Name: Yumlembam Hemajit Singh
Supervisor: Prof. Seemi Farhat Basir
Co-supervisor: Dr. Shandar Ahmad
Department of Biosciences

Title: "Solvent Accessibility Studies on Globular and Transmembrane Proteins"

Abstract

Solvent Accessibility or Accessible surface area (ASA) is the degree to which an amino acid residue is exposed to the solvent. The applications of solvent accessibility has been used for identification of active site residues, DNA binding sites, functional residue in membrane proteins and modeling side chain conformations in proteins. Thus, knowing the solvent accessibility of amino acid residue of proteins will also assist us in understanding many biological processes. In this work, we explore a novel method to broaden the scope of sequence-based predictions of solvent accessibility or accessible surface area (ASA) to the atomic level. An analysis of ASA distribution of all 167 heavy atoms from the 20 types of amino acid residues in different proteins has been performed and rotamer-style libraries have been developed. We observe that the ASA of some atomic groups (e.g., backbone C and N atoms) can be estimated from the sequence environment within a mean absolute error of 2–3Å2. However, some side chain atoms such as CG in Pro, NH1 in Arg and NE2 in Gln show a strong variability making it more difficult to estimate their ASA from sequence environment. In general, the prediction of ASA becomes more difficult for atomic positions at the side chain extremities of long amino acid residues (aromatic side chain terminals being the exception). Several atomic groups are frequently exposed to solvent. Some of them have a bimodal distribution, suggesting two stable conformations in terms of their solvent exposure. Moreover, we also explored a novel method to determine the role of normalization in prediction of solvent accessibility (ASA) of amino acid residues. Typically normalization is attained by assuming the highest exposure state based on extended state of that residue when it is surrounded by Ala or Gly on both sides i.e. Ala-X-Ala or Gly-X-Gly solvent exposed area. However, the sequence context, the folding state of the residues, and the actual environment of a folded protein impose additional constraints on the highest *possible* (or highest *observed*) values of

ASA, which are currently ignored. Here, we analyze the statistics of such observations and examine how renormalizing absolute ASA values of residues using context-dependent Highest Observed ASA (HOA) instead of context-free extended state ASA (ESA) of residues can influence the performance of sequence-based prediction of solvent accessibility. Neural networks were retrained to predict both relative ASA from evolutionary information. We found that neural networks trained with HOA-normalized data do outperform ones trained with ESA-normalized values. Furthermore, a novel method for the prediction of Lipid solvent accessibility (LSA) of amino acid residues in transmembrane (α-helix and β-barrel) proteins using support vector regression approach was explored. We could predict LSA of transmembrane α-helix with ~ 13 % MAE and 0.68 correlations using evolutionary information. And for the transmembrane β-barrel, we could predict LSA with ~ 11 % and 0.69 correlations using evolutionary information.