JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 19, 943-952 (2003)

Short Paper

Finding Regularity in Various Types of Secondary Protein Structures

YEN-WEI CHU AND JINN-MOON YANG*
Department of Computer and Information Science
**Department of Biological Science and Technology & Institute of Bioinformatics*
National Chiao Tung University
Hsinchu, 300 Taiwan
E-mail: ywchu@cis.nctu.edu.tw
E-mail: moon@cc.nctu.edu.tw

Protein function is structurally determined, but analyzing the structure of a protein is both difficult and time-consuming. The authors look at the natural instincts of protein secondary structures, which used to be analyzed in terms of their statistical relationship with a single amino acid. Several schemas are offered for identifying regular patterns among various types of secondary protein structures. The schemas employ genetic algorithms based on a steady-state strategy. Two disjunctive data sets were used to verify fitness. The paper concludes with some illustrations of significant schemas produced as part of this study, with brief explanations of their significance.

Keywords: secondary protein structure, schema, steady-state genetic algorithm, rule extraction, sequence alignment, substitution matrix, voting mechanism, sequential pattern

1. INTRODUCTION

The latest version of the Protein Information Resource (PIR) database (updated on June 3, 2002) contains 233,236 protein sequences. In comparison, the Protein Data Bank (PDB) only contains 18,455 protein structures since they are much more difficult to determine. The secondary structures of proteins are now considered crucial to understanding their tertiary structures [1-5]; however, even though secondary structure data is often used in protein recognition and protein structure prediction [6-11], 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 certain degree, be determined by sequential amino acid interaction [13]. Here we will propose a representative schema for amino acid interactions as an aid in analyzing the relationship between them and various protein secondary structures.

Received November 1, 2002; accepted June 5, 2003.
Communicated by Wen-Lian Hsu.

A schema can be regarded as a sequential pattern. According to its general definition, a sequential pattern is a frequently occurring pattern 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 a priori-like (statistics-based) algorithms , with different researchers using different parameter settings and constraints [15-18]. Traditional statistical methods identify significant patterns or rules according to their frequencies in data sets; in contrast, a schema also considers distinguishability.

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 powerful (because of its fitness function) and flexible (because of its problem encoding capability) [20-24]. 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.

The paper is organized as follows. Section 2 gives an overview of our approach. In section 3, we describe the methodology in our experiment and analyze our results. Finally, we discuss some issues related to our study in section 4.

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] while using schemas to reflect regularity. A schema is a bit string in which a bit is either an amino acid or an asterisk that represents any amino acid. Fig. 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. We only focused on schemas that are 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 shown in Fig. 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 was 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 varies, and c) a measure is needed to evaluate the schema's quality.

2.2 Algorithm

A decision was made to use steady-state Genetic Algorithms (SSGAs) to search for possible schemas because they are frequently used in rule-based systems [26] and schemas can be considered types of rules. Each schema that evolved was used to classify the secondary structure of a protein sequence. As shown in Fig. 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.

Fig. 2. Methodology used in this research.

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

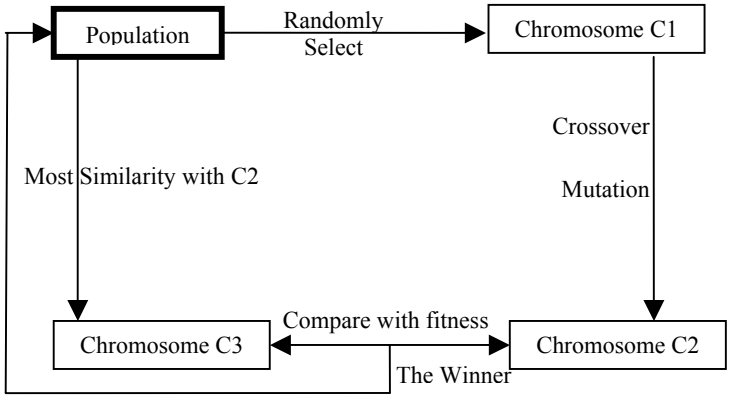

Fig. 3. SSGA flowchart.

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

To reduce the computational complexity, instead of using a single huge population of chromosomes, we initialized a population for each amino acid. In a particular population for an amino acid, e.g., *R*, each chromosome represented a potential schema nine amino acids in length with the amino acid, *R*, fixed at the center position, while the others were randomly determined. The reason we fixed *R* at the center of the schema in this case was that we wanted to model the interactions between the center *R* and the other neighboring amino acids. During the evolutionary process, the genetic operations, i.e., mutation and crossover, were applied to all the positions except the center in order to maintain the specificity showing the schemas in each particular population. An illustration of the 20 populations studied here is presented in Fig. 4. Each schema was associated with a specific secondary structure determined by its fitness.

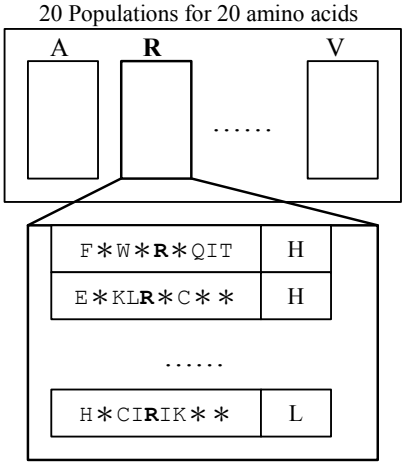

Fig. 4. A sample population for amino acid R.

The design of our fitness function was 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 can first measure its tendency under each secondary structure, defined as follows:

$$Tendency(s)_{SS} = \frac{1}{SA_{SS_{\max}} - SA_{SS_{\min}}} \sum_{SA_{SS_i} > threshold} SA_{SS_i} \tag{1}$$

where $tendency(s)_{ss}$ is the tendency of schema *s* to have a particular secondary structure *ss* and SA_{ssi} is the alignment score of schema *s* with a secondary structure ss_i in the training set. SA_{ssmax} and SA_{ssmin} are the maximum and minimum alignment scores, respectively. Note that we only considered those alignments having scores above a specified threshold. Alignments with low scores were considered noise.

We preferred schemas with greater discrimination power, that is, a good schema should have a strong tendency toward a particular secondary structure. Given all the tendencies toward various structures, we defined the fitness as

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

where $Tendency(s)_{highest}$ and $Tendency(s)_{second}$ represent the highest and the second highest tendencies of schema s, respectively.

We adopted a steady-state selection mechanism to choose candidate schemas to participate in the evolutionary process. Standard genetic operators, such as uniform crossover and multi-point mutation, were applied to generate new populations. The same evolutionary process was repeated until the fitness values of the schemas did not improve. After convergence was achieved, from all twenty populations, we combined those schemas having high fitness values to form the final significant schema set. These schemas could then be used to classify the secondary structures of new protein sequences.

3. EXPERIMENTS

3.1 Methodology and Data Sets

Our experiments had two purposes. First, we wanted to verify the positive predictive value of our system; second, we wanted to validate the fitness function. As our system was within the supervised learning paradigm, we prepared the training set and testing set, respectively. The training set consisted of 124 protein sequences, each of which was more than 80 amino acids in length, and the pairwise similarity was below 25%. They were used to train the SSGA to find significant schemas associated with various protein secondary structures. The 124 proteins are listed in Table 1. To obtain a positive predictive value, we tested the SSGA on the nr-PDB data set created by NCBI after removing those sequences used for training. A positive predictive value was defined as

$$\text{positive predictive value} = \frac{\text{number of correct classifications}}{\text{number of schema matches}} \quad (3)$$

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 1ifc 1ipd 1le4 1mgn 1mup 1ndk 1ofv 1omp 1osa 1phh 1plc 1pyp 1rhd 1rnd 1s01 1sgt 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 4gcr 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

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 could indicate that the fitness function we used could approximate real biological meanings, thus justifying its use.

3.2 Results

The statistics of the amino acids and secondary structures in the non-redundant Protein Data Bank are 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 numbers of occurrences and the percentages of the secondary structures H, E, and L within the nr-PDB, respectively.

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

	Num	**%**	**H 354429**	**H% 35.9%**	**E 210513**	**E% 21.3%**	**L 42111**	**L% 42.7%**
A	82743	8.39%	41219	**4.18%**	13582	1.38%	27942	2.83%
C	10701	1.09%	3398	0.34%	3095	0.31%	4208	**0.43%**
D	57508	5.83%	18068	1.83%	6736	0.68%	32704	**3.32%**
E	65288	6.62%	31741	**3.22%**	9616	0.98%	23931	2.43%
F	38874	3.94%	13778	**1.40%**	11807	1.20%	13289	1.35%
G	73432	7.45%	12872	1.31%	10714	1.09%	49846	**5.06%**
H	22508	2.28%	7438	0.75%	4818	0.49%	10252	**1.04%**
I	56906	5.77%	20985	**2.13%**	20959	**2.13%**	14962	1.52%
K	57486	5.83%	23243	2.36%	9637	0.98%	24606	**2.50%**
L	88394	8.96%	41502	**4.21%**	20762	2.11%	26130	2.65%
M	22057	2.24%	9477	**0.96%**	4944	0.50%	7636	0.77%
N	43029	4.36%	12045	1.22%	5784	0.59%	25200	**2.56%**
P	45803	4.65%	8320	0.84%	4076	0.41%	33407	**3.39%**
Q	37829	3.84%	17031	**1.73%**	6345	0.64%	14453	1.47%
R	50134	5.08%	21199	**2.15%**	9545	0.97%	19390	1.97%
S	57626	5.84%	16779	1.70%	10797	1.09%	30050	**3.05%**
T	57004	5.78%	15774	1.60%	14590	1.48%	26640	**2.70%**
V	71239	7.22%	22665	2.30%	28564	**2.90%**	20010	2.03%
W	13325	1.35%	5150	**0.52%**	3670	0.37%	4505	0.46%
Y	34173	3.47%	11745	1.19%	10472	1.06%	11956	**1.21%**

After the evolutionary process terminated, we checked each of the twenty converged populations to determine the most frequent secondary structures for each amino acid. We summarize the results in Table 3. It shows that most of the natural correlations between amino acids and preferred structures were also found in the converged populations with the exception of amino acid Y. Note that all the initial populations were randomly generated. The finding of similar correlations between amino acid preferences and particular structures in the final converged populations certainly provides some confidence supporting application of the fitness function to SSGA.

Table 3. Tendencies of various amino acids to have particular types of secondary structures.

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
Result	H	L E	L	H	H	L	L	H E	H L	H	H	L	L	H	H	L	L	E	H L	***E***

The learned schemas from the training set were later tested on the nr-PDB test set to measure their positive predictive values. Some of the most significant schemas identified in the study are shown in Table 4.

Table 4. Sample schemas with high fitness.

Schema	Secondary Structure	No. of schema occurrences in nr-PDB	Positive predictive value
＊＊＊A＊＊LAE	Helix	81	97.53%
＊＊＊＊PP＊＊＊	Loop	2049	95.17%
＊P＊＊＊PT＊＊	Loop	129	91.47%
＊＊＊G＊PS＊＊	Loop	201	89.05%
＊＊VVI＊＊＊＊	Sheet	348	80.46%
＊＊＊E＊LLR＊	Helix	58	89.66%
＊＊＊＊＊P＊＊S	Loop	2777	79.87%
＊＊R＊N＊P＊＊	Loop	305	78.69%
K＊＊＊E＊L＊D	Helix	160	76.25%
＊＊A＊E＊＊＊K	Helix	461	75.49%
＊＊VVL＊S＊＊	Sheet	93	75.27%

4. DISCUSSION

Instead of getting in a horse-race with current approaches to protein secondary structure prediction, we have attempted to study protein secondary structures from a different point of view by extracting regularity between sequence patterns and various structures. This regularity could be used as new features and fed into other prediction systems. In this way, SSGA could be used as a preprocessor.

There are several directions for our future work. First, though sequence schemas currently are 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 higher-order models, e.g., HMM, to reflect the relationships 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 and to verify whether the learned schemas are effective.

Third, in this paper, GA was applied to find regularity in various protein secondary structures, and we have described the learned regularity in terms of sequence patterns. However, applying GA and using sequence patterns inevitably incurs process bias and representation bias. These biases can either lead to 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.

ACKNOWLEDGEMENT

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

REFERENCES

1. Y. Yu, "Coiled-coils: stability, specificity, and drug delivery potential," *Advanced Drug Delivery Reviews*, Vol. 54, 2002, pp. 1113-1129.
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," *Current Protein Peptide Science*, Vol. 3, 2002, pp. 513-530.
3. NK. Nagradova, "Three-dimensional domain swapping in homooligomeric proteins and it's functional significance," *Biochemistry*, Vol. 67, 2002, pp. 839-849.
4. Y. Kaizhi and A. D. Ken, "Constraint-based assembly of tertiary protein structures from secondary structure elements," *Protein Science*, Vol. 9, 2000, pp. 1935-1946.
5. E. S. Robert and M. T. Janet, "Prediction of strand pairing in antiparallel and parallel β-sheets using information theory," *Proteins: Structure, Function, and Genetics*, Vol. 48. 2002, pp. 178-191.
6. B. Rost, R. Schneider, and C. Sander, "Protein fold recognition by prediction-based threading," *Journal of Molecular Biology*, Vol. 270. 1997, pp. 471-480.
7. Y. An and R. A. Friesner, "A novel fold recognition method using composite predicted secondary structures," *Proteins: Structure*, *Functions*, *and Genetics*, Vol. 48, 2002, pp. 352-366.
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*, Vol. 49, 2002, pp. 104-113.
9. B. Rost, "Review: Protein secondary prediction continues to rise," *Journal of Structural Biology*, Vol. 134, 2001, pp. 204-218.
10. S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of Molecular Biology*, Vol. 308, 2001, pp. 397-407.
11. J. K. Rainey and M. C. Goh, "A statistically derived parameterization fro the colla-

gen triple-helix," *Protein Science*, Vol. 11, 2002, pp. 2748-2754.
12. C. K. Mathews, K. E. Van Holde, and K. G. Ahern, *Biochemistry*, 3rd edition, Addison Wesley Longman, 2000.
13. B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, Vol. 232, 1993, pp. 584-599.
14. R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of 11th International Conference on Data Engineering*, 1995, pp. 3-14.
15. M. S. Chen, J. S. Park, and P. S. Yu, "Efficient data mining for path traversal patterns," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 10, 1998, pp. 209-221.
16. K. Hatonen, M. Klemettinen, P. Mannila, H. Ronkainen, and H. Toivonen, "Knowledge discovery from telecommunication network alarm databases," in *Proceedings of Second International Conference on Data Engineering*, 1996, pp. 115-122.
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," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, 1998, pp. 1-12.
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," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, 1994, pp. 115-125.
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, 2002, pp. 1-5.
21. H. S. Yoon and B. R. Moon, "An empirical study on the synergy of multiple crossover operators," *IEEE Transactions on Evolutionary Computation*, Vol. 6, 2002, pp. 212-223.
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*, 1987, pp. 14-21
23. J. M. Yang and C. Y. Kao, "Combined simulated evolutionary algorithm for real parameter optimization," *IEEE International Conference on Evolutionary Computation*, 1996, pp. 732-737.
24. W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, Vol. 22, 1983, pp. 2577-2637.
25. C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Press, New York, 1991.

Yen-Wei Chu (朱彥煒) received the B.S. and M.S. degree in Computer Information Science from National Chiao Tung University. From 1998, he has been a Ph.D. candidate in Department of Computer Information Science, National Chiao Tung University. His research interests include bioinformatics, genetic algorithms and learning systems.

Jinn-Moon Yang (楊進木) received the M.S. degree from National Central University, Chung-Li, Taiwan, in 1994, the M.B.A. degree from Tamkang University, Taipei, Taiwan, and the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001. In 1985-1998, he worked for PanChiao Telecommunications, Ministry of Transportation and Communications. In 1998-2001, he worked for Chunghwa Telecom Training Institute, Taipei, Taiwan. He has been an assistant professor with Department of Biological Science and Technology Computer Science & Institute of Bioinformatics, National Chiao Tung University, Taipei, Taiwan since 2001. He has published more than 35 technical papers in various journals and conference records. His research interests include evolutionary computation, bioinformatics, structural biology, rational drug-design design, and machine learning.