Poster I-43

Novel Protein Structural Classification by Discovering Coherence in Hydrophobicity Scales

Authors:
Sumeet Dua (*College of Engineering and Science, Louisiana Tech University*)
Pradeep Chowriappa (*College of Engineering and Science, Louisiana Tech University*)
Ramakrishnan Rajagopalan (*College of Engineering and Science, Louisiana Tech University*)

Short Abstract: In this work, we present a novel feature extraction and classification computational paradigm that discovers the points of spectral coherence among different hydrophobicity scales for proteins and employ them for structural class identification of proteins, yielding a significant increase in prediction accuracy over previous results in this area.

Long Abstract:
Background: The endeavor to decipher the structure and function of a protein from its amino acid (AA) sequence has provided an enduringly interesting challenge. Due to the sheer quantity of the existing protein data, this challenge naturally presents itself as a complex computational problem requiring the deployment of novel data mining techniques. Most of the existing research in this area [1,2] performs structural classification using the following six physical and streochemical properties: AA composition, predicted secondary structure (PSS), hydrophobicity, normalized van der Waals volume, polarity, and polarizability. Feature vectors generated from these property values are used to predict the secondary class by constructing a machine-learning classifier. Evidence demonstrates that each one of these parameters makes a significant toward predicting the fold-class of a protein.

Motivation: It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment is reflected in the sequence and chemical properties of the residues in the protein [3]. Consequently, several qualitative, algorithmic, and quantitative techniques have been introduced to model and detect the periodic variation of chemical properties along the protein sequence that establish characteristic secondary structural features. Hydrophobicity and hydrophilicity are two such incontrovertible physico-chemical properties. Their usefulness arises from the natural and predictable differences in inter-molecular forces between the individual amino acids and water (or any other medium). Thus, the concept of hydrophobicity provides a clearer understanding of how amino acids interact within proteins as well as providing a basis upon which to predict the structural properties of proteins from sequence information. It follows that the most hydrophobic sequences in a protein will be found in the interior of the native structure while the most hydrophilic sequences will be found on the exterior [4]. Since the hydropathic character of the amino acids is one of the primary factors involved in protein folding, which in turn determines the conformation of a protein, structure can be directly correlated with hydropathy. Therefore, it is presumed that hydropathy provides a strong, valid descriptor for structural classification and prediction. In order to exploit this prediction with greatest success, the most accurate evaluations of the hydrophobicity and hydrophilicity of each

amino acid side-chain should be formulated [4]. Both hydrophobicity and hydrophilicity should be considered for predicting structural class. Failure to consider both properties will neglect half of the potential information they provide [4].

The Data set: The dataset we employed is that used in the study conducted by Aik Choon Tan et al. [2], available publicly from http://www.nersc.gov/~cding/protein/. The original dataset consists of independent training and testing protein sets. The training set was extracted from PDB-select and consisted of 408 proteins distributed across 25 fold classes. These proteins were randomly chosen from SCOP 1.61 and Astral 1.61 with an overall sequence similarity of less than 40%. Likewise, the testing set consisted of 174 randomly chosen proteins. This resulted in a total dataset of 582 proteins from 25 different fold classes and 5 structural classes, each of variable sizes.

Methodology: In this work, we propose a novel methodology for computing the feature vector using only one physico-chemical property—hydrophobicity—thereby producing appreciable improve-ments in the accuracy of secondary structural class identification. The information from seven different hydrophobicity scales are explored; their features are determined by orthonormal transformation and principal component analysis. Next, points of spectral coherences among the hydrophobicity scales are discovered and assembled to produce the resultant coherence vector characterized by principal components and orthonormal coefficients. These features are subjected to a random tree classifier to achieve multi-class classification. Similar classification is performed using support vector machine, and the performances of each method are compared to evaluate the strength of the proposed feature vectors and algorithms in terms of true-positives and false-negatives for individual structural classes. In another set of experiments, we append the feature vectors for other stereochemical properties (as mentioned above), to the computed hydrophobicity features and study the change in specificity and sensitivity of classification on an incremental basis. In summary, extensive experimental results are provided to demonstrate that the treatment of hydrophobicity in previous literature has suffered from inadequate feature representation of hydrophobicity scales. In our approach, points of spectral coherence are discovered among different hydrophobicity scales and exploited for multi-class classification using random forest and support vector machine-based classification.

Results: Using only the principal components of coherence derived from different hydro-phobicity scales, we obtained a peak accuracy of 65.51% on a dataset of 582 proteins spanning five structural classes with less than 40% overall sequence similarity. Previous research achieved prediction accuracy of only 36.5% for the same dataset (again, using hy-drophobicity alone. Furthermore, by including two other parameters (AA composition, PSS) we realized a classification accuracy of 83.33%. By comparison, the best result prior to our work provides a peak accuracy of 74.2% when including all six of the above size parameters in the feature vector [2].

Conclusion: Although the elucidation of the absolute contribution of hydrophobicity to protein folding is far from complete, we believe that these extensive empirical results will open novel directions for scientific presentation and debate.