Characterization and Classification of Local Protein Surfaces Using Self-Organizing Map

Lee Sael, Purdue University, USA
Daisuke Kihara, University, USA

ABSTRACT
Annotating protein structures is an urgent task as increasing number of protein structures of unknown function is being solved. To achieve this goal, it is critical to establish computational methods for characterizing and classifying protein local structures. The authors analyzed the similarity of local surface patches from 609 representative proteins considering shape and the electrostatic potential, which are represented by the 3D Zernike descriptors. Classification of local patches is done with the emergent self-organizing map (ESOM). They mapped patches at ligand binding-sites to investigate how they distribute and cluster among the ESOM map. They obtained 30-50 clusters of local surfaces of different characteristics, which will be useful for annotating surface of proteins.

Keywords: 3D Zernike Descriptors, Affinity Propagation Clustering, Local Protein Surface Matching, Physicochemical Properties, Protein Active Site Self-Organizing Map, Spherical Harmonics

INTRODUCTION
The importance and the impact of computational work in biology is ever increasing since large scale biological data of various types were accumulated in the past decade, which include genome sequences, protein-protein interaction networks, metabolomes, genome-scale transcriptions, and gene expression patterns. Those data contain key information for understanding orchestrated behavior of molecules in biological systems that is essential to sustain life. It is expected that bioinformatics play significant roles in analyzing such data, as computational techniques, e.g. clustering, feature characterization, data mining, and modeling, are indispensable in the analyses.

Of a particularly important and interesting problem is function prediction of proteins from the tertiary structures, as the structural genomics projects (Burley, 2000; Westbrook et al., 2003; Zhang & Kim, 2003) have been solving an increasing number of protein structures of un-
known function. Indeed, as of May 2009, there are over 2800 proteins of unknown function in the Protein Structure Databank (PDB) (Berman et al., 2000). These proteins are remained as unknown function because so far no one has yet conducted experiments to characterize their function, and moreover, conventional sequence comparison based methods (Hawkins & Kihara, 2007), e.g. homology searches (Altschul et al., 1990; Altschul et al., 1997; Pearson & Lipman, 1988), functional motif (Hulo et al., 2006), and domain searches (Coggill et al., 2008), did not find significant similarity against protein sequences of known function. Ongoing efforts for better function prediction include development of sequence-based methods which are more sensitive and accurate than the conventional methods (Chitale et al., 2009a; Hawkins et al., 2008). For example, our group has recently developed two sequence-based methods, named the automated Protein Function Prediction (PFP) method (Hawkins et al., 2006; Hawkins et al., 2009) and the Extended Similarity Group (ESG) method (Chitale et al., 2009b), which efficiently and accurately mine function information from PSI-BLAST searches.

Alternatively, one can use the tertiary structure information for capturing similarity to proteins with known function that are stored in PDB (Thornton et al., 2000). Potential advantages of using structure information are two folds: Firstly, evolutionarily more distantly related proteins to a query protein could be identified because the global structure is better conserved than the primary sequence (Chothia & Lesk, 1986; Kihara & Skolnick, 2004). Secondly, physical features of functional local sites of proteins can be directly compared where interactions with ligand molecules or other proteins take place (Laskowski et al., 2005). A number of methods have been proposed which use local structure as a key feature for predicting function of proteins. Since small ligand molecules usually bind to a protein at its surface pocket regions, simply identifying pockets in the protein surface can identify active sites of enzyme in most of the cases (Li et al., 2007; Laskowski et al., 1996; Kawabata & Go, 2007). Programs which identify pockets include Visgrid (Li et al., 2007), POCKET (Levitt & Banaszak, 1992), LIGSITE (Hendlich et al., 1997; Huang & Schroeder, 2006), SURFNET (Laskowski, 1995), and PocketDepth (Kalidas & Chandra, 2008). An identified pockets can be further compared with known ligand binding pockets in a database to make prediction of the type of ligand that binds to it (Kahraman et al., 2007; Tseng et al., 2009; Kihara et al., 2009; Chikhi & Kihara, 2009; Binkowski & Joachimiak, 2008; Kalidas & Chandra, 2008; Yeturu & Chandra, 2008). In these methods, pockets are characterized by geometrical shapes, amino acid residues at pockets, and physicochemical properties such as the electrostatic potentials and hydrophobicity.

A fundamental limitation of these methods is that they only deal with pocket regions, i.e. prediction of ligand binding at pockets in protein surface. Although binding small ligand molecules is a very important class of protein function, it only occurs in a subset of proteins, mainly enzymes and proteins that use co-factors. Moreover, on average pocket regions in an enzyme share only about 5% of the entire surface of the protein (Li et al., 2007). Hence, many of proteins in a genome and most of surface region of proteins are left out from applicability of these methods. Ideally, methods should be more generalized so that they can describe and compare any protein surface regions and annotate local regions with its potential functions.

Toward this goal, here, we classified local surface regions of proteins to see how diverse local surface patches are and how many different types of surface patches exist. We characterized a surface patch by two features, geometric shape and the electrostatic potential. Among shape descriptors that have been studied (Sael et al., 2009; Tangelder & Veltkamp, 2008), we have chosen the 3D Zernike descriptors, a projection-based approach, because they have already been shown to be very effective in describing the shape and other physicochemical properties of proteins (Kihara et al., 2009; Sael et al., 2008b; Sael et al., 2008a).
Research, Leadership, and Resource-Sharing Initiatives: The Role of Local Library Consortia in Access to Medical Information
Reysa Alenzuela (2017). *Library and Information Services for Bioinformatics Education and Research* (pp. 199-211).
[www.igi-global.com/chapter/research-leadership-and-resource-sharing-initiatives/176145?camid=4v1a](www.igi-global.com/chapter/research-leadership-and-resource-sharing-initiatives/176145?camid=4v1a)

Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain
[www.igi-global.com/article/hierarchical-density-based-clustering-white/49547?camid=4v1a](www.igi-global.com/article/hierarchical-density-based-clustering-white/49547?camid=4v1a)