31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Geometry of protein shape and its evolutionary pattern for function prediction and characterization

Jie Liang

Abstract— **Proteins contain thousands or more atoms and have complex shapes. We discuss here the computation of protein packing defects, in the form of voids and pockets, from experimentally resolved protein structures, and the nature of their distribution and scaling behavior, as well as their origin. We further discuss how evolutionary selection pressure due to biological function unaltered by selection pressure due to constraints from folding and stability can be isolated and estimated, and how such information can be used to predict protein function and characterize binding properties of enzymes.**

Typical proteins contain thousands or more atoms and have complex shapes. Understanding how they pack is an important question, as it helps us to gain insight on important biological questions such as how proteins function. Packing defects in the form of voids and pockets in experimentally resolved protein structure can be computationally identified and measured. We discuss the overall distributions of voids and pockets in proteins, as well as the scaling properties of packing related measures with protein size, and findings on the origin of packing defects and the role played by evolution. Finally, we describe how protein binding activities and biological functions can be predicted for the important class of enzyme proteins based on geometric computation and evolutionary analysis of voids and pockets.

Geometric models. We use Voronoi diagram, Delaunay triangulation , and alpha shape to characterize protein structures. The Voronoi region of an atom ball is the set of points closest to this ball by the power distance definition [1–4]. The power distance, denoted as $\pi_x(y)$, of a point $y \in \mathbb{R}^3$ from an atom ball $b(x,r)$ centered at $x \in \mathbb{R}^3$ with radius r is defined as $\pi_x(y) = \|y-x\|^2 - r^2$. The collection of Voronoi regions and their boundaries form the *weighted Voronoi diagram*, or the power diagram of the molecule. 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 [1].

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

J. Liang is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60612, USA jliang@uic.edu

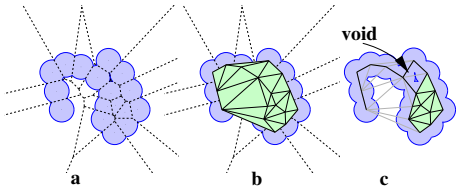

Fig. 1. Geometry of a simplified molecule in two-dimensional space for illustration. (a) The molecule formed by the union of atom disks. Voronoi diagram is in dashed lines. (b) The shape enclosed by the boundary polygon is tessellated by the *Delaunay triangulation*. (c) The alpha shape of the molecule 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 [5]).

The alpha shape of a molecule is formed by a subset of the simplices in the weighted Delaunay triangulation [2]. It captures the connectivity of the convex Voronoi regions in the form of a *dual complex*, denoted as $\mathcal{K}_0$: $\mathcal{K}_0 = \{\sigma = \text{conv}x_B \,|\, \bigcap V_B \cap \bigcap B \neq \emptyset\}$, 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 convx_B is the same as the simplex formed by the convex hull of the atom centers, denoted as x_B. Details of the geometric model for protein structure can be found in [1, 2, 4, 5].

Distribution of voids and pockets. Protein cores are often considered to be solid-like [6], as proteins have high packing densities [7] and low compressibilities. Analysis of Voronoi diagrams of protein structures showed that the average packing density in a protein is as high as that inside crystalline solids [8, 9]. Sometimes protein is compared to an assemble jigsaw puzzle [10].

However, there exists unfilled spaces inside proteins, in the form of 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 [11–14].

The prevalence of voids and pockets in proteins can be assessed using the pocket algorithm described in [15]. For a set of proteins representative of all known protein structures based on the PDBSELECT database of proteins of different fold, 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

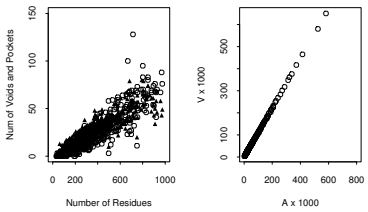

Fig. 2. The scaling behavior of geometric properties of proteins. (a) Voids and pockets for a set of 636 proteins representing most of the known protein folds. The number of voids and pockets is linearly correlated with the number of residues in a protein. Solid triangles and empty circles represent 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. (Adpated from [16])

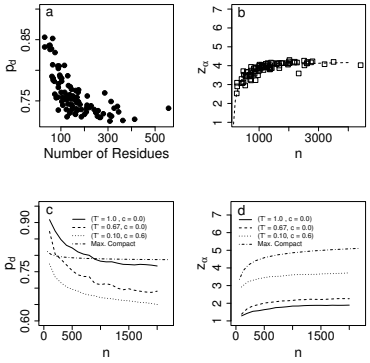

Fig. 3. Comparison of scaling behavior of packing density and coordination number of proteins and compact chain polymers. (a) Packing density p_d of proteins of different lengths; (b) Scaling behavior of coordination number Z_α calculated based on alpha contact and protein chain length. (c) Packing density p_d and (d) coordination number 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 (Adapted from [20]).

(Fig 2a) [16]. 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 probe, so they are large enough to contain at least one water molecule. This finding suggest that voids and pockets are quite common in protein structures.

Scaling behavior. 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: $V \propto A^{3/2}$. In contrast, the volume of proteins scales linearly with the surface areas of proteins (Fig 2b). This linear relationship is also what is observed in models for disordered materials [17, 18].

For randomly packed spheres, when the packing density p_d 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 [17, 19]. At this percolation threshold p_c, the volume V of a cluster of random spheres scale with the length R of the cluster as $V \propto R^D$, with a characteristic exponent $D = 2.5$ in three-dimensional space [17, 18]. In proteins, it was found that $\ln V \propto D \ln R$, with a fractal dimension $D = 2.47 \pm 0.04$ (Figure **??**) [16]. This suggests that packing in proteins behaves like random spheres near their percolation threshold.

Origin of voids and pockets in protein structure. Using geometric algorithms, packing density p_d can be readily computed [4, 13], and the scaling relationship of p_d and protein chain length N is shown in Fig 3a [16]. To answer the question that whether the scaling behavior of p_d with chain length is unique to proteins, we can study voids and packing in generic model chain polymers that are not proteins [20]. For this purpose, one needs to generate self-avoiding walks (SAW) of chain polymers.

One technical challenge is that it is very difficult to generate long chain SAWs. This can be overcome by using the chain-growth based sequential Monte Carlo method [21], which keeps proper weights for samples generated by growth. extensive well-designed resampling. Thousands of SAWs in three dimensional space at any specified intervals of compactness can be successfully generated [20]. The scaling behavior of p_d and chain length for these randomly generated SAWs is very similar to that observed in protein (Fig 3). These 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 [20].

Voids and pockets important for protein functions and their evolution. The abundance of random voids and pockets poses a significant challenge, namely, how can we distinguish those pockets and voids that are important for biological functions [14, 22] from those formed by random chance?

One approach is to decide if a void or a pocket on a protein structure is strongly similar to a void or a pocket on another protein structure, and the biological roles of the latter are known. If so, we can infer that the protein with the void or pocket under investigation is likely to have similar biological functions as the second protein. 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. Voids and pockets performing similar functions but on different structures have strong resemblance. Fig 4 provides an illustration.

This approach was implemented in a software called PVSOAR for detecting related binding pockets for protein function inference [23, 24]. A library of concatenated sequence fragments (> 2 million) of residues located on the wall of a void or pocket is constructed. The sequence fragment of the pocket on the query protein is then used to search for similar pocket sequence fragment through a standard dynamic programming algorithm. Further details such as the statistical model for assessing significance of detected similarity and the alternative measure of oRMSD

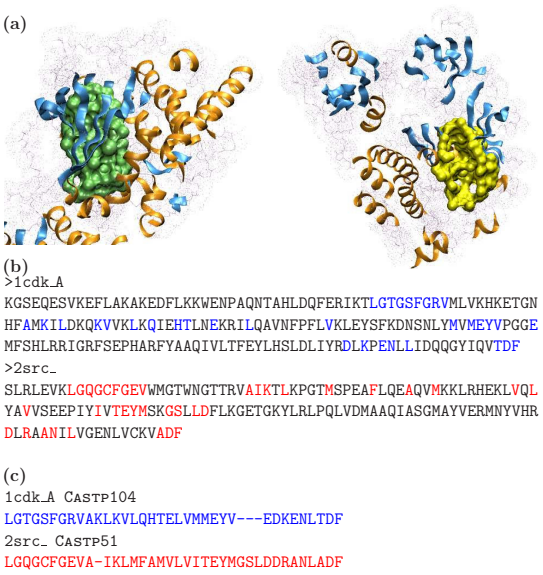

Fig. 4. Functional surfaces on the catalytic domains of cAMP-dependent protein kinase (1cdk) and tyrosine protein kinase (2src) are very similar. (a) In both cases, the active sites are computed as surface pockets. (b) Residues defining the pockets are well dispersed throughout the primary sequences (full sequence identity = 16%). It would be difficult to detect that these two proteins have similar function by examining only the global sequences of these two proteins. (c) The identity of their surface sequence patterns is much higher (51%) (Adpated from [29]).

for assessing shape similarity can be found in reference [23]. With this approach, numerous previously unrecognized protein binding surfaces are found to be related [23].

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. However, the widely used matrices (such as the PAM matrix and the BLOSUM matrix) have implicit parameters whose values were determined from precomputed analysis of large quantities of sequences, while the information of the protein of interest has limited or no influence. A more effective approach is to employ an explicit model for residue substitution based on a continuous time Markov process and a phylogenetic tree of this specific protein [25–27]. 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. It is also 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 [27].

The pattern of residue substitutions on protein functional

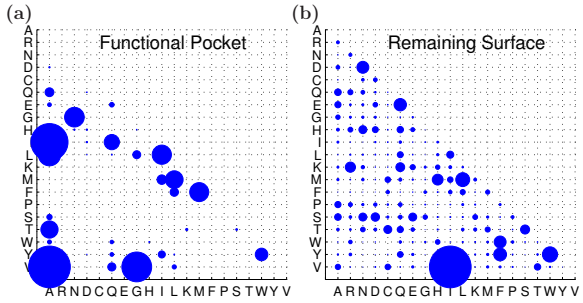

Fig. 5. Patterns of substitution rates of residues in the functional binding surface and the remaining surface of alpha-amylase (pdb 1bag) are very different. (a) Substitution rates of the functional binding surface. (b) substitution rates of the remaining surface on 1bag. It is clear that the selection pressures for residues located in the functional site and for residues on the rest of the protein surface are very different (Adapted from [27]).

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 5, along with that of the remaining surface residues of the protein.

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 [28]. 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 [29] (Fig. 6).

Bsed on estimated residue substitution rates on the surface of the binding pocket, scoring matrices for assessment of similarity to this binding surface can be calculated [27]. 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. 6a and 6b, 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. 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 [29, 30].

Summary. The atomic structures of protein molecules provide a wealth of information for understanding the biological roles of proteins. With geometric characterization, we can gain important insight on the structural basis of

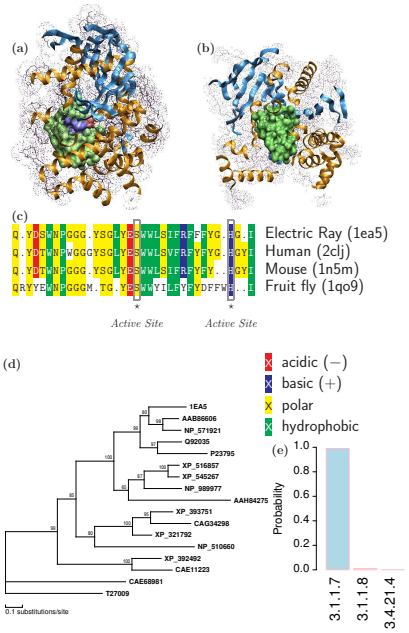

Fig. 6. Predicting biochemical functions of acetylcholinesterase (E.C. 3.1.1.7) by comparison of binding pockets. (a) The functional pocket (CASTP id = 79) on a structure of acetylcholinesterase (1ea5) was identified. It includes two residues from the catalytic triad: Ser200 (red) and His440 (blue). (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 $\pi_1 \approx 0.99$) (Adpated from [29]).

proteins. By directly estimate the evolutionary pattern of residue substitituon for voids or pockets, we can separate selection pressure due to biological role from that due to the need to maintain protein structure and folding stability. The 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 generate new knowledge and lead to important innovation in computational tools for furthering our understanding of biology.

REFERENCES

[1] H. Edelsbrunner, "The union of balls and its dual shape," *Discrete Comput. Geom.*, vol. 13, pp. 415–440, 1995.

[2] H. Edelsbrunner and E. Mücke, "Three-dimensional alpha shapes," *ACM Trans. Graphics*, vol. 13, pp. 43–72, 1994.

[3] 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*, vol. 5. Los Alamitos, California: IEEE Computer Scociety Press, 1995, pp. 256–264.

[4] 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*, vol. 33, pp. 1–17, 1998.

[5] J. Liang, *Computational algorithms for protein structure prediction*. Springer, 2006, ch. Computation of protein geometry and its applications: Packing and function prediction.

[6] C. Chothia, "Principles that determine the structure of proteins." *Ann. Rev. Biochem.*, vol. 53, pp. 537–572., 1984.

[7] F. M. Richards, "Areas, volumes, packing, and protein structures." *Ann. Rev. Biophys. Bioeng.*, vol. 6, pp. 151–176, 1977.

[8] C. Chothia, "Structural invariants in protein folding." *Nature*, vol. 254, pp. 304–308, 1975.

[9] M. Gerstein and C. Chothia, "Packing at the protein-water interface," *Proc. Natl. Acad. Sci. USA.*, vol. 93, pp. 10 167–10 172, 1996.

[10] F. M. Richards and W. A. Lim, "An analysis of packing in the protein folding problem." *Q. Rev. Biophys.*, vol. 26, pp. 423–498, 1994.

[11] H. Edelsbrunner, "The union of balls and its dual shape," *Discrete Comput Geom*, vol. 13, pp. 415–440, 1995.

[12] H. Edelsbrunner, M. Facello, and J. Liang, "On the definition and the construction of pockets in macromolecules." *Discrete Applied Math.*, vol. 88, pp. 83–102, 1998.

[13] 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*, vol. 33, pp. 18–29, 1998.

[14] 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*, vol. 7, pp. 1884–1897, 1998.

[15] H. Edelsbrunner, M. Facello, and J. Liang, "On the definition and the construction of pockets in macromolecules," *Disc. Appl. Math.*, vol. 88, no. 83–102, 1998.

[16] J. Liang and K. A. Dill, "Are proteins well-packed?" *Biophys. J.*, vol. 81(2), pp. 751–766, 2001.

[17] 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.*, vol. 26, pp. 4711–4722, 1993.

[18] D. Stauffer, *Introduction to percolation theory*. London: Taylor & Francis, 1985.

[19] R. Meester, R. Roy, and A. Sarkar, "Nonuniversality and continuity of the critical covered volume fraction in continuum persolation," *J. Stat. Phys.*, vol. 75, pp. 123–134, 1994.

[20] 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.*, vol. 118, pp. 6102–6109, 2003.

[21] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044, 1998.

[22] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton, "Protein clefts in molecular recognition and function," *Protein Sci.*, vol. 5, pp. 2438–2452, 1996.

[23] T. A. Binkowski, L. Adamian, and J. Liang, "Inferring functional relationship of proteins from local sequence and spatial surface patterns," *J. Mol. Biol.*, vol. 332, pp. 505–526, 2003.

[24] T. 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.*, vol. 32, pp. W555–558, 2004.

[25] Z. Yang, R. Nielsen, and M. Hasegawa, "Models of amino acid substitution and applications to mitochondrial protein evolution." *Mol Biol Evol*, vol. 15(12), pp. 1600–11, 1998.

[26] 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.*, vol. 18(5), pp. 691–699, 2001.

[27] 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.*, vol. 23(2), pp. 421–436, Feb 2006.

[28] S. F. Altschul and W. Gish, "Local alignment statistics," *Methods Enzymol.*, vol. 266, pp. 460–480, 1996.

[29] 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.*, vol. In press, p. doi:10.1016/j.jmb.2008.12.072, 2009.

[30] 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*, vol. 75, 2008.