The University of Nottingham
Malaysia Campus

School of Computer Science

Intelligent Decision Support Systems

Bioinformatics and Protein Research

Application of Granular Computing Methods in the Classification of Predicted Protein Structures

Introduction

Bioinformatics is a relatively new field and involves research into many different areas. One of the these areas is the study of proteins and their functions. Protein files in the Protein Data Bank (PDB) are increasing in numbers and computational tools to help in the classification of these new proteins are urgently needed.

Background

Proteins are molecules formed from the concatenation of different amino acid types. Today there are more than 50,000 sequences in PDB and methods for classification of these data are still being developed and tested. The functions of a protein are dependent on the folding of the protein as well as the inner elements. As such the structure of a protein is an important aspect whereby the functions of the protein may be predicted based on it.

Motivation

This research is inspired by the need to develop a robust and efficient algorithm for extracting the most significant functional information from raw protein data. This should combine the sequence and the folding details in a biologically meaningful way. Diseases like Alzheimer's are caused by misfolded proteins and by using a good analysis tool, drug discovery may be accelerated.

Processing of PDB files

↓

Protein Structure Analysis

↓

Description of the Surface of the Protein

↓

Classification of Protein Files

Processing of PDB Files

A pre-processing stage is carried out on the PDB file of a protein. A PDB file contains various information on the chosen protein, including the name, experimental remarks as well as discrete information on each atom making up the protein. Here we would like to extract useful information for the construction of the protein.

```
HETATM   16  C5' 5CM A   1      19.545  17.019  19.320  1.00 11.69           C
HETATM   17  O5' 5CM A   1      19.586  17.809  18.122  1.00 12.09           O
ATOM     18  P    DG A   2      18.205  12.781  18.304  1.00 14.13           P
ATOM     19  OP1  DG A   2      19.047  12.916  17.082  1.00 14.49           O
```

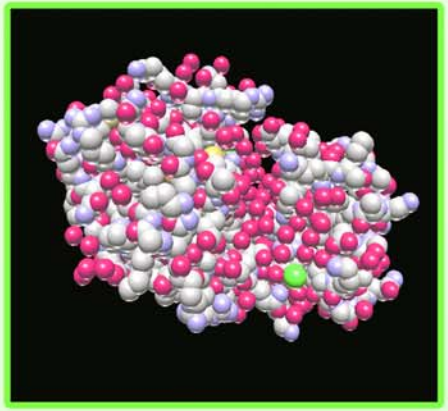
By using the above values, the protein can then be represented and projected into 3D space.

Protein Structure Analysis

The structure of a protein provides vital imformation about the characteristics and functionalities of a protein. Analysis is carried out onto the structures constructed from the PDB files. It is known that proteins are of self-similarity, and structures like the alpha-helix and beta-sheet can be easily identified within the protein, each arranged in some way.

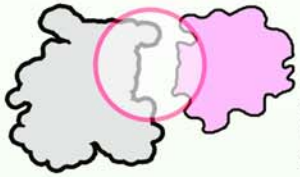
Alpha Helix

Beta Sheet



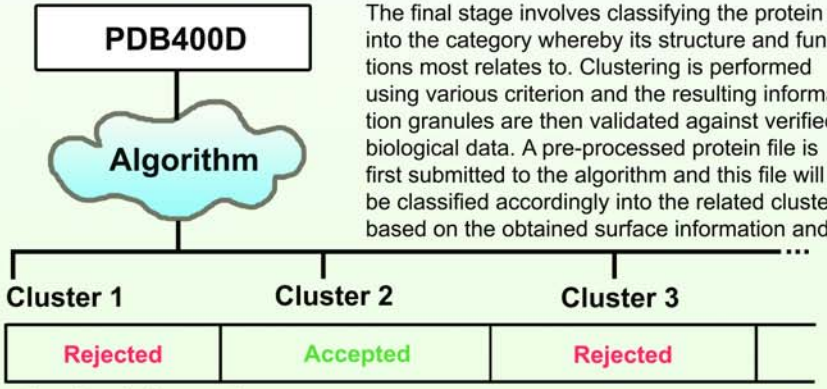
Description of the Surface of the Protein

The surface of a protein gives an idea on the functions of the protein. Therefore it is important to gain a good description of the crevices and concaves. These specific areas act as binding sites and are useful in drugs development.



Circled area show one of the binding sites

Classification of Protein Files



The final stage involves classifying the protein into the category whereby its structure and functions most relates to. Clustering is performed using various criterion and the resulting information granules are then validated against verified biological data. A pre-processed protein file is first submitted to the algorithm and this file will be classified accordingly into the related clusters based on the obtained surface information and other chemical properties.

Once clustering is completed, analysis is conducted upon the final set of data whereby the functions and identity of the protein may be predicted and determined using its membership function within each cluster - it should be noted that each cluster represents different qualities and functions that may be associated with the protein.

A multi-resolution processing of the clustered protein data is performed to obtain as much information as possible on the protein. There are no fine distinctions in the clustering process as each protein may or may not possess functions belonging to different clusters. Applying a strong cut-off point or threshold sometimes result in losing vital information that may appear in other clusters.

In this research, we look at the global properties of the protein instead of delving into the local properties. The global representation should give a good idea of the protein especially regarding the docking sites. These are the areas of concern as they react to other components like other proteins or drugs. Perhaps experimental bio-computations may be conducted in the future for the in-vitro testing of protein binding to verify the data obtained from this research.

Research Student : Lee Ling Wei
Email : Ling-Wei.Lee@nottingham.edu.my Tel: +60389248601

First Supervisor : Andrzej Bargiela
Email : Andrzej.Bargiela@nottingham.ac.uk Tel: +60389248137

Second Supervisor : Natalio Krasnogor
Email : nxk@cs.nott.ac.uk Tel: (0115)8467592