DOI : https://doi.org/10.56452/7-2-556

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

An Hybrid of Ensemble Model and Feature Extraction Algorithm for Classification of Protein Structural Class Prediction

Kiran Kumain

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

The structural class of the protein reveals crucial details regarding the general folding type. Understanding its structure is essential to comprehending how the protein works. Because there aren't any better terminology, structural information is typically categorized as either secondary, tertiary, or quaternary structure. Such projections still need a workable, all-encompassing answer. The majority of the work done so far has been focused on heuristics that are generally effective. Many of the improvements are based on pattern recognition techniques. The current method uses feature extraction that is both sequence- and physicochemical-based. The suggested system uses a quick correlation-based filter technique for feature extraction. The categorization is then done using the consensus characteristics. To demonstrate the efficacy of protein structure class prediction, various classifiers are utilized.

1. INTRODUCTION

The research into biological, behavioural, and social systems is done through the creation and implementation of computational simulation tools, mathematical modelling, and data-analytical and theoretical methodologies. Applied mathematics, animation, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, ecology, evolution, anatomy, neurology, also visualisation are all included in its broad definition, which also encompasses computer science underpinnings. The field of bioinformatics involves using biological data to create relationships and algorithms between diverse biological systems. The human genome has been sequenced with the aid of computational biology, and the human brain has been accurately modelled in addition to other biological systems. A branch of computational biology called evolutionary computation develops algorithms predicated upon hypotheses of interspecies evolution.

Computational genomics, which aspires to create a collection of data from the sequencing of the entire human genome, includes the Human Genome Project as one example. This creates the opportunity for customised medicine, which would include treating patients according to their unique genetic profiles. The genomes of all other forms of life, including bacteria, plants, and animals, are also being sequenced by researchers. The primate genome may need to be included in homology, a tool for comparing the genomes of related species and mRNA sequences, in order to improve cutting-edge technology in personalised gene therapy. The field of computational neuroscience attempts to model the brain so as to explore particular sorts of elements of the neurological system. It studies how the brain functions in terms of the information processing capabilities of the neural structures.

Realistic brain models are one of the many sorts of models of the brain; these models have the highest margin of error but also the most information about the brain. The implementation that requires the greatest computation power and money are realistic brain models.

Computational neuroscientists are working on Simplifying Brain Models to enhance the techniques and data structures now employed to speed up such computations. Computational pharmacology is the study of how genetic data may be used to evaluate pharmacological data and identify associations between particular genotypes and disorders. Computational pharmacology is required because the pharmaceutical business has hit the so-called Excel barrier. Computational techniques are created by scientists and researchers to examine these enormous data sets and enable the creation of more precise pharmaceuticals. Due to the need for more trained analysts of the massive data sets needed for developing new pharmaceuticals, leading pharmaceutical companies are urging doctoral candidates in computational biology to consider employment as opposed to post-doctoral positions, in industries.

In many respects, computational biology has helped the science of evolutionary biology, such as by evaluating the evolution of species across time using DNA data. Cancer computational biology is a topic that uses an algorithmic method of data analysis to predict potential cancer mutations in the future. The multidisciplinary area of bioinformatics creates and enhances techniques for conserving, retrieving, organising, and interpreting biological data. Artificial intelligence, soft computing, data mining, image processing, and simulation algorithms all rely on theoretical underpinnings like discrete mathematics, control theory, system theory, information theory, and statistics. More quickly than previously, complex machines are being utilised to read in biological data. The word "bioinformatics" was created by Paulien Hogeweg to describe the study of information processing in biotic systems.

Protein microarrays with high throughput mass spectrometry (HT-MS) can give a quick overview of proteins found in biological samples. Making sense of protein microarray and high-throughput mass spectrometry data requires bioinformatics, this involves doing extensive statistical analyses on sample data where many, incomplete peptides from each protein are found, comparing enormous amounts of mass data against anticipated masses from protein sequence databases, and so on. Massive sequencing projects are utilised to find point mutations in several genes that were previously unknown in cancer. Novel physical detection techniques are used, such as single-nucleotide polymorphism arrays for detecting known point mutations and oligonucleotide microarrays for identifying chromosomal gains and losses.

When employed at high-throughput to analyse thousands of samples, these detection techniques simultaneously assess hundreds of thousands of sites across the genome, producing terabytes of data for each experiment.

In order to infer actual copy number changes, approaches like Change-point analysis along with the Hidden Markov model are being designed since it is frequently discovered that the data contains a great deal of unpredictability or noise. Analysis of lesions discovered to be recurring among numerous cancers is another sort of data that necessitates the creation of innovative informatics. Establishing the correlation between genes and other genomic properties in various species is the goal of comparative genome analysis. It helps to identify the evolutionary pathways that led to the divergence of two genomes.

The parallels between human haemoglobin and that found in legumes are the most crucial information in this text (leghemoglobin). These proteins have almost similar protein structures and fulfil the same function in the organism—transporting oxygen. To explore interactions between proteins, ligands, and peptides, such as protein-ligand, protein-protein, and protein-peptide interactions, computational techniques are utilised. In the past two decades, X-ray crystallography and protein nuclear magnetic resonance spectroscopy have helped identify the three-dimensional structures of tens of thousands of proteins (protein NMR). Is it practicable to make predictions about potential protein-protein interactions only based on these 3D forms, rather than doing protein-protein interaction experiments?

Although several strategies have been established to address this issue, more effort remains. For a wide range of bioinformatics applications, SOAP- and REST-based interfaces have been created, enabling a programme operating on a computer in one region of the world to access algorithms, information, and computational resources on servers in other regions of the world. A particular type of workflow

management system called a bioinformatics workflow management system is created to construct and carry out a workflow, which is a set of computational or data manipulation activities, in a bioinformatics application. They offer application scientists need a user-friendly environment to create their own workflows, and scientists need interactive tools to carry them out as well as examine findings in real-time, and the ability for scientists to share and reuse workflows.

2. LITERATURE SURVEY

This study examines the viability of structural class prediction/assignment in accordance with one-dimensional secondary structure, where each residue is assigned a secondary structure without understanding its spatial layout. The author also examines the viability of automating the assignment of classes based on the 1D secondary structure using early definitions of structural classes that were created prior to the creation of the SCOP database. The assigned (based on the tertiary structure) and predicted (from the protein sequence) 1D secondary structures are used to achieve the aforementioned objectives. SSAsc (secondary structure-based assignment of structural classes) and PSSAsc are two novel assignment models that we suggest (predicted secondary structure-based assignment of structural classes). After that, these two models were contrasted with current assignment techniques (objective 2) to see if the structural class could be correctly allocated using 1D secondary structure predicted from the protein sequence. Without having any prior knowledge of the structure, such an assignment model would provide a workable method for automating the assignment of structural classes. The aforementioned results were then evaluated against typical techniques that predict structural classes based on protein sequences [1].

The identification of protein structural class is a very important issue in protein research since it may provide vital information about a protein's overall structure if known a priori. Yet, it is time-consuming and expensive to do so based only on experimental procedures given the fast growth in newly discovered protein sequences entering databanks. Consequently, it is crucial to create a computer approach for quickly and reliably identifying the protein structure class. This research offers a dual-layer support vector machine (SVM) fusion network that is characterized by adopting a separate pseudo-amino acid composition to address the problem (PseAA). The PseAA in this instance provides a wealth of data pertaining to a protein's sequence order as well as how many hydrophobic amino acids are found in its chain. For the two benchmark data sets, the exacting jackknife cross-validation test was conducted that Zhou created as a demonstration. The observed considerable increase in success rates suggests that the proposed strategy may be an effective addition to other currently used techniques in this field [2].

Ada Boost has a bigger average impact than bagging, yet their operating characteristics are distinct. Although bagging tends to simply decrease the error term's volatility, Ada Boost often lowers the bias as well as variance parts of error. The prediction error of a committee is less affected by each new member than by any of its predecessors. Combining the two may have an even bigger impact since an ensemble's accuracy may be improved by raising the variance in the committee members' forecasts without impacting their individual error rates. The variety of the committee's membership should rise with the use of several methods for selecting its members, which should reduce prediction error [3].

When describing the all in all topological folding type of a protein or its domains, the structural class is a crucial component. Several researchers have concentrated their attention and efforts on the prediction of protein structure categorization. To address this issue, the Ada Boost Learner, an unique predictor, was presented in this study. The core idea behind the Ada Boost Learner is that a collection of several "weak" learning algorithms, each of which outperforms a random guessing algorithm by a little margin, will result in a "strong" learning algorithm. Ada Boost fared better than other predictors like SVM and has the ability to enhance the quality of other protein characteristics like sub-cellular location and receptor type. In this sense, it may function as a supplement to already-existing algorithms [4].

To solve this challenge, a conventional neural network model is used. The SCOP database, which is based on domains of known structure, evolutionary relationships, and the rules that control their 3-D structure, was used as the basis for the neural network approach used in this study. The explanation for why the data utilised in the present investigation are more logical is given. As a consequence, high jackknife test and self-consistency rates were attained. This suggests that a protein's structural class and

the makeup of its amino acids have a strong correlation, as well as the neural network approach can be a helpful method for forecasting a protein's structural class [5].

3. PROPOSED SYSTEM

The identification of protein structural class is a very significant subject in protein research since previous understanding of a protein's structural class may provide useful details regarding the protein's entire structure. Yet, relying only on experimental procedures would be time-consuming and expensive given the fast growth in newly discovered protein sequences that are being added to databanks. Hence, the creation of a rapid and accurate computational approach for identifying the protein structure class is crucial.

The feature extraction and classification are employed in the strategy we've suggested. Rapid Correlation Based Filter method is employed for feature extraction. These characteristics are precisely retrieved using this method. Every of the characteristics' symmetrical uncertainty is computed. The fcbf algorithm's efficacy and efficiency are demonstrated before the classifications are used. Fig. 1 depicts the suggested approach's overall system design. Following is a list of the several benefits of the suggested approach:

- The categorization uses the best attributes. The quick correlation-based filter method makes it feasible.
- The protein structure prediction made using this approach is also quite precise.

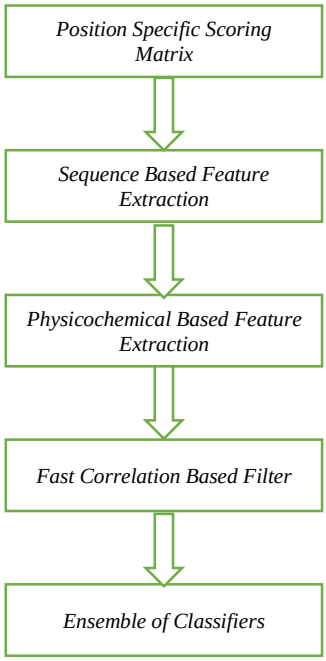

Fig 1: Flow Diagram

The next section provides an explanation of the many processes that are involved in putting the suggested technique into practice:

1. Position Specific Scoring Matrix

The dataset is transformed into score matrices that are position-specific. It comprises of the 20 amino acids and the protein's length. This position-specific scoring matrix is used to compute feature extraction. Both physicochemical feature extraction and sequence-based feature extraction are employed.

2. Sequence Based Feature Extraction

There are two kinds of sequence-based feature extraction. The first is the evolutionary-based composition feature group, which is derived from the frequency with which each amino acid occurs in a specific protein sequence. Second is Evolutionary-based Auto Covariance Feature Group: This feature group calculates the auto covariance of each amino acid substitution score along a protein sequence.

3. Physicochemical Based Feature Extraction

Consensus sequence is created from the initial protein sequence. The consensus sequence is used to extract physical and chemical characteristics. We employ Overlapping Segmented Autocorrelation and Overlapping Segmented Distribution Method. In accordance to both physicochemical-based and feature extraction techniques, the extracted feature groupings.

4. Fast Correlation Based Filter

The technique identifies a collection of dominating features Sbest for the class idea given N characteristics and a class C. Based on the predetermined threshold, the SU value for each feature picks pertinent characteristics into the S0 list and arranges them in decreasing order by SU value. Filtering away features those are inferior to fp and include fp with the rest of their redundant peers is always possible when fp has previously been identified to be a significant feature. Starting with the first entry in the S0 list, the iteration moves on as follows. If fp is a redundant peer to any of the remaining features, fq will be deleted from the S0 list.

5. *Ensemble of Classifiers*

Utilizing an ensemble of many classifiers works effectively in order to anticipate the structural class of proteins. AdaBoost. The many classifiers employed include M1, Logit-Boost, Naive Bayes, SVM, also MLP. They are employed to investigate the predicted protein structure class.

4. RESULTS

To predict protein structure class, various classifiers, consensus features, quick correlation based filter algorithms, pattern recognition, and feature extraction from sequences and physicochemical data are all employed. Accurate feature extraction and classification are achieved using the Fast Correlation Based Filter technique. The algorithm's efficiency and efficacy are demonstrated by the calculation of the symmetrical uncertainty. The following screenshots demonstrate how the best attributes are utilised for protein categorization and structure prediction.

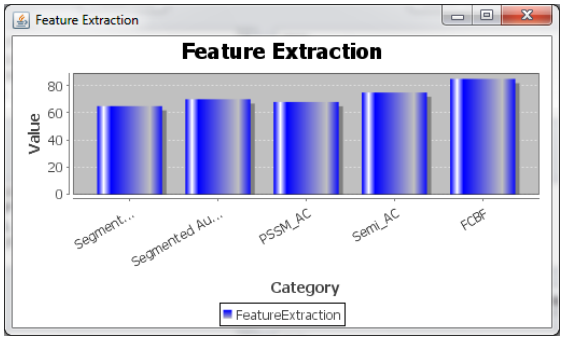

Fig 2: Feature Extraction

DOI : https://doi.org/10.56452/7-2-556

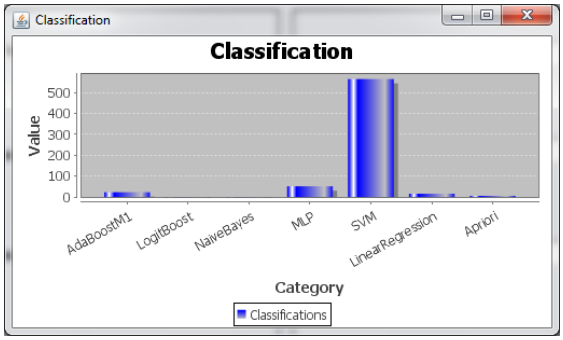

Fig 3: Classification

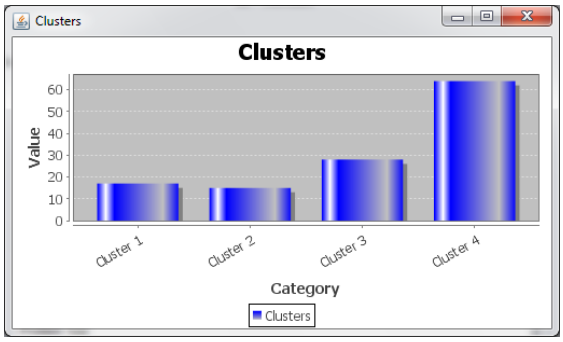

Fig 4: K-Mean Clusters

5. CONCLUSION

The protein's structural class provides vital information on the overall folding type. Knowing the protein's structure is crucial to understanding how it functions. Structural information is often classed as secondary, tertiary, or quaternary structure since there is no better language. Such forecasts still require a practical, comprehensive response. Pattern recognition techniques are the foundation of many of the advancements. The current approach employs feature extraction that is physicochemically and sequence-based. The recommended approach extracts features quickly using a correlation-based filtering method. The agreed qualities are then used to categorise the information. To correctly extract and categorise information, the Fast Correlation Based Filter method is utilised. The algorithm's usefulness and efficiency are demonstrated by computing the symmetrical uncertainty. For categorization, the best characteristics are applied, and protein structure predictions are precise.

REFERENCE

[1] M. Levitt and C. Chothia, “Structural patterns in globular proteins,”Nature, vol. 261, no. 5561, pp. 552–558, 1976.

[2] G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “Scop: Astructural classification of proteins database for the investigationof sequences and structures,” Journal of Molecular Biology, vol. 247,no. 4, pp. 536–540, 1995.

[3] M. Mizianty and L. A. Kurgan, “Modular prediction of proteinstructural classes from sequences of twilight-zone identity withpredicting sequences,” BMC Bioinformatics, vol. 10, no. 1, p. 414,2009.

[4] Z. C. Li, X. B. Zhou, Y. R. Lin, and X. Y. Zou, “Prediction ofprotein structure class by coupling improved genetic algorithmand support vector machine,” Amino Acids, vol. 35, no. 3, pp.581–590, 2008.

[5] L. A. Kurgan and L. Homaeian, “Prediction of structural classesfor protein sequences and domains - impact of prediction algorithms,sequence representation and homology, and test procedureson accuracy,” Pattern Recognition, vol. 39, pp. 2323–2343,2006.

[6] T. Liu, X. Geng, X. Zheng, R. Li, and J.Wang, “Accurate predictionof protein structural class using auto covariance transformationof psi-blast profiles,” Amino Acids, vol. 42, pp. 2243–2249, 2012.

[7] L. A. Kurgan, T. Zhang, H. Zhang, S. Shen, and J. Ruan,“Secondary structure-based assignment of the protein structuralclasses,” Amino Acids, vol. 35, pp. 551–564, 2008.

[8] J. Y. Yang, Z. L. Peng, and X. Chen, “Prediction of proteinstructural classes for low-homology sequences based on predictedsecondary structure,” BMC Bioinformatics, vol. 11, no. Suppl 1,p. S9, 2010.

[9] K. Y. Feng, Y. D. Cai, and K. C. Chou, “Boosting classifierfor predicting protein domain structural class,” Biochemical andBiophysical Research Communications, vol. 334, no. 1, pp. 213–217,2005.

[10] Niu, Y. D. Cai, W. C. Lu, G. Z. Li, and K. C. Chou, “Predictingprotein structural class with adaboost learner,” Protein and PeptideLetters, vol. 13, no. 5, pp. 489–492, 2006.