Regularity of Protein Secondary Structures and Its prediction

Yen-Wei Chu [1,3], Chuen-Tsai Sun[3], Chung-Yuan Huang [2,3]
[1)] Department of Information Management
[2)] Department of Computer Science and Information Engineering
Yuanpei Institute of Science and Technology
306 Yuan Pei Road
Hsinchu, Taiwan 30015, ROC
[3)] Department of Computer Science
National Chiao Tung University
1001 Ta Hsueh Road
Hsinchu, Taiwan 300, ROC
{ywchu, ctsun, gis89802}@cis.nctu.edu.tw http://www.cis.nctu.edu.tw/~ctsun

Abstract: - We define a schema representation for visualizing the relationship between primary and secondary protein structures. In the low sequence similarity training set, the steady-state genetic algorithm outperforms the association rule mining to find those high discrimination and confidence schemata. These found schemata not only can be provided to biologists for the regularity of protein secondary structures but also applied to predict the protein secondary structures. Because of the poor Q3 accuracy in the previous study, we offer a clustering method to the steady-state genetic algorithm. The clustering method plays two important roles: one is to generate parts of initial chromosomes in genetic algorithms and another one is to assist schemata in predicting secondary protein structures. In accordance with our tests, the new approach improves 12% of Q3 accuracy by comparing to previous efforts. We also raise some new examples of schemata with the interesting biological meaning to do some discussions.

Key-Words: - protein secondary structure, genetic algorithms, clustering, data mining, knowledge discovery

1 Introduction

Determining protein structure in a laboratory is much more difficult than identifying protein sequence, which explains why as of September 23, 2005 the Protein Information Resource (PIR) database contained 2,203,641 protein sequences while the Protein Data Bank (PDB) contained only 32,727 protein structures. Accordingly, independent researchers and an organization known as the Critical Assessment of Techniques for Protein Structure Prediction support the practice of predicting protein structures from previously known sequences [1, 2]. A typical approach is to predict secondary protein structures from a sequence, then use a combination of the secondary information and various biological heuristic functions to improve predictive algorithms [3]. Protein secondary structures also play an important role in protein function discovery, protein classification, and establishing phylogenetic trees. For this reason, we decided to take a closer look at the natural instincts of protein secondary structures and their potential for assisting in protein secondary structure prediction.

Most secondary protein structure prediction methods are incapable of clearly identifying observable regularity. In light of the low Q3 values currently reported by researchers [4], we proposed a schema generated by a steady-state genetic algorithm (SSGA), which is known to outperform association-rule mining methodology in RS130 data sets for these kinds of schemata [5]. In this paper, our schema discovery approach combines SSGA and clustering to identify high confidence schemata and to improve Q3 accuracy by at least 10 percent [6].

1.1 Schema Definition

Protein secondary structures are designated as H (alpha helix, 3/10 helix, pi helix), E (beta bridge, beta ladder), or L (turn, bend) [7]. The regularity of secondary structures (which consist of amino acids and one secondary structure) are usually discussed in terms of factors that cause amino acids to combine in order to form a specific secondary structure. An amino acid that plays a role in certain secondary structures are affected by neighboring amino acids, while secondary structure sheets often require extra consideration for remote amino acids. In the same manner that many researchers de-emphasize the effect of remote amino acids on protein secondary structure [8], we decided to underplay the remote effect in order to simplify schema design.

1.2 Representation

We modified Holland's (1975) one-dimensional schema format

schema $s \in \{1, 0, *\}^l$

Table 1. S Statistics for 20 amino acids in the PDB_select chain set. % is the percent of each amino acid in the PDB_select. %H, E% and L% is the percent of each secondary structure respectively in the PDB_select

Amino Acid	Num.	%	H 87690	H% 35.2%	E 55134	E% 21.1%	L 106160	L% 42.6%
A	18937	7.60%	9278	49%	3216	17%	6443	34%
R	12469	5.01%	5234	42%	2585	20.7%	4650	37.3%
N	11335	4.55%	3093	27.3%	1579	13.9%	6663	58.8%
D	14300	5.74%	4441	31.1%	1629	11.4%	8230	57.6%
C	4497	1.81%	1260	28%	1293	28.8%	1944	43.2%
Q	16934	6.80%	7855	46.4%	2823	16.7%	6256	37%
E	9989	4.01%	4658	46.6%	1643	16.4%	3688	37%
G	17764	7.13%	2952	16.6%	2553	14.4%	12259	69%
H	5857	2.35%	1978	33.8%	1254	21.4%	2625	44.8%
I	14136	5.68%	5247	37.1%	5485	38.8%	3404	24.1%
L	21635	8.69%	10053	46.5%	5188	24%	6394	29.6%
K	15587	6.26%	6050	38.8%	2837	18.2%	6700	43%
M	5550	2.23%	2373	42.8%	1174	21.2%	2003	36.1%
F	10109	4.06%	3641	36%	3201	31.7%	3267	32.3%
P	11238	4.51%	1960	17.4%	1122	9.98%	8156	72.6%
S	15481	6.22%	4193	27.1%	2924	18.9%	8364	54%
T	13623	5.47%	3684	27%	3576	26.2%	6363	46.7%
W	3705	1.49%	1339	36.1%	1115	30.1%	1251	33.8%

(where l is a fixed length and * is either 0 or 1) into a two-dimensional format:
schema s$\in${an amino acid, *} $^{(l-1)/2}$ X {an amino acid} X {an amino acid, *} $^{(l-1)/2}$ $\rightarrow$ {H, E, L| one kind of secondary structures},
where l is a fixed length (an odd number) and * is don't care.

According to our proposed schema, the central amino acid plays a role that corresponds to a specific secondary structure due to non-asterisk amino acids on each of its two sides. In Figure 1, amino acid A is found in the first and last positions and amino acid L is in the center position. Amino acid L is eventually categorized as having a H protein secondary structure—in other words, L is only affected by the first position amino acid on its left side and fourth position amino acid on its right. The other asterisk positions (which have no affect on L) can consist of any amino acid. We focused on the 9 windows in the front part of the schema, since that length is long enough to contain sufficient local structural information for analysis [9].

$$A{*}{*}{*}L\ {*}{*}{*}A \rightarrow H$$

Figure 1. Schema example.

2 Preprocessing the Raw Data

We established a data set according to the PDB_select protein chain list because it is representative of PDB chain identifiers that help researchers save considerable time and effort. The PDB_select protein chain list allows for introductory browsing, protein architecture analysis, prediction method development, and model building via modular construction [10].

2.1 PDB_select constraints

There are many versions, from which no two proteins have more than 25% sequence identity to 95%, in the PDB_select list. Furthermore, it excludes chains according to the following criteria:

- length less than 30 residues;
- number of non-standard amino acid residues (including chain breaks) exceeds 5 percent of chain length;
- resolution exceeds 3.5 angstroms;
- R-factor exceeds 30 percent;
- some chains are known to be of inferior quality;
- number of residues without side chain coordinates < 90 percent chain length;
- number of residues without backbone coordinates < 90 percent chain length;
- content of ALA plus GLY exceeds 40 percent of chain length; and
- data on resolution or R-factor (i.e., NMR-structures) are not available.

2.2 Constraints

We separated the data set into two independent sets (training and testing) and used the most stringent 25% PDB_select list (2,485 chains with 388,067 residues). Next, we located the secondary structures of proteins in the 25% PDB_select list from the Database of Secondary Structure in Proteins (DSSP) of secondary structure assignments for all PDB protein entries. However, due to problems with DSSP secondary structure information, we eliminated some chains from the 25% list for the following reasons:

- incorrect PDB identification in the 25% list;
- no information in the DSSP files;
- broken chains; or
- inclusion of an unknown symbol X.

Our data set consisted of 1,600 chains with 248,984 residues. We randomly selected 1,200 chains for use as a training set for mining schemata; the remainders were used for testing.

2.3 Data Set Analysis

It was assumed that the distribution characteristics of the data set would affect the experimental results. We used the data in Table 1 to inspect a) whether a relationship exists between the amount of a schemata and the percentage of each amino acid in the data set, and b) the individual tendencies of all amino acids in the data set. Data in the first column of Table 1 are for 20 amino acids and second and third column data represent the number of occurrences for each amino acid and their respective percentages. The final column contains data on the corresponding amino acids, number of occurrences, and percentage of secondary helix (H), sheet (E), and Coil (L) structures. The first row presents information on the number of occurrences and percentages of each secondary structure in the data set.

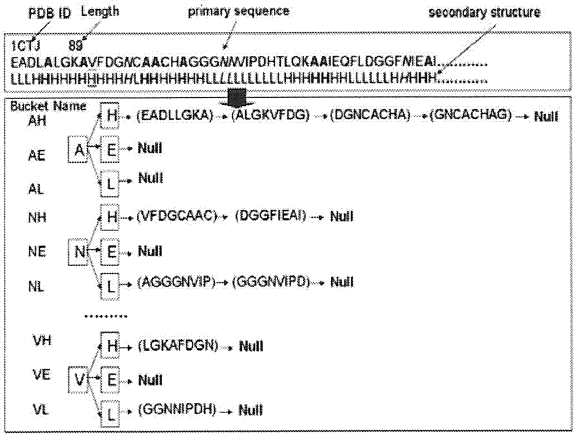

Figure 2. An example of using sequence 1CTJ to make a training set.

2.4 Making Training Sets

For every protein sequence, each amino acid can be viewed as a central amino acid in a schema. We defined amino acids on both sides of a central amino acid as a "neighbor pattern." According to our size choice of 9 windows, neighbor pattern length = 8, or 4 amino acids on each side. To create the training set we placed the neighbor pattern into a corresponding bucket according to the central amino acid and secondary structure; a partially assigned training set is shown in Figure 2. A complete training set consists of 20*3 buckets. Using the fifth amino acid in the 1CTJA protein sequence as an example, the neighbor pattern EADLLGKA should be put into bucket AH, since the central amino acid is A and its secondary structure is H.

3 Cluster-based Genetic Algorithm

Average Q3 accuracy in studies of protein secondary structure prediction using genetic algorithms is only 46 percent. Three issues are considered central to this problem: data set selection, solution search space, and fitness function design. At first, for the data set in previous studies, RS130 cannot represent so far the whole known proteins. Moreover, the number of similarities among DSSP protein families is considered too high. These kinds of problems are not associated with PDB_select.

Based on the 9-window size of the schema we applied, search space size is 20*3*21*8. To reduce search time, the very important thing is let genetic algorithm can search from good start. Therefore, once clustering was completed, we placed cluster centers as chromosomes into the initial population (Fig. 3).

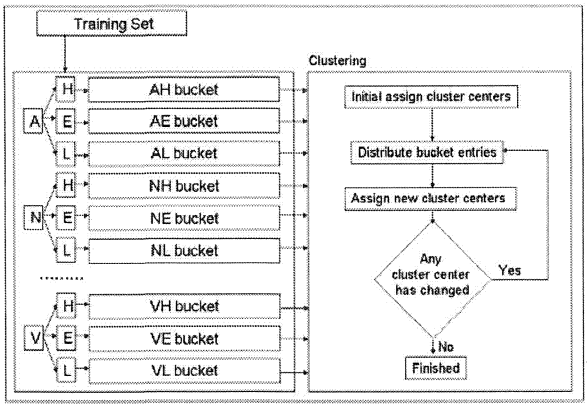

Figure 3. Our proposed clustering strategy.

The fitness function gives evolutionary direction to chromosomes [11]. When designing our fitness function, we assumed that a good schema should have a strong tendency toward a certain secondary structure. Furthermore, our fitness function states that increased chromosome confidence in the training set also increases Q3 accuracy in the protein secondary structure prediction.

As shown in Figure 4, our model includes evolutionary and application phases. With the exception of standard GA steps, during the evolutionary phase we generated some initial chromosomes by clustering. The evolutionary process makes use of a steady-state strategy. In each generation we placed certain high fitness chromosomes into our schemata set. Chromosomes placed in the set were removed from the population; the population consequently generated new chromosomes at random.

For protein secondary structure predictions we cut the sliding windows (9 window lengths) to use as protein sequence patterns for testing. Each pattern aligns with all schemata in the schemata set. After alignment, the secondary structure of the most similar schema was selected as the predictive result. When the fitness of the most similar schema was insufficient, the pattern was aligned with the neighbor patterns of cluster centers in the training set. The final predictive result was the secondary structure that the most similar cluster center belonged to. Our approach uses blosum62 as a substitution matrix for alignment purposes.

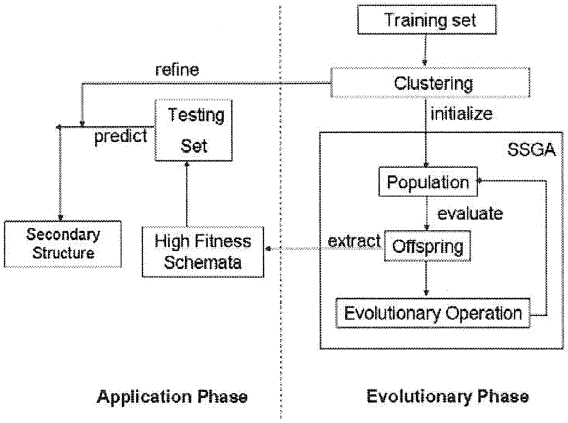

Figure 4. Our cluster-based genetic algorithm for mining schemata and its application for predicting protein secondary structures.

3.1 Population and Evaluation

Our approach uses 20 populations for each amino acid. Each chromosome includes a neighbor pattern and a secondary structure. Initial populations take on the neighbor pattern of the cluster center; all other chromosomes are randomly generated.

To evaluate a chromosome, we used its neighbor pattern for alignment with neighbor patterns in all secondary structure buckets. Alignment scores that exceeded a certain threshold were labeled as one hit. nH, nE, and nL are the respective hit numbers in the H, E, and L buckets. Chromosome secondary structure is determined according to the maximum hit number.

In the following equation,

confidence=nSS/(nH+nE+nL) (1),

nSS is defined as the maximum hit number among nH, nE, and nL. Confidence is relative to Q3; one of our goals was to find schemata with distinct tendencies toward certain secondary structures. We defined the discrimination rate (DR) as

DR=(nHighest-nSecond)/(nH+nE+nL) (2),

where nHighest is equal to nSS and nSecond is the second highest score among nH, nE, and nL. As a result,

fitness=confidence*DR (3)

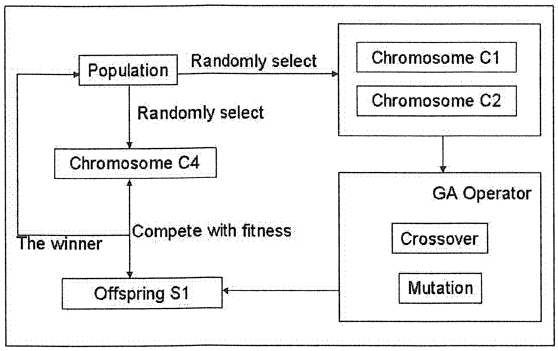

Figure 5. Steady-state strategy for our cluster-based genetic algorithm.

3.2 Steady-state Strategy

The initial step in the steady-state strategy shown in Figure 5 is to randomly select two chromosomes, C1 and C2. Two offspring are generated by one-point crossover and multi-point mutations of C1 and C2; a single S1 offspring is randomly selected from these two offspring. Another chromosome (C4) is selected from the population for comparison with the S1 offspring in terms of fitness. The best chromosome is used to replace C4 in the population.

4 Experimental Results

Association rule mining method is often used to analyze the relationship among items in data mining. We try to use this method to acquire schema-like patterns in our training set and then compare them with the results generated from steady-state genetic algorithm.

4.1 SSGA Compare with Association Rule Mining (ARM)

The training set consists of 124 protein sequences each of which has more than 80 amino acids in length, and the pairwise similarity is below 25% (similar to RS130 [12]). They were used to train SSGA to find significant schemas associated with various protein secondary structures. To obtain the confidence and support value, we tested SSGA on the nr-PDB data set created by NCBI after removing those sequences used for training. If $A \Rightarrow B$ is the form of rules, and $P(A \cup B)$ is a probability of both A and B. The confidence and support value are defined as

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{number of correct classifications}}{\text{number of schema matches}} \quad (4)$$

Table 2. Test Results of ARM30, ARM60 and SSGA (in nr-PDB)

Method	Total Mined Schema Number	confidence											support
		%	0 | 10	10 | 20	20 | 30	30 | 40	40 | 50	50 | 60	60 | 70	70 | 80	80 | 90	90 | 100	Avg. (%)	Avg. (%)
ARM30	11	Partial Mined Schema Number	11	0	0	0	0	0	0	0	0	0	0	0
ARM60	27		0	0	7	17	3	0	0	0	0	0	34.59	0.718
SSGA	904		166	16	20	33	60	60	92	120	74	263	61.51	8.364

$$\text{support}(A \Rightarrow B) = P(A \cup B) = \frac{\text{number of correct classifications}}{\text{number of secondary structure matches}} \quad (5)$$

To reduce time complexity, we adopt FP-growth algorithm for association rule mining to avoid generating candidates from the frequent itemsets [13]. Before using the ARM method for schema finding, we need to set two criteria (confidence and support). In our training set, 124 protein sequences could be further sampled into 23,448 transactions (obtained through sliding window sampling within the protein sequence, window size=9). The support value in the worst case is 4.264e-5 (1/23448). In order to discover more possible patterns, the support value could be set as 5e-5 in this experiment

A higher confidence value schema means it has a higher relationship between sequence and structure (like the form shown in figure 1) within the training data. Thus we assume that such schema could have higher confidence in testing data. The result of this assumption will be explained in the subsequent experiment. We run ARM with two different confidence values. The confidence value of ARM30 is 30% and ARM60 is 60% in the training set. Table 3 illustrates the performance of ARM30 and ARM60 under the testing set (nr-PDB). All 11 schemas of ARM30 fall within the bracket (0%-10%). However, ARM60 has a higher and broader confidence range (20%-50%).

After the evolutionary process terminated, we checked each of the twenty converged populations to get the most frequent secondary structures for every amino acid. We summarize the results in Table 2. It shows that most of the natural correlations between amino acids (statistics from nr-PDB) and the preferred structures were also found in the converged populations (evolved by SSGA) with one exception of amino acid Y. Note that all the initial populations were randomly generated. The finding of similar correlations between amino acid preferences toward particular structures in the final converged populations certainly provides some confidence of the fitness function applied in SSGA.

The learned schemas from the training set were later tested on the nr-PDB test set to measure their confidence and support values. Finally, there are 904 total possible rules to be found. The average confidence value is 61.51% and nearly half of mined rules are over 70%. Table 2 is the testing results of ARM30, ARM60 and the SSGA approach. It could be divided into three parts, the left-hand column shows the total mined schema number from compared methods; the central part shows the number of schemas mined from different confidence ranges (10% increments); and the right-hand part shows the aver-age of confidence and support value. Hence, table 3 clearly shows that the average value of confidence and support from the SSGA approach are significantly higher than the ARM method.

If the average support value of the significant schemas is 1%, then we need approximately 9861 (986059*1%) significant schemas to handle all known proteins. So the number of schemas are not enough to predict secondary structure in our results.

Table 3. Tendencies of various amino acid secondary structure types

Amino acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
nr-PDB	H	L	L	H	H	L	L	H E	H L	H	H	L	L	H	H	L	L	E	H L	H L
SSGA Population	H	L E	L	H	H	L	L	H E	H L	H	H	L	L	H	H	L	L	E	H L	E

4.2 Clustering-based SSGA

Since our approach uses a clustering strategy for the initial population, we ran several trials using cluster numbers between 20 and 70 to predict protein secondary structures; results are shown in Figure 6.

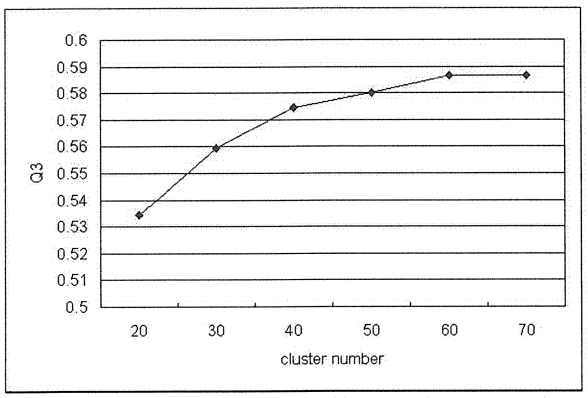

Figure 6. Q3 accuracy in different cluster numbers using our approach.

At 70 clusters our Q3 accuracy was 58.7 percent—approximately 12 percent better than predictive results from studies using genetic algorithms only.

4.3 Illustrate Some Interesting Schemata

Table 4 presents a comparison of our Table 1 results with nr-PDB. Several differences are observed when K, W, and Y are in both PDB_select and nr-PDB. This underscores the importance of selecting a suitable data set.

Selected schemata with interesting biological meaning and high fitness are displayed in Table3. The central amino acid in the first schema is P; when its neighbor pattern is D***P**N, the central amino acid plays an L role in the secondary structure. Note that L is the tendency for D, P, and N in Table 5.

Table 4. Secondary structure tendencies for each amino acid in nr-PDB and PDB_select chain sets.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
nr-PDB	H	H	L	L	L	H	H	L	L	H E	H	H L	H	H	L	L	L	H L	H L	E
PDB_select	H	H	L	L	L	H	H	L	L	H E	H	L	H	H	L	L	L	H	H E L	E

5 Conclusion and Discussion

In a previous study we reported that our steady-state genetic algorithm outperformed association rule mining in finding schemata for describing relationships between protein primary and secondary structures. The identified schemata provided biologists with sufficient data for studying protein secondary structure, but they were insufficient for predicting secondary structure. In this paper we addressed the issues of finding high-fitness schemata and improving secondary structure prediction. Although we were able to increase Q3 accuracy by approximately 12 percent, we acknowledge that Q3 accuracy is still inadequate due to the insufficient number of found schemata. Two main reasons for this approach can not find sufficient schemata are the huge search space and incomplete status of current protein structure databases.

Our future plans are to reduce search space by considering some protein evolution information—for example, HMM profile or PSSM. On the other hand, these schemata can be applied to a consensus strategy for secondary structure prediction. When other methods (e.g., SVM, PSIPRED, or PROF) are not reliable for predicting certain protein structures and when exit found schemata can be aligned with these corresponding amino acids, it is possible to determine these protein secondary structures.

Table 5. Sample schemata of biological interest.

Schema	The tendency of secondary structure
D***PP**N->L	D, P and N are all L
TS**NP**K->L	T, S, P, and K are all L
K***DP**C->L	K, P, and C are all L
****G*P*N->L	P and N are all L
G***AP**P->L	G and P are all L
*F**A*L**H->H	F, L, and H are all H
*EQMRQ*L*->H	E, Q, M, and L are all H
E***E***Q->H	E and Q are all H
I*V*V***Y->E	I, V, and Y are all E
*Y**V*I**E->E	Y, I, and E are all E

Acknowledgement:

This work was supported in part by National Science Council, Taiwan, Republic of China under grant NSC 92-2524-S-009-004 and NSC 93-2520-S-009-003.

References:

[1] Hua, S. and Sun, Z., A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach, Journal of Molecular Biology, 308, 2001, pp. 397-407.

[2] Rost, B., Review: Protein Secondary Prediction Continues to Rise, Journal of Structural Biology, 134, 2001, pp. 204-218.

[3] Lin, H.N., Chang, J.M., Wu, K.P., Sung, T.Y. and Hsu, W.L., HYPROSP II—A Knowledge-based Hybrid Method for Protein Secondary Structure Prediction Based on Local Prediction Confidence, Bioinformatics, 21, 2005, pp. 3227-3233.

[4] Chu, Y.W. and Yang, J.-M., Finding Regularity in Various Types of Secondary Protein Structures, Journal of Information Science and

Engineering, 19, 2003, pp. 943-952.
[5] Chu, Y.W. and Sun, C.T., Regularity of Secondary Protein Structures: A Genetic Algorithm Approach, Proceedings of the World Congress on Intelligent Control and Automation, 3, 2004, pp. 2104-2108.
[6] Huang, H. C., An Evolutionary Approach to Finding Schemas for 3-Class Protein Secondary Structure Prediction, Proceedings of the Computational Systems Bioinformatics, 2003
[7] Branden, C. and Tooze, J., Introduction to Protein Structure, Garland Press, 1991.
[8] Yi, T.M. and Lander, E.S., Protein Secondary Structure Prediction Using Nearest-neighbor Methods, Journal of Molecular Biology, 232, 1993, pp. 1117-1129.
[9] Bystroff, C., Simons, K.T., Han, K.F. and Baker, D., Local Sequence-structure Correlations in Proteins, Current Opinions in Biotechnology, 7, 1996, pp. 417–421.
[10] Hobohm, U. and Sander, C., Enlarged Representative Set of Protein Structures, Protein Science, 3, 1994, pp. 522.
[11] Mitchell, M., An Introduction to Genetic Algorithms, MIT Press, 1996.
[12] Rost, B. and Sander, C. Prediction of pro-tein secondary structure at better than 70% accuracy. J. Mol. Biol, 1993, 232, 584-599
[13] Han,J. and Kamber,M. Data Mining: Concepts and Techniques, 2001, Morgan Kauf-mann Publishers.