Biocomputing: analysis of protein sequence with application of data-mining tasks

Deepalakshmi R[1], Jothi Venkateswaran C[2,*]

[1]Research Scholar, Department of Computer Science, Presidency College, Chennai-600005, Tamil Nadu, India, Email: deepakugan@yahoo.com

[2]Research Supervisor, Reader & Head, Dept. of Computer Science, Presidency College, Tamil Nadu, India, Email: jothivenkateswaran@yahoo.co.in

[*]***Corresponding author**:* Research Supervisor, Reader & Head, Dept. of Computer Science, Presidency College, Tamil Nadu, India, Email: jothivenkateswaran@yahoo.co.in

Abstract

In Biocomputing, Data mining (DM) techniques are widely used for prediction of protein structure. Interpreting voluminous Biological data is complex and the need for Data mining concepts is significant. Molecular data such as DNA/Protein sequence, level of genetic expression, biochemical pathways, biomarkers and protein structures constitute a major part of biological data. In this paper, an attempt is made to discuss how standard data mining techniques such as extraction of protein data, segregation by clustering, association and visualization on a protein sequence dataset.

Keywords: Biocomputing; Molecular data; Protein sequences; Data mining; Clustering.

Abbreviations: DM-Data mining; DNA-Deoxyribonucleic acid; NCBI-National centre for biotechnology information; KDD-Knowledge discovery in database; aa-amino acid; RNA-Ribonucleic acid; SCOP-Structural Classification of Proteins.

Introduction

The basic building blocks of all organisms are Proteins derived from the Greek word ‘protos’ which means first or primal. It helps in regulation of all biological process. Usually it is made of a combination of 20 different amino acids. A protein’s function is determined by its structure. Different proteins fold to different structure based on their sequence. Thus it has been widely accepted that a protein sequence determines its structure which in turn determines its function. Therefore, it is possible to analyse protein sequence to understand its function. Clustering and classification identifies homologous and heterogeneous protein sequences.

Methods and Results

Molecular data

Molecular data such as DNA or protein sequences can be analysed and interpreted in numerous ways. Most of the features discussed elsewhere concerning editing and analysis of general categorical data also apply to molecular data; here we focus on features specifically designed for sequence data. Example for protein sequence as downloaded from NCBI is shown below,

>gi|63428|emb|CAA43320.1| lysozyme [Gallus Gallus]

MLGKNDPMCLVLVLLGLTALLGICQGGTGC
YGSVSRIDTTGASCRTAKPEGLSYCGVRASRT
IAERDLGSMNKYKVLIKRVGEALCIEPAVIAG
IISRESHAGKILKNGWGDRGNGFGLMQVDKR
YHKIEGTWNGEAHIRQGTRILIDMVKKIQRKF
PRWTRDQQLKGGISAYNAGVGNVRSYERMD
IGTLHDDYSNDVVARAQYFKQHGY

The above sequence is a combination of 20 amino acids. This protein is present in Chicken [Gallus Gallus]. 63428 refer to the accession number. We have downloaded 100 lysozyme protein sequences from NCBI which has been used in our analysis throughout.

Statistical methods in Data-mining

Statistics is defined as "turning data into information" (Bowley Arthur, 2000). It comprises of data collection and management by inferring results from numerical facts. In cases where raw data is voluminous, it is not possible to extract information directly. Hence statistical methods are used to fit the data into a model which aids in understanding the information (knowledge), (Fig.1).

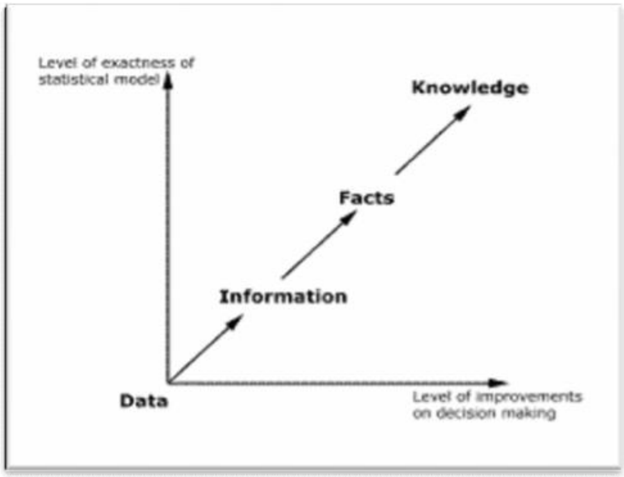

Level of exactness of statistical model
Knowledge
Facts
Information
Data
Level of improvements on decision making

Fig. 1. Process of extracting Knowledge from Data

Data Mining

Data mining is defined as finding interesting patterns from a huge collection of data (Han and Kamber, 2002). Data mining falls under one of the steps of KDD as described in Fig 2. Data mining as defined by Fayyad is the process of identifying valid, novel, potentially useful, and ultimately comprehensible understandable patterns or models in data to make crucial business decisions (Han & Kamber, 2002). "Valid" means that the patterns hold in general, "novel" that we did not know the pattern beforehand, and "understandable" means that we can interpret and comprehend the patterns.

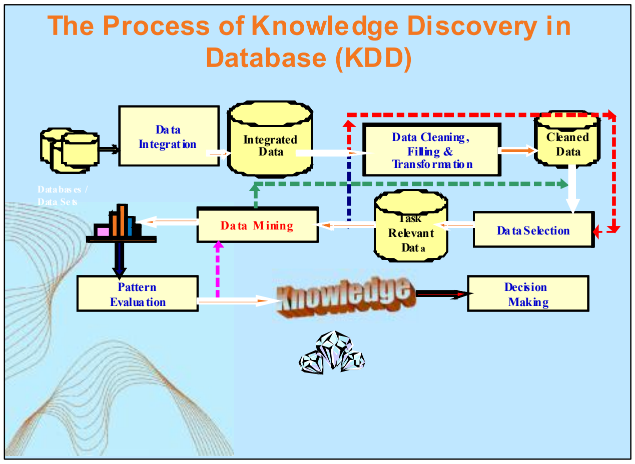


Fig. 2. Steps involved in Data

Problem narration is the important part of data mining. Since the data in the real world is dirty, data must be *cleaned* before it is going to be used. Analyzing the data is the critical part and the correct software must be used for matching patterns and similarities for the data set. Repository of data or the data which is *integrated* from various sources is what we call as *Data Warehouse.* It allows the enterprise to remember what it has noticed.

A. Data mining tasks

The main tasks of data mining which are covered in this paper are:

- *Classification* is to examine features of the object and to assign it correctly in one of predefined set of classes
- *Estimation* is to come up with continuous variable for given input data
- *Prediction:* here the records are classified according to some predicted future behaviour or estimated future value
- *Affinity grouping or association rules* determines correlation between data, also known as dependency modelling, e.g. in a shopping cart at the supermarket - market basket analysis
- *Clustering* is grouping of related items together
- *Description and visualization* is visual form of data mining.

Most of the above mentioned tasks are well-suited for data mining, which imparts extracting knowledge from the given data. We have tested the following data mining tasks on a set of protein data set.

DM Techniques in Molecular Data

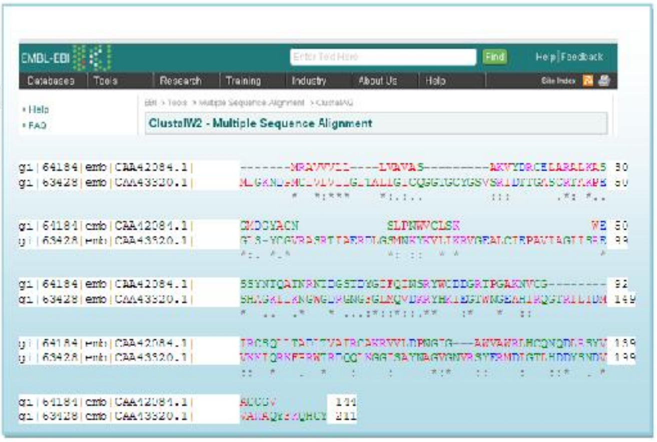
A. Clustering

This is used to group related sequences or organisms together. A group of lysozyme dataset is taken and multiple sequence alignment is performed using ClustalW2. ClustalW2 is a general purpose multiple sequence alignment program for

DNA or proteins. Aligned sequences are represented using same color and the sequence which doesn't match is denoted by hyphen (--). This algorithm is constructed with affine gap penalties. The gap cost model penalizes insertion and deletion with a linear function where one term is length dependent and other is length independent.

Gap penalty = Gapopen+Len * Gapextend (Hert and Stormo, 1999)

Fig.3 is an example showing multiple sequence alignment of pasted protein sequence using ClustalW2 tool. The number in the right hand side denotes sequence length. * Represents identical amino acid (aa), double dots (..) indicates 2 or more aa is identical and a single dot (.) indicating only single aa is placed.



ClustalW2 - Multiple Sequence Alignment

```
gi|64184|emb|CAA42084.1|      -------MRAVVVLL----LVAVAS---------AKVYDRCELARALKRS 30
gi|63428|emb|CAA43320.1|      MLGKNDPMCLVLVLLGLTALLGICQGGTGCYGSVSRIDTTGASCRTAKPE 50
                                     *  *:***   *:.:..          :::     .*: *..

gi|64184|emb|CAA42084.1|      GMDGYACN                SLPNWVCLSK             WE 50
gi|63428|emb|CAA43320.1|      GLS-YGGVRASRTIAERDLGSMNKYKVLTKRVGEALCTRPAVTAGILSRR 99
                              *:.  *.*                *: ::  * *             *

gi|64184|emb|CAA42084.1|      SSYNTQATNRNTDGSTDYGIFQINSRYWCDDGRTPGAKNVCG-------- 92
gi|63428|emb|CAA43320.1|      SHAGKLLKNGNGDPGNGsGLNQVDKRYHKLEGTWNGEAHIRQGTRILIDM 149
                              *   ..   .*    *  ...:*::*::.**    :*    *   ::

gi|64184|emb|CAA42084.1|      IRCSQLLTADLTVAIRCAKRVVLDPNGTG---AWVAWRLHCQNQDLRSYV 139
gi|63428|emb|CAA43320.1|      VKKIQRNFERWTRDCGTNGGTSAYNAGVGNVRSYFRMDIGTLHDDYSNDW 199
                              ::   *    .   *     :     :      *:*    ::    :   ::*   .*

gi|64184|emb|CAA42084.1|      ACCGV        144
gi|63428|emb|CAA43320.1|      VAKAQRSKQHCY 211
```

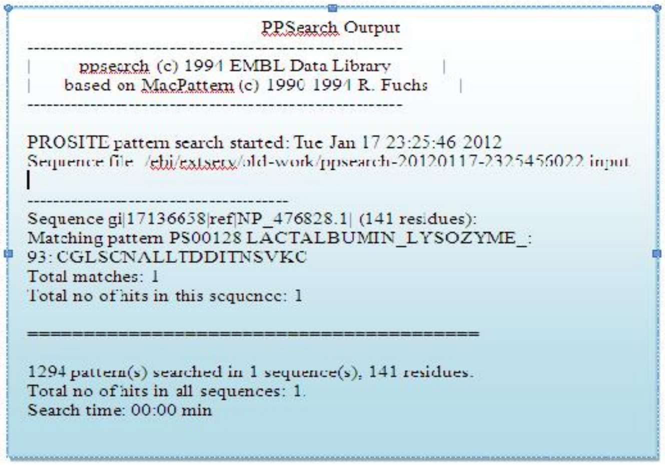
Fig.3. Multiple sequence Alignment using ClustalW2

B. Classification

Assigning to a set of predefined class classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Pattern recognition tools like Prosite Pattern Search, Scan Prosite, Scop and Phylogenetic tree uses the technique of Classification. The dataset is experimented with each of these expasy tools. Some amino acid residues are important for biological function. Patterns are description of biological relevant residues.

1. Prosite Pattern Search

This tool searches for a conserved region (motif) in protein sequences. Any protein sequence can be searched against the pattern stored in Prosite to determine its function based on the motif region. This tool requires a protein sequence as input, but DNA/RNA may be translated into a protein sequence using transeq and then queried. We have pasted the protein sequence of Drosophila melanogaster (Common fruit fly) in prosite pattern search. Motif is found from the 93rd aa and the pattern is displayed. In one pattern around 141 residues are also found. The output is shown in Fig.4. This is used to search motifs present in the protein sequence. We have pasted the protein sequence in prosite pattern search. The output is shown in Fig.4.



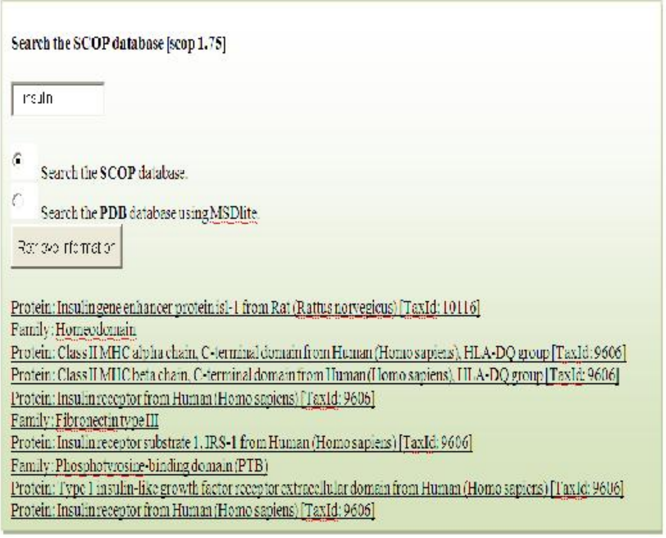
```
                      PPSearch Output
-------------------------------------------------------
|       ppsearch (c) 1991 EMBL Data Library          |
|    based on MacPattern (c) 1990 1991 R. Fuchs          |
-------------------------------------------------------

PROSITE pattern search started: Tue Jan 17 23:25:46 2012
Sequence file  /ebi/extserv/old-work/ppsearch-20120117-232545602 2 input
|
--------------------------------------
Sequence gi|17136658|ref|NP_476828.1| (141 residues):
Matching pattern PS00128 LACTALBUMIN_LYSOZYME_:
93: CGLSCNALLTDDITNSVKC
Total matches: 1
Total no of hits in this sequence: 1

==============================================

1294 pattern(s) searched in 1 sequence(s), 141 residues.
Total no of hits in all sequences: 1.
Search time: 00:00 min
```

Fig .4. Output of PPsearch showing no. of hits in the sequence

2. SCOP

Structure of a protein is a basic key for enzyme function. The primary structure is the order of different amino acids in a protein chain (Murzin *et al*., 1995). The database SCOP (Structural Classification of Proteins) has become a major resource in bioinformatics and protein science (Lo Conte *et al*., 2002). A particular strength of SCOP is the flexibility of its rules enabling the preservation of the many details spotted by experts in the classification process. Here we endow classic

SCOP Families with quantified structural information and comment on the structural diversity found in the SCOP hierarchy. Keyword search for SCOP entries Fig.5 shows the part of different classification of Insulin which is present in SCOP database.


Search the SCOP database [scop 1.75]

insulin

● Search the SCOP database.

○ Search the PDB database using MSDlite.

Random information

Protein: Insulin gene enhancer protein isl-1 from Rat (Rattus norvegicus) [TaxId: 10116]
Family: Homeodomain
Protein: Class II MHC alpha chain, C-terminal domain from Human (Homo sapiens), HLA-DQ group [TaxId: 9606]
Protein: Class II MHC beta chain, C-terminal domain from Human (Homo sapiens), HLA-DQ group [TaxId: 9606]
Protein: Insulin receptor from Human (Homo sapiens) [TaxId: 9606]
Family: Fibronectin type III
Protein: Insulin receptor substrate 1, IRS-1 from Human (Homo sapiens) [TaxId: 9606]
Family: Phosphotyrosine-binding domain (PTB)
Protein: Type 1 insulin-like growth factor receptor extracellular domain from Human (Homo sapiens) [TaxId: 9606]
Protein: Insulin receptor from Human (Homo sapiens) [TaxId: 9606]

Fig.5. Output showing the classification of Insulin

3. Phylogeny

The tree showing the evolutionary relationships among various biological species or other entities that has a common ancestor of the descendents is a phylogenetic tree. Phylogeny is a guide tree which is used for both Classification and Clustering. This contains the information for building the cladogram or phylogram. This tree is saved as .dnd file which describes the phylogenetic tree. A Phylogram is a branching diagram (tree) assumed to be an estimate of a phylogeny; branch lengths are proportional to amount of evolutionary change inferred.

These options are now in control with new buttons in the output file as well as a pop-up menu, which is available by right-clicking on the applet. The buttons on the page include "Show as Phylogram Tree", "Show as Cladogram Tree" and "Show Distances" (Antonis Rokas, 2011). This tree is used to find the evolutionary relationship between organisms. The tree in Fig.6 is a phylogeny tree showing the relationship between our lysozyme data. This also shows the distance which is denoted.

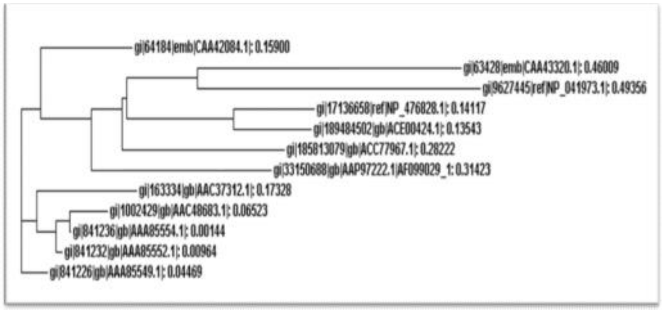

gi|64184|emb|CAA42084.1|: 0.15900
gi|63428|emb|CAA43320.1|: 0.46009
gi|9627445|ref|NP_041973.1|: 0.49356
gi|17136658|ref|NP_476828.1|: 0.14117
gi|189484502|gb|ACE00424.1|: 0.13543
gi|185813079|gb|ACC77967.1|: 0.28222
gi|33150688|gb|AAP97222.1|AF099029_1: 0.31423
gi|163334|gb|AAC37312.1|: 0.17328
gi|1002429|gb|AAC48683.1|: 0.06523
gi|841236|gb|AAA85554.1|: 0.00144
gi|841232|gb|AAA85552.1|: 0.00964
gi|841226|gb|AAA85549.1|: 0.04469

Fig.6. Phylogeny tree showing nucleotide difference between

Description and visualization

Many tools help in visualization of similarities between protein sequences, like Dotmatcher, Dotpath and Dottup.

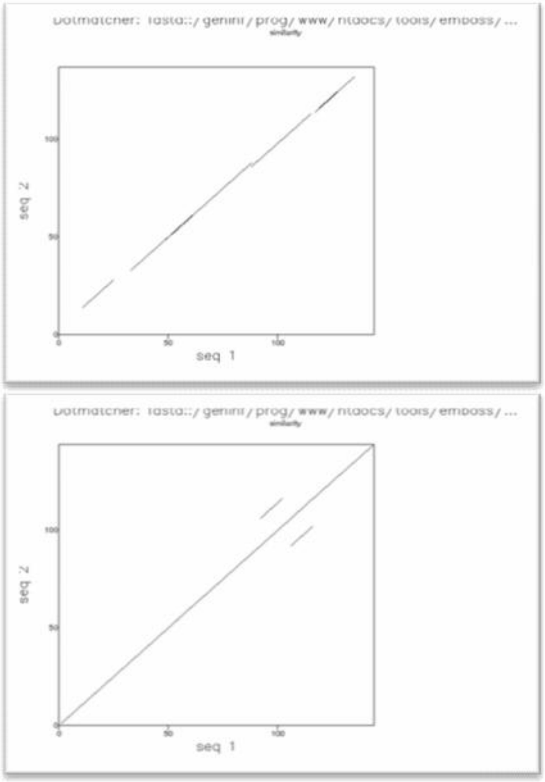


Fig.7. Figure showing the similarities between two sequences where the perfect diagnol line indicating the more similar sequence

Dotmatcher

Graphical representation of the match among the sequences is viewed by Dotmatcher. Here two sequences are compared for the similarity, the small lines at the side indicating the similar patterns (Fig.7).

Input section

BLAST

Visualization Basic Local Assignment Search Tool to show the graphical representation of how alignment has started (Fig.8). Putative conserved domains have been detected; click on the Fig.8 for detailed results.

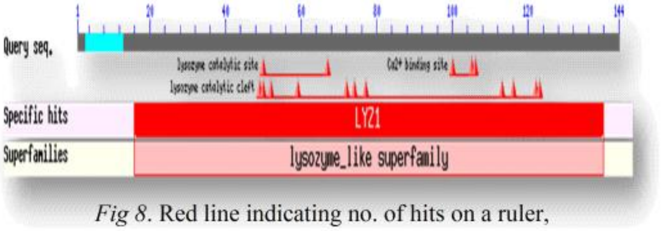


Fig 8. Red line indicating no. of hits on a ruler, distribution of 100 blast hits on the query sequence

Association rule

Identifying co-occurring gene sequences and linking genes to different stages of disease development. Association analysis methods can be used to help determine the kind of genes that are likely to co-occur in target samples. This analysis would facilitate the group of genes and the study of interaction and relationships between them (Jiawei Han, 2002). Heat maps are used to visualize the patterns of concurrence of genes in an effective manner.

Heat maps

In papers reporting microarray data analyses 'heat maps' are produced. The heat map presents a grid of colored points where each color represents a gene expression value in the Fig.9 and the sample is taken from a recent paper using expression levels for cancer classification. The grid coordinates correspond to the sample by gene combinations. In this case, the columns (samples) are tumors, some from patients who have relapsed and some from patients who have not relapsed. The rows represent 348 genes found to distinguish the patients according to their relapse status. In the heat map colors at a particular point (i.e., row by column coordinate) are assigned to represent the level of expression for that gene (row) in the sample (column) with red corresponding to high expression, green corresponding to low expression and black corresponding to an intermediate level of expression. The ordering of the rows and columns was determined using hierarchical clustering and the associated dendrogram for the samples shown at the top of the Fig.9. In this example, six relapsing patients were clustered together to the left and the non-relapsed patients clustered to the right. The heat map gives an overall view that the 348 genes have low expression in relapse patients as indicated by green color in the left-hand columns under the relapse patients.

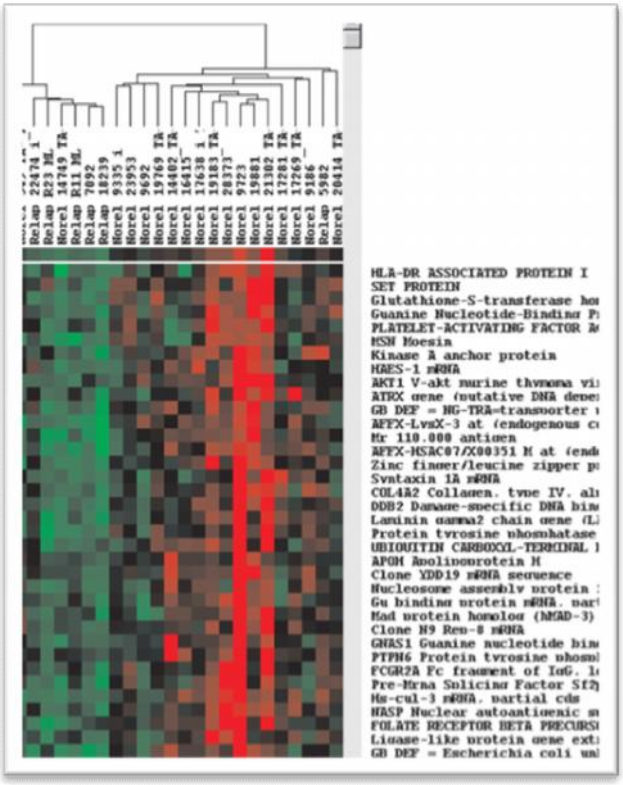


Fig 9. Heat map showing the gene with low, high

Conclusion

Protein classification is a varied landscape. Results which are generated using different approaches are quite impressive. Classification of protein sequences is enormous. Phylogeny can be used both with clustering and classification. At a glance, we cannot match the patterns and we cannot find the similarity among sequences, hence computational work with data mining tasks minimizes the work of sequence comparison. For more number of genes, heat maps produced a very impressive result by associating normal genes with diseased genes. Data mining techniques are very effectively used in protein data set which produces better results. Structure based comparison along with protein interaction and prediction can be a future work.

Reference

1. Antonis Rokas (2011) Phylogentic Analysis of Protein Sequence Data Using the RAXML Program. CORD Conference Proceedings, Chap. 19, Unit 19.11.
2. Bowley Arthur (2000) Elementary Statistics, 4th edition.
3. Hert GZ and Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8), 563-77.
4. Jiawei Han (2002) How Can Data Mining Help Bio-Data Analysis?. BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference), Department of Computer Science, University of Illinois at Urbana-Champaign.
5. Jiawei Han and Micheline Kamber (2002) Data Mining: Concepts and Techniques, Simon Fraser University.
6. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C & Murzin AG (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30(1), 264–267.
7. Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247**,** 536–540.