parameter derived from native protein structure can correlate well with measured folding rate, the extent to which native structure determines folding rate remains unclear. An experimental study showed that a designed artificial protein with no homologous sequence in nature that adopts the same structure as a natural protein can fold 4,000 times faster [21]. Such observations contradict with the view that different sequences for the same protein structural fold would all have very similar folding rates.

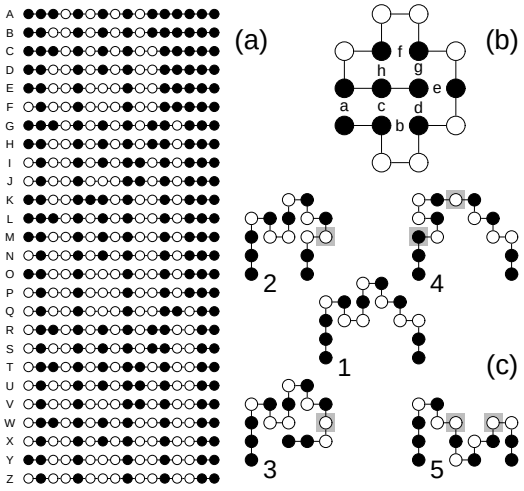



Fig. 3: Protein-like sequences and the set of primitive moves. The largest protein family contains (a) 26 sequences that all fold into (b) the same structure. Here filled circles are H residues. (c) The move set used to study folding dynamics includes: among (1, 2, and 3), single point moves rotate around a single point; between (1 and 4), generalized corner moves reflect around a diagonal axis connecting any two residues; between (1 and 5), generalized crankshaft moves reflect around a horizontal or vertical axis. Points of rotation are on gray background. For a given conformation, all possible point moves at different positions, all possible generalized corner moves and crankshaft moves at all pairs of positions are exhaustively searched (Adapted from [22]).

Mechanistic understanding of folding dynamics can be gained by analyzing folding of model protein molecules [22]. Two-dimensional hydrophobic and polar (HP) lattice model [23] has been widely used for studying protein folding, for example, on collapse and folding transitions [24–28], influence of packing on secondary structure and void formation [29–32], and the roles of mutation and recombination in the evolution of protein thermodynamics [33, 34]. In this model, contacts between spatially neighboring sites that are not sequence neighbors can be regarded as equivalent to the geometric contacts computed from alpha shape in real protein structures. This simple toy model exhibits complex protein-like behavior. For example, by evaluating a simple energy scheme for all $2^{16}$ HP sequences of 16-mers on all enumerated $802,075$ conformations, it is found that there are 1,539 protein-like foldable sequences with unique ground state conformations, and 456 conformations that are the unique

ground state for 1 or more foldable sequences [22]. There are 26 sequences that all fold into the same ground state conformation (Fig. 3a and 3b). This set of sequences forms the largest protein family, where each sequence adopts the same conformation, and all are connected by (a series of) point mutation [22].

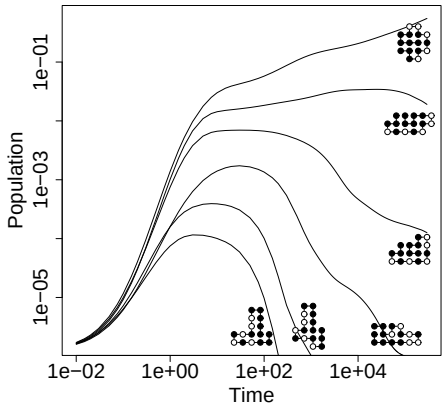



Fig. 4: The time evolution of the native state and several local minima states. The probability of occupation of native state conformation (top) increases monotonically through a time span of $10^{-2} - 10^{5}$, but local minima conformations go through transiently accumulating intermediate states [22]).

By employing a set of physically realizable move set (Fig. 3c) and solving a master equation describing the probability of transition between any two different conformations through realizable moves, the folding dynamics of all $802,075$ conformations of the model proteins spanning 7 decades of time scale can be followed (Fig. 4). It was found that folding rates vary enormously for sequences of the same length, same energy, same energy gap, and even for the 26 sequences with the same ground state conformation [22]. Furthermore, thermodynamic parameters such as collapse cooperativity are found to be weak predictors of folding rates. Instead, properties of the kinematic landscape (defined by the conformations and the physical moves connecting them), such as the number of local minima on this connected landscape, provide excellent correlation with folding rates [22].

### 2.3 Remark

The physical basis for the observed correlation of folding rate and $N_\alpha$ is likely that proteins fold via a mechanism of zipping and assembly, where contacts among monomers that are more widely separated in the sequence are slower

to form because their conformational search is more costly in chain entropy [35–40]. It is expected that further studies on correlating specific types of contact interaction of residues pair may shed further light on the nature of the determinants of protein folding rates.

Another area where geometric analysis could contribute is the evaluation of the compactness of three dimensional conformations. An important aspect of protein structure prediction is to recognize native-like conformation from a large number of decoy structures that are far from the native structure of protein. Although various force fields can model physical interactions in details and often can be used to sample a large number of conformations, they often have limited discrimination in selecting protein-like compact structures from loosely packed but low energy structures. It is likely that simple geometric contact analysis similar to what was applied in folding rate study can help to resolve this problem.

## 3 Scoring function for predicting protein structure

The observation of spontaneous refolding of a denatured protein indicates that the sequence of amino acids of the protein contains all of the information needed to specify its three-dimensional native structure [41–43]. A fundamental problem in molecular biology therefore is to the predict three-dimensional structures of proteins from sequences. As sequences of most proteins are now available, while the number of proteins with known structures lags far behind, protein structure prediction can be valuable for gaining understanding on how protein molecules work.

In protein structure prediction, often a large number of candidate conformations are generated, and a scoring function or energy potential is used to select the correct conformations, called the native or near-native conformations, from an ensemble of alternative conformations (called decoys) [44]. The discrimination of native and near native conformations is a stringent requirement for a scoring function. An effective approach for developing scoring functions is to empirically generate parameters based on statistical analysis of geometric features of protein structures [44–50]. These scoring functions are also called *knowledge-based potentials* or *empirical potentials*.

Geometric analysis of protein structures can be used to design very effective knowledge-based scoring functions, as they reflect the shape and contact interactions of proteins with high accuracy. Potential functions based on geometric constructs include those that employ the Voronoi diagram [49, 51], the Delaunay triangulation [52–57], and the alpha shape [11, 50, 58, 59] of the protein molecules. These geometry based scoring functions have achieved significant successes. For example, a scoring function based on the Voronoi diagram of proteins structures is among one of the best performing atom-level scoring functions [49].

Because the alpha shape of a protein structure contains rich topological, combinatorial, and metric information [6], we discuss a scoring functions based on alpha edges in more detail below as an example of this class of geometric scoring function. More details can be found in [12]

### 3.1 Principle of knowledge-based scoring function for protein structures.

The main assumption in developing empirical scoring functions is that the frequencies of structural features observed in a protein structural database follows the Boltzmann distribution under a potential function. The probability of each structural feature in native conformations is assumed to be independent. By estimating the probability of the occurrence of these structural features and assuming Boltzmann distribution, one can reconstruct empirically a potential function as the scoring function. A widely used class of structural feature is the number counts of various types of amino acid residue pairs in contact interactions.

*Statistical probability and empirical energy potential function.*

Denote the collection of number counts of various structural features observed in a conformation as a vector $\boldsymbol{c}$, and the sequence of amino acid residues of the protein as $\boldsymbol{a}$. The Boltzmann distribution connects a scoring function $H(\boldsymbol{c})$ interpreted as an energy potential function for a conformation represented by $\boldsymbol{c}$ to its probability of occurrence $\pi(\boldsymbol{c})$:

$$\pi(\boldsymbol{c}) = \exp[-H(\boldsymbol{c})/kT]/Z(\boldsymbol{a}), \tag{1}$$

where $k$ and $T$ are the Boltzmann constant and the absolute temperature measured in Kelvin, respectively. The partition function is $Z(\boldsymbol{a}) \equiv \sum_{\boldsymbol{c}} \exp[-H(\boldsymbol{c})/kT]$, which sums over all possible feature count vectors $\boldsymbol{c}$ obtained for all possible conformations for the sequence $\boldsymbol{a}$.

*Reference state and collection of non-interacting pairs.*

In order to derive an energy potential function that encodes interactions specific to proteins, one has to remove the occurrence of structural features due to random chances. For this purpose, a reference state is constructed that models the background random probability $\pi'(\boldsymbol{c})$ of the occurrence of structural features. In this reference state, the frequencies of structural features (such as pairwise residue interactions) are of random nature and are independent of the sequence and structure of the protein [60].

Denote the energy potential function that would result from these random occurrence as $H'(\boldsymbol{c})$. A potential energy $\Delta H(\boldsymbol{c})$ is then obtained as:

$$\begin{aligned} \Delta H(\boldsymbol{c}) &= H(\boldsymbol{c}) - H'(\boldsymbol{c}) \\ &= -kT \ln[\frac{\pi(\boldsymbol{c})}{\pi'(\boldsymbol{c})}] - kT \ln[\frac{Z(\boldsymbol{a})}{Z'(\boldsymbol{a})}], \end{aligned} \tag{2}$$

Here $-kT\ln(Z(\boldsymbol{a})/Z'(\boldsymbol{a}))$ is a constant that does not depend on the conformation and the vector $\boldsymbol{c}$. If one assumes that $Z(\boldsymbol{a}) \approx Z'(\boldsymbol{a})$ [60], the effective potential energy can be calculated as:

$$\Delta H(\boldsymbol{c}) = -kT\ln[\frac{\pi(\boldsymbol{c})}{\pi'(\boldsymbol{c})}] \tag{3}$$

Since we the probability distribution of each structural feature is assumed to be independent, we have $\pi(\boldsymbol{c})/\pi'(\boldsymbol{c}) = \prod_i c_i[\frac{\pi_i}{\pi_i'}]$, where $c_i$, $\pi_i$, and $\pi_i'$ are the number count of a single $i$-th type of structural feature, the probability of $i$-type feature in the database of native proteins and in the reference state, respectively. We then have

$$\Delta H(\boldsymbol{c}) = -kT\ln[\frac{\pi(\boldsymbol{c})}{\pi'(\boldsymbol{c})}] = -kT\sum_i c_i\ln[\frac{\pi_i}{\pi_i'}]. \tag{4}$$

We can now decompose $\Delta H(\boldsymbol{c})$ into basic energetic terms associated with each structural feature that can be linearly summed up:

$$\Delta H(\boldsymbol{c}) = \sum_i \Delta H(c_i) = -kT\sum_i c_i w_i. \tag{5}$$

where we have:

$$w_i = \ln[\frac{\pi_i}{\pi_i'}]., \tag{6}$$

as the distribution of each of the $i$-th feature is assumed to be independent. Eqn(4) is just a manifestation of the statement that the probability of each structural feature follows the Boltzmann distribution. Analysis of the distributions of many protein structural features, including residues between the surface and interior of globules, the occurrence of various $\phi, \psi, \chi$ angles, *cis* and *trans* prolines, ion pairs, and empty cavities in protein globules, all are found to follow the Boltzmann distribution [61].

### 3.2 Alpha contact scoring function

The structural feature in scoring functions are often number counts of various types of interacting amino acid residue pairs. Similar to the study of protein folding rate, one can improve the description of protein contact interactions by carrying out geometric analysis. Similar to the study of protein folding rate, one can define that contact occurs if atoms from non-bonded residues share a Voronoi edge, and this edge is at least partially contained in the body of the molecule through the computation of the alpha shape [10, 62],

This condition models the requirement that atoms must be in physical nearest neighbor contact.

*Probabilistic model for pairwise contact interactions.*

Specifically, the probability $\pi_{(i,j)}$ for residue of type $i$ interacting with residue of type $j$ can be derived from observed pairwise alpha contacts between atoms involving both residue types. The number count of such observed contacts from different proteins in the entire database of many selected diverse protein structures are pooled together, and $\pi_{(i,j)}$ is calculated as:

$$\pi_{(i,j)} = \frac{c_{(i,j)}}{\sum_{i',j'} c_{(i',j')}}$$

Here $c_{(i,j)}$ is the count of atomic contacts between residue type $i$ and residue type $j$, and $\sum_{i',j'} c_{(i',j')}$ is the total number of all atomic contacts.

*Model for reference state.*

We can compute the random probability $\pi'_{(i,j)}$ where residue $i$ and $j$ interact randomly by adopting the model that a pair of contacting atoms is picked from a residue of type $i$ and a residue of type $j$. Here these atoms are chosen randomly and independently [63]. We have:

$$\pi'_{(i,j)} = N_i N_j \cdot (\frac{n_i n_j}{n(n-n_i)} + \frac{n_i n_j}{n(n-n_j)}), \quad \text{when} \quad i \neq j$$

and

$$\pi'_{(i,j)} = N_i N_{i-1} \cdot \frac{n_i n_i}{n(n-n_i)}, \quad \text{when} \quad i = j$$

where $N_i$ is the number of interacting residues of type $i$, $n_i$ is the number of atoms residue of type $i$ has, and $n$ is the total number of interacting atoms.

*Energy evaluation of a protein.*

The weight coefficient $w_i$ of the scoring function is calculated as:

$$w_{(i,j)} = -\ln[\frac{\pi_{(i,j)}}{\pi'_{(i,j)}}] \tag{7}$$

The overall energy of a protein structure is calculated as:

$$E = \sum_{(i,j)} c_{(i,j)} \cdot w_{(i,j)},$$

where the summation is over all contacts between different residue pairs.

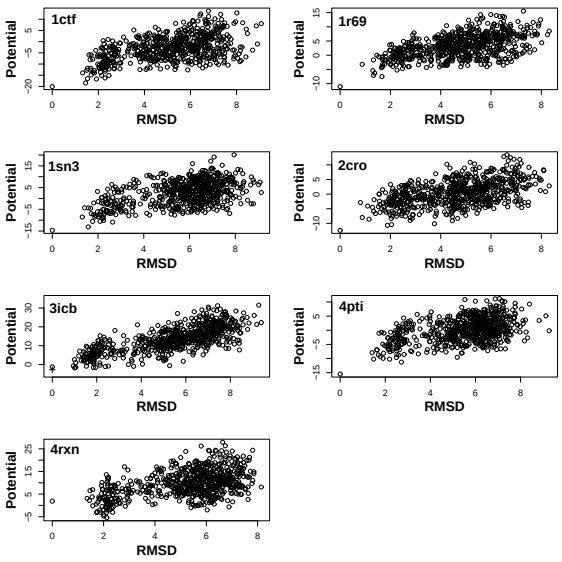



Fig. 5: Energy evaluated by alpha contact potential plotted against the RMSD to native structures for conformations in the Park & Levitt Decoy Set. For vitamin D-dependent calcium-binding protein (`3icb`), a structure with better resolution (`4icb`) has the lowest energy (denoted by "+"). (Adapted from [65]).

### 3.3 Evaluation of scoring functions for protein structure prediction

To recognize typically only a few protein conformations similar to the native conformation from an ensemble of a very large number of decoy conformations ($10^6$) a scoring function needs to be very discriminative. How such discriminations work can be demonstrated through the example of the Park & Levitt decoy set. This decoy test set contains native and near-native conformations of seven sequences, along with about 650 misfolded decoy structures for each sequence. The positions of $C_\alpha$ of these decoys were generated to mimic realistically the geometry of real proteins. Conformations in the decoy sets all have low energy by a number of potential functions, and have low RMSD to the native structure [64].

The results of energy calculation using the alpha contact potential are shown in Fig 5. For five of the seven proteins, the native structures have lowest energy by alpha contact potential. For protein `3icb` and `4rxn`, the native structures have the 5th and 51st lowest energy values, respectively. For all proteins, decoys with the lowest energy are within 2.5 Å RMSD to the native structure. These results show that alpha scoring function works well for this class of decoys. Studies of alpha contact potential with other decoys can be found in [65]

### 3.4 Remark

In this section, we showed how accurate geometry can help in developing effective scoring function for protein structure prediction. Using the edge simplices in the alpha shapes of protein structures, a scoring function based on the statistics of edge simplices is shown to work well [65]. In further analysis, it was found that the geometric representation of accurate contact interaction is very important, although the specific details of the residue types are often less so [65].

## 4 Nature and origin of voids and pockets in protein structures

### 4.1 Voids and pockets in protein structures

Protein cores are often considered to be solid-like [66, 67], as proteins have high packing densities [1] and low compressibilities [68]. Analysis of Voronoi diagrams of protein structures showed that the average packing density in a protein is as high as that inside crystalline solids [69–71]. Sometimes protein is compared to an assembled jigsaw puzzle [2].

However, there exists unfilled spaces both inside and on the surface of proteins. Gerstein *et al* calculated the volume associated with atoms on the protein surface using Voronoi diagram and water molecules generated by molecular-dynamics simulation [72]. It was found that nonpolar atoms on the protein surface and their associated water are less tightly packed, and charged atoms and water are more tightly packed. The large volume fluctuation associated with atoms at the protein-water interface was further studied for understanding the compressibilities of protein and solvent atoms [72]. By placing water molecules randomly on a grid surrounding the molecule, Charavarty *et al* carried out calculations using Voronoi diagram to identify empty spaces located on protein surfaces [73]. By correlating volume of empty spaces created after replacement of amino acid residues with measured stability change of the protein, these authors were able to estimate the strength of hydrophobic forces that drive protein folding [73].

**Distribution of voids and pocket.** Geometrically, the unfilled spaces can be formally classified as voids, pockets, and depressions. *Voids* are unfilled spaces inside the protein that are fully enclosed by atoms. *Pockets* are caverns that open to the outside of the protein through *mouths* that are small relative to cavern dimensions but big enough that the probe ball has access to the outside of the molecule. The mouth of a pocket is narrower than at least one cross section of the interior of the pocket. *Depressions* are concave regions on protein surfaces that have no constriction at the mouth [74–77].

The prevalence of voids and pockets in proteins can be assessed using the pocket algorithm described in [79]. For a set of 636 proteins representative

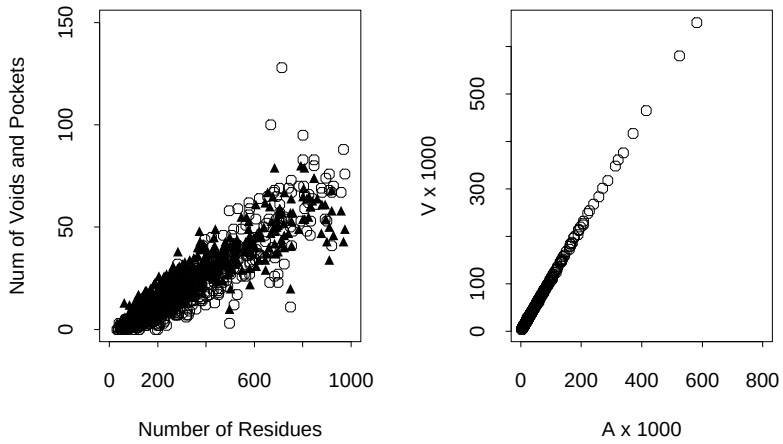



Fig. 6: The scaling behavior of the geometric properties of proteins. (a) Voids and pockets for a set of 636 proteins representing most of the known protein folds. The number of voids and pockets detected with a 1.4 Å probe is linearly correlated with the number of residues in a protein. Solid triangles and empty circles represent the pockets and the voids, respectively. (b) The van der Waals ($vdw$) volume and van der Waals area of proteins scale linearly with each other. Here the van der Waals volume is the volume of the union of overlapping atom balls adopting van der Waals radii (Adapted from [78]).

of all known protein structures of different backbone fold families [80], it was found that the numbers of pockets and voids are approximately linearly correlated with the number of residues in each protein, namely, the size of the protein (Fig 6a) [81]. Roughly speaking, for every additional 100 residues, a protein has about an additional 7–8 voids and 7–8 pockets. These spaces are found by a 1.4 Å spherical probe, therefore are large enough to contain at least one water molecule. This finding shows that voids and pockets are quite common in protein structures.

**Scaling behavior.** A useful way to characterize protein interior packing and to understand whether proteins are packed more like jigsaw or random objects as in disordered material is through surface/volume relationships. For a perfectly solid three-dimensional sphere of radius $r$, the relationship between volume $V = 4\pi r^3/3$ and surface area $A = 4\pi r^2$ is: $V \propto A^{3/2}$. The volume of proteins, however, is found to scale linearly with the surface areas of proteins (Fig 6b).

A model for disordered materials is clusters of random uncorrelated spheres, which has a characteristic scaling behavior [82]. Monte Carlo stud-

ies show that the volume $V$ of clusters of random spheres, of either uniform radius or of mixtures of different radii, scales linearly with the surface area $A$ of the cluster: $V \propto A$ [82, 83]. The same scaling behavior is found in lattice models of simple clusters [83]. This linear relation of $V$ with $A$ is also what is observed in proteins (Fig 6b). Similarly, there is a linear correlation between the volume and the number of atoms $N$ (or the number of residues) of proteins, the same as observed in random packed spheres [82].

A key property of randomly packed spheres is the so-called *percolation threshold*. In randomly packed spheres, when the packing density $p$ is greater than a threshold density $p_c$, clusters become connected to each other, and the size of the largest cluster approaches the size of the whole system [82, 84, 85]. At this threshold $p_c$, the volume $V$ of a cluster of random spheres is known to scale with the length $R$ of the cluster as $V \propto R^D$, with a characteristic exponent $D = 2.5$ in three-dimensional space [82, 83]. The same exponent is also found in lattice models of clusters [86]. The size $R$ of a cluster of spheres can be calculated as the maximum extent of the cluster along the coordinate axes: $R = \frac{1}{2d}\sum_{j=1}^{d}(x_{j,max} - x_{j,min})$, where $d = 3$ in $\mathbb{R}^3$ [82]. For random packed spheres, $D = 2.5$ at $p = p_c$, but no scaling behavior is known for $p > p_c$. Based on three-dimensional lattice studies, it is expected that $D$ will cross over from $D = 2.5$ if $p \approx p_c$ to $D = 3$ if $p \gg p_c$ [83].

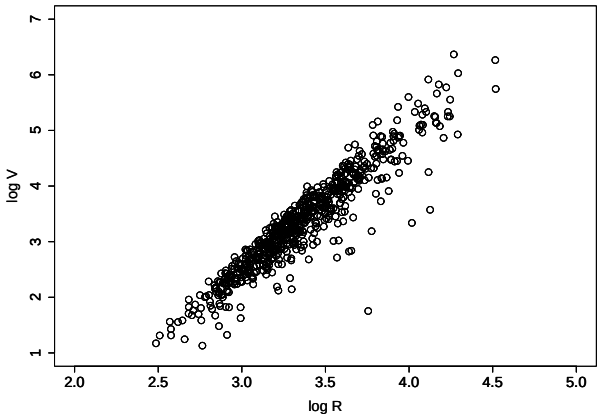



Fig. 7: The logarithm of protein molecular surface volume $\log V$ scales with the logarithm of the length of the protein $\log R$ with the characteristic slope $d$ of 2.47 (Adapted from [81]).

In proteins, it is found that $\ln V \propto D \ln R$, with a fractal dimension $D = 2.47 \pm 0.04$ (Figure 7) [81]. This suggests that packing in proteins behaves

like random spheres near their percolation threshold. The surface areas $A$ of proteins also scale with $R$ with a fractal exponent $D = 2.26$. Therefore, both $V$ and $A$ scale with $R$ with a similar fractal dimensionality. This is consistent with the direct linear correlation we observed between $V$ and $A$. A recent study indicates that buried residues of a protein often contribute significantly to its overall side chain entropy [87], consistent with the postulation of packing in protein interior behaves like random spheres.

### 4.2 Origin of voids and pockets in protein structure

Another useful parameter describing packing is the packing density $p_d$ [2, 88]. $p_d$ can be thought of as the physical volume $v_{vdw}$ occupied by the union of van der wall atoms, divided by the volume of an envelope $v_{env}$ that tightly wraps around the body of atoms: $p_d \equiv v_{vdw}/v_{env}$ [81]. Voids contained within the molecule will not be part of the van der Waals volume $v_{vdw}$, but will be included in $v_{env}$. Using geometric algorithms, $v_{vdw}, v_{env}$ and $p_d$ can be readily computed for protein structures [10, 76]. The scaling relationship of $p_d$ and chain length $N$ obtained from such calculation is shown in Fig 8a [81], along with the coordination number $Z_\alpha$ calculated from alpha shapes.

To answer the question that whether the scaling behavior of $p_d$ with chain length is unique to proteins, one can study voids and packing in generic model chain polymers that are not proteins [78]. For this purpose, one needs to generate self-avoiding walks (SAW) to model chain polymers, namely, monomer beads connected by a self-avoiding chain. These chain polymers have little resemblance to proteins, other than the imposed condition that they have similar compactness as that of proteins. Specifically, one needs to generate samples of long chain SAWs from the target distribution of uniformly distributed all SAWs satisfying protein-like compactness requirement.

One technical challenge is that it is very difficult to generate long chain SAWs. This is the well-known attrition problem: the success rate for generating SAW of length 48, regardless of its compactness, is only 0.79%, and this rate deteriorates rapidly when chain length increases [89]. This difficulty can be overcome by using the chain-growth based sequential Monte Carlo method [90], which keeps proper weights for samples generated by growth. Using a discrete 32-state off-lattice model and sequential Monte Carlo, one can successfully generate thousands of SAWs in three dimensional space at any specified intervals of compactness [78]. Void and packing density and coordination number computation for these randomly generated SAWs show that the scaling behavior of $p_d$ and chain length is very similar to that observed in protein (Fig 8c and d). These results suggest that protein retain the same packing property of generic compact chain polymers, and they are unlikely to be optimized by evolution to eliminate voids [78].

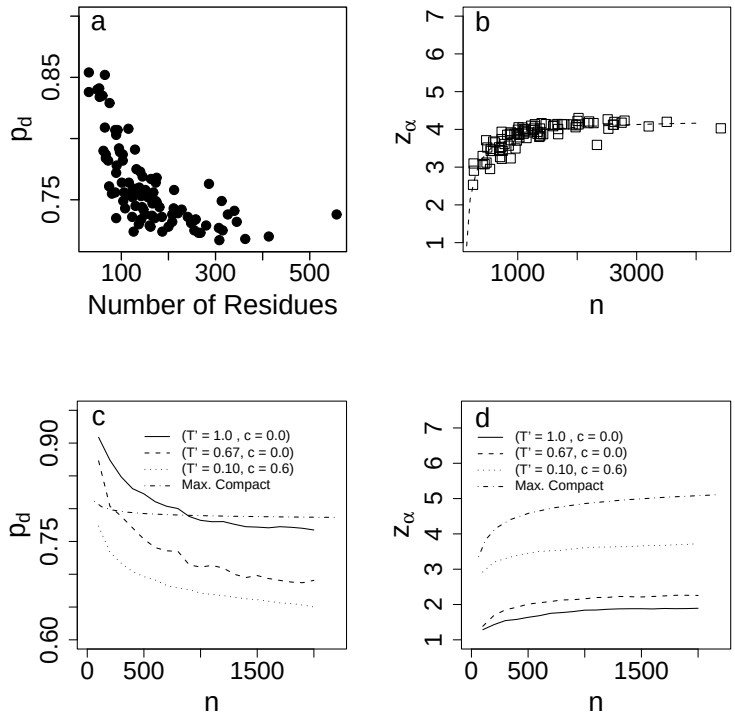
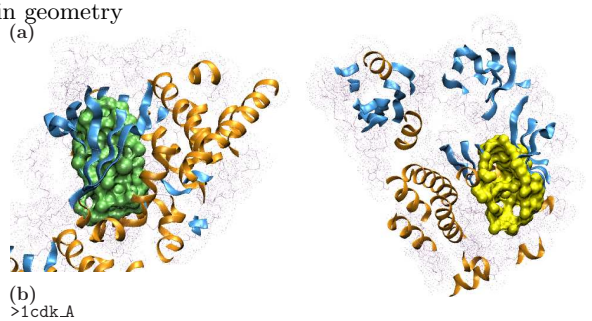



Fig. 8: Comparison of scaling behavior of packing density and coordination number of proteins and compact chain polymers. (a) Packing density $p_d$ of proteins of different lengths; (b) Scaling behavior of coordination number $Z_\alpha$ calculated based on alpha contact and protein chain length. (c) Packing density $p_d$ of randomly generated homopolymer of different lengths. Different curves reflect models generated using different parameters $(T, C)$ that adjust the importance of compactness, number of neighbors, and distance to neighbor (see ref [78] for details.) (d) Scaling behavior of coordination number $Z_\alpha$ of random chain polymers (Adapted from [78]).

### 4.3 Remark

Proteins are found not to be packed like solid, rather, there are numerous voids and pockets. The scaling behavior of volume, area, and cluster size all suggest that proteins are packed more like random spheres than like jig-saw puzzles. The origin of voids and pockets is the requirement of packing of random chain in compact space, and evolution has played overall little role in this.



(b)

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

(c)

```
1cdk_A Castp104
LGTGSFGRVAKLKVLQHTELVMMEYV---EDKENLTDF
2src_ Castp51
LGQGCFGEVA-IKLMFAMVLVITEYMGSLDDRANLADF
```

Fig. 9: Functional surfaces on the catalytic domains of cAMP-dependent protein kinase (1cdk) and tyrosine protein kinase (2src) are very similar. (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%) (Adapted from [92]).

## 5 Biochemical function prediction from protein geometry

### 5.1 Voids and pockets important for protein functions

The abundance of random voids and pockets poses a significant challenge, namely, how can we distinguish those few that are important for biological functions [77, 91] from those formed by random chance?

One approach is to determine if a void or a pocket on a protein structure is strongly similar to a void or a pocket involved in binding on another protein structure, and if the biological roles of the latter are known. For proteins carrying out similar functions such as binding similar substrates and catalyzing similar chemical reactions, their binding surfaces experience similar physical and chemical constraints. If a matching surface pocket or void from a protein structure with known functions is found, one can infer that the protein under investigation is likely to have similar biological functions as well. This is reminiscent of inferring protein functions by sequence alignment, but with a significant difference. Because key residues important for protein function are often sparsely located in diverse regions of the primary sequence of a protein, methods based on sequence similarity do not work well. Geometric analysis

is required in this case, as these key residues fold together to form spatially a pocket or a void.

Fig 9 provides an illustration. The overall sequence identity between the catalytic domains of cAMP-dependent protein kinase (pdb `1cdk.A`) and tyrosine protein kinase (`2src`) is low (16%), and it would be difficult to detect that these two proteins have similar function by examining only the global sequences of these two proteins, as reliable transfer of functional annotation requires an overall sequence identity of >60-70% [93, 94]. However, with the structural information of these two proteins, we can compute the surface pockets on these two proteins and identify the residues that are involved in substrate ATP binding [95]. As can be seen in Fig 9b, residues defining the pockets are well dispersed throughout the primary sequences, and it would be very difficult to select exactly these pocket residues important for binding from sequence, without the knowledge of protein structures and the application of geometric computation. Once these pocket residues are identified, the identity of the sequence fragments obtained from concatenation of the residues is much higher (51%, Fig 9c), suggesting a strong relationship of these two binding surfaces.

An implementation of this method called pvSoar can be used to detect related binding pockets for protein function inference [96, 97]. A library ($> 2$ million) of concatenated sequence fragments of residues located on the wall of a void or pocket is constructed and is used as the basis to search for functionally related proteins. Specifically, the sequence fragment of the pocket on the query protein is used to search against this library for detection of similar pocket sequence fragment. This can be done through a standard dynamic programming algorithm. Further details such as the statistical model for assessing significance of detected similarity, and the alternative measure of orientational root mean square distance (oRMSD) for assessing shape similarity can be found in reference [96]. With this approach, many previously unrecognized protein binding surfaces are found to be related [96].

### 5.2 Evolutionary pattern of binding surface of voids and pockets

Success in detecting similarity between sequence fragments of binding surface residues depends on the use of a scoring matrix, which is used to quantify the similarity of two sequence fragments. In general, scoring matrix for assessing sequence similarity is often derived from the evolutionary history of proteins sharing similar functions. As proteins of similar functions are under the constraints of binding to the same substrate and carry out similar reactions, the environment provided by the binding surface must maintain certain properties. This is reflected in specific pattern of amino acid residues substitution, as during evolution certain residues are free to mutate but others are constrained and cannot tolerate any mutations. However, the widely used matrices (such as the Pam matrix [98] and the Blosum matrix [99]) have implicit parameters whose values were determined from precomputed analysis of large quantities

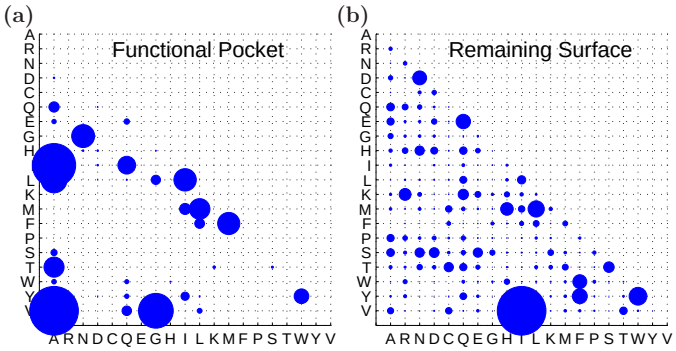



Fig. 10: Patterns of substitution rates of residues in the functional binding surface and the remaining surface of alpha-amylase (pdb 1bag) are very different. (a) Substitution rates of the functional binding surface. (b) substitution rates of the remaining surface on 1bag (Adapted from [102]).

of sequences, while the information of the protein of interest has limited or no influence. A more effective approach for studying residue substitutions is to employ an explicit model for residue substitution based on a continuous time Markov process and a phylogenetic tree of the protein [100–102]. In this case, residue substitution patterns are estimated from data specific to the protein of interest [103]. In addition, by focusing on residues located in binding surface, the selection pressure due to biological function can be clearly separated from the selection pressure on residues in other locations due to structural or folding requirement. In addition, it is easy to incorporate phylogenetic information in this model, which is important when sample sequences are unbalanced, *i.e.*, sequences from branches of the phylogenetic trees that have not diverged far will not skew the estimation. A Bayesian Monte Carlo method has been developed that can estimate accurately the substitution rates of amino acid residues located in a specific binding pocket, using a phylogenetic tree, a set of multiple-aligned sequences, and computed pocket/void as input data [102].

The pattern of residue substitutions on protein functional surfaces is often different from that of the remaining part of the surface. As an example, the substitution rates for residues on the functional surface of alpha amylase (pdb `1bag`) are shown in Figure 10, along with that of the remaining surface residues of the protein. It is clear that the selection pressures for residues located in the functional site and for residues on the rest of the protein surface are very different.

### 5.3 Function prediction by detecting similar binding surfaces.

The estimated substitution rates can be converted into scoring matrices for assessing similarity of residues in binding pockets [104, 105]. The utility of these

scoring matrices can be tested by examining if one can discover functionally related proteins, namely, whether one can identify protein structures that have similar binding surfaces and carry out similar biological functions. This can be demonstrated by the example of acetylcholinesterase [92]. Acetylcholinesterase (Enzyme Commission number $E.C.3.1.1.7$) is found in the synapse between nerve cells and muscle cells. It breaks down acetylcholine molecules into acetic acid and choline upon stimuli. Using a template structure (pdb `1ea5`), one seeks to identify other structures that are also acetylcholinesterase with the same E.C. number of all four digits and to locate the surface regions that are involved in enzyme activities. E.C. numbers represent a progressively finer classification of an enzyme, with the first digit about the basic reaction, and the last digit often about the specific functional group that is cleaved during reaction.

In this example, all pockets on a template structure of acetylcholinesterase (protein data bank name `1ea5`) are first exhaustively computed [77, 106]. Based on annotation derived from experimental literature, a pocket containing 32 residues is determined as the functional pocket (Fig. 11a), which contains the Ser and His residues of the active site triad [92]. A set of 17 sequences homologous to the template protein are used to build a phylogenetic tree (Fig. 11d) [92, 107]. The residue substitution rates on the surface of the binding pocket are estimated, and scoring matrices for assessment of similarity to this binding surface are then calculated [102]. Using these scoring matrices, a total of 70 protein structures are found to have similar functional surfaces as that of the query template `1ea5`, and hence are predicted as acetylcholinesterase. Indeed, all of them have the same $E.C.3.1.1.7$ label as that of `1ea5`. The query protein and an example of matched protein surface is shown in Fig. 11a and 11b, respectively. There are 71 PDB entries with enzyme class label $E.C.3.1.1.7$ in the Enzyme Structures Database (ESD, Version Oct. 2005, `www.ebi.ac.uk/thornton-srv`). This approach successfully identified 70 of them.

### 5.4 Remark

As there are many voids and pockets in protein structures, a challenging problem is to distinguish those that are important for biological functions from those formed by random chance. By identifying pockets or voids that are similar to binding surfaces on protein structures with known biochemical function, one can infer the function of the protein structure under investigation. A key element for this approach to work is the ability to capture subtle selection pressure on binding surfaces due to biological function and to separate it from selection pressure due to protein structure and folding. This can be achieved by estimating the substitution rates of residues on binding voids or pockets. A Bayesian Monte Carlo method based on a continuous time Markov process can be used for this task [102]. This idea has been carried out further for computing the binding profiles of enzymes, which characterizes

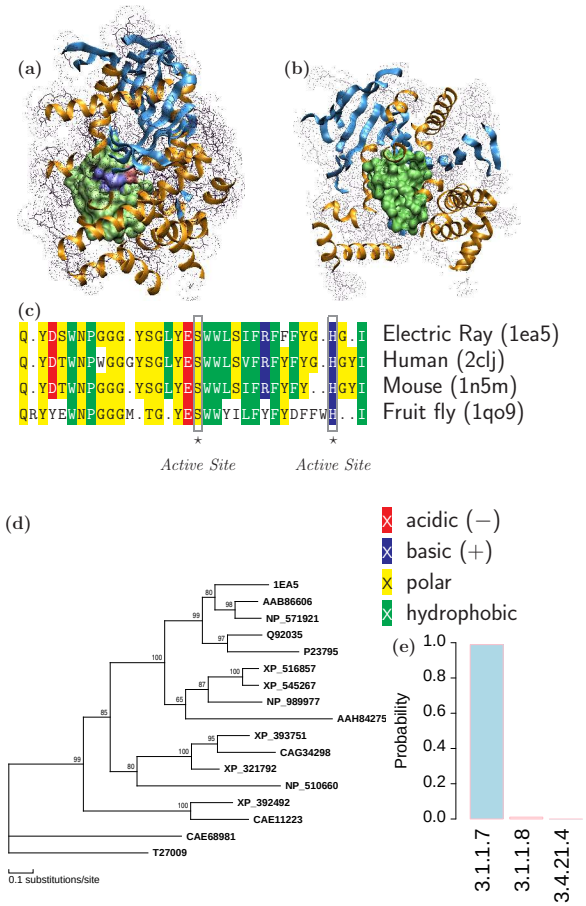



Fig. 11: Predicting biochemical functions of acetylcholinesterase (E.C. 3.1.1.7) by comparison of binding pockets. (a) The functional pocket (CASTP id = 79) on a structure of acetylcholinesterase (1ea5 ). It contains 32 resides and has a molecular volume of 986.3Å$^3$. Two residues from the catalytic triad are shown: Ser200 (red) and His440 (blue). (b) A matched binding surface on a human protein structure (2clj, CASTP id = 96), with 34 residues and a molecular volume of 981Å$^3$. (c) The multiple sequence alignment of several orthologous sequence fragments of residues located in the binding pockets. The two triad residues Ser200 and His440 are conserved. (d) The phylogenetic tree consisting of 17 sequences of acetylcholinesterase is used for estimating substitution rates of residues at the binding pocket. (e) The structure 1ea5 is predicted to be an acetylcholinesterase (E.C. 3.1.1.7, with a probability $\pi_1 \approx 0.99$) (Adapted from [92]).

enzyme substrate specificity and promiscuity [92, 108]. It was shown that this approach can be used to predict enzyme functions accurately. In a large scale test of 100 enzyme families with thousands of structures, at the specificity level of 99.98% (namely, few mistakes are made among predictions), enzyme functions can be correctly predicted for 80.55% of the proteins. This approach can also be applied to the challenging problems of inferring functions of orphan protein structure, whose biochemical roles are uncharacterized. More details can be found in [92, 108].

## 6 Summary

The atomic structures of protein molecules provide a wealth of information for understanding the how proteins work. With geometric characterization, we can gain important insight on the structural basis of protein folding behavior, develop effective empirical potential function for protein structure prediction, understand and characterize the prevalence of the geometric features of voids and pockets, as well as explore their origin. By directly estimating the evolutionary substitution rates of residues located on voids or pockets functionally important, we can separate selection pressure due to biological role from selection pressure due to the need to maintain protein structure and folding stability. The estimated evolutionary pattern can be used to predict and characterize protein functions. It is likely that continued geometric and topological studies of protein structures and their interplay will continue to generate new knowledge and lead to important innovation in computational tools important for furthering our understanding of biology.

## 7 Acknowledgment

This work is supported by grants from NSF (DBI-0646035 and DMS-0800257), NIH (GM079804, GM081682, and GM086145) and ONR (N00014-09-1-0028).

## References

1. F. M. Richards. Areas, volumes, packing, and protein structures. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.
2. F. M. Richards and W. A. Lim. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.*, 26:423–498, 1994.
3. M. Gerstein and F. M Richards. *Protein Geometry: Distances, Areas, and Volumes*, volume F, chapter 22. International Union of Crystallography, 1999.
4. A. Poupon. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr Opin Struct Biol*, 14(2):233–41, 2004.
5. M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

6. H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.*, 13:415–440, 1995.
7. H. Edelsbrunner and E.P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.
8. H. Edelsbrunner, M. Facello, P. Fu, and J. Liang. Measuring proteins and voids in proteins. In *Proc. 28th Ann. Hawaii Int'l Conf. System Sciences*, volume 5, pages 256–264, Los Alamitos, California, 1995. IEEE Computer Scociety Press.
9. M. A. Facello. Implememntation of a randomized algorithm for delaunay and regular triangulations in three dimensions. *Computer Aided Geometric Design*, 12:349–370, 1995.
10. J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computing of macromolecules I: Molecular area and volume through alpha-shape. *Proteins*, 33:1–17, 1998.
11. X. Li, C. Hu, and J. Liang. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, 53:792–805, 2003.
12. J. Liang. *Computational algorithms for protein structure prediction*, chapter Computation of protein geometry and its applications: Packing and function prediction. Springer, 2006.
13. F. Aurenhammer. Voronoi diagrams - a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23:345–405, 1991.
14. B. Gillespie and K.W. Plaxco. Using protein folding rates to test protein folding theories. *Annu Rev Biochem*, 73:837–59, 2004.
15. K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 227:985–994, 1998.
16. J. R. Bienkowska, R. G. Rogers, and T. F. Smith. Filtered neighbors threading. *Proteins*, 37:346–359, 1999.
17. J.R. Bienkowska, R.G. Rogers, and T.F. Smith. Filtered neighbors threading. *Proteins*, 37:346–59, 1999.
18. W.R. Taylor. Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.*, 269:902–43, 1997.
19. Z. Ouyang and J. Liang. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci*, 17(7):1256–63, 2008.
20. K.W. Plaxco, K.T. Simons, I. Ruczinski, and D. Baker. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry*, 39:11177–83, 2000.
21. M. Scalley-Kim and D. Baker. Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. *J. Mol. Biol.*, 338:573–583, 2004.
22. S. Kachalo, H.M. Lu, and J. Liang. Protein folding dynamics via quantification of kinematic energy landscape. *Phys Rev Lett*, 96(5):058106, 2006.
23. Hue Sun Chan and Ken A. Dill. Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.*, 100(12):9238–9257, 1994.
24. A. Šali, E. I. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.
25. N. D. Socci and J. N. Onuchic. Folding kinetics of proteinlike heteropolymer. *J. Chem. Phys.*, 101:1519–1528, 1994.

26. I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan, and J. R. Banavar. Lattice model for rapidly folding protein-like heteropolymers. *Proc. Natl. Acad. Sci. U.S.A*, 92:9206–9209, 1995.
27. D. K. Klimov and D. Thirumalai. Criterion that determines the foldability of proteins. *Phys. Rev. Lett.*, 76:4070–4073, 1996.
28. R. Mélin, H. Li, N. Wingreen, and C. Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. Chem. Phys.*, 110:1252–1262, 1999.
29. H. S. Chan and K. A. Dill. Compact polymers. *Macromolecules*, 22:4559–4573, 1989.
30. H. S. Chan and K. A. Dill. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.*, 92:3118–3135, 1990.
31. J. Liang, J. Zhang, and R. Chen. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. *J. Chem. Phys.*, 117:3511–3521, 2002.
32. J. Zhang, Y. Chen, R. Chen, and J. Liang. Importance of chirality and reduced flexibility of protein side chains: A study with square and tetrahedral lattice models. *J. Chem. Phys.*, pages 592–603, 2004.
33. Y. Xia and M. Levitt. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci U S A*, 99(16):10382–7, 2002.
34. Y. Xia and M. Levitt. Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol*, 14(2):202–7, 2004.
35. K. M. Fiebig and K. A. Dill. Protein core assembly processes. *J. Chem. Phys.*, 98:3475–3487, 1993.
36. K. A. Dill, K. M. Fiebig, and H. S. Chan. Cooperativity in protein-folding kinetics. *Proc. Natl Acad. Sci.*, 90:1942–1946, 1993.
37. T.R. Weikl and K.A. Dill. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.*, 329:585–98, 2003.
38. T.R. Weikl and K.A. Dill. Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.*, 332:953–63, 2003.
39. T.R. Weikl, M. Palassini, and K.A. Dill. Cooperativity in two-state protein folding kinetics. *Protein. Sci.*, 13:822–9, 2004.
40. C. Merlo, K.A. Dill, and T.R. Weikl. Phi values in protein-folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci. U. S. A.*, 102:10171–5, 2005.
41. C. Anfinsen, E. Haber, M. Sela, and F. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci.*, 47:1309–1314, 1961.
42. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
43. R. Janicke. Folding and association of proteins. *Prog. Biophys. Mol. Biol.*, 49:117–237, 1987.
44. C. Keasar and M. Levitt. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.*, 329:159–174, 2003.
45. S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

46. R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275:895–916, 1998.
47. H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44:223–232, 2001.
48. H. Y. Zhou and Y. Q. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, 11:2714–26, 2002.
49. B. J. McConkey, V. Sobolev, and M. Edelman. Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. U.S.A.*, 100(6):3215–20, 2003.
50. A. Zomorodian, L. Guibas, and P. Koehl. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. *Computer Aided Geometric Design*, 23:531–544, 2006.
51. L. Wernisch, M. Hunting, and S. J. Wodak. Identification of structural domains in proteins by a graph heuristic. *Proteins.*, 35(3):338–52, 1999.
52. R. K. Singh, A. Tropsha, and I. I. Vaisman. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol*, 3(2):213–221, 1996.
53. W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. A new approach to protein fold recognition based on delaunay tessellation of protein structure. *Pac Symp Biocomput*, pages 486–497, 1997.
54. C. W. Carter Jr., B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.*, 311(4):625–638, 2001.
55. B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12):1540–8, 2003.
56. W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In R.B. Altman, A.K. Dunker, L. Hunter, and T.E. Klein, editors, *Pacific Symposium on Biocomputing'97*, pages 486–497, Singapore, 1997. World Scientific.
57. B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12):1540–8, 2003.
58. X. Li and J. Liang. Geometric cooperativity and anti-cooperativity of three-body interactions in native proteins. *Proteins*, in press, 2005.
59. X. Li and J. Liang. Computational design of combinatorial peptide library for modulating protein-protein interactions. *Pacific Symposium of Biocomputing*, pages 28–39, 2005.
60. M. J. Sippl. calculation of conformational ensembles from potentials of the main force. *J. Mol. Biol.*, 213:167–180, 1990.
61. A. V. Finkelstein, A. Ya. Badretdinov, and A. M. Gutin. Why do protein architectures have boltzmann-like statistics? *Proteins.*, 23(2):142–50, 1995.
62. H. Edelsbrunner and E.P. Mücke. Three-dimensional alpha shapes. *ACM Trans Graphics*, 13:43–72, 1994.
63. L. Adamian and J. Liang. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.*, 311:891–907, 2001.

64. Park B. and Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.
65. X. Li, C. Hu, and J. Liang. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, 53:792–805, 2003.
66. C. Chothia. Principles that determine the structure of proteins. *Ann. Rev. Biochem.*, 53:537–572., 1984.
67. K. P. Murphy and S. J. Gill. Solid model compounds and the thermodynamics of protein unfolding. *J. Mol. Biol.*, 222:699–709, 1991.
68. B. Gavish, E. Gratton, and C. J. Hardy. Adiabatic compressibility of globular proteins. *Proc. Natl. Acad. Sci. USA*, 80:750–754, 1983.
69. C. Chothia. Structural invariants in protein folding. *Nature*, 254:304–308, 1975.
70. Y. Harpaz, M. Gerstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.
71. M. Gerstein and C. Chothia. Packing at the protein-water interface. *Proc. Natl. Acad. Sci. USA.*, 93:10167–10172, 1996.
72. M. Gerstein, J. Tsai, and M. Levitt. The volume of atoms on the protein surface: calcualted from simulation, using Voronoi polyhedra. *J. Mol. Biol.*, 249:955–966, 1995.
73. S. Chakravarty, A. Bhinge, and R. Varadarajan. A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J Biol Chem*, 277(35):31345–53, 2002.
74. H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput Geom*, 13:415–440, 1995.
75. H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discrete Applied Math.*, 88:83–102, 1998.
76. J. Liang, H. Edelsbrunner, P. Fu, P. V. Sudhakar, and S. Subramaniam. Analytical shape computing of macromolecules II: Identification and computation of inaccessible cavities inside proteins. *Proteins*, 33:18–29, 1998.
77. J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, 7:1884–1897, 1998.
78. J. Zhang, R. Chen, C. Tang, and J. Liang. Origin of scaling behavior of protein packing density: A sequential monte carlo study of compact long chain polymers. *J. Chem. Phys.*, 118:6102–6109, 2003.
79. H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Disc. Appl. Math.*, 88(83–102), 1998.
80. U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.
81. J. Liang and K. A. Dill. Are proteins well-packed? *Biophys. J.*, 81(2):751–766, 2001.
82. B. Lorenz, I. Orgzall, and H-O. Heuer. Universality and cluster structures in continuum models of percolation with two different radius distributions. *J. Phys. A: Math. Gen.*, 26:4711–4722, 1993.
83. D. Stauffer. *Introduction to percolation theory.* Taylor & Francis, London, 1985.

84. R. Meester, R. Roy, and A. Sarkar. Nonuniversality and continuity of the critical covered volume fraction in continuum persolation. *J. Stat. Phys.*, 75:123–134, 1994.
85. S. C. van der marck. Network approach to void percolation in a pack of unequal spheres. *Phys. Rev. Lett.*, 77:1785–1788, 1996.
86. J. Adler, Y. Meir, A. Aharony, and A. B. Harris. Series study of percolation moments in general dimension. *Physic. Rev. B*, 41:9183–9206, 1990.
87. J. Zhang and J.S. Liu. On side-chain conformational entropy of proteins. *PLoS Comput Biol*, 2(12):e168, 2006.
88. P. L. Privalov. Intermediate states in protein folding. *J. Mol. Biol.*, 258:707–725, 1996.
89. J. S. Liu. *Monte Carlo strategis in scientific computing.* Springer, New York, 2001.
90. J. S. Liu and R. Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.
91. R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Sci.*, 5:2438–2452, 1996.
92. Y-Y Tseng, J. Dundas, and J. Liang. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.*, 387(2):451–64, 2009.
93. B. Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, 2002.
94. W. Tian and J. Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*, 333(4):863–82, 2003.
95. T. A. Binkowski, S. Naghibzadeh, and J. Liang. CASTp: Computed atlas of surface topography of proteins. *Nucleic Acids Res.*, 31:3352–3355, 2003.
96. T. A. Binkowski, L. Adamian, and J. Liang. Inferring functional relationship of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.*, 332:505–526, 2003.
97. T.A. Binkowski, P. Freeman, and J. Liang. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nuc. Aci. Res.*, 32:W555–558, 2004.
98. M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. Atlas of protein sequence and structure. 5 suppl. 3:345, 1978.
99. S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.
100. Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15(12):1600–11, 1998.
101. S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5):691–699, 2001.
102. Y.Y. Tseng and J. Liang. 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.*, 23(2):421–436, Feb 2006.
103. S. Whelan, P. Liò, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genet.*, 17(5):262–272, 2001.

104. S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.*, 87:2264–2268, 1990.
105. S. F. Altschul and W. Gish. Local alignment statistics. *Methods Enzymol.*, 266:460–480, 1996.
106. T.A. Binkowski, S. Naghibzadeh, and J. Liang. CASTp: Computed atlas of surface topography of proteins. *Nuc. Aci. Res.*, 31(13):3352–3355, 2003.
107. J. Adachi and M. Hasegawa. A computer program package for molecular phylogenetics. ver 2.3, 1996.
108. J. Liang, Y-Y. Tseng, J. Dundas, A. Binkowski, A. Joachimiak, Z. Ouyang, and L. Adamian. Predicting and characterizing protein functions through matching geometric and evolutionary patterns of binding surfaces. *Advances in protein chemistry*, 75:107–141, 2008.