Modeling probabilistic data into a flexible surface

Motivation for this problem: Interaction of two proteins depends on shape complimentarity of their surface (as well as many other chemical factor including electrostatics, hydrogen bonding, disulphide bonding etc). In proteins the surface is actually flexible which calls for schemes to model flexible surfaces. In proteins, the flexibility arises from discrete rotations about certain bonds leading to rotamers.

Abstraction of this problem: A protein is usually represented by a set of spheres corresponding to each atom. Surface of a protein can be found by purely geometrical means from those spheres. One popular way is to the volume sweeped by a probe sphere which can touch the atom spheres but not penetrate into them. Boundary of this sweeped volume can be thought of as the suraface of protein which does depend on the radius of the probe sphere.

This simple rigid model of protein and its surface can be easily extended to a flexible one. In stead of having a unique center for an atom sphere, we have a discrete probability distribution of centers. In general, the events that a particular atom is at a particular location are not independent of each other. We abstract it by assuming that only a subset of the atoms are flexible. This is supported by the biological observation that sidechains are much more flexible than the backbone of a protein. We also assume that, the locations of atoms in a sidechain are influenced by only the backbone which is fixed and not by other sidechain. This is only a simplifying assumption under which, we completely specify the probability distribution of the protein by the joint probability distribution of each sidechains.

Our simplified problem now is as follows. Given discrete probability distribution of the centers of a set of spheres with known radii, can we find out the probability distribution of its surface for a given probe radius? We are looking for a representation of a probabilistic surface which will assign a real number at each point on the surface represting its flexibility at that point. The surface may be defined as a mesh of triangles and we may look for flexibility only at verticies of those triangles. We may assume that the flexibility can be interpolated linearly in a mesh triangle. We may also require that three flexibility values at three verticies of a triangle should not differ more than a prescribed value and refining the mesh (possibly locally) adaptively to ensure this.

Direction of local changes in a surface is normal to the surface, so this also calls for normal estimation in the mesh. As our ongoing research suggests, the estimated normals should be continuous and if we take any two orthogonal directions on the surface, the partial deriavatives in those direction of the components should be compatible.

Validation of solution: To validate the usefulness of our abstraction we can use it in predicting protein interaction. We plan to validate our algorithm to compute the flexibibity mesh by doing error analysis as used in numerical methods.

A Simple example: Let us look at one dimensition lower, so to model flexibility of a curve. Let us assume the curve is defined by a real function defined on [0,1]. We have probability distribution of its value at 0 and 1. Let us assume linear interpolation is used to construact the function. It can be shown that, if the distributions at end points are independant, a representative curve is simply the line joining mean values at the end points. We tacitly assume a distance function among curves, namely the L2-norm between the underlying functions on [0,1]. Under this distance assumption, the representative curve still remains a straight line when the end values are dependant on each other and we have a joint probability distribution.

As we know, a better distance function between curves is Hausdorff distance and even better, Fretchet distance. Finding the representative curve under these metrics becomes interesting. As we recall, a representative curve is one from which the sum of squared distances to other probable curves weighted by their probabilities is minimized. In L2-norm, this is same as the curve obtained from mean function.

Issues: The distribution of centers were computed by looking at the protein structures present in protein data bank (PDB) at that time. Before doing a validation of our assumptions, we must compute theoretical confidence to make sure enough data were present at that time.