

# A Schema Finding Approach for the Regularity of Secondary Protein Structure

Yen-Wei Chu<sup>1</sup> AND Chuen-Tsai Sun<sup>2</sup>

*Department of Computer and Information Science, National Chiao Tung University,  
Hsinchu, 300 Taiwan*

<sup>1</sup>*E-mail: ywchu@cis.nctu.edu.tw*

<sup>2</sup>*E-mail: ctsun@cis.nctu.edu.tw*

*Tel: +886-3-5712121 Ext. 56601~56604*

*Fax: +886-3-5721490*

**Keywords:** *secondary protein structures; machine learning; steady-state genetic algorithm; pattern finding; association rule mining*

## ABSTRACT

*Instead of putting too much focus on current approaches to protein secondary structure prediction, the authors will look at the natural instincts of protein secondary structures, and propose a schema representation which are offered for identifying regular patterns among various types of secondary protein structures. The schemas employ genetic algorithms base on a steady-state strategy and two disjunctive data sets will be used to verify fitness function for our approach. In this study, 904 schemas were found, and nearly half of the said schemas reached confidence of 70% and higher. Finally, the paper concludes with some illustrations of significant schemas produced as part of this study, with brief explanations of their significance.*

## 1. INTRODUCTION

The latest version of the Protein Information Resource (PIR) database (updated on Dec, 9 2002) contains 283,269 protein sequences. In comparison, the Protein Data Bank (PDB) only contains 19,551 protein structures, since they are much more difficult to determine. The secondary structures of proteins are now considered crucial to understanding their tertiary structures [1, 2, 3, 4, 5]; however, even though secondary structure data is often used in protein recognition and protein structure prediction [6, 7, 8, 9, 10, 11], few attempts have been made to determine shared secondary structure patterns. Based on studies describing statistical regularity between single amino acids and various secondary structures [12], some researchers have suggested that secondary structure formation may, at least to a cer-

tain degree, be determined by sequential amino acid interaction [13]. Here we will propose a representative schema for amino acid interactions as an aid for analyzing the relationship between them and various protein secondary structures.

A schema can be regarded as a sequential pattern. In its general definition, a sequential pattern means frequently occurring patterns related to time or other sequences, and schema differences are often expressed in terms of positions. Agrawal and Srikant introduced the concept of mining sequential patterns from a set of market-basket data [14]. Sequential pattern mining methods make use of variations in Apriori-like (statistics-based) algorithms, with different researchers using different parameter settings and constraints [15, 16, 17, 18]. But there is still no related research on our problem yet. On the other hand, association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. A pattern in the Association rule mining dose not follow any specific sequential order, which differs from sequential pattern mining. Thus, more general patterns can be found with association rule mining. Traditional statistical methods identify significant patterns or rules according to their frequencies in data sets; in contrast, a schema also considers distinguishability. Hence, we will use association rule mining to find some patterns in our data set for comparison. By this, we will not only get a comparison with our method but also emphasize the significance and necessity of distinguishability in a schema.

In the absence of a widely applied data mining

method, we adopted a genetic algorithm based on a steady-state strategy. This approach is based on genetic algorithms as described by John Holland [19], whose work is associated with natural selection principles. The computational aspect of this method, which entails a great deal of random searching, is considered both effective (because of its fitness function) and flexible (because of its problem encoding capability) [20, 21, 22, 23, 24]. We adopted a genetic algorithm approach for two reasons: a) it allows for the design of a fitness function that considers frequency and distinguishability (as opposed to traditional data mining methods that emphasize frequency only); and b) unlike traditional data mining methods (which lack crossover and mutation operators), genetic algorithms are more useful in determining regularity over a training set.

## 2. METHODS

### 2.1 Schema

Protein secondary structures are generally designated as H (alpha helix, 3/10 helix, pi helix), E (beta bridge, beta ladder), and L (turn, bend) [25]. Biologists acknowledge that the behavior of any amino acid in a protein sequence is susceptible to adjacent amino acids, but little work has been done to identify the regularity of these interactions. To address this problem, we applied Holland's schema theory [19], with the use of schemas to reflect the regularity. A schema is a bit string in which a bit is either an amino acid or an asterisk that represents any amino acid. Figure 1 shows an example of a schema in which the first and last positions are both amino acid A, and amino acid L is in the center. Currently, we only focus on schemas of nine amino acids in length.

A\*\*\*L\*\*\*A

H

Fig. 1. Schema example

Secondary structures are thought to be related to molecular interaction; a schema represents the most stable molecular configuration in terms of Van Der Waal's forces and hydrogen bonds. The schema in Figure 1 could be associated with the helical structure, which may be determined by interactions among the amino acids A (first position), L (middle position), and A (final position). Our goal is to identify significant schemas that can be used to characterize various protein secondary structures. This is a non-trivial task because a) the number of necessary schemas is unknown, b) the schema length is variable, and c) a measure is needed to evaluate the schema's quality.

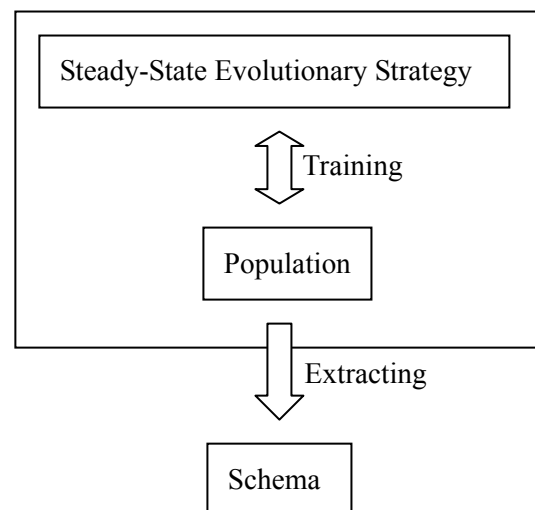


Fig. 2 Methodology used in this research

### 2.2 Algorithm

A decision was made to use steady-state Ge-

netic Algorithms (SSGAs) to search for possible schemas because they are frequently used in rule-based systems [26] and schemas can be considered a form of rules. Each schema that evolved was used to classify the secondary structure of a protein sequence. As shown in Figure 2, the schemas were encoded into the SSGA population. During the evolutionary process, schema that matched the established criterion were organized into a schema set and used to analyze protein secondary structure regularity in terms of its primary sequence patterns.

The framework of a SSGA in our study is illustrated in Figure 3. As shown in Figure 3, a chromosome C1 is first randomly selected from a population. C1 will be either mutated or crossed over a second randomly selected chromosome to yield C2. A chromosome C3 that is most similar to C2 is then taken from the population for comparison. The one with better fitness will survive to the next generation.

There are several components in our system. They are chromosome encoding, population initialization, fitness function and genetic operators.

To reduce the computational complexity, in-

stead of using a single huge population of chromosomes, we initialize a population for each amino acid. In a particular population for amino acid, e.g. R, each chromosome represents a potential schema of nine amino acids in length with the amino acid, R, fixed at the center position, while the others randomly determined. The reason we fix R at the center of the schema in this case is that we try to model the interactions between the center R and the other neighboring amino acids. During the evolution process, the genetic operations, i.e., mutation and crossover, apply to all positions except the center to maintain the specificity of the schemas in each particular population. An illustration of the 20 populations under consideration is presented in Figure 4. Each schema is associated with a specific secondary structure determined by its fitness.

The design of our fitness function is based on the fact that there is a correlation between the primary sequence and the secondary structure it forms. To evaluate the fitness of a schema,  $s$ , we first measure its tendency toward each secondary structure, defined as follows.

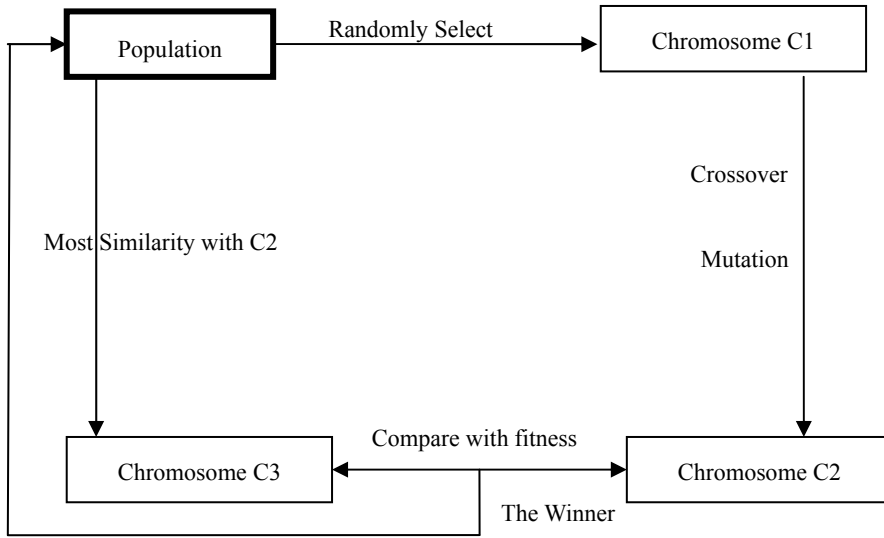


Fig. 3. SSGA flowchart

$$Tendency(s)_{ss} = \frac{1}{SA_{ss_{max}} - SA_{ss_{min}}} \sum_{SA_{ss_i} > threshold} SA_{ss_i} \quad (1)$$

where  $tendency(s)_{ss}$  is the tendency of schema  $s$  toward a particular secondary structure  $ss(H, E$

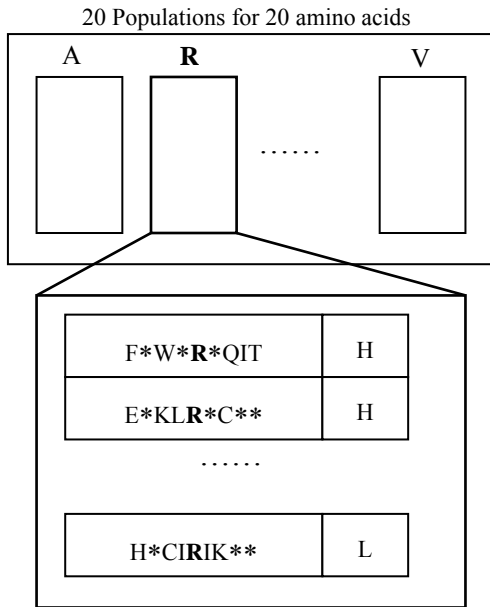


Fig. 4. A sample population for amino acid R

or L).  $SA_{ss_i}$  represents the alignment score between schema  $s$  with a secondary structure  $ss$  and the  $i^{th}$  schema of  $ss$  in the training set.  $SA_{ss_{max}}$

and  $SA_{ss_{min}}$  mean the maximum and minimum alignment scores respectively. Note that we only

Table 1. Training set of 124 proteins used for learning schemas

1aaj	1aba	1add	1ads	1apa	1aps	1btc	1c5a	1caj	1ccr
1cdb	1cde	1cgt	1cid	1crl	1cyo	1dog	1eco	1ede	1ezm
1fdd	1fha	1fhb	1gal	1gpb	1hbq	1hmy	1hra	1lfc	1lipd
1le4	1mgn	1mup	1ndk	1ofv	1omp	1osa	1phh	1plc	1pyp
1rhd	1rmd	1s01	1sqt	1snc	1spa	1ten	1tlk	1trb	1ula
1vqb	2aak	2abh	2abk	2ayh	2cbp	2cdv	2cp4	2cpl	2cro
2cts	2cyp	2fox	2liv	2nrd	2pfl	2phy	2sga	2sim	2snv
2spo	3adk	3dfr	3gbp	3grs	3pgk	3pgm	3tgl	4enl	4fgf
4ger	4xis	5nn9	6taa	8abp	8acn	8ilb	9rnt	1291	1aep
1arb	1bw3	1dhr	1eaf	1gky	1gof	1lis	1nar	1poa	1poc
1ppn	1rcb	1sbp	1tml	1utg	2baa	2cas	2cmd	2ctc	2dri
2end	2mhr	2mnr	2omf	2pgd	2pia	2por	2rn2	2sas	2stv
2tgi	3chy	3cla	5p21						

consider those alignments with scores above a specified threshold. Alignments with low scores

are considered noise.

where  $Tendency(s)_{highest}$  and  $Tendency(s)_{second}$

Table 2. Statistics for 20 amino acids in the nr-PDB chain set

	Num	%	<b>H</b> 354429	<b>H%</b> 35.9%	<b>E</b> 210513	<b>E%</b> 21.3%	<b>L</b> 421117	<b>L%</b> 42.7%
<b>A</b>	82743	8.39%	41219	<b>4.18%</b>	13582	1.38%	27942	2.83%
<b>C</b>	10701	1.09%	3398	0.34%	3095	0.31%	4208	<b>0.43%</b>
<b>D</b>	57508	5.83%	18068	1.83%	6736	0.68%	32704	<b>3.32%</b>
<b>E</b>	65288	6.62%	31741	<b>3.22%</b>	9616	0.98%	23931	2.43%
<b>F</b>	38874	3.94%	13778	<b>1.40%</b>	11807	1.20%	13289	1.35%
<b>G</b>	73432	7.45%	12872	1.31%	10714	1.09%	49846	<b>5.06%</b>
<b>H</b>	22508	2.28%	7438	0.75%	4818	0.49%	10252	<b>1.04%</b>
<b>I</b>	56906	5.77%	20985	<b>2.13%</b>	20959	<b>2.13%</b>	14962	1.52%
<b>K</b>	57486	5.83%	23243	2.36%	9637	0.98%	24606	<b>2.50%</b>
<b>L</b>	88394	8.96%	41502	<b>4.21%</b>	20762	2.11%	26130	2.65%
<b>M</b>	22057	2.24%	9477	<b>0.96%</b>	4944	0.50%	7636	0.77%
<b>N</b>	43029	4.36%	12045	1.22%	5784	0.59%	25200	<b>2.56%</b>
<b>P</b>	45803	4.65%	8320	0.84%	4076	0.41%	33407	<b>3.39%</b>
<b>Q</b>	37829	3.84%	17031	<b>1.73%</b>	6345	0.64%	14453	1.47%
<b>R</b>	50134	5.08%	21199	<b>2.15%</b>	9545	0.97%	19390	1.97%
<b>S</b>	57626	5.84%	16779	1.70%	10797	1.09%	30050	<b>3.05%</b>
<b>T</b>	57004	5.78%	15774	1.60%	14590	1.48%	26640	<b>2.70%</b>
<b>V</b>	71239	7.22%	22665	2.30%	28564	<b>2.90%</b>	20010	2.03%
<b>W</b>	13325	1.35%	5150	<b>0.52%</b>	3670	0.37%	4505	0.46%
<b>Y</b>	34173	3.47%	11745	1.19%	10472	1.06%	11956	<b>1.21%</b>

We prefer schemas with greater discrimination power, that is, a good schema should have a strong tendency, a significant difference in value between the highest and the second highest tendencies, toward a particular secondary structure. Given all the tendencies toward various structures, we define the fitness as:

$$Fitness(s) = \log \frac{Tendency(s)_{highest}}{Tendency(s)_{second}} \quad (2)$$

represent the highest and the second highest tendency of schema  $s$  respectively.

We adopt a steady-state selection mechanism to choose candidate schemas to participate in evolutionary process. Standard genetic operators such as uniform crossover and multi-point mutation are applied to generate new populations. The same evolutionary process is repeated until the fitness values of schemas do not improve. After the convergence, from all the twenty populations,

Table 3. 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

we combine those schemas with high fitness values to form the final significant schema set. These schemas can then be used to classify the secondary structures of new protein sequences.

### 3. EXPERIMENTS

#### 3.1 Methodology and Data sets

There are two purposes of our experiments. First, we want to verify the confidence value of our system; second, we want to validate the fitness function. As our system is within the supervised learning paradigm, we prepared the training set and testing set respectively. 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 [13]). They were used to train SSGA to find significant schemas associated with various protein secondary structures. The 124 proteins are listed in Table 1. 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

$$\text{confidence}(A \Rightarrow B) = P(B | A) = \frac{\text{number of correct classifications}}{\text{number of schema matches}} \quad (3)$$

$$\text{support}(A \Rightarrow B) = P(A \cup B) = \frac{\text{number of correct classifications}}{\text{number of secondary structure matches}} \quad (4)$$

From large databases, biologists have found that there exists some preference of secondary structures for each amino acid. We thus looked into the finally converged twenty populations for similar correlations. The similar correlations may suggest that the fitness function we used can approximate real biological meanings so as to justify its usage.

#### 3.2 Results

The statistics of amino acids and secondary structures in the non-redundant Protein Data Bank is summarized in Table 2. The first two columns present the number of occurrences of each amino acid and its percentage in the nr-PDB, and the remaining columns show the number of occurrences and the percentage of secondary structure H, E and L within the nr-PDB respectively.

Table 4. 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	H	H	H	L	L	H	H	L	L	E	H	H
								E	L										L	L
SSGA	H	L	L	H	H	L	L	H	H	H	H	L	L	H	H	L	L	E	H	<i>E</i>
Population		E						E	L											L

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 our proposed approach.

### 3.2.1 Association rule mining (ARM)

To reduce time complexity, we adopt FP-growth algorithm for association rule mining to avoid generating candidates from the frequent itemsets [27]. 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%).

The following section will describe the result of our proposed approach and its comparisons with the ARM method.

### 3.2.2 Our SSGA Approach

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 4. 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 meas-

Table 5. Sample schemas with high fitness

Schema	Secondary Structure	No. of schema occurrences in nr-PDB	confidence (%)	support (%)
***A**LAE	Helix	81	97.53	0.0228
****PP***	Loop	2049	95.17	6.1334
*P***PT**	Loop	129	91.47	0.0306
***G*PS**	Loop	201	89.05	0.0477
**VVI****	Sheet	348	80.46	1.8321
***E*LLR*	Helix	58	89.66	0.0163
****P**S	Loop	2777	79.87	0.6594
**R*N*P**	Loop	305	78.69	1.2603
K***E*L*D	Helix	160	76.25	0.5041
**A*E***K	Helix	461	75.49	1.4523
**VVL*S**	Sheet	93	75.27	0.4479

ure 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 3 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 average 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.

Some of the most significant schemas identified in the study are shown in Table 5. 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.

#### 4. DISCUSSION

Instead of getting in a horse race with current approaches to protein secondary structure prediction, we attempt to open a different view of the protein secondary structures by extracting regularity between sequence patterns and various structures. The regularity could be used as new features and fed into other prediction systems. In a way, SSGA could be considered a preprocessor.

There are several directions in our future work. First, though the sequence schemas are currently treated independently, they can be combined to better characterize particular secondary structures. We plan to either apply different composition operators, e.g. Boolean connectives, to combine schemas or use a higher-order models, e.g. HMM to reflect the relation among different schemas more realistically.

Second, we can apply SSGA to widely-used

protein data sets to generate useful schemas as new features for other protein secondary structure prediction tools to verify whether the schemas learned are effective.

Third, in the paper GA was applied to find the regularity in various protein secondary structures, and we described the learned regularity in terms of sequence patterns. Applying GA and using sequence patterns inevitably incur the process bias and the representation bias. These biases can either make useful inductive leaps or hinder the learning/mining process. We plan to evaluate different types of biases, and measure their usefulness in various protein domains.

## 5. ACKNOWLEDGEMENTS

The author would like to thank Yuh-Jyh Hu, Jenn-Kang Hwang, Jinn-Moon Yang, Shian-Shyong Tseng, Dai-Yi Wang, Chun-Chen Chen, and Ching-Yao Wang at National Chaio Tung University for their guidance and feedback.

## 6. REFERENCES

- [1] Y. Yu. Coiled-coils: stability, specificity, and drug delivery potential. *Adv. Drug. Deliv. Rev.* 54(8), pp. 1113.-1129, 2002.
- [2] M. Cianfriglia, C. Cenciarelli, S. Barca, M. Tombesi, M. Flego, and ML. Dupuis. Monoclonal Antibodies as a Tool for Structure-Function Studies of the MDR1-P-Glycoprotein. *Curr. Protein Pept. Sci.* 3(5). pp. 513-530, 2002.
- [3] NK. Nagradova. Three-dimensional domain swapping in homooligomeric proteins and it's functional signifiacne. *Biochemistry* 67(8). pp. 839-849, 2002.
- [4] Y. Kaizhi and A. D. Ken. Constraint-based assembly of tertiary protein structures from secondary structure elements. *Protein Sci.* 9. pp. 1935-1946, 2000.
- [5] E. S. Robert and M. T. Janet. Prediction of Strand Pairing in Antiparallel and Pararallel  $\beta$ -Sheets Using Information Theory. *Proteins: Structure, Function, and Genetics* 48. pp. 178-191, 2002.
- [6] B. Rost, R. Schneider, and C. Sander. Protein fold Recognition by Prediction-based Threading. *J. Mol. Biol.* 270. pp. 471-480, 1997.
- [7] Y. An, and R. A. Friesner. A Novel Fold Recognition Method Using Composite Predicted Secondary Structures. *Proteins: Structure, Functions, and Genetics* 48. pp. 352-366, 2002.
- [8] M. Cieplak, T. X. Hoang, and M. O. Robbins. Thermal Folding and Mechanical Unfolding Pathways of Protein Secondary Structure. *Proteins: Structure, Functions, and Genetics* 49. pp. 104-113, 2002.
- [9] B. Rost, Review: Protein Secondary Prediction Continues to Rise. *J. Structural Biology* 134. pp. 204-218, 2001.
- [10] S. Hua and Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.* 308, pp. 397-407, 2001.
- [11] JK. Rainey and MC. Goh, A statistically derived parameterization fro the collagen triple-helix. *Protein Sci.* 11(11). pp. 2748-2754, 2002.
- [12] C. K. Mathews, K. E. Van Holde, and K. G. Ahern. *Biochemistry* third edition, Addison Wesley Longman, 2000
- [13] B. Rost and C. Sander, Prediction of protein

- secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, pp. 584-599, 1993.
- [14] R. Agrawal and R. Srikant. Mining Sequential Patterns, Proc. 11th Int'l Conf. Data Eng. 1995.
- [15] M. S. Chen, J. S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. *IEEE Trans. Knowledge and Data Eng.* Vol. 10 no. 2. pp. 209-221, 1998.
- [16] K. Hatonen, M. Klemettinen, PMannila, H. Ronkainen, and H. Toivonen, Knowledge Discovery from Telecommunication Network Alarm Databases. Proc. Second Int'l Conf. Data Eng. pp. 115-122, 1996.
- [17] D. Tsur, J.R. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Query Flocks: A Generalization of Association-Rule Mining. Proc. ACM SIGMOD Int'l Conf. Management of Data. pp. 1-12, 1998.
- [18] J.T. L. Wang, G. W. Chirn, T.G. Marr, B. Shapiro, D. Shasha, and K. Zhang. Combinatorial Pattern Discovery for Scientific Data: Some Preliminary Results. Proc. ACM SIGMOD Int'l Conf. Management of Data. pp. 115-125, 1994.
- [19] J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press. 1975.
- [20] C. Y. Lee. Entropy-Boltzmann Selection in the Genetic Algorithms. *Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*. Issue: 99. pp. 1-5, 2002.
- [21] H. S. Yoon and B. R. Moon. An empirical study on the synergy of multiple crossover operators. *IEEE Transaction on Evolutionary Computation*. Volume: 6, Issue: 2. pp. 212-223, 2002.
- [22] J. E. Baker. Reducing Bias and Inefficiency in the Selection Algorithm. *Genetic Algorithm and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms*. Erlbaum. 1987.
- [23] J.M. Yang and C.Y. Kao, Combined Simulated Evolutionary Algorithm for Real Parameter Optimization. *IEEE Int. Conf. on Evolutionary Computation*. pp. 732-737, 1996.
- [24] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22(12). pp. 2577-2637, 1983.
- [25] C. Branden, and J. Tooze, *Introduction to Protein Structure*. GARLAND press, New York, 1991.
- [26] M. Michell, and M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1998.
- [27] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.