BioGeometry NEWS
October 2004
biogeometry.cs.duke.edu

Research

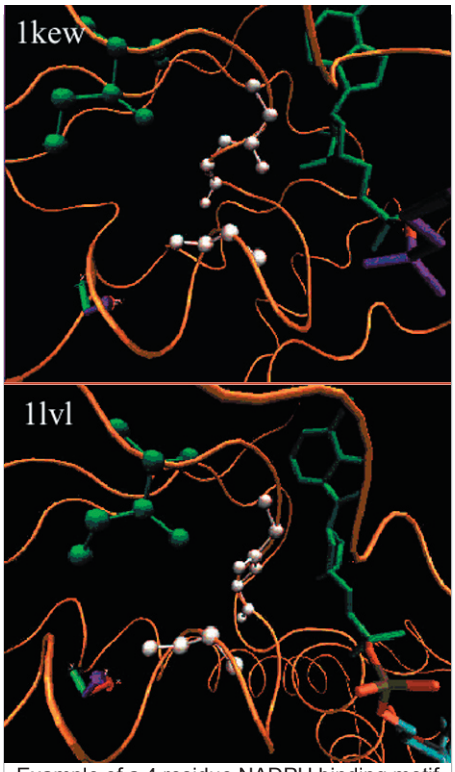
Mining spatial motifs in SCOP families by Luke Huan and Deepak Bandyopadhyay

Biogeometry group students Luke Huan (supervised by data mining specialist Wei Wang) and Deepak Bandyopadhyay (supervised by Jack Snoeyink) continue to work together on applying subgraph mining techniques to geometric graphs of proteins to find spatial motifs.

In work with faculty Wei Wang, Jans Prins, Alex Tropsha, and Jack Snoeyink, they tested different graph representations of protein structures, including Delaunay tessellation graphs, contact maps, and chose almost-Delaunay graphs as being robust in the presence of coordinate perturbations. They developed a fast method to generate maximal frequent subgraphs directly using a tree representation, to mine larger structures and databases significantly faster than exhaustive enumeration of all subgraphs by depth-first search. By finding family fingerprints, which are frequent subgraphs within a protein family that are rare outside the family, they are able to classify proteins into families based on structure motifs. (e.g., a two-class classification with over 90% accuracy between the SCOP prokaryotic and eukaryotic serine protease families.)

Their current work derived relationships between two families from a comparison of their family fingerprints, and applies it to find remote structural similarities between protein families from SCOP. The idea is to consider the appearance of spatial motifs (recurring residue-packing patterns) in protein families that have different folds and low sequence similarity. This is an important addition to the two prevelant approaches to classification by global fold structure (e.g. CATH) or sequence similarity, since the same local structure can be present in different folds or in proteins with little sequence similarity.

As a test on a known example, in the SCOP database there are two groups (superfamilies) of NADPH binding proteins: FAD/NAD(P)-binding domain (SCOPID: 51905) and NAD(P)-binding Rossmann-fold domains (SCOPID: 51735), between which no sequence and fold similarity are shared. After finding up to 1,045 family fingerprints for the first family, they found 10 that are enriched in another family, mostly in the Rossmann-fold domain. The figure shows the first of these; others are similar. The remarkable local commonality between the two remote SCOP families and the clear interaction between the motif and the ligand gives confidence that the method can reveal hidden yet biological significant patterns by comparing distinct protein families.



Example of a 4 residue NADPH binding motif that is significantly enriched in two SCOP superfamilies, drawn in white in proteins 1kew (chain a), and1lvl. The DALI z-score of the two protein structures is 4.5 and the pairwise sequence identity is 16%. The USC local alignment server reveals no sequence similarity in the region of the spatial motif.

The method is unsupervised, making possible the detection of unexpected patterns in protein families that could be as powerful and discriminating as patterns from known functional sites used by others. The only supervision is in the choice to use SCOP families as they share overall structure, have been well studied and characterized, and are usually known to be functionally and evolutionarily related. No information from the NADPH molecule was included in the search and the result that they obtain such a motif is due to their strong structural conservation among proteins in a SCOP superfamily. A publication is in preparation for submission to RECOMB.

People

Homme Hellinga, Professor of Biochemistry and Biophysics at Duke and a PI on the BioGeometry project, was awarded the Foresight Institute Feynman Prize this month in recognition of his leadership in nanotechnology research. According to the Institute, Hellinga was recognized with the experimental award, "for his achievement in the engineering of atomically precise devices capable of precise manipulation of other molecular structures. Building on a broad base of achievement in computationally directed protein engineering, he has extended this work to the construction of an enzyme. This achievement demonstrates an innovative blend of techniques, applying computational design to reengineer a structure found in nature into a novel one with a different function. This work breaks new ground in engineering devices that transform molecular structures."

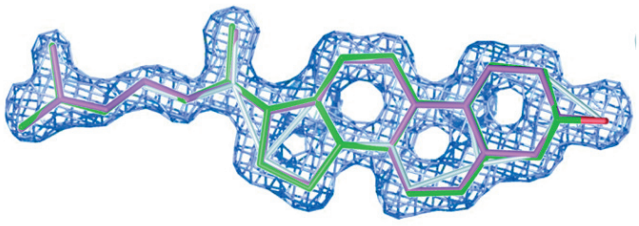
Foresight Institute is a leading think tank and public interest organization focused on nanotechnology. Foresight Feynman prizes are given to researchers whose recent work has advanced the construction of atomically-precise products through the use of molecular machine systems.

BioGeometry collaborator **Brian Kuhlman**, Assistant Professor of Biochemistry and Biophysics at UNC, was awarded the Feynman Prize in the theory category along with David Baker of the University of Washington for their work in the development of RosettaDesign, a software program that has had a high success rate in designing stable protein structures.

Student Profile: Daniel Russel

Daniel came to Stanford after completing his B.S. in Computer Science at Princeton in 1998 and immediately became interested in geometric computation and its potential impact in structural molecular biology. Working with his advisor, Leonidas Guibas, he has been exploring a variety of problems related to the structure and motion of biological macromolecules, and developing efficient geometric algorithms motivated by these problems.

Soon after his arrival, in a collaboration with Prof. Axel Brunger's Lab, Daniel developed techniques for fitting molecular structures into electron density maps by exploiting recent geometric developments allowing the efficient extraction of the medial axis of a surface. The surface in question here is an isosurface of the density map of a protein, or a protein-ligand complex, obtained using X-ray diffraction crystallography. The instable nature of the medial axis and the noise in the experimental data lead to medial axes that need significant pruning and cleaning before they can be useful. Nevertheless this approach was used successfully to fit highly flexible ligands into known protein structures and will be incorporated into the well-known CNS software package (Crystallography and NMR System). An example output from the system is shown below.



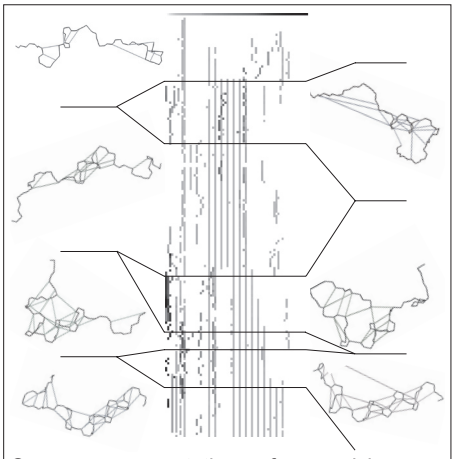
Automatic extraction of ligand structure data from x-ray data.

The figure shows the simplified and cleaned medial axis (gray), the structure selected by the software (pink), and the manually fit structure (green) for a cholesterol ligand. Structure fitting starts by matching the medial axis against the known bond structure of the ligand. The result provides a number of estimates of ligand orientation. These estimates are locally refined to better match the known electron density and energy constraints, and then the resulting matches are clustered and ranked by how well they match the electron density.

In the summer of 2002, Daniel interned in a small biotech startup, Protein Mechanics, where he worked on algorithms for computing the solvent accessible area of protein molecules. Computation of this area and its derivatives are fundamental tools in implicit solvent computations -- molecular dynamics simulations that avoid the explicit representation of thousands of water molecules with their associated costs. Delaunay triangulations of point sets, and their generalizations to balls (power diagrams) play a fundamental role in this computation. Motivated by this experience, Daniel studied and compared a variety of algorithms for updating Delaunay triangulations efficiently, as the underlying elements (points or balls) move. This was done within the kinetic data structures framework as well as in certain variations that better model motion in physical simulations. This led to a paper at the 2004 ACM Symposium on Computational Geometry (SoCG).

While coding these algorithms, Daniel also developed a general programming framework for implementing kinetic data structures in a software library. This framework is currently being incorporated into the European CGAL Geometric Algorithms Library, as a basic tool for dealing with geometric objects in motion.

Most recently, Daniel became interested in studying and comparing macromolecules in motion, such as protein folding trajectories. In work that will be presented in the Pacific Symposium on Biocomputing in 2005, Daniel exploited the idea of geometric spanners as a tool for visualizing and comparing molecular trajectories. A spanner compactly captures the key proximitites in a deforming molecule and allows us to abstract and "combinatorialize" the continuous deformation into a set of discrete events that capture the significant changes in the conformation of the molecule during the motion. The spanner history can be visualized as a 2D image, further aiding the intuitve understanding of the motion. An example is shown below.



Spanner representations of an evolving protein structure

The figure illustrates a spanner-based technique for visualizing protein trajectories. Each row of the figure is one frame from the trajectory as the protein (BBA5) folds from an extended state to its native state. The shade of gray at the i-th square from left to right in a row is the length of the longest spanner edge originating at the i-th backbone atom. Secondary structure shows up as regular patterns of edges. Tertiary structure can be seen as longer range (darker color) spanner edges.

- Profile by Leonidas Guibas

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.