DOE Early Career Principal Investigator Award
DOE Award No. DE-FG02-06ER25735

Feature Extraction, Characterization, and Visualization for Protein Interaction
via
Geometric and Topological Methods
Final Technical Report — (09/2006 – 10/2009)

Yusu Wang
Department of Computer Science and Engineering
The Ohio State University
yusu@cse.ohio-state.edu

1 Summary

Shape analysis plays an important role in many applications. In particular, in molecular biology, analyzing molecular shapes is essential to the fundamental problem of understanding how molecules interact. This project aims at developing efficient and effective algorithms to characterize and analyze molecular structures using geometric and topological methods. Two main components of this project are (1) developing novel molecular shape descriptors; and (2) identifying and representing meaningful features based on those descriptors. The project also produces accompanying (visualization) software. Results from this project (09/2006–10/2009) include the following publications. We have also set up web-servers for the software developed in this period, so that our new methods are accessible to a broader scientific community. The web sites are given below as well.

In this final technical report, we first list publications and software resulted from this project. We then briefly explain the research conducted and main accomplishments during the period of this project.

2 Publications resulted from the project

References

- [1] L. J. Guibas and Y. Wang. Towards unsupervised segmentation of semi-rigid low-resolution molecular surfaces. *Algorithmica*, 48(4):433–448, August 2007.
- [2] A. Sacan, O. Ozturk, H. Ferhatosmanoglu, and Y. Wang. LFM-Pro: A tool for detecting significant local structural sites in proteins. *Bioinformatics*, 23(6):709–716, 2007.
- [3] H. Sun, F. Ferhatosmanoglu, and Y. Wang. An enhanced partial order curve comparison over multiple protein folding trajectories. In *Proc. Intl. Conf. Computational Systems Bioinformatics*, 299–310, October, 2007.

- [4] Y. Wang. Approximating nearest neighbor among triangles in convex position. *Information Processing Letters*, 108 (6): 379–385, 2008.
- [5] M. Belkin, J. Sun, and Y. Wang. Discrete laplace operator for meshed surfaces. *ACM Sympos. Comput. Geom.*, 278–287, 2008.
- [6] V. Natarajan, P. Koehl, Y. Wang, and B. Hamann. Visual analysis of biomolecular surfaces. In L. Linsen, H. Hagen, and B. Hamann (editors), *Mathematical Methods for Visualization in Medicine and Life Science*. Springer Verlag, Mathematics and Visualization, 237–256, 2008.
- [7] H. Sun, F. Ferhatosmanoglu, M. Ota, and Y. Wang. An enhanced partial order curve comparison algorithm and its application to analyzing protein folding trajectories. *BMC Bioinformatics* 9: 344, 2008.
- [8] K. Buchin, M. Buchin and Y. Wang. Exact algorithms for partial curve matching via the Frechet distance. In *SIAM/ACM Sympos. Discrete Alg.*, 645–654, 2009.
- [9] M. Belkin, J. Sun, and Y. Wang. Constructing Laplace operator from point clouds in R^d . In *SIAM/ACM Sympos. Discrete Alg.*, 1031–1040, 2009.
- [10] C. Luo, I. Safa, and Y. Wang. Approximating gradients for meshes and point clouds via diffusion metric. *Comput. Graph. Forum* 28(5): 1497–1508, 2009.
- [11] Tamal K. Dey, Jian Sun, Yusu Wang. Approximating loops in a shortest homology basis from point data. *CoRR abs/0909.5654*, 2009.
- [12] H. Sun, A. Sacan, F. Ferhatosmanoglu, and Y. Wang. Smolign: Spatial Motifs Based Multiple Protein Structure Alignment. Manuscript, to be submitted to Bioinformatics, 2009.
- [13] W. Harvey and Y. Wang. Generating and exploring a collection of topological landscapes for visualization of scalar-valued functions. Manuscript, to be submitted to IEEE EuroVis, 2009.
- [14] T. K. Dey, K. Li, C. Luo, P. Ranjan, I. Safa, and Y. Wang. Persistent heat signature for pose-oblivious matching of incomplete models. Manuscript, 2009.

Softwares developed. We have set up web-servers for software developed in this projects. The first software identifies structure motifs from multiple proteins based on shape descriptors computed from the topological features of distance field [2]. The second one is an algorithm to align multiple, possibly high dimensional, curves. The software has been used to analyze both protein folding trajectories and protein structures [3, 7]. The third one is a novel software suite to capturing common motifs from *multiple* protein structures [12]. Results are returned in both text and graphical formats. The URLs are:

[S1] **LFM-Pro**: <http://www.ceng.metu.edu.tr/~ahmet/Wiki/bioinfo/lfmpro/lfmpro/>

[S2] **EPO**: <http://bio.cse.ohio-state.edu/EPO/>

[S3] **Smolign**: <http://bio.cse.ohio-state.edu/Smolign/>

We have also set up a preliminary web server for visualizing geometric shape descriptors defined on molecular structures at:

[S4] **MolViz**: http://digitalunion.osu.edu/r2/summer07/bal_nasca/.

3 Technical progress and accomplishments

To understand molecular structures and study molecular interaction, our research focuses on developing novel *geometric and topological* methods to describe [2, 4, 5, 9, 10, 11], to extract [1, 2, 7, 8, 12], and to visualize [6, 13] shape features for molecular data.

3.1 Shape descriptors

Convex distance. Given a surface M , we have previously developed the Elevation function to capture protrusions and cavities. While the Elevation function can identify meaningful features, it is rather time consuming to compute. One descriptor function we proposed in the project proposal is to measure for each point on M , its closest distance to the convex cover of M — the *convex cover* of M is the portion of its convex hull that are not part of M , and intuitively, it covers the cavities of M . This simple function works well in practice to measure concavities. In [4], we developed an efficient algorithm to approximate this distance for all points in M in near linear time, which is a significant improvement over the original quadratic-time algorithm.

Distance field. A protein performs its functions by interacting with its environment, which is the three-dimensional space around it. However, most descriptors focus only on molecular surfaces. In [2], we described the space around a molecule P by the distance field $d_P : \mathbb{R}^3 \rightarrow \mathbb{R}$. In particular, let P denote the backbone of the molecule. For any $x \in \mathbb{R}^3$, $d_P(x)$ is defined as the nearest distance from x to any point in P . We then identify interesting features of P by considering a sphere around each critical point (i.e. minimum, maximum, and saddle) of d_P . Intuitively, we model the influence of the protein P to its environment through the distance field, and potential features occur at the singular positions (critical points) of this field. It was demonstrated in [2] that features constructed from the distance field correspond to important functional sites. See the right figure for an example, where the most common features (represented by the spatial neighborhoods) for all proteins in the Serine Proteases family correspond to the catalytic triad, as well as the disulfide bridge that is crucial in the stabilization of the triad.

Laplace-based features. The Laplace-Beltrami operator is a fundamental geometric object associated to a Riemannian manifold and has many desirable properties. In particular, eigenfunctions of the Laplacian form a natural basis for square integrable functions on the manifold analogous to Fourier harmonics for functions on a circle (i.e. periodic functions). Therefore, computing eigenfunctions of a Laplacian allows one to construct a basis reflecting the geometry of the surface. One of the goals in this project is to develop a complete and diverse set of shape descriptors both for the molecular surfaces, and for the space around a molecule (i.e. volumetric data). To this end, it is natural to use Laplace eigenfunctions as shape descriptors. Since in practice, we are usually given piecewise-linear (PL) manifolds (such as a mesh) as input, the first step is to develop a *discrete* version of the Laplace operator for PL manifolds. In [5, 9], we developed the first discrete Laplace operators for general discrete input with a theoretical guarantee.

We then investigated several directions to develop a concise family of shape descriptors based on discrete Laplace eigenfunctions. One idea we explored was to locally check which Laplacian eigenfunctions are “dominating” (i.e. the combination of them produces the local geometry). The goal was to associate each point on a molecular surface with a set of indices of dominating eigenfunctions, and then consider two

points to be similar if they share similar dominating eigenfunctions. We can then segment molecular surface based on such a similarity clustering to hopefully aid the automatic extraction of molecular binding sites. The second idea we investigated was to see whether we can identify feature points based on Laplacian eigenfunctions. Unfortunately, both directions failed for molecular surfaces. For the first direction, the “dominating” eigenfunctions are rather discontinuous as we move from one point to another, which causes segmentation based on such information rather unstable. For the second direction, we can indeed define robust shape features based on Laplacian eigenfunctions (more specifically, based on how well the first few Laplacian eigenfunctions reconstruct the coordinate functions). Such features proved to be quite useful for general shapes used in graphics, such as models for animals and other objects. However, molecular shapes are very homogeneous (locally similar everywhere), and such features unfortunately are not very discriminative for molecular structures. Nevertheless, such investigation produced useful insights and results that eventually contributed to publications for graphics applications [10, 14].

Shortest system of loops. In general there is no consensus on what “features” are. Intuitively, they represent “critical” and “essential” structures of an input shape. Hence topological methods are natural tools to capture and describe such singular entities. One particular interesting family of topological structures are generating cycles, which contain rich both topological and geometric information of a given shape. For example, the loops in the right figure characterize the hole contained in this torus shape. However, to make such “loop”-features more canonical, we wish to identify the shortest system of loops — intuitively, the solid loop in the right is better than the dotted one in this sense as it “hugs” the hole of torus tightly. In [11], we presented an efficient algorithm to compute such shortest system of loops for general domains. Such loops describe both the topology (being non-trivial loops) and the geometry (being shortest) of the input domain, and we can potentially compare two objects by comparing their shortest system of loops.

3.2 Feature identification and applications

Features under hinge-motion. A protein may change its conformation significantly when interacting with other molecules. With current structure determination technologies, it is possible to obtain some “snapshots” of this deformation process, such as the beginning and ending conformations. In order to understand the entire interaction, it is necessary to find correspondences between these two snapshots and reconstruct how one deforms to the other. In [1], we consider the following problem. Assume that given two arbitrarily oriented surfaces A and B (corresponding to two conformations of the same protein) undergoes some hinge motion, which is one of the most popular type of macromolecular motions. (See the right figure for an example.) We wish to find features that are invariant with respect to the hinge motion, and further identify the correspondences between the two surfaces (i.e, it is a non-rigid registration problem). We implemented a novel feature identification method using both the Elevation function developed earlier by us, and a landmark-based virtual coordinate idea. This results in robust features on A and B that helps to produce the desirable correspondences them.

Common feature extraction. It is often desirable to extract common substructures (motifs) from a family of proteins sharing some function(s) as features, as potentially they correspond to the crucial common functional sites. In particular, since a protein is usually modeled as a polygonal curve to represent its

backbone for efficiency, the goal now is to identify common sub-curves from a set of backbone curves as potential features. This multiple curve alignment problem is a very hard problem and most previous methods focus on curves that have *high* similarity. However, a set of multi-functional and multi-domain proteins may share only *small partial* similarity corresponding to their common functional sites. Based on the topological features of the distance field described above, we developed LFM-Pro [2] that can identify common motifs from a family of similar proteins. We then developed a novel EPO (Enhanced Partial Order) algorithm, that aims at detecting low level of similarities among multiple curves (protein backbones). By combining a clustering idea, a novel two-level scoring function, together with a merging post-process, the EPO algorithm significantly improved the sensitivity of the previous multiple curve alignment algorithms in detecting local common subcurves [3, 7]. (The development of EPO also leverages our theoretical understanding on curve matching [8].) We have recently applied the EPO algorithm to protein structures, leading to our software Smolign [12], and obtained results significantly better than those returned by previous popular methods (such as MultiProt, Mustang, and Posa).

Features from Multiple functions. Given that there may be multiple descriptor functions defined over an input molecular surface, it is desirable to identify features induced by several descriptors, instead of by a single one. Together with researchers in the Lawrence Livermore National Lab (LLNL) and one of the PI's Ph.D students, we investigated an approach of measuring the correlations between multiple functions based on the overlay of the so-called Morse-Smale (MS) complexes. Roughly speaking, given a function on the input surface M , its MS complex decompose M into regions within each of which this function is monotone. Given multiple functions, the overlay of their MS complexes partition M into meaningful regions such that within each region, every function is monotone. One can then compute the (e.g, statistical) correlation for input functions within each such patch. We built a hierarchical representation of the overlay of MS complexes, so that the correlation of input functions can be studied at different resolution. Unfortunately, after we investigated the segmentation resulted from such overlap of multiple MS complexes, we were not able to observe characterization of molecular binding sites that have meaningful biophysical interpretation.

Visualization and software development. The project produced insights on visualizing biomolecular surfaces [6]. We have also started to investigate building visualization tool for a collection of molecular structures. In particular, together with one of the PI's Ph.D students, we used the topological landscape idea, originally developed in LLNL and LBNL by Weber, Bremer and Pascucci in their IEEE Vis'07 paper, and investigated how to use such a terrain metaphor for visualizing molecular energy landscapes [13].

Finally, there are several software developed in the project as described in Section 2.