

Predicting residue contact numbers in proteins using support vector regression.

Mukta Phatak, University of Cincinnati

Predicting 3D structure of a protein from its primary amino acid sequence remains one of the main challenges in computational biology. One important intermediate step towards that bigger goal is the prediction of residue contact numbers. Residue contact information may help deciphering protein structure by elucidating how residues are arranged in 3D.

Here, two residues are said to be in contact if the geometric centers of their side chains are closer than a specific cutoff distance. Contact number for a given residue of a protein is defined by the number of residues within the sphere around the geometric center of a side chain of that residue.

The prediction of residue contact number can be cast as a 2 state classification problem where residues are considered buried or exposed depending upon the number of contacts. However, this requires choosing an arbitrary threshold and adds additional uncertainty to the model. In this work we develop a novel method for the prediction of the residue contact number from a protein sequence using support vector regression (SVR). In particular we propose a novel representation that utilizes predicted relative solvent accessibility (RSA) of an amino acid residue in a protein. RSA is an inverse measure with respect to residue contact number. We compare the new representation to multiple sequence alignment (MSA) based representation using cross validation on representative non redundant set of 425 protein chains and 88260 protein residues. The MSA based predictor yields correlation coefficient of 0.58 between predicted and actual contact numbers. On the other hand, when using the predicted RSA and secondary structure (obtained from SABLE) a significantly improved accuracy is achieved with correlation coefficient of 0.63

The results suggest that RSA prediction can be used to enhance prediction of contacts between residues. Furthermore SVR provides flexible approach to the prediction of the residue contact number that allows one to model expected levels of error for different types of residues.