Development of the Property Encoded Shape Distributions and their application to protein binding site comparison and protein-ligand binding affinity prediction

by

Sourav Das

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

DOCTOR OF PHILOSOPHY

Major Subject: Chemistry

The original of the complete thesis is on file
In the Rensselaer Polytechnic Institute Library

Examining Committee:

Prof. Curt Breneman, Thesis Adviser

Dr. Dominic Ryan, Member

Prof. Mark Wentland, Member

Prof. Steven Cramer, Member

Prof. Wilfredo Colon, Member

Rensselaer Polytechnic Institute
Troy, New York

March, 2010
(For Graduation May, 2010)

ABSTRACT

Patterns in shape and property distributions on the surface of binding sites are often conserved across functional proteins without significant conservation of the underlying amino-acid residues. To explore similarities of these sites in terms of the physico-chemical environment experienced by a ligand, a sequence and fold-independent method was created to rapidly and accurately compare binding sites. Within this paradigm, signatures for property-mapped Gauss-Connolly surfaces of binding sites were generated by calculating their Property-Encoded Shape Distributions (PESD). PESD represent the probability that a particular property will be at a specific distance to another on the molecular surface. Similarity between the signatures can then be treated as a measure of similarity between binding sites. As postulated, the PESD method rapidly detected high level of similarity in binding site surface characteristics even in cases where there was very low similarity at the sequence level. In a screening experiment involving each member of the PDBBind 2005 data set as a query against the rest of the set, the PESD method was able to retrieve a binding site with identical Enzyme Commission numbers as the top match in 79.5% of cases. The ability of the method in detecting similarity in binding sites with low sequence conservation was compared to state-of-the-art binding site comparison methods. The method was further validated on a diverse set of non-redundant proteins and on a set of kinases. A server applying the PESD method for rapid comparison of all ligand-bound sites in the Protein Data Bank (PDB) was also implemented.

The PESD algorithm was extended to binding affinity prediction: PESD signatures together with standard support vector machine (SVM) techniques were used to produce validated models that could predict the binding affinity of a large number of protein ligand complexes. This “PESD-SVM” method with no subjective feature selection had performance comparable to the scoring function, SFCscore. SFCscore was previously shown to perform better than 14 other scoring functions. For most complexes with a dominant enthalpic contribution to binding, a good correlation between true and PESD-SVM predicted affinities was observed.