國立交通大學

資訊工程系

博 士 論 文

蛋白質二級結構的規則性及其應用

Finding Protein Secondary Structure Regularity and Related Applications

研 究 生：朱彥煒

指導教授：孫春在　教授

楊進木　教授

中 華 民 國 九 十 五 年 六 月

蛋白質二級結構的規則性及其應用
Finding Protein Secondary Structure Regularity and Related Applications

研 究 生：朱彥煒　　Student：Yen-Wei Chu

指導教授：孫春在　教授　　Advisor：Prof. Chuen-Tsai Sun

楊進木　教授　　Prof. Jinn-Moon Yang

國 立 交 通 大 學

資 訊 工 程 系

博 士 論 文

A Dissertation

Submitted to

Department of Computer Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

蛋白質二級結構的規則性及其應用

學生：朱彥煒　　指導教授：孫春在教授
楊進木教授

國立交通大學資訊工程學系

摘　　要

本論文從序列的觀點，討論蛋白質二級結構的規則性。我們定義一個模式以表示蛋白質二級結構的規則性，並提出一個分群-穩態基因演算法來找尋符合蛋白質二級結構規則性的模式。在方法的驗證上，針對所提的演算法利用消去研究法則，証明加入分群的概念是有效果的；並與資料探勘上常用的關聯規則和決策樹等方法做比較，本方法的確在蛋白質資料集中，有相對優異的表現。在應用上，我們分析 PSIPRED 和 PROF 這二種方法在做蛋白質二級結構的預測時，有某些區域都是無法猜對的，但利用我們所提出來的模式，可改進此區域約 40% 至 60% 左右的預測正確率。進一步，我們將所找到的二級結構模式結合 PSIPRED 和 PROF 的預測結果，可改進目前二級結構的預測。此外，我們亦以此實驗提出一個教案，以符合問題導向式的學習在生物資訊課程上的教學。

Finding Protein Secondary Structure Regularity and Related Applications

Student：Yen-Wei Wu　　Advisors：Dr. Chuen-Tsai Sun

Dr. Jinn-Moon Yang

Department of Computer Science
National Chiao Tung University

ABSTRACT

The author explores protein secondary structure regularity from the perspective of sequences. Regularity is defined in terms of a schema discovered by a cluster-based genetic algorithm. Two steps taken to validate the algorithm were a) finding the weightiness of cluster and b) comparing the approach with data mining methods. Schemata were used to address secondary structure predictions for residues that PSIPRED and PROF could not predict. The results indicate that the proposed schemata can improve prediction accuracy for these residues by approximately 40% and 60% for the CB513 and RS126 data sets, respectively. Furthermore, schemata combine the prediction results of PSIPRED and PROF to improve secondary structure prediction. A bioinformatics teaching plan using a problem-based approach is discussed.

Acknowledgements

I am indebted to several mentors, especially my advisor, Dr. Chuen-Tsai Sun, who showed concern and support for my efforts. He taught me both the proper approach to and attitude toward problem solving. Professor Jinn-Moon Yang taught me the importance of completeness in a study, professor Yuh-Jyh Hu helped me overcome research bottlenecks many times, and professor Jenn-Kang Hwang gave me many valuable suggestions.

I also wish to thank several classmates. Chin-Sheng Yu kindly provided advice on biological principles and bioinformatics tools; we met frequently via MSN whenever I came to a dead end. Ching-Yao Wang gave me many useful suggestions regarding methods. Dai-Yi Wang, Hsu-Chih Wu, and other fellow lab members also assisted me in developing this research.

Finally, this dissertation could not have been completed without the endless love and support from my father, mother, and the rest of my family. I owe all of my achievements to them and send them my love and gratitude.

Contents

List of Tables

List of Figures

Chapter 1.

Introduction

1.1 Motivation

Protein sequences consist of different combinations of a four-letter DNA alphabet (A, G, C and T) that is used to create a 20-word vocabulary of native amino acids. Genes are considered the blueprint or library of life, and proteins the machinery. Proteins are macromolecules that perform all-important tasks in organisms, including the catalysis of biochemical reactions, nutrient transport, and signal recognition and transmission.

Protein function is determined via a three-dimensional structure [1]. Researchers know that determining protein structure in a laboratory is much more difficult than identifying protein sequence. This explains why as of March 6, 2006 the Protein Information Resource (PIR) database contained 2,826,393 protein sequence records while the Protein Data Bank (PDB) contained only 35,343 protein structure records [2, 3].

Independent researchers and an organization known as the Critical Assessment of Techniques for Protein Structure Prediction (CASP) currently support the practice of predicting protein structure from previously known sequences [4, 5, 6, 7, 8].

Protein secondary structure is very valuable information for predicting 3D protein structure. In many applications (e.g., identifying protein functions, classifying proteins, establishing phylogenetic trees), protein structure knowledge requires information on protein secondary structure. [9, 10, 11]. The present research—analyzing the natural instincts of protein secondary structure and its potential for assisting in protein secondary structure prediction—was motivated by the bottleneck that secondary structure researchers are currently dealing with [12].

1.2 Study Importance

Protein secondary structure is considered crucial to understanding protein tertiary structure [13, 14, 15, 16, 17]. However, even though secondary structure data is often used for protein recognition and protein structure prediction [18, 19, 20, 12, 21, 22], 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 [23], some researchers are suggesting that secondary structure formation may (at

least to a certain degree) be determined by sequential amino acid interaction [24]. At the center of this thesis is a proposed representative schema for amino acid interactions as an aid for analyzing their relationship with various protein secondary structures.

One challenge is uncovering schema details—that is, the regularity of protein secondary structures. To avoid predictive deviation in the learning stages of various methods, data sets such as RS126 or CB513 usually have low sequence identity for protein secondary structure. The proposed solution to this problem described in this thesis involves a cluster-based genetic algorithm, since traditional data mining methods (e.g., arm and decision trees) cannot be used with such kinds of data sets.

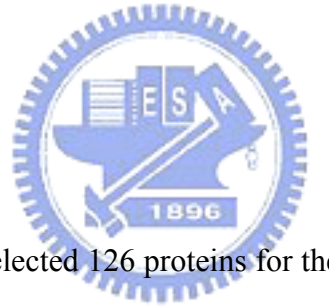
1.3 Thesis Organization

A review of related studies is presented in Chapter 2. The chapter will also include a discussion concerning the construction of a data set from pdb_select list (except for RS126 or CB513), clustering methods, protein secondary structure prediction, and problem-based learning. Details on the defined schema and the steady-state strategy that was incorporated into the genetic algorithm are presented in Chapter 3, along with an analysis of the proposed cluster-based genetic algorithm. Two applications (predicting protein secondary structure and creating a teaching plan for bioinformatics) will be described in Chapters 4

and 5, respectively. Conclusions and suggestions for future research will be given in Chapter 6.

Chapter 2.

Related Work



2.1 Data Sets

Rost and Sander (1993) selected 126 proteins for the training and testing of secondary structure prediction algorithms [24]. Their definition of non-redundancy states that no two proteins in a set share more than 25% sequence identity over a length of more than 80 residues. Unfortunately, the RS126 set contains protein pairs that are very similar in terms of sequence according to methods considered more sophisticated than sequence identity percentage. Cuff and Barton's CB513 dataset [25], consisting of 513 chains with low similarity, has been used to evaluate classifier accuracy. Almost all sequences found in the RS126 set are included in the CB513 set. Both are non-homologous, but CB513 homology measurement is more strict than for the RS126 set.

In addition to RS126 and CB513, we established a data set based on the PDB_select protein chain list. The chain list is a representative of PDB chain identifiers that researchers use in order to 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 [26].

2.2 Clustering (K-means)

K-means is one of the simplest unsupervised learning algorithms capable of solving the well-known clustering problem [27]. Its main idea is to define one k centroid for each cluster. Care must be taken with centroid placement because different locations will lead to different results. The best approach is to place them as far away from each other as possible. The next step is to take each point belonging to a given data set and forge an association between it and the nearest centroid. When no points are pending, the first step is completed and an early groupage is performed. At this point it is necessary to re-calculate k new centroids as barycenters of clusters produced in the previous step. The

appearance of k new centroids means that more binding must be performed between the same data set points and the new set of nearest centroids. This generates a loop that allows for the step-by-step observation of changes in k centroid locations until no more changes are required (i.e., the centroids stop moving). This algorithm minimizes the chosen distance between a data point and cluster center.

The algorithm consists of four steps:

1. Place K points into the space represented by the objects to be clustered. These points represent initial group centroids.

2. Assign each object to the group containing the closest centroid.

3. After all objects have been assigned, recalculate the K-centroid positions.

4. Repeat steps 2 and 3 until the centroids stop moving. This produces groups for calculating the metric to be minimized.

Although the procedure always terminates at some point, the k-means algorithm does not necessarily find the most optimal configuration that corresponds to a minimum global objective function. The algorithm is also significantly sensitive to the initial cluster centers that are randomly selected—though it can be run multiple times to reduce this effect. For

this reason, the k-means algorithm has been adapted for use with many problem domains [28, 29, 30, 31, 32].

2.3 Genetic Algorithms

Holland's original genetic algorithms [33] included a well-known heuristic algorithm inspired by Darwin's theory of evolution ("survival of the fittest"). Later efforts by Goldberg and others have allowed genetic algorithms to be applied to optimization and search problems in many fields [34, 35, 36, 37, 38, 39, 40]. Genetic algorithms do not always find optimal solutions, but in large search spaces they are more efficient than most exhaustive search techniques in attaining near-optimal solutions.

For any given problem, genetic algorithms alternate between working on coding space and solution space [41]. Coding space work involves the need to know how to transfer real problems into chromosomes and to work with chromosomal evolution. These chromosomes are evaluated in the solution space. The major parts of simple genetic algorithm operations are shown in Figure 2.1.

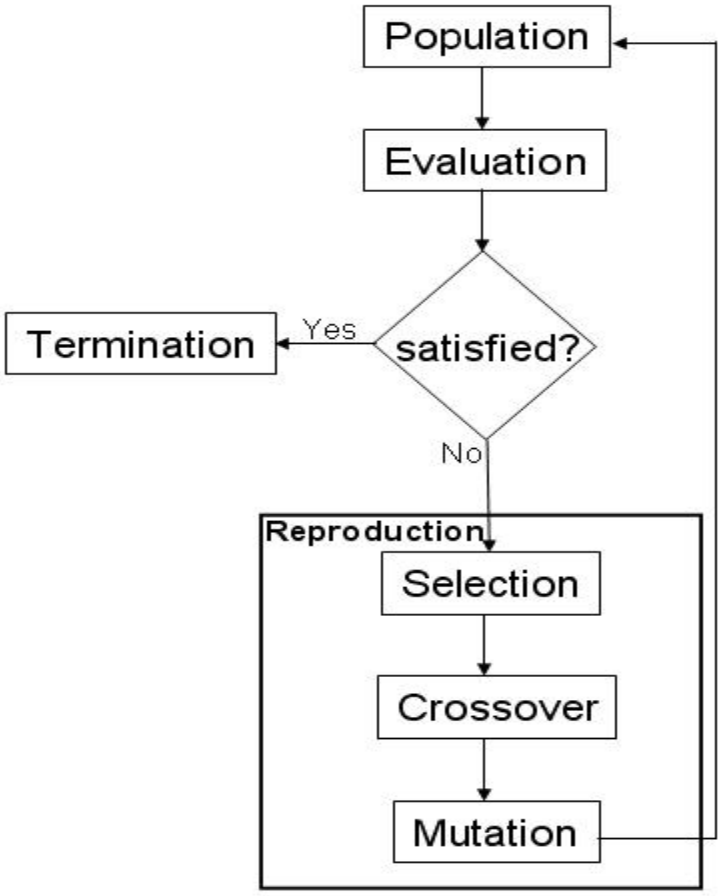


Figure 2. 1: Genetic algorithm flowchart.

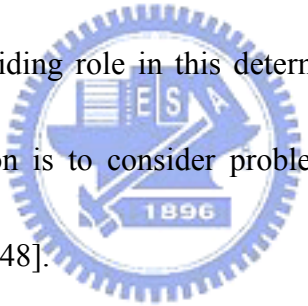
2.3.1 Initializing the Population

A population consists of a set number of chromosomes, with each chromosome serving as a candidate solution. A chromosome consists of genes, with each gene serving as a feature of a problem. The feature called genotype in a gene and phenotype in a

problem. At the beginning of the evolutionary process, a binary code or character is randomly assigned to each gene in a chromosome. Through competition among chromosomes in a population, either one or a set of chromosomes eventually satisfies pre-established requirements.

2.3.2 Fitness Function

For a given problem, a specific fitness function must be designed to determine whether a chromosome is a good candidate for survival [42, 43, 44]. In a genetic algorithm, the fitness function plays a guiding role in this determination—in other words, the dual purposes of the fitness function is to consider problem characteristics and to assemble domain knowledge [45, 46, 47, 48].



2.3.3 Selection

Each chromosome has a fitness value (score) that is determined by the fitness function. Chromosomes with higher fitness values are considered more fit for survival, have a higher probability of producing offspring, and tend to dominate other chromosomes in a

population. However, higher scores do not guarantee that a chromosome contains good genes only, nor do low scores indicate a complete lack of genes for positive characteristics. Accordingly, the presence of niche chromosomes must be taken into account when designing a genetic algorithm [49, 50, 51, 52, 53].

2.3.4 Crossover

Each pair of chromosomes has what is called a crossover rate—that is, a probability for proceeding crossover. Based on a pre-assigned crossover rate, two chromosomes randomly exchange their genetic information [54, 55]. One-point or two-point crossovers entail cutting and exchanging genes, whereas uniform crossover genes are exchanged according to a random template. Examples of these crossover types (all commonly found in genetic algorithms [56, 57, 58] are shown in Figure 2.2.

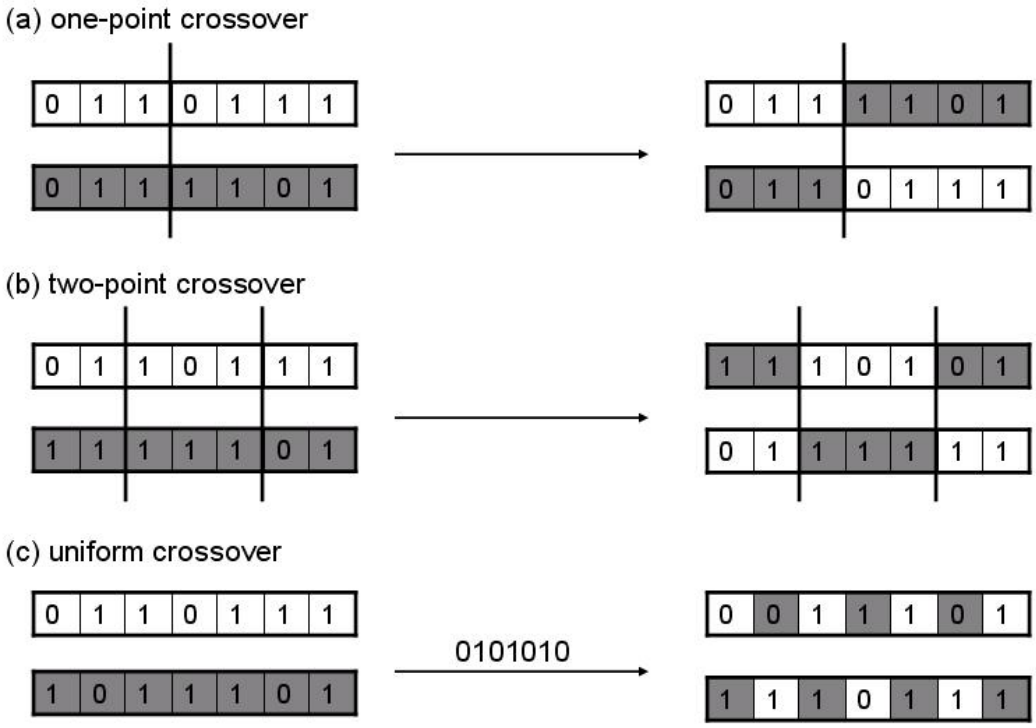


Figure 2. 2: Three crossover examples.

2.3.5 Mutation

Each chromosome has a mutation probability called a mutation rate. Based on pre-assigned mutation rates, individual genes are randomly chosen to change their value from 0 to 1 or from 1 to 0 (Fig. 2.3a) [59, 60, 61]. An example of multi-point mutation is shown in Figure 2.3b. In that figure, P1, P2, and P3 are three pre-assigned probabilities. P1

is much larger than the others and P2 is bigger than P3. In addition to avoids falling into the local optima area, mutations also maintain chromosome diversity [62, 63].

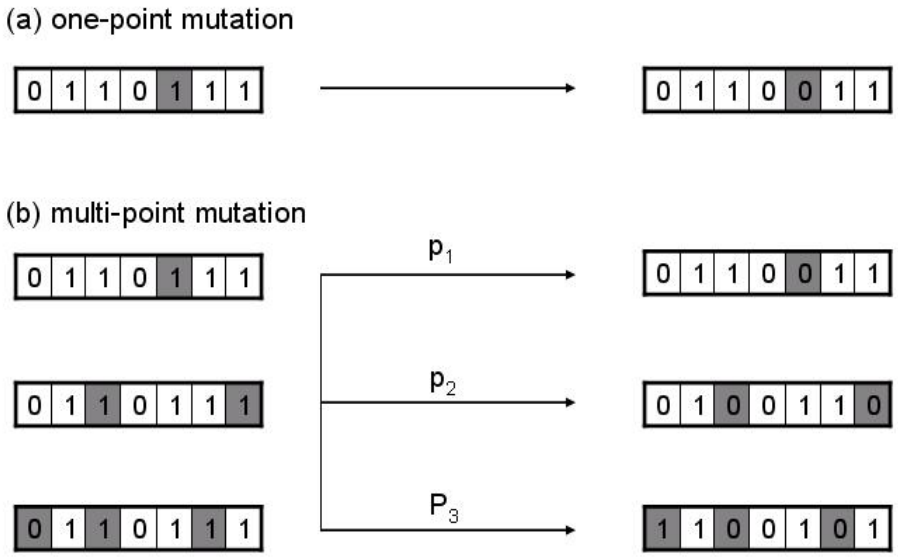


Figure 2. 3: Two mutation examples.

2.4 Protein Secondary Structures

In 1951, biologists Linus Pauling and Robert Corey proposed two kinds of periodic protein structures: alpha helix and beta sheet (Fig. 2.4) [64, 65]. In 1957 their proposal was

confirmed via x-ray diffraction [66, 67], which describes the chemical structure of a protein based on the primary structure. Later research determined that protein secondary structures express local spatial structure in certain linear segments.

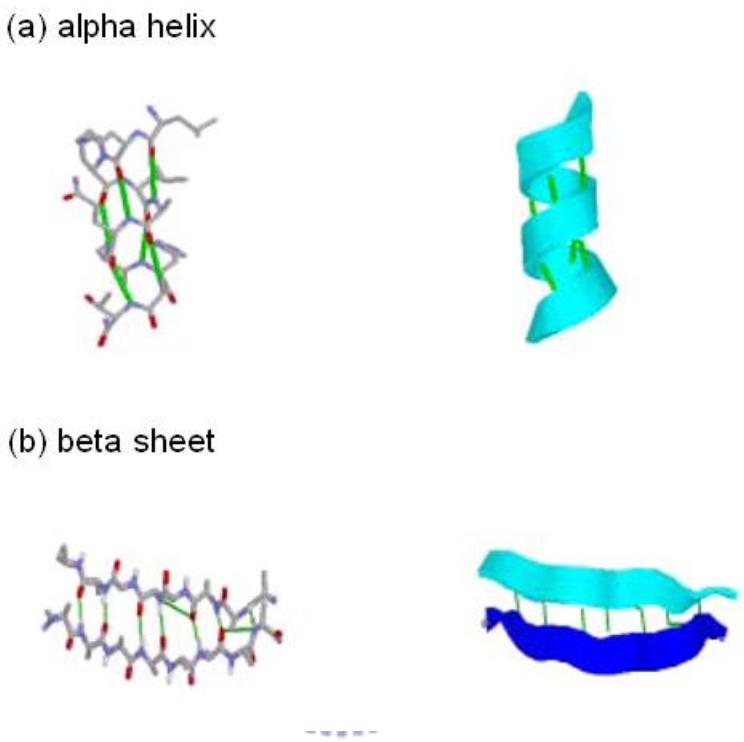


Figure 2. 4: Illustrations of alpha helix and beta sheet.

A randomly generated protein chain may have a loop structure. Achieving a stable conformation requires a large number of weak bonds (e.g., hydrogen bonds, salt bridges and van der waal interactions). Stable conformations are called protein secondary structures. So far, there are 90% residues be located in alpha helix or beta sheet in the database.

2.4.1 Classification

Protein secondary structures have many classifications. The three most common are DSSP, STRIDE, and DEFINE [68, 69, 70]. DSSP (Database of Secondary Structure in Protein), a widely applied classification for protein secondary structure, includes a computer program for defining various features of a protein via a PDB protein structure file. DSSP files include data on secondary structure, molecular properties, and solvent accessibility. Seven DSSP codes for protein secondary structures are shown in Table 2.1.

Table 2. 1: DSSP codes and their meanings

DSSP Code	Protein Structure
H	alpha helix
B	residue in isolated beta-bridge
E	extended strand, participates in beta ladder
G	3-helix (3/10 helix)
I	5-helix (pi helix)
T	hydrogen bonded turn
S	bend

Protein secondary structures are usually predicted using three of the seven DSSP codes: H (helix), E (sheet) and L (loop; this is sometimes referred to as C, coil) [71, 72, 73].

The five categories for the three kinds of DSSP codes are shown in Table 2.2; it is important to note category choice has an important effect on protein secondary structure prediction accuracy [71]. Jones [74] has shown that the fifth category in Table 2.2 performs best for protein secondary structure prediction, but the first category is more commonly used for comparisons with the PHD approach. In 1999, Baldi proposed three new categories: H (H, G, I), E (B, E) and C (T, S) [75].

Table 2. 2: Five categories of merged codes for the three DSSP codes.

	H	E	L
1	H, G, I	E	B, T, S
2	H, G	E, B	I, T, S
3	H, G	E	I, B, T, S
4	H	E, B	G, I, T, S
5	H	E	G, I, B, T, S

2.4.2 Prediction

Most secondary structure prediction methods make use of the fact that segments of consecutive residues have preferences for certain secondary structure states [76, 77]. The

prediction problem is thus transformed into a pattern-classification problem that can be addressed by pattern recognition algorithms, with the guiding goal being to predict whether the residue at the center of a segment of 13-21 adjacent residues has a helix, strand, or no regular secondary structure (loop or coil).

Before the protein secondary structure hypothesis was proven and accepted, biologists tried a variety of approaches to predict protein secondary structure, including the use of protein sequences [78]. All of these methods can now be placed in three categories based on their original assumptions [12]. These categories can also be described in terms of generations.

Secondary structure prediction methods in the first generation focused on four types of residues: helix, sheet, loop former and breaker. Protein secondary structure segments were predicted by considering the characteristics of a single residue [79]. These methods assume that when an amino acid forms a secondary structure, the amino acid acts independently. However, we now know that amino acids are affected by their adjacent amino acids, therefore, accuracy for this method is approximately 50-60%. Method names include Chou & Fasman, GOR1, and Lim [79, 80, 81].

Second generation methods consider local information in residues 3-51, using a fixed window size for a protein sequence and a sliding window for cutting several segments.

Secondary structures are retrieved from these segments. Second-generation method accuracy is only about 60-65% due to a lack of long-distance information—for example, information on the effect of hydrogen bonds between two amino acids separated by a long distance. Method names include GOR3[82], Levin et al. [83], Nishikawe and Ooi [84], Qian and Sejnowski [85], Holley and Karplus [86], Asai et al.[87], and Yi and Lander [88].

Third generation methods added evolution information to the second generation concept [12]—that is, gene mutation occurs as part of the evolutionary process, meaning that one amino acid can be replaced by another. Accordingly, proteins with similar structures may have different amino acids in the same position. Almost all third generation methods take into account multiple sequence alignment results when inputting data into a learning model such as neural networks or SVM. The best-known third generation method, PHD, can reach 70% accuracy or higher for Q3 predictions and over 80% for helix predictions.

Method names include Zvelebil et al. [89], PREDATOR [90, 91], NNSSP [92], DSC [93], PHD [24], Jnet [94], PSIPRED [74], Baldi et al. [75, 95] and HMMSTR [96].

2.4.3 Evaluation

Rost and Sander's (1993) arrangement of evaluative methods for protein secondary structure prediction is shown as Table 2.3. Its evaluative method parameters have been placed in a 3x3 matrix (for three kinds of secondary structures).

Table 2. 3: Matrix for nine parameters of evaluative methods.

A_{HH}	A_{HE}	A_{HC}
A_{EH}	A_{EE}	A_{EC}
A_{CH}	A_{CE}	A_{CC}

In the matrix, A_{ij} is the number of those residues that belong to secondary structure i but are predicted for secondary structure j.

To sum up each element in the column, a_i, is the predictive number for each secondary structure.

$$a_i = \sum_{\forall j} A_{ji} \text{ , for } i = H, E, C$$

To sum up each element in the row, b_i, is the number for each secondary structure.

$$b_i = \sum_{\forall j} A_{ij} \text{ , for } i = H, E, C$$

To sum up all elements in the matrix, b, is the number of residues.

$$b = \sum_{\forall i} a_i = \sum_{\forall i} b_i$$

For examples, the secondary structure H has ($A_{HH} + A_{HE} + A_{HC}$) residues, and there are ($A_{HH} + A_{EH} + A_{CH}$) residues predicted to H.

Overall 3-state accuracy, Q3, is a score for secondary structure prediction [97, 85, 12, 88, 98, 99]. It is the most popular evaluative method and shown as follows,

$$Q_3 = \frac{\sum_{\forall i} A_{ii}}{b} x100$$

On the other hand, we can simply discuss the evaluation for each secondary structure. There are two kinds of evaluative methods for predictive accuracy discussed. One show the predictive accuracy of secondary structure i,

$$Q_i = Q_i^{obs} = \frac{A_{ii}}{b_i} x100 \text{ , for } i = H, E, C$$

The other show the percentage that how many residues are predicted correctly in the predictive number of secondary structure *i*.

$$Q_i^{pre} = \frac{\sum_{\forall i} A_{ii}}{a_i} x100 \text{ , } for\ i = H, E, C$$

Matthew's correlation coefficient, C, is also usually discussed when measure the accuracy of secondary structure shown as follows [100].

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}} \text{ , } for\ i = H, E, C$$

p_i is those who residues are belong to secondary structure *i*, and the predictive result is also *i*.

$$p_i = A_{ii} \text{ , } for\ i = H, E, C$$

n_i is those who residues are not belong to secondary structure *i*, and the predictive result is not *i*.

$$n_i = \sum_{\forall j \neq i} \sum_{\forall k \neq i} A_{jk} \text{ , for } i = H, E, C$$

u_i is those who residues are belong to secondary structure i, and the do not be predicted to i.

$$u_i = \sum_{\forall j \neq i} A_{ij} \text{ , for } i = H, E, C$$

o_i is those who residues are not belong to secondary structure i, but the predictive result is i.

$$o_i = \sum_{\forall j \neq i} A_{ji} \text{ , for } i = H, E, C$$

Chapter 3.

Materials and Methods

3.1 Process data set

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 [26].

3.1.1 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.

3.1.2 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:

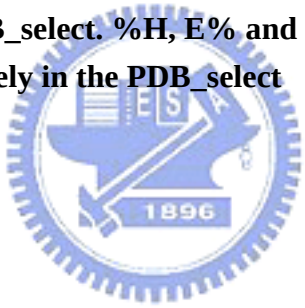
- incorrect PDB identification in the 25% list;
- no information in the DSSP files;
- broken chains; or
- inclusion of an unknown symbol X.

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.

3.1.3 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 3.1 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 3.1 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.

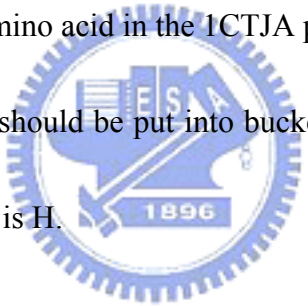
Table 3. 1: Statistics for 20 amino acids in the PDB_select chain set. % is the percent of each amino acid in the PDB_select. %H, E% and L% is the percent of each secondary structure respectively in the PDB_select



	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%
Y	8799	3.53%	2936	**33.4%**	2959	**33.6%**	2904	**33%**
V	17039	6.84%	5465	32.1%	6978	**41%**	4596	27%

3.1.4 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 3.1. 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.



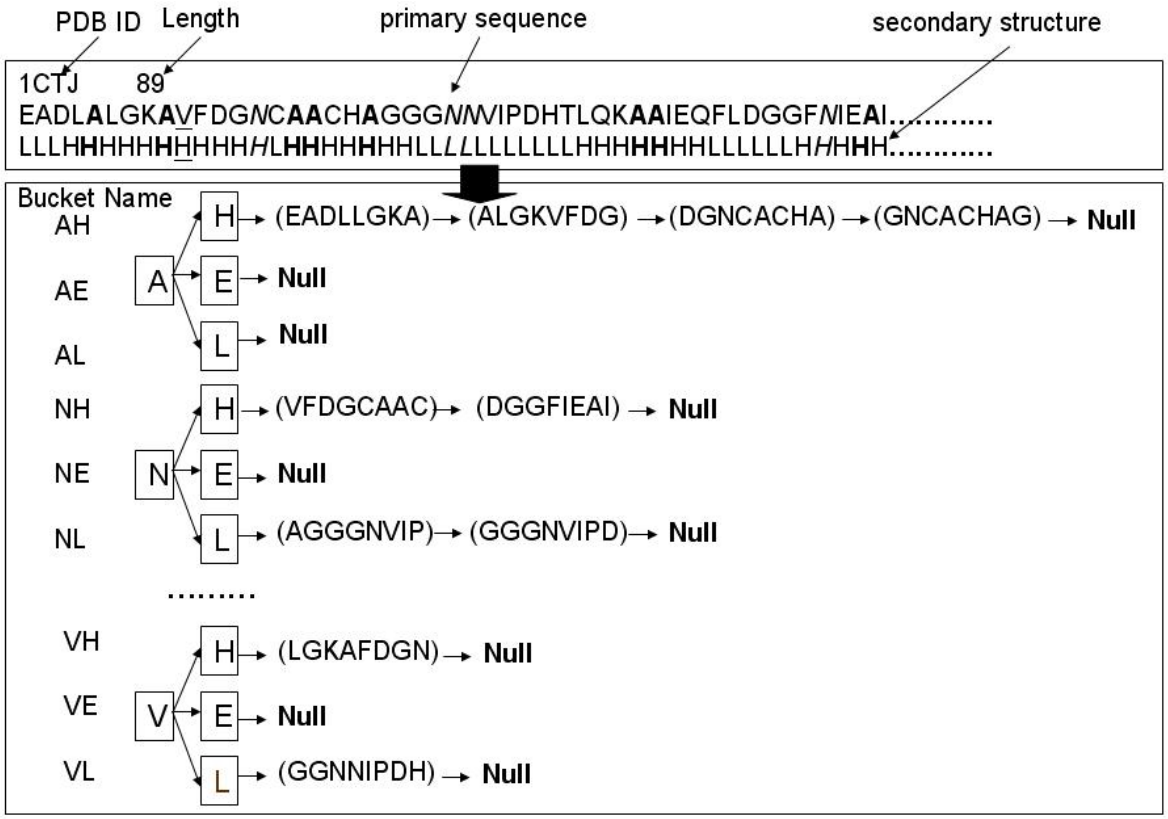


Figure 3. 1: An example of using sequence 1CTJ to make a training set.

3.2 Schema

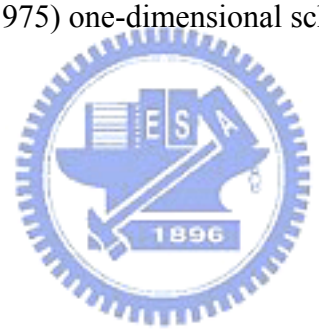
Protein secondary structures are designated as H (alpha helix, 3/10 helix, pi helix), E (beta bridge, beta ladder), or L (turn, bend) [76]. The regularity of secondary structures (which consist of amino acids and one secondary structure) are usually discussed in terms of factors that cause amino acids 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 [88], we decided to underplay the remote effect in order to simplify schema design.

Representation

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

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



(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 3.2, 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 an *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 [101].

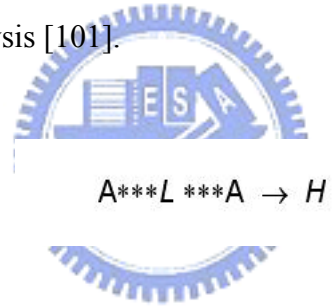


Figure 3. 2: Schema example.

3.3 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. 3.3).

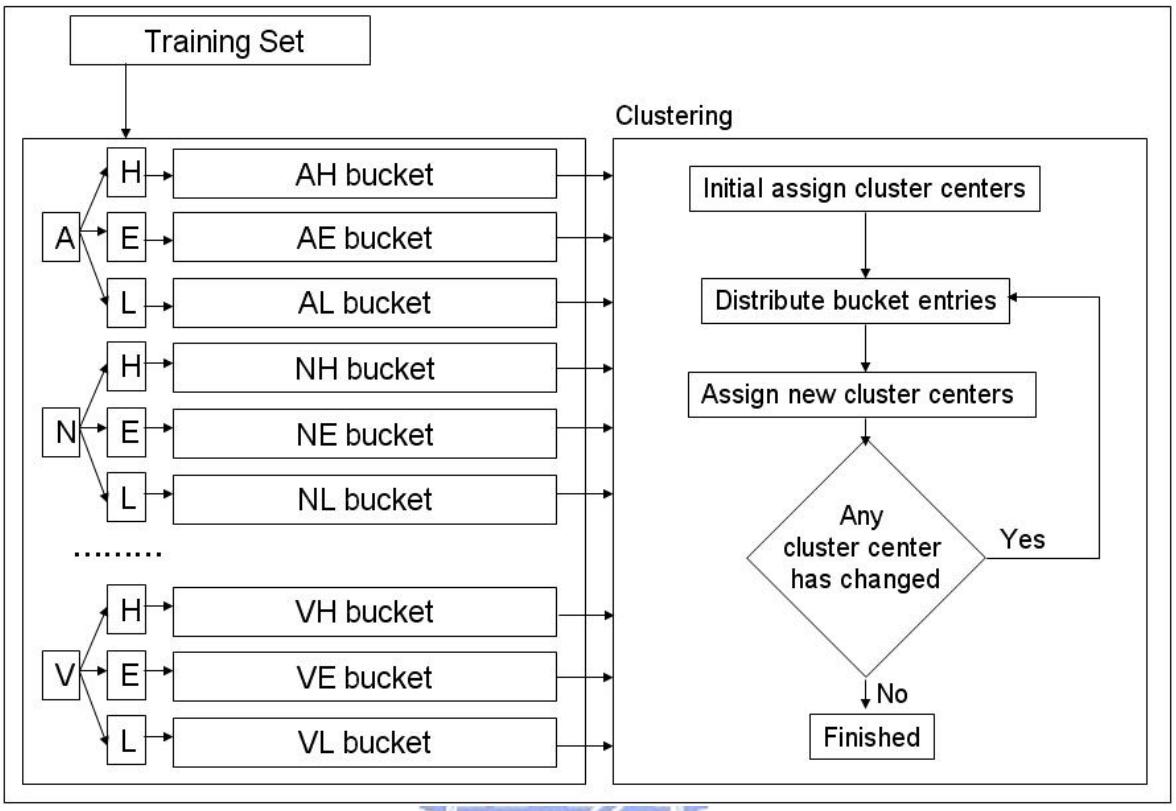


Figure 3. 3: Our proposed clustering strategy.

The fitness function gives evolutionary direction to chromosomes [102]. 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 3.4, 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.

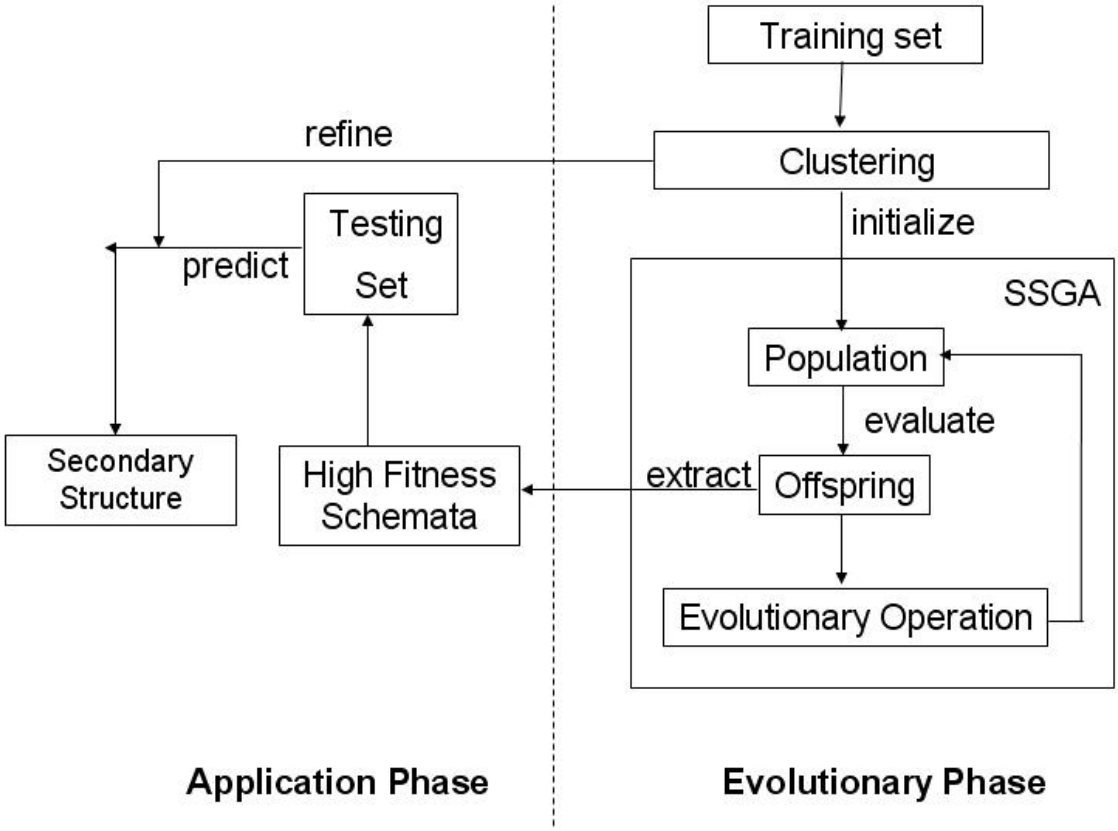


Figure 3. 4: Our cluster-based genetic algorithm for mining schemata and its application for predicting protein secondary structures.

3.3.1 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,

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

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

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

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

$$fitness = confidence * DR \quad (3)$$

3.3.2 Steady-state Reproduction

The initial step in the steady-state strategy shown in Figure 3.5 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.

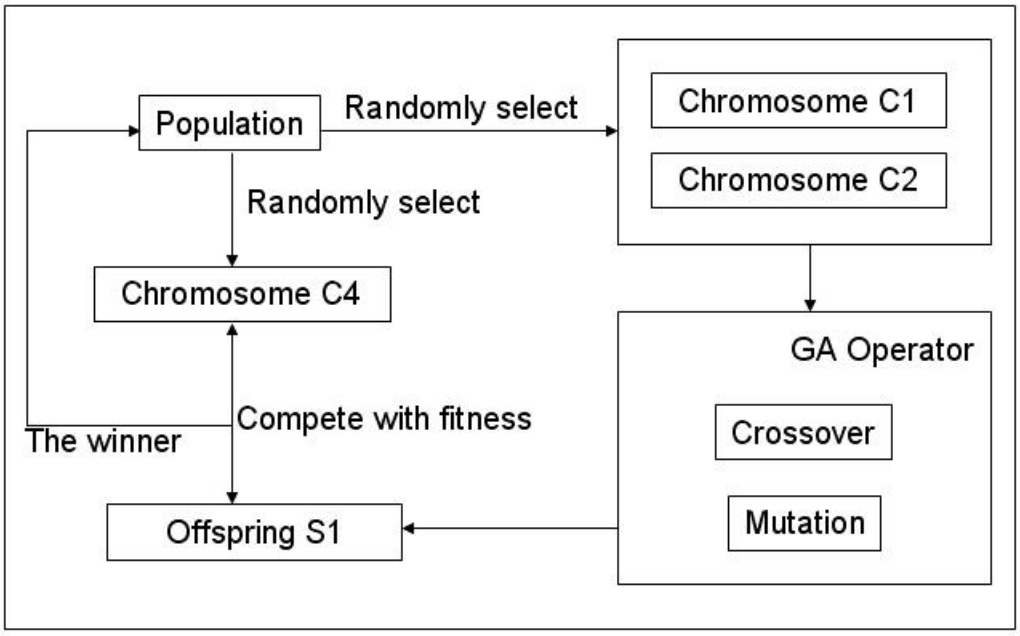


Figure 3. 5: Steady-state strategy for our cluster-based genetic algorithm.

3.4 Compare with Associate Rule

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 [24]). 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

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

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

To reduce time complexity, we adopt FP-growth algorithm for association rule mining to avoid generating candidates from the frequent itemsets [103]. 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 3.2) 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 3.2 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%).

Table 3. 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

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 3.3. 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.

Table 3. 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*

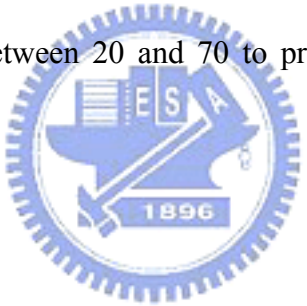
The learned schemas from the training set were later tested on the nr-PDB test set to measure their confidence and support values. Finally, there are 904 total possible rules to be found. The average confidence value is 61.51% and nearly half of mined rules are over 70%. Table 3.2 is the testing results 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 3.3 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.

3.5 Experimental Results

3.5.1 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 3.6.



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

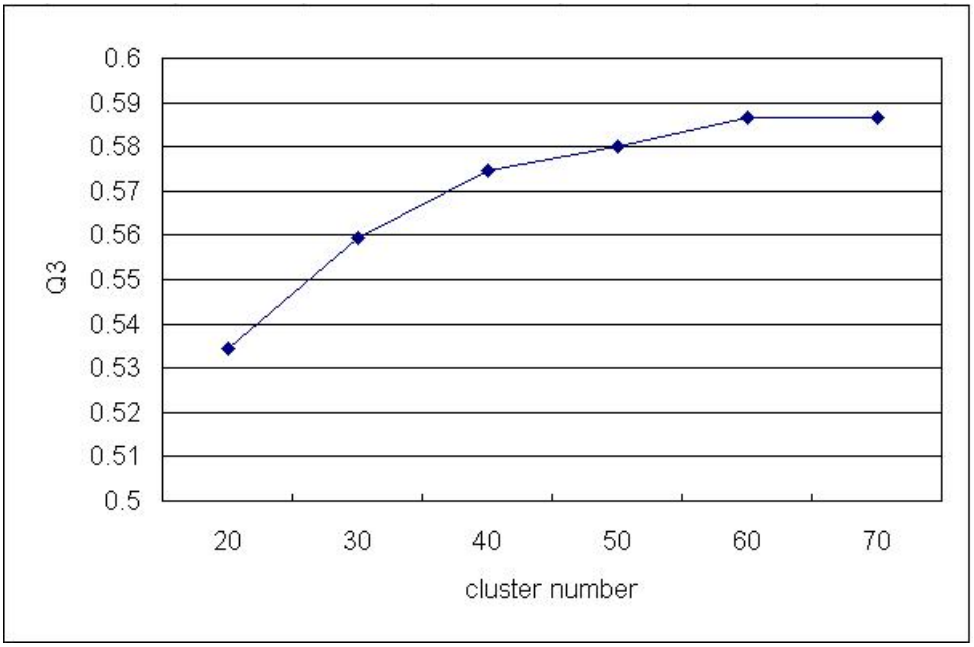


Figure 3. 6: Q3 accuracy in different cluster numbers using our approach.

3.5.2 Illustrate Some Interesting Schemata

Table 3.4 presents a comparison of our Table 3.1 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.

Table 3. 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

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 3.5.

Table 3. 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

Chapter 4.

Predictive Tools for Protein Secondary Structure

Even though the protein folding process may require catalysts [104], it is widely accepted that the three-dimensional structure of a protein is associated with its amino acid sequence [105]. This implies the possibility of predicting protein structure from a sequence. However, with the increasing number of amino acid sequences generated by large-scale sequencing projects and the continuing shortage of data on crystallized homologous structure, the need for reliable structural prediction methods is greater than ever.

Making accurate comparative assessments of different secondary structure prediction methods is difficult because they use different learning process datasets and different secondary assignments [106]. Still, a number of authors have designed methods

with accuracies above the 70% threshold by taking advantage of multiple sequence alignments [24, 92, 93, 107, 108] or selected alignment fragment pairs [91]. Most methods do not take the long-distance (beta sheet) effect into consideration because it is difficult to incorporate this feature into a model. Accordingly, secondary structure prediction accuracy appears to have reached its current limits. Analyses of several predictive tools indicate that approximately 12% of data set residues (dead areas) cannot be predicted. The complete schemata for all proteins have not yet to be identified because of a need for additional protein information. However, tests indicate that the schemata described in this paper can improve dead area prediction accuracy by 40% to 60%.

4.1 EVA

EVA (EValuation of Automatic protein structure prediction) is a plan for evaluating protein structure predictive tools [109]. Its users can evaluate tools associated with secondary structure, comparative modeling, and threading. EVA constantly downloads the latest protein structure data from PDB. Structures are added to mySQL databases; after sequences are extracted for each protein chain, they are sent to prediction servers via META-PredictProtein (META-PP), which collects the results and sends them to EVA. Each week EVA runs alignment programs for sequence searches and structure

databases to determine homologues. Secondary structure predictions, inter-residue contact predictions, and comparative modeling are evaluated by personnel at EVA satellites (Columbia University, Rockefeller University, and CNB Madrid). Employees at the central EVA site at Columbia University collect all assessments from the other two centers as well as results from database searches, then publishes the information on its main web site. Mirror web sites are maintained at the other EVA satellite locations.

EVA has evaluated at least 10 types of secondary structure predictive methods. Two of these methods, PSIPRED and PROF, were selected for this experiment, based on their proven predictive abilities and their accessibility in terms of downloads.

4.2 PSIPRED and PROF

A two-stage neural network has been used to predict protein secondary structure based on position-specific scoring matrices generated by PSI-BLAST. This approach, proposed by Jones in 1999, is called PSIPRED. PSIPRED used a new test set based on 187 unique folds and three-way cross-validation based on structural similarity criteria rather than previously favored sequence similarity criteria. Its predictive accuracy achieved an average Q3 of 76.5% to 78.3%, depending on the definition of observed secondary structure.

The three stages of this prediction method are generating a sequence profile, predicting an initial secondary structure, and filtering the predicted structure (Fig. 4.1). The dual goals are to generate sequence profiles and to predict secondary structure. Standard approaches to generating sequence profiles are considered cumbersome and time-consuming. The PSI-BLAST method uses profiles as direct input to secondary structure prediction rather than extracting sequences and creating an explicit multiple sequence alignment as a separate step. The time-consuming multiple sequence alignment task is eliminated by using PSI-BLAST profiles directly. The final position-specific scoring matrix from PSI-BLAST is used as neural network input. The matrix has 20 x M elements, with M representing the target sequence length and each element representing the log-likelihood of a particular residue substitution at a template position based on a weighted average of BLOSUM62 matrix scores for the given alignment position.

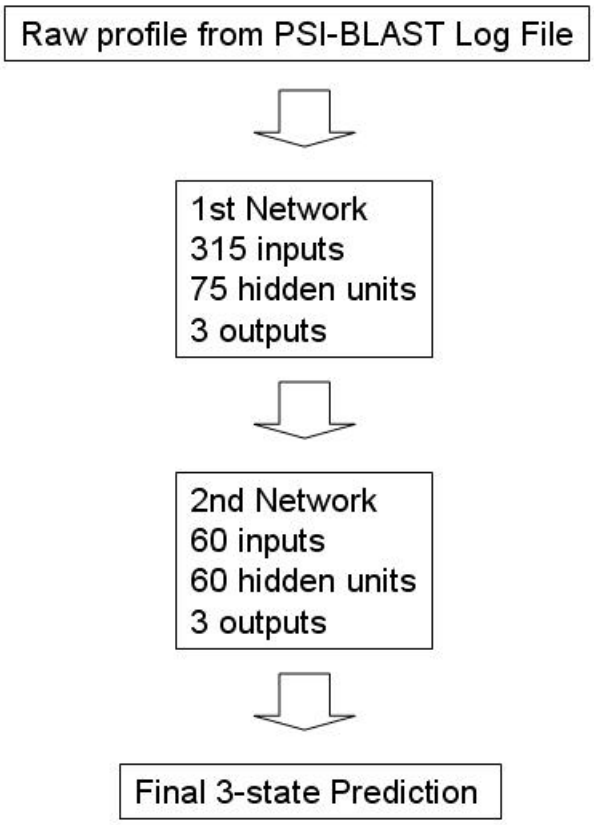


Figure 4. 1: PSIPRED flowchart.

PSIPRED utilizes a standard feed-forward back-propagation network architecture [110] with a single hidden layer. A window of 15 amino acid residues (producing an overall Q3 score of 80.1%) is considered optimal, therefore the final input layer consists of 315 input units divided into 15 groups of 21 units each. A large hidden layer of 75 units was used, with another three units (representing the three states of secondary structure—helix, strand or coil) being used to create the output layer. As with previous neural network secondary structure prediction methods [24], a second network is used to

filter successive outputs from the main network. Since only three inputs are necessary for each amino acid position, this network has an input layer of only 60 units divided into 15 groups of equal size. In this project, a smaller hidden layer of 60 units was used for this network.

PROF is a method proposed by Rost [111]. However, the author has created a downloadable version for predicting secondary structures. PROF is described as an improved version of PHDsec—a profile-based neural network predictor of protein secondary structure.

4.3 Experiment and Results

The two purposes of this experiment were to locate the shared bottleneck of the three generation methods in predicting protein secondary structures—in other words, determining if some residues exist that neither PSIPRED nor PROF can predict. The region that contains these residues, known as the “dead area,” is shown in Figure 4.2. The second purpose was to activate the dead area by inserting the proposed schemata.

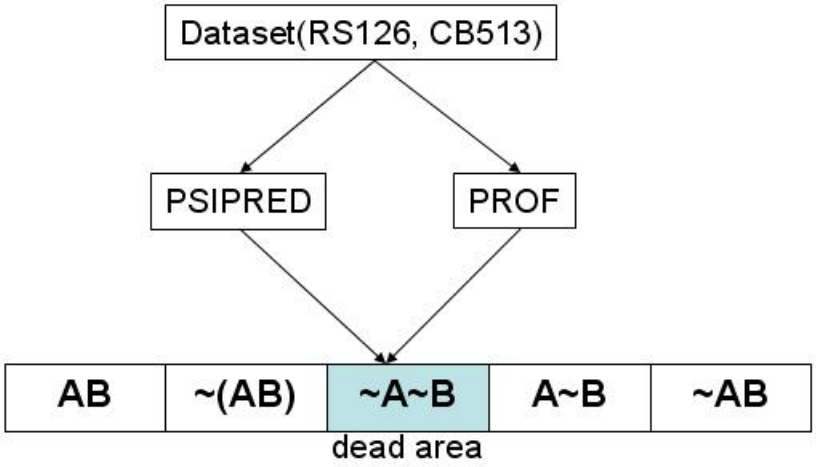


Figure 4. 2: Flowchart for generating AB, ~AB, ~A~B, A~B, and ~AB classifications.

PSIPRED and PROF predictive results are shown in Table 4.1. The results were used to define the following symbols:

A: successful PSIPRED prediction area,

~A: failed PSIPRED prediction area,

B: successful PROF prediction area,

~B: failed PROF prediction area.

PSIPRED and PROF predictive results were observed simultaneously and divided according to five classifications:

AB: areas where PSIPRED and PROF produced the same successful prediction,

~(AB): areas where PSIPRED and PROF produced the same failed prediction,

~A~B: areas where PSIPRED and PROF produced different predictions, both of them failed,

A~B: areas that PSIPRED predicted successfully but PROF did not, and

~AB: areas that PROF predicted successfully but PSIPRED did not.

Table 4. 1: PSIPRED and PROF prediction accuracy percentages for the two data sets.

	RS126	CB513
PSIPRED	80.9%	80.5%
PROF	76.3%	76.1%

The percentages of these five classifications for data sets RS126 and CB513 are shown in Table 4.2. The data indicate type AB percentages that exceed 70% for both sets, meaning that third-generation secondary structure predictive methods that include evolution information can improve accuracy to 70%. The percentage of the type A~B classification was double that of type ~AB, meaning that PSIPRED outperformed PROF. Furthermore,

the two methods made an identical but incorrect prediction—type ~(AB)—less than 1% of the time, indicating a 98% prediction confidence when the same result was predicted by both methods. The last type (~A~B) represents the dead area, which neither was able to predict, but with different results; coverage for this area was 12%. Accordingly, the upper boundary for secondary structure prediction accuracy for third generation methods is approximately 88%.

Table 4. 2: Percentages of each prediction classification for the two data sets.

	AB	~(AB)	~A~B	A~B	~AB
RS126	70.5%	1.0%	12.4%	10.3%	5.8%
CB513	70.4%	1.0%	12.8%	10.1%	5.7%

The proposed schemata were applied to dead areas for the purpose of improving secondary structure predictions. A schemata experiment flowchart is shown in Figure 4.3. In the first part of the experiment, predictions were generated by PSIPRED and PROF for the RS126 and CB513 data sets. The two predictive results were compared for the purpose of defining the dead area. The second part of the experiment focused on using the proposed cluster-based genetic algorithm to derive schemata from PDB_select. Each case was run

several times using different cluster numbers to predict RS126 and CB513 secondary structures.

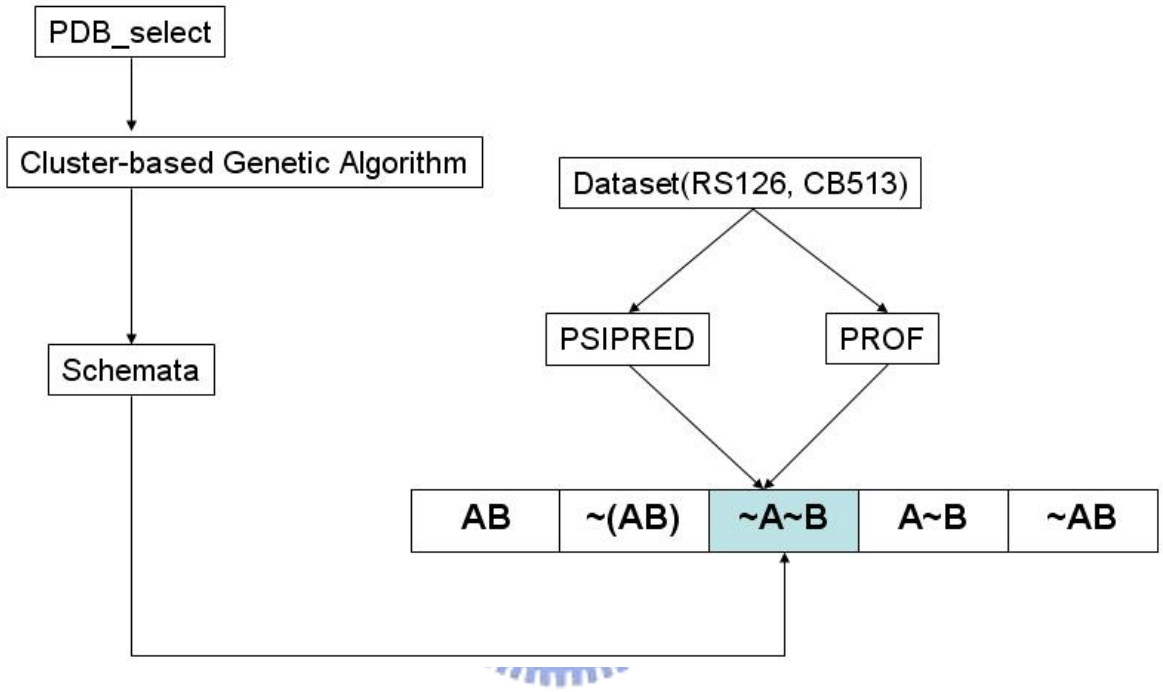


Figure 4. 3: Schemata-generating flowchart for addressing dead areas.

Although the predictive ability of the proposed schemata did not surpass that of the third-generation prediction methods, it did produce balanced predictive results according to the five classifications described above. It is therefore suggested that the proposed schemata can be used to assist PSIPRED and PROF in predicting secondary structures in dead areas. We observe the accuracy of all data set and dead area only in the different

parameter value of cluster number. Predictive accuracies for all RS126 and CB513 sequences produced by the proposed schemata are shown in Figure 4.4. The highest prediction accuracy figures for RS126 (73%) and CB513 (60%) were achieved when cluster number equaled 70. PSIPRED and PROF were capable of 80.9% and 80.5% accuracy for RS126 and CB513, respectively, but neither method was capable of correctly predicting any residues in dead areas—in other words, their predictive accuracy for dead areas was 0%. Dead area prediction accuracies using the proposed schemata were 58% for RS126 when the cluster number was 70 and 38% for CB513 when the cluster number was 60 (Fig. 4.5). Figure 4.6 presents data for when the proposed schemata were used to predict all sequences and dead areas in RS126. As shown, in each case accuracy increased. However, for CB513 the dead area prediction accuracy increased slowly as the predictive accuracy for all sequences increased (Fig. 4.7).

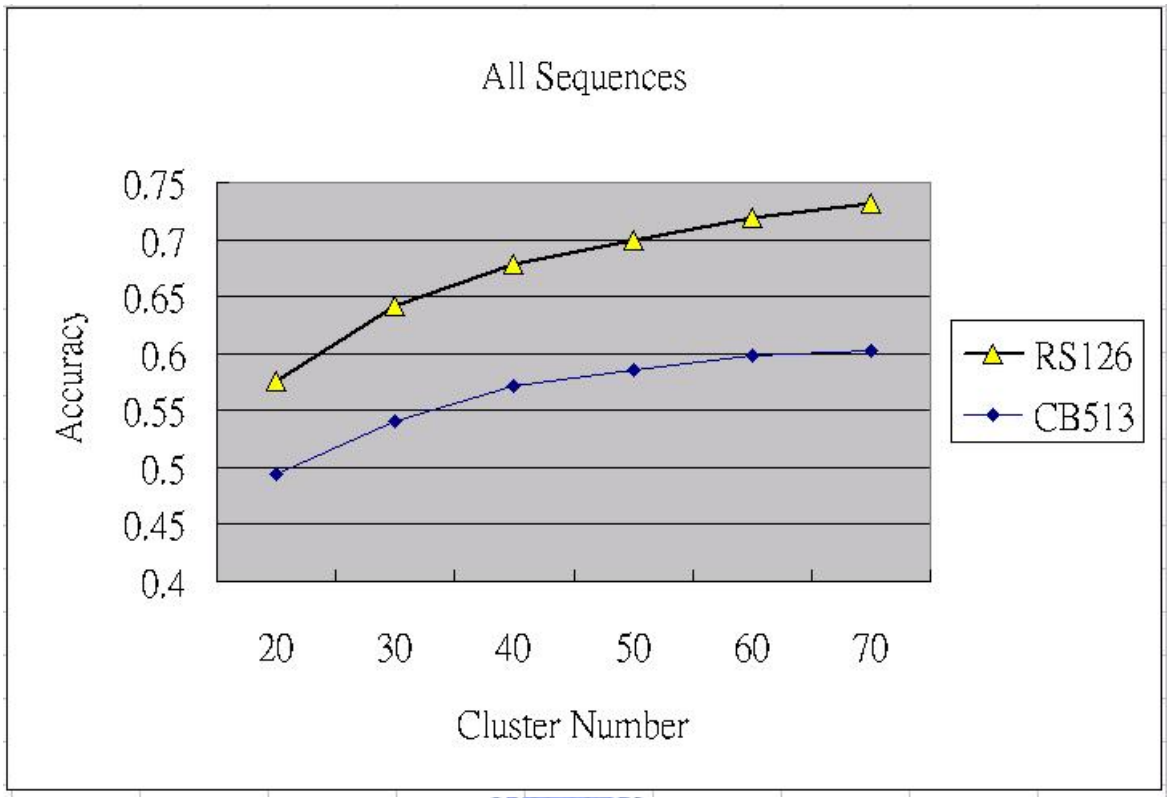


Figure 4. 4: Accuracy data for all sequences of the data sets at different cluster numbers.

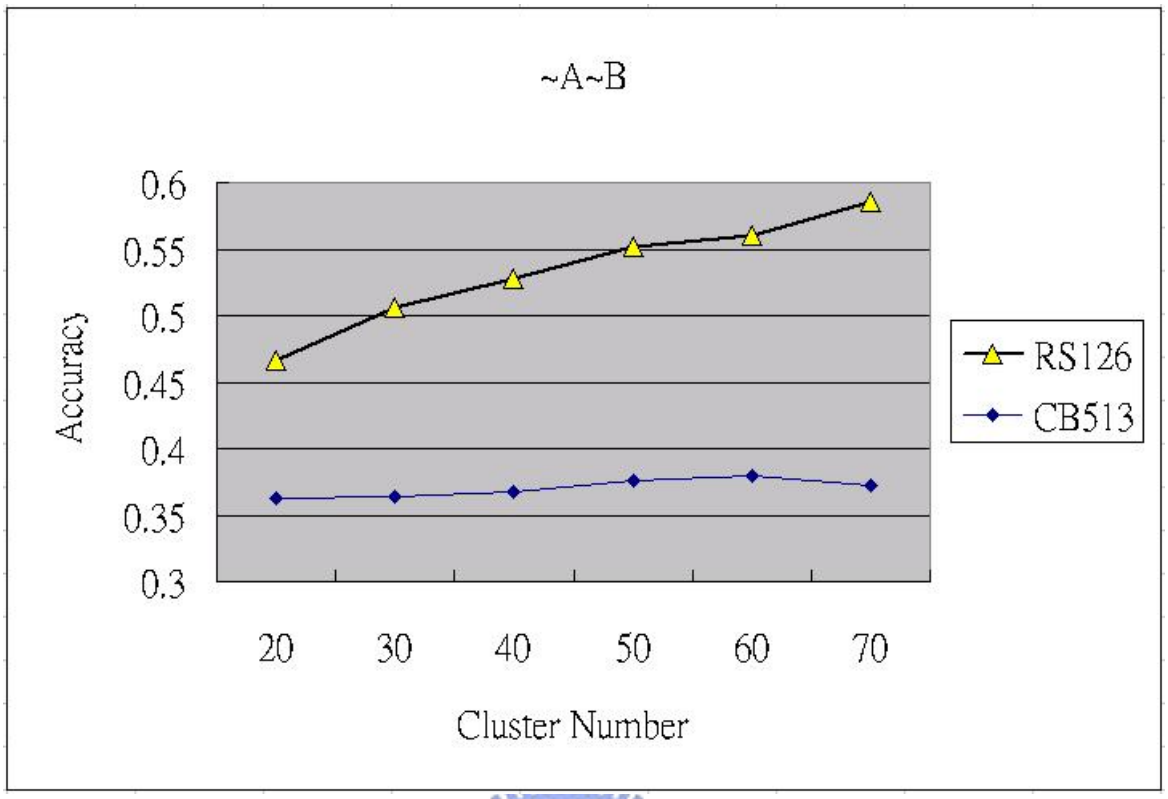


Figure 4. 5: Accuracy data for the ~A~B classification of the data sets at different cluster numbers.

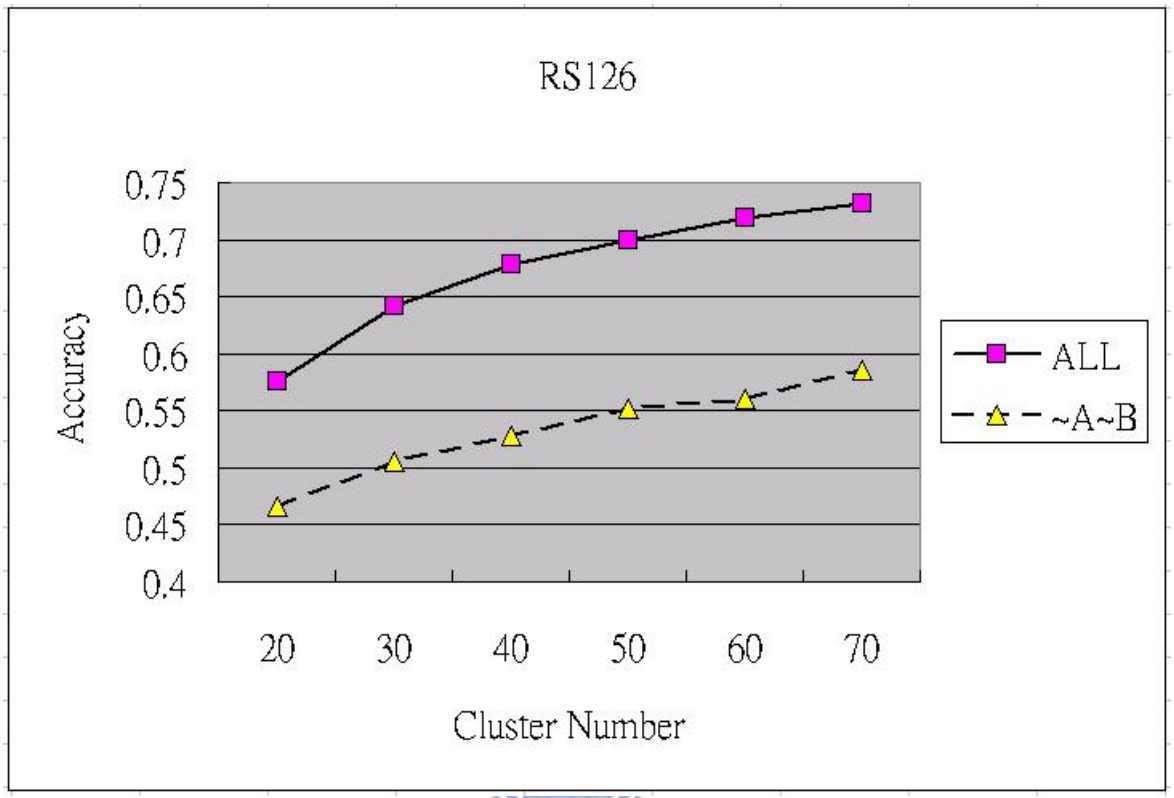


Figure 4. 6: Accuracy data for all sequences and ~A~B classification for data set RS126.

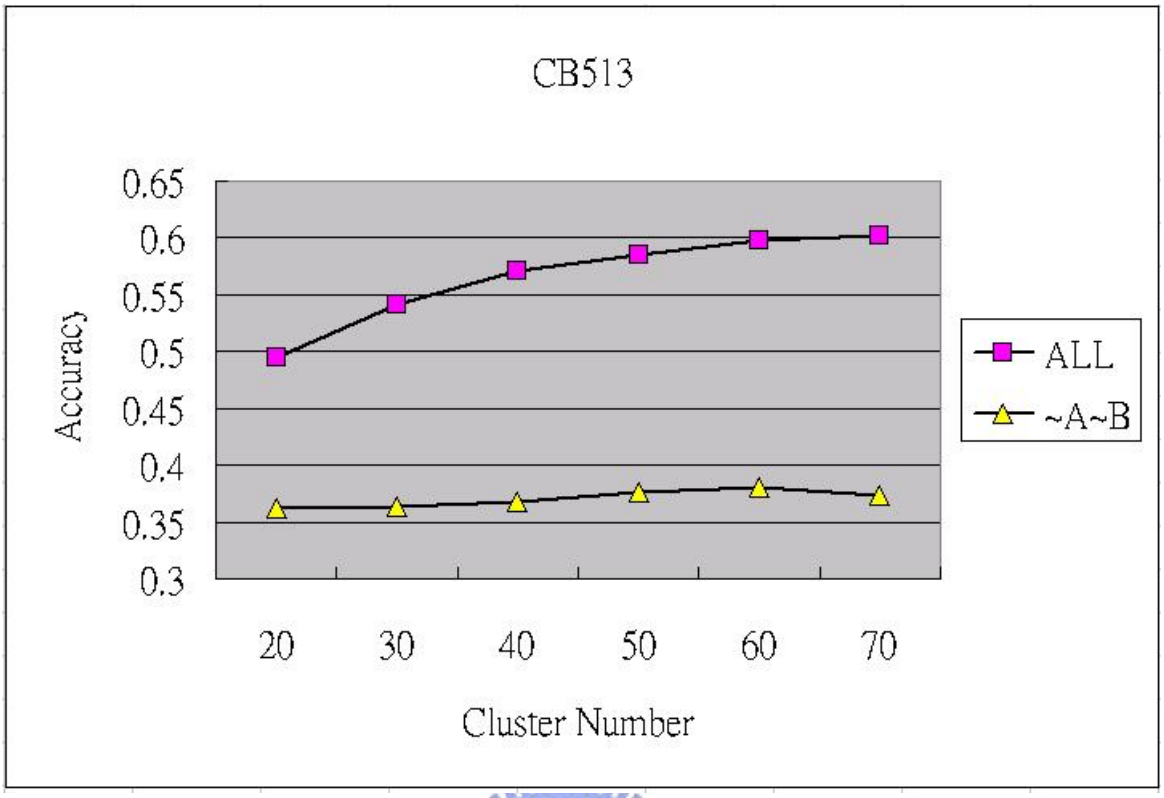
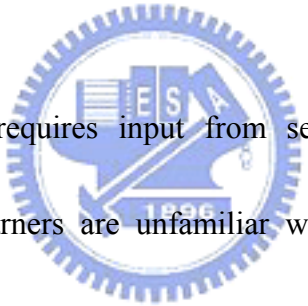


Figure 4. 7: Accuracy data for all sequences and ~A~B classification for data set CB513.

Chapter 5.

A Teaching Plan for Bioinformatics



Bioinformatics research requires input from several different domains, but the majority of bioinformatics learners are unfamiliar with specific biological issues. We propose an approach that combines problem-based learning and concept map methodology to realize and construct the biological problems. As part of the problem-solving process, learners must gather materials and identify essential knowledge—thus creating a scenario conducive to learner training. We believe this approach will be of great use to non-biologist learners in the bioinformatics field.

The human genome project has attracted a large number of information science researchers to work in the area of bioinformatics. Of particular interest to these researchers is the development and refinement of algorithms for culling meaningful information from

large bodies of data. However, information science experts have little understanding of biology, and only a handful of biologists understand information algorithm requirements.

In this section, we will propose a problem-based learning approach that makes use of concept maps for bioinformatics learning. Our goals are to a) create a process through which information specialists can easily identify the core issues of biology problems, and b) reduce research costs associated with applying information theory to biology problems.

5.1 Introduction

5.1.1 Bioinformatics

In 1989, the U.S. National Institutes of Health invited James D. Watson—best known for describing the double-helix structure of DNA—to establish a human genome research center. The guiding objective for researchers from the United States and 17 other countries has been to identify over 3 billion DNA sequences that make up the human genetic code. The project has generated an enormous amount of data that needs to be organized and

analyzed. This has lead to an explosion in research in the field of bioinformatics, which combines the domains of information science and biology. Communication among researchers in the two fields is critical to achieving research success.

5.1.2 Problem-Based Learning

Problem-based learning—an idea that originated in medical education in the 1960s—is learner-centered rather than instructor-centered [112, 113, 114, 115]. It is considered not only a curriculum organizing method, but also an instructional strategy and learning process for dealing with poorly structured real world problems [116, 117]. According to Wegner et al. (1998), the process involves a) defining the problem, b) determining whether information is lacking, c) collecting and categorizing related information, d) identifying content and learning targets, e) examining methods for solving the problem, and f) finding optimal solutions [118].

Learners must train themselves in problem solving and communication skills in order to manage and apply learning information [119]. Instructors are viewed as partners, consultants, advisors, or trainers.

5.1.3 Concept Maps

Novak used the meaningful learning theory of American cognitive psychologist David Ausubel to establish a concept map instructional strategy [120]. The method emphasizes the integration of old and new concepts into newer concept skeletons.

5.2 Instructional Design

The five categories of bioinformatics applications are a) establishing and integrating databases, b) analyzing sequences, c) analyzing structure and function, d) analyzing experimental data, and e) managing knowledge [121]. Bioinformatics knowledge has four properties: a) a database for storing raw or processed data from a biology experiment, b) a

simulation that embodies molecules for easy observation and analysis, c) one or more tools for solving specific problems, and d) a package in which related tools are integrated.

The primary goal of a problem-based learning approach is actively transmitting information in a manner that encourages knowledge construction. It is an approach that is well suited to teaching scientific principles and properties [122]. Learners construct meaningful knowledge on their own. Cognition helps in terms of adaptability—the integration of new data with previous experiences instead of the discovery of specific entities. In other words, individuals build knowledge through an adaptation process [123, 124]. When constructing knowledge in interactive environments, learners must address and resolve cognitive conflicts based on past experiences that have received repeated confirmation.

Barrows (1985) lists the five primary characteristics of problem-based learning as:

1. Using problems as the starting point of learning.

2. Using problems that are not well structured and without standard answers.

3. Regarding problems as learning content.

4. Valuing small group learning over a teacher-centered approach.

5. Helping learners understand that they must accept responsibility for their learning [125]. Teachers serve as coaches who help learners practice cognitive skills.

Figure 5. 1 presents the process of our problem-based bioinformatics instruction approach based on these characteristics are listed in Table 5.1.

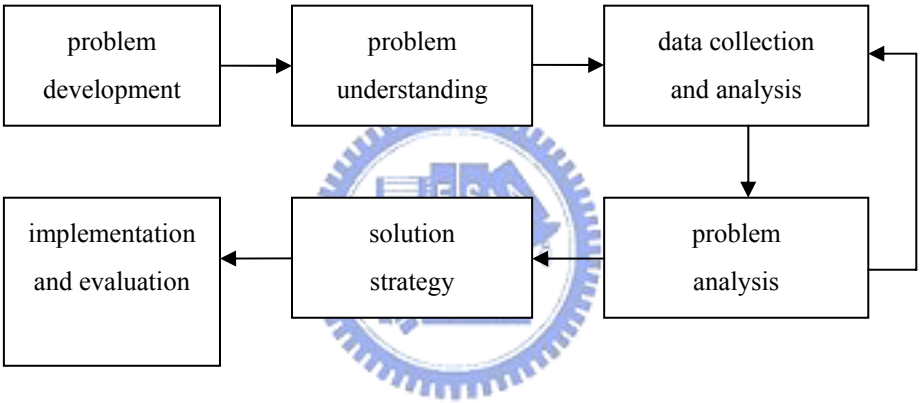


Figure 5. 1: Implementation flowchart for problem-based approach to teaching bioinformatics.

Table 5. 1: Implementation table for problem-based approach to teaching bioinformatics.

STEP	STEP NAME	ELABORATION
1	Problem development	1.1 Problem design: open-ended and poorly structured on a biological topic.
2	Problem understanding	2.1 Hypothesis: pose and ponder question. 2.2 Construct concept maps: determine knowledge needed to solve problem.
3	Data collection and analysis	3.1 Data sources: networks, books, magazines, specialists, and CDs. 3.2 Sharing: small group discussion and evaluation of sources and data.
4	Problem analysis	4.1 Thinking: Who, What, When, Where, Why and How.
5	Solution strategy	5.1 Evaluation: from correct and useful information.
6	Implementation and evaluation	6.1 Display concept maps: construct knowledge relationships and propose problem strategy. 6.2 Propose result for biologists to evaluate and analyze.

We adopted three types of concept maps for our approach:

1. Spider-web Maps

In spider web maps, links connect minor types of major concepts; each minor concept can be extended in a manner that leads to a more complex map. The major concept in the example presented in Figure 5.2 is protein structure, and each of its four minor concepts represents one structure type.

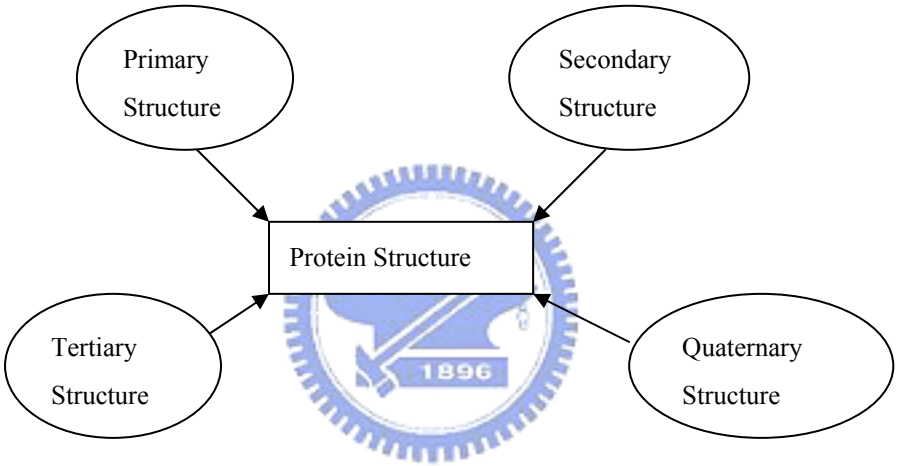


Figure 5. 2: An example of a spider-web map.

2. Chain Maps

Each link in a chain map either leads to or enables next concept.. For example, a PHD algorithm generates the predictive result of the secondary structure shown in Figure 5.3.

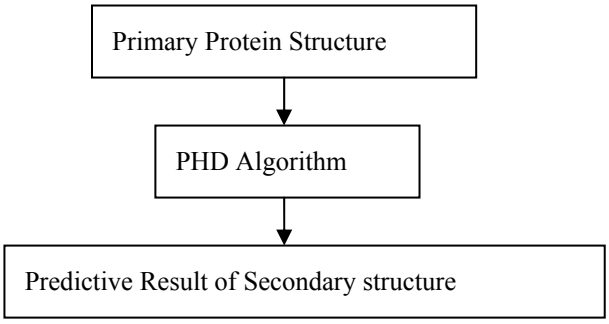


Figure 5. 3: An example of a chain map.

3. Hierarchy Maps

Hierarchy maps are usually viewed as the means by which knowledge is organized in the human cerebrum. A hierarchy map of structure alignment applications is shown in Figure 5.4.

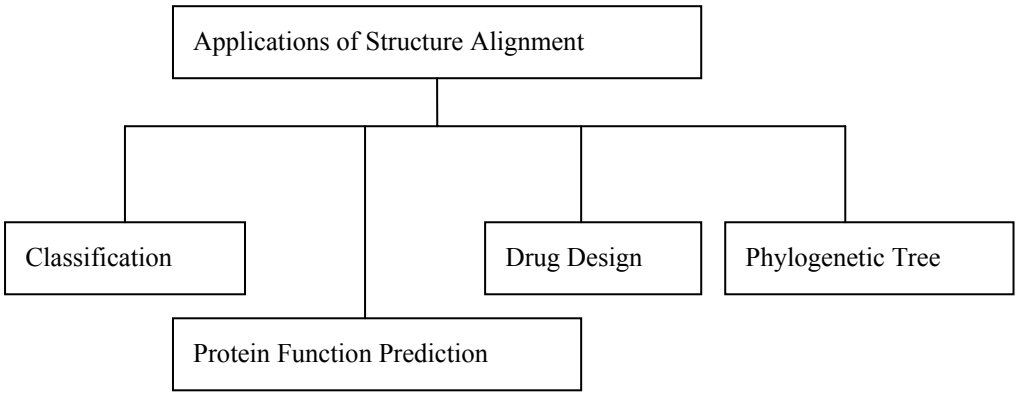


Figure 5. 4: An example of a hierarchy map.

5.3 A Bioinformatics Teaching Plan

The teacher may propose a biological question related on life and learners discuss that question by a succession of group discussion in the experiment or the media. While discussing, learner carries on the cooperative learning with others and develops his analysis ability.

Objective: To build an understanding of the definition of four protein structures.

Guidance Question: How do the following physiological reactions occur: enzyme catalysis, protein transportation and storage, immunoreactions, nerve impulse generation and propagation, and growth and differentiation?

First, learners will be guided to information on the importance of protein structure and secondary protein structure prediction. They will rehearse the protein structure prediction problem by using neural networks to design original solutions (Table 5.2).

Table 5. 2: Teaching plan design using a problem-based approach for secondary protein structure prediction.

Topic	Secondary Protein Structure Prediction.
Object	Learn the four primary types of protein structure.
Keywords	Protein structure, secondary structure prediction, neural networks.
Introduction	Proteins play a prominent role in all biological reactions. Their main functions include enzyme catalysis, transportation and storage, immunoreactions, nerve impulse generation and propagation, and growth and differentiation control.
Guidance	How to identify protein structure? If it cannot be obtained from a biological experiment, it can be predicted by its primary structure.
Goal	Propose an algorithm for protein secondary structure prediction.
Practice	1. Difficulties involved in determining protein structure from a biological experiment. 2. Understanding relationships between secondary and tertiary protein structures. 3. Understanding relationships between secondary and primary protein structures. 4. Perform a web-based protein structure search.

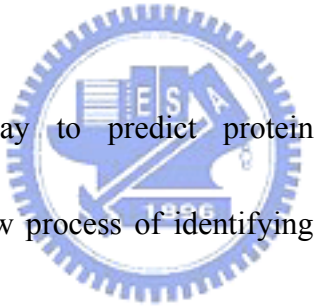
	5. Train and test datasets for neural networks. 6. Observe the capability and characteristics of neural networks for predicting secondary protein structures. 7. Refine the neural networks approach.
Method	Video media, small group discussion, brainstorming, problem solving.
Activity	1. Problem understanding. a. Use key points for topic discussion. b. Propose questions. c. Ponder the problem. 2. Data search and analysis. a. Gain deeper understanding of problem. b. Display search results and identify references. c. Share knowledge with other group members. 3. Problem analysis. a. Brainstorm to check data and opinions for correctness. b. Who, What, When, Where, Why and How. 4. Solution strategy. a. Create strategy as a team. 5. Conclusion. a. Identify final solution strategy. b. Perform evaluation.
Reference Teaching Materials	Bioinformatics / Oxford University Press Bioinformatics: The Machine Learning Approach / Baldi, Pierre. / Brunak, Soren. / NetLibrary, Inc. / MIT Press
Website Reference	Protein Structure: NCBI: http://www.ncbi.nlm.nih.gov/Structure/ Protein Database: PDB: http://www.rcsb.org/pdb/ DSSP: http://www.cmbi.kun.nl/gv/dssp/

Bioinformatics combines information science and biology—two fields with forms of logic that are difficult to negotiate. Here we proposed a hybrid bioinformatics teaching approach that uses problem-based learning techniques and concept maps. Problem-based learning can be regarded as a knowledge development and learner guidance system based

on well-constructed questions; and concept map construction can be used to make learning meaningful. Using this approach, learners can construct biology knowledge and identify important topics and the best potential solutions to a problem.

Chapter 6.

Conclusions and Future Research



Identifying the best way to predict protein secondary structure is not a winner-take-all race, but a slow process of identifying ways to extract regularity among sequence patterns. The contribution of this research is to add a clustering feature to a steady-state genetic algorithm. Clustering not only generates initial genetic algorithm populations, but also provides a solution when low-confidence schema cannot be applied to a problem. The protein secondary structure prediction problem was used to test a lesion study. By adding the clustering schema, predictive accuracy was improved by approximately 12%. As part of the competitive study, an associate rule and decision tree were also used to find schemata, but the cluster-based genetic algorithm is more capable of

finding a larger number of schemata. The schemata found by the two methods can be included with schemata identified by a cluster-based genetic algorithm.

Regarding schemata applications, results from third-generation methods of secondary structure prediction were used to define dead areas that PSIPRED and PROF are not capable of predicting. The schemata improved dead area prediction accuracy by between 40 and 60%. Accordingly, this model was used as an example for designing a teaching plan using a problem-based approach.

There are several possible directions for future research. First, although sequence schemata are currently treated independently, they can be combined for the purpose of characterizing specific secondary structures. One goal is to apply different composition operators (e.g., Boolean connectives) to combine schemata, or use higher-order models (e.g., HMM) to establish more realistic representations of relationships among different schemata.

A second goal is to apply cluster-based genetic algorithms to commonly used protein data sets to generate useful schemata as a feature of other protein secondary structure prediction tools for verifying the effectiveness of learned schemata.

Another goal involves the paper cluster-based genetic algorithm that was used in this research to find regularity in various protein secondary structures in terms of sequence patterns. However, the application of a cluster-based genetic algorithm and the use of

sequence patterns inevitably incur process and representation biases that can either help identify useful inductive leaps or hinder the learning/mining process. These biases must be evaluated in terms of usefulness for various protein domains.

Bibliography

[1] B. Rost. Protein structure prediction in 1D, 2D, and 3D. Encyclopedia of Computational Chemistry, John Wiley: Sussex, pp. 2242-2255, 1998.

[2] Protein Information Resource (PIR), http://pir.georgetown.edu/

[3] Protein Data Bank (PDB), http://www.rcsb.org/pdb/

[4] Critical Assessment of Techniques for Protein Structure Prediction (CASP), http://predictioncenter.org/

[5] K. Ginalski, N.V. Grishin, A. Godzik, and L. Rychlewski. Practical lessons from protein structure prediction. Nucleic AcidsRes. 33, 1874–1891, 2005.

[6] M. A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. andSali. Comparative protein structure modeling of genes andgenomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325, 2000.

[7] R. Sanchez and A. Sali. Advances in comparative proteinstructuremodelling. Curr. Opin. Struct. Biol. 7, 206–214, 1997.

[8] D. Petrey and B. Honig. Protein structure prediction: inroads to biology. Mol Cell. 22;20(6):811-9. 2005.

[9] M. Jia and L. Luo. The relation between mRNA folding and protein structure. Biochem Biophys Res Commun. Online version, 2006.

[10] S. Navea, R. Tauler, E. Goormaghtigh, A. de Juan. Chemometric tools for classification and elucidation of protein secondary structure from infrared and circular dichroism spectroscopic measurements. Proteins. Online version, 2006.

[11] L. N. Kinc, K. Ginalski, L. Rychlewski, NV. Grishin. Identification of novel restriction endonuclease-like fold families among hypothetical proteins. Nucleic Acids Res. 33(11):3598-605, 2005.

[12] B. Rost. Review: protein secondary structure prediction continues to rise. J Struct Biol. 134(2-3):204-18, 2001.

[13] Y. Yu. Coiled-coils: stability, specificity, and drug delivery potential. Adv. Drug. Deliv. Rev. 54(8), pp. 1113.-1129, 2002.

[14] M. Cianfriglia, C. Cenciarelli, S. Barca, M. Tombesi, M. Flego, and ML. Dupuis. Mono-clonal Antibodies as a Tool for Struc-ture-Function Studies of the MDR1-P-Glycoprotein. Curr. Protein Pept. Sci. 3(5). pp. 513-530, 2002.

[15] N. K. Nagradova. Three-dimensional domain swapping in homooligomeric proteins and it's functional signifiance. Biochemistry 67(8). pp. 839-849, 2002.

[16] Y. Kaizhi and A. D. Ken. Constraint-based assembly of tertiary protein structures from secondary structure elements. Protein Sci. 9. pp. 1935-1946, 2000.

[17] E. S. Robert and M. T. Janet. Prediction of Strand Pairing in Antiparallel and Pararallel β-Sheets Using Information Theory. Proteins: Structure, Function, and Genetics 48. pp. 178-191, 2002.

[18] B. Rost, R. Schneider, and C. Sander. Protein fold Recognition by Prediction-based Threading. J. Mol. Biol. 270. pp. 471-480, 1997.

[19] Y. An, and R. A. Friesner. A Novel Fold Recognition Method Using Composite Pre-dicted Secondary Structures. Proteins: Struc-ture, Functions, and Genetics 48. pp. 352-366, 2002.

[20] 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.

[21] S. Hua and Z. Sun, A novel method of pro-tein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, pp. 397-407, 2001.

[22] J. K. Rainey and MC. Goh, A statistically de-rived parameterization fro the collagen tri-ple-helix. Protein Sci. 11(11). pp. 2748-2754, 2002.

[23] C. K. Mathews, K. E. Van Holde, and K. G. Ahern. Biochemistry third edition, Addison Wesley Longman, 2000

[24] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, pp. 584-599, 1993.

[25] J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for prote in secondary structure prediction. Proteins: structure, function, and genetics, 34,508–519, 1999.

[26] U. Hobohm. and C. Sander. Enlarged Representative Set of Protein Structures. Protein Science, 3, pp. 522, 1994.

[27] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.

[28] D. L. Shrestha and D. P. Solomatine. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. Online version, 2006.

[29] P. L. Grimaldi. Medicare Part C means more choices. Nurs Manage. 28(11):30-3, 1997.

[30] C. T. Zhang, K. C. Chou and G. M. Maggiora. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Eng. 8(5):425-35, 1995.

[31] A. E. Reinberg and I. E. Ashkenazi. Interindividual differences in chronopharmacologic effects of drugs: a background for individualization of chronotherapy. Chronobiol Int. 10(6):449-60, 1993.

[32] E. A. Gale, T. Bennett, I. A. Macdonald, J. J. Holst, J. A. Matthews. The physiological effects of insulin-induced hypoglycaemia in man: responses at differing levels of blood glucose. Clin Sci (Lond). 65(3):263-71, 1983.

[33] J. H. Holland. Adaptation in Natural and Artificial Systems, Michigan University Press, 1975.

[34] D. E. Goldberg. Genetic Algorithm in Search, Optimization and Machine Learning, Addison Wesley, 1989.

[35] G. Golino. Improved genetic algorithm for the design of the optimal antenna division in sub-arrays: a multi-objective genetic algorithm. IEEE International Conference on Radar. pp. 629 - 634, 2005.

[36] L. Djerid. and M. C. Portmann. Genetic algorithm operators restricted to precedent constraint sets: genetic algorithm designs with or without branch and bound approach for solving scheduling problems with disjunctive constraints. IEEE International Conference on Systems, Man, and Cybernetics. vol 4, pp. 2922 - 2927, 1996.

[37] S. Salcedo-Sanz, A. Gallardo-Antolin, J. M. Leiva-Murillo, and C. Bousono-Calzon. Offline Speaker Segmentation Using Genetic Algorithms and Mutual Information. IEEE Transactions on Evolutionary Computation. Volume 10, Issue 2, pp. 175 - 186, 2006.

[38] R. J. Wai and K. H. Su. Supervisory Control for Linear Piezoelectric Ceramic Motor Drive Using Genetic Algorithm. IEEE Transactions on Industrial Electronics. Volume 53, Issue 2, pp. 657 - 673. 2006.

[39] J. T. Tsai, J. H. Chou and T. K. Liu. Tuning the Structure and Parameters of a Neural Network by Using Hybrid Taguchi-Genetic Algorithm. IEEE Transactions on Neural Networks. Volume 17, Issue 1, pp. 69 - 80, 2006.

[40] Li. Ling, D. H. Werner, J. A. Bossard and T. S. Mayer. A model-based parameter estimation technique for wide-band interpolation of periodic moment method impedance matrices with application to genetic algorithm optimization of frequency selective surfaces. IEEE Transactions on Antennas and Propagation. Volume 54, Issue 3, pp. 908 - 924, 2006.

[41] C. Runwei, G. Mitsuo and T. Yasuhiro. A Tutorial Survey of Job-Shop Scheduling Problems Using Genetic Algorithm--I. Representation. Computers and Industrial Engineering. vol. 30, no. 4, pp. 983-997(15), 1996.

[42] V. Petridis, S. Kazarlis and A. Bakirtzis. Varying fitness functions in genetic algorithm constrained optimization: the cutting stock and unit commitment problems. IEEE Transactions on Systems, Man and Cybernetics, Part B. Volume 28, Issue 5, Page(s):629 - 640, 1998.

[43] S. Takahashi, H. Kita, H. Suzuki, T. Sudo and S. Markon. Proceedings of 2003. CEC '03 Congress on Evolutionary Computation. Volume 3, Page(s):1582 - 1587, 2003.

[44] J. A. Lima, N. Gracias, H. Pereira and A. Rosa. Fitness function design for genetic algorithms in cost evaluation based problems. Proceedings of IEEE International Conference on Evolutionary Computation. Page(s):207 - 212, 1996.

[45] D. Bhandari, S. K. Pal, M.K. Kundu. Image enhancement incorporating fuzzy fitness function in genetic algorithms. Proceedings of Second IEEE International Conference on Fuzzy Systems. vol.2, Page(s):1408 - 1413, 1993.

[46] G. J. Gray, Y. Li, D. J. Murray-Smith and K.C. Sharman. Specification of a control system fitness function using constraints for genetic algorithm based design methods. Proceedings of (GALESIA) First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications. Page(s):530 - 535, 1995.

[47] A. Walairacht, C. Thanapandi, S. Ohara, O. Wongwirat and I. Burintramart. A learning mechanism for adaptive fitness function in auto 3D graphics layout using genetic algorithm. Proceedings of 2001. Canadian Conference on Electrical and Computer Engineering. Volume 1, Page(s):219 - 224, 2001.

[48] M. T. Hoque, C. Madhu and L. S. Dooley. Partially computed fitness function based genetic algorithm for hydrophobic-hydrophilic model., Proceedings of HIS 2004. Fourth International Conference on Hybrid Intelligent Systems. Page(s):291 - 296, 2004.

[49] S. Sundhararajan and A. Pahwa. Optimal selection of capacitors for radial distribution systems using a genetic algorithm. IEEE Transactions on Power Systems. Volume 9, Issue 3, Page(s):1499 - 1507, 1994.

[50] K. Jae-Kwang, C. Dong-Hyeok, J. Hyun-Kyo and L. Cheol-Gyun. Niching genetic algorithm adopting restricted competition selection combined with pattern search method. IEEE Transactions on Magnetics. Volume 38, Issue 2, Part 1, Page(s):1001 - 1004, 2002.

[51] L. Chang-Yong. Entropy-Boltzmann selection in the genetic algorithms. IEEE Transactions on Systems, Man and Cybernetics, Part B. Volume 33, Issue 1, Page(s):138 - 149, 2003.

[52] S. Mashohor, J. R. Evans, and T. Arslan. Elitist selection schemes for genetic algorithm based printed circuit board inspection system. Proceedings of The IEEE Congress on Evolutionary Computation. Vol. 2, Page(s):974 - 978, 2005.

[53] T. Okabe, J. Yaochu and B. Sendhoff. A New Approach to Dynamics Analysis of Genetic Algorithms without Selection. Proceedings of The IEEE Congress on Evolutionary Computation. Vol. 1, Page(s):374 - 381, 2005.

[54] M. Srinivas and L. M. Patnaik. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Transactions on Systems, Man and Cybernetics. Volume 24, Issue 4, Page(s):656 - 667, 1994.

[55] A. Prugel-Bennett. Modeling crossover-induced linkage in genetic algorithms. IEEE Transactions on Evolutionary Computation. Volume 5, Issue 4, Page(s):376 - 387, 2001.

[56] L. Fan, L. Qi-He, M. Fan and Y. Guo-Wei. A New Crossover Operator Based on the Rough Set Theory for Genetic Algorithms. Proceedings of 2005 International Conference on Machine Learning and Cybernetics. Volume 5, Page(s):2907 - 2912, 2005.

[57] M. Liang, C. Yiu-Ming and W. Yu-Ping. A dynamically switched crossover for genetic algorithms. Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Volume 5, Page(s):3254 - 3257, 2004.

[58] C. F. Huang. The role of crossover in an immunity based genetic algorithm for multimodal function optimization. CEC '03 Congress on Evolutionary Computation. Volume 4, Page(s):2807 - 2814, 2003.

[59] D. W. Boeringer and D. H. Werner. Adaptive mutation parameter toggling genetic algorithm for phase-only array synthesis. Electronics Letters. Volume 38, Issue 25, Page(s):1618 - 1619, 2002.

[60] P. Jianxun, L. Kang and S. Thompson. A combined adaptive bounding and adaptive mutation technique for genetic algorithms. Proceedings of WCICA Fifth World Congress on Intelligent Control and Automation. Volume 3, Page(s):2154 - 2158, 2004.

[61] H. Chen and J. Zhang. Multiple faces detection based on chaos-mutation niche genetic algorithms. Proceedings of WCICA Fifth World Congress on Intelligent Control and Automation. Volume 5, Page(s):4199 - 4202, 2004.

[62] R. Tinos, C. de and A. C. P. L. F. A genetic algorithm with gene dependent mutation probability for non-stationary optimization problems., Proceedings of CEC2004. Congress on Evolutionary Computation. Volume 2, Page(s):1278 - 1285, 2004.

[63] G. W. Greenwood. Adapting mutations in genetic algorithms using gene flow principles. Proceedings of CEC '03 Congress on Evolutionary Computation. Volume 2, Page(s):1392 - 1397, 2003.

[64] L. Pauling and R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheet. Proc. Natl. Acad. Sci. USA 37, 729-740, 1951.

[65] L. Pauling, R. B. Corey and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA 37, 205-234, 1951.

[66] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. J. Hart, D. R. Davies, and D. C. Phillips. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. Nature. 185, 422-427, 1960.

[67] M. F. Perutz, M. G. Rossmann, A. F. Cullis, G. Muirhead, G. Will and A. T. North. Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. Nature. 185, 416-422, 1960.

[68] W. Kabsch. and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. vol. 22, no. 12, pp. 2577-2637, 1983.

[69] D. Frishman. and P. Argos. Knowledge-based protein secondary structure assignment. Proteins. Vol. 23, no. 4, pp. 566-579, 1995.

[70] F. M. Richards. and C. E. Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins. Vol. 3, pp. 71-84, 1998.

[71] J. A. Cuff. and G. J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins. Vol. 40, No. 3, pp. 502-511, 2000.

[72] H. Kim. and H. Park. Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng. 16(8), pp. 553-60, 2003.

[73] G. Pollastri, D. Przybylski, B. Rost. and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins. 47(2), pp. 228-35, 2002.

[74] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. Vol. 292, no. 2, pp. 152-202, 1999.

[75] P. Baldi, S. Brunak, P. Frasconi, G. Soda and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics. Vol. 15, pp. 937-946, 1999.

[76] C. Brandon. and J. Tooze. Introduction to protein structure. Garland Publishing Inc., New York, 1991.

[77] B. Rost. PHD: predicting 1D protein structure byprofile based neural networks. Meth. Enzymol. Vol. 266, pp. 525-539, 1996.

[78] A. G. Szent-Györgyi and C. Cohen. Role of proline in polypeptide chain configuration of proteins. Science. Vol. 126, pp. 697, 1957.

[79] P. Y. Chou and G. D. Fasman. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry. Vol. 13, pp. 211-222, 1974.

[80] J. Garnier, D. J. Osguthorpe and B. Robson. Analysis of the accuracy and implications of simple methods for predictiong the secondary structure of globular proteins. J. Mol Biol. Vol. 120, pp. 97-120, 1978.

[81] V. I. Lim. Algorithm for prediction of α-helical and β-structural regions in globular proteins. J. Mol. Biol. Vol. 88, pp. 873-894, 1974.

[82] J. F. Gibrat. B. Robson. and J. Garnier. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol. Vol. 198, pp. 425-443, 1987.

[83] J. Levin, B. Robson and J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett. Vol. 205, pp. 303-308, 1986.

[84] K. Nishikawa and T. Ooi. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. Biochim. Biophys. Acta. Vol. 871, pp. 45-54, 1986.

[85] N. Qian and T. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. Vol. 202, pp. 865-884, 1998.

[86] L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. Proc Natl. Acad. Sci. USA. Vol. 86, pp. 152-156, 1989.

[87] K. Asai, S. Haymizu and K. Handa. Prediction of protein secondary structure by the hidden Markov model. CABIOS. Vol. 2, pp. 141-146, 1993.

[88] T. M. Yi and E. S. Lander. Protein secondary structure prediction using nearest-neighbor methods. J. Mol. Biol. Vol. 232, pp. 1117-1129, 1993.

[89] M. J. J. M. Zvelebil, G. J. Barton, W. R. Taylor and M. J. E. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. Vol. 195, pp. 957-961, 1987.

[90] D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino-acid sequence. Prot. Eng. Vol. 9, pp. 133-142, 1996.

[91] D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. Proteins. Vol, 27, pp. 329-335, 1997.

[92] V. V. Solovyev and A. A. Salamov. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. Vol. 247, pp. 11-15, 1995.

[93] R. D. King. and M. J. E. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Prot Sci. Vol. 5, pp. 2298-2310, 1996.

[94] J. A. Cuff and G. J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins. Vol. 40, pp. 502–511, 2000.

[95] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri and G. Soda. Bidirectional dynamics for protein secondary structure prediction. IJCAI99 workshop on neural, symbolic, and reinforcement methods for sequence learning, 1999.

[96] C. Bystroff, V. Thorsson and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J. Mol. Biol. Vol. 301, pp. 173-190, 2000.

[97] D. G. Kneller, F. E. Cohen and R. Langridge. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. Vol. 214, pp. 171-182, 1990.

[98] V. V. Solovyev, and A. A. Salamov. Predicting α-helix and β-strand segments of globular proteins. CABIOS. Vol. 10, pp. 661-669, 1994.

[99] 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. Vol. 308, pp. 397-407, 2001.

[100] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. Vol. 405, pp. 442-451, 1975.

[101] C. Bystroff, K. T. Simons, K. F. Han and D. Baker. Local Sequence-structure Correlations in Proteins, Current Opinions in Biotechnology, 7, pp. 417–421, 1996.

[102] M. Mitchell, An Introduction to Genetic Algorithms, MIT Press, 1996.

[103] J. Han and M. Kamber. Data Mining: Concepts and Techniques. *Morgan Kauf-mann Publishers*, 2001.

[104] T. J. P. Hubbard and C. Sander. The role of heat shock and chaperone proteins in protein folding: Possible molecular mechanisms. *Protein Eng* 4:711–717, 1991.

[105] R. L. Baldwin and G. D. Rose. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem Sci* 24:26–33, 1999.

[106] J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 4:508–519, 1999.

[107] A. A. Salamov and V. V. Solovyev. Protein secondary structure prediction using local alignments. *J Mol Biol* 268:31–36, 1997.

[108] J. M. Levin. Exploring the limits of nearest neighbor secondary structure prediction. *Protein Eng* 10:771–776, 1997.

[109] EValuation of Automatic protein structure prediction (EVA), http://www.rostlab.org/eva/index.html

[110] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323, 533-536, 1986.

[111] B. Rost: unpublished. http://cubic.bioc.columbia.edu/predictprotein

[112] M. C. LaPlaca, W. C. Newstetter and A. P. Yoganathan, A. P. Problem-Based Learning in Biomedical Engineering Curricula. *31th ASEE/IEEE Frontiers in Education Conference*, Session F3E, 16-21, 2001.

[113] D. L. Maskell and P. J. Grabau. A multidisciplinary cooperative problem-based learning approach to embedded system design. *IEEE Transaction—On Education* 41(2), 101-103, 1998

[114] F. Ronteltap and A. Eurelings. Activity and interaction of students in an electronic learning environment for problem-based learning. *Distance Education* 23(1), 11-22, 2002.

[115] P. Soderberg, P. An examination of problem-based teaching and learning in population genetics and evolution using EVOLVE, a computer simulation. *International Journal of Science Education* 25(1), 35-55, 2003.

[116] H. S. Barrows and A. Kelson. Problem-based Learning: A Total Approach to Education. *Illinois: STU School of Medicine*, Department of Medical Education Springfield, 1998.

[117] L. Torp and S. M. Sage. Problems as Possibilities: Problem-based Learning for K-12 Education. *Virginia: Association for Supervision and Curriculum Development,* 1998.

[118] S. B. Wegner, K. Holloway and A. Crader. Utilizing a problem-based approach on the World Wide Web. *ERIC Document Reproduction* No. ED414262, 1998.

[119] Illinois Mathematics and Science Academy. Problem-based learning. Retrieved July 13, 2004 from http://www2.imsa.edu/programs/pbl/problem.html

[120] J. D. Novak. Concept maps and Vee diagrams: Two metacognitive tools to facilitate meaningful learning. *Instructional Science*, 19(1), 29-52, 1990.

[121] Academia Sinica Computing Center. Introduction to bioinformatics. *Computing Center Newsletter* 19(20),1-6, 2003.

[122] S. A. Gallagher and W. J. Stepien. Implementing problem-based learning in science classrooms. *Science and Mathematics* 95(3), 136-146, 1995.

[123] S. Lerman. Constructivism, mathematics and mathematics education. *Education Studies in Mathematics* 20, 211-223, 1989.

[124] E. Von Glasersfeld. Constructivism in education. In T. Husen & N. Postlewaite (Eds.), *International Encyclopedia in Education* (supp 1.1), 162-163. Oxford, UK: Pergamon Press, 1989.

[125] H. S. Barrows. How to design a problem-based curriculum for the preclinical years. *New York: Springer*, 1985.

Appendix A

Schemata From PDB_select with cluster number: 70

A 217	PSV***VDH	T****TIKL	IRS***Q*H	T****TIKL	FC***VP*H
IRS***Q*H	IRS***Q*H	*TS***KMH	VER****DE	*TS***KMH	YH***IKYH
T****TIKL	VER****DE	**E***RGH	F*T***SNE	**E***RGH	PSV***VDH
IRS***Q*H	F*T***SNE	FC***VP*H	Y****KIKE	FC***VP*H	IRS***Q*H
VER****DE	Y****KIKE	PSV***VDH	VA***TEHL	PSV***VDH	VER****DE
Y****KIKE	VA***TEHL	IRS***Q*H	T****TIKL	IRS***Q*H	F*T***SNE
VA***TEHL	T****TIKL	VER****DE	*TS***KMH	VER****DE	Y****KIKE
T****TIKL	*TS***KMH	F*T***SNE	**E***RGH	F*T***SNE	DH***AVWL
*TS***KMH	**E***RGH	Y****KIKE	FC***VP*H	Y****KIKE	VA***TEHL
IRS***Q*H	PSV***VDH	VA***TEHL	PSV***VDH	VA***TEHL	T****TIKL
VER****DE	IRS***Q*H	T****TIKL	IRS***Q*H	T****TIKL	*TS***KMH
F*T***SNE	VER****DE	*TS***KMH	VER****DE	*TS***KMH	**E***RGH
Y****KIKE	F*T***SNE	**E***RGH	F*T***SNE	**E***RGH	FC***VP*H
VA***TEHL	Y****KIKE	FC***VP*H	Y****KIKE	FC***VP*H	YH***IKYH
T****TIKL	VA***TEHL	PSV***VDH	VA***TEHL	YH***IKYH	PSV***VDH
*TS***KMH	T****TIKL	IRS***Q*H	T****TIKL	PSV***VDH	IRS***Q*H
E*RGH	*TS***KMH	VER****DE	*TS***KMH	IRS***Q*H	VER****DE
IRS***Q*H	**E***RGH	F*T***SNE	**E***RGH	VER****DE	F*T***SNE
VER****DE	FC***VP*H	Y****KIKE	FC***VP*H	F*T***SNE	Y****KIKE
F*T***SNE	PSV***VDH	VA***TEHL	PSV***VDH	Y****KIKE	DH***AVWL
Y****KIKE	IRS***Q*H	T****TIKL	IRS***Q*H	DH***AVWL	IMM****SL
VA***TEHL	VER****DE	*TS***KMH	VER****DE	VA***TEHL	VA***TEHL
T****TIKL	F*T***SNE	**E***RGH	F*T***SNE	T****TIKL	T****TIKL
*TS***KMH	Y****KIKE	FC***VP*H	Y****KIKE	*TS***KMH	*TS***KMH
E*RGH	VA***TEHL	PSV***VDH	VA***TEHL	**E***RGH	**E***RGH

FC***VP*H	T****TIKL	FR***MR*E	FV***NFYL	IMR***EWE	IMR***EWE
YH***IKYH	*TS***KMH	*KE****NL	IA***E*CL	*AA***TVE	*AA***TVE
PSV***VDH	**E***RGH	IA***E*CL	GPL*****H	*KE****NL	*D***RRNE
IRS***Q*H	FC***VP*H	GPL*****H	C*K***A*H	FV***NFYL	*KE****NL
VER****DE	YH***IKYH	C*K***A*H	PA***A**H	IA***E*CL	FV***NFYL
F*T***SNE	PSV***VDH	PA***A**H	NM***TK*H	LF***FEHL	IA***E*CL
Y****KIKE	IRS***Q*H	NM***TK*H	FR***MR*E	GPL*****H	LF***FEHL
DH***AVWL	VER****DE	FR***MR*E	IMR***EWE	L*Y***MHH	GPL*****H
IMM****SL	F*T***SNE	*KE****NL	*KE****NL	C*K***A*H	L*Y***MHH
VA***TEHL	Y****KIKE	IA***E*CL	FV***NFYL	PA***A**H	C*K***A*H
T****TIKL	WVA***C*L	GPL*****H	IA***E*CL	NM***TK*H	PA***A**H
*TS***KMH	DH***AVWL	C*K***A*H	GPL*****H	FR***MR*E	NM***TK*H
E*RGH	IMM****SL	PA***A**H	C*K***A*H	IMR***EWE	FR***MR*E
FC***VP*H	VA***TEHL	NM***TK*H	PA***A**H	*AA***TVE	IMR***EWE
YH***IKYH	EGT****WL	FR***MR*E	NM***TK*H	*D***RRNE	*AA***TVE
PSV***VDH	T****TIKL	*KE****NL	FR***MR*E	*KE****NL	*D***RRNE
IRS***Q*H	*TS***KMH	IA***E*CL	IMR***EWE	FV***NFYL	*KE****NL
VER****DE	**E***RGH	GPL*****H	*KE****NL	IA***E*CL	FV***NFYL
F*T***SNE	FC***VP*H	C*K***A*H	FV***NFYL	LF***FEHL	Q*H***LWL
Y****KIKE	YH***IKYH	PA***A**H	IA***E*CL	GPL*****H	IA***E*CL
DH***AVWL	PSV***VDH	NM***TK*H	LF***FEHL	L*Y***MHH	LF***FEHL
IMM****SL	IRS***Q*H	FR***MR*E	GPL*****H	C*K***A*H	GPL*****H
VA***TEHL	VER****DE	*KE****NL	C*K***A*H	PA***A**H	L*Y***MHH
EGT****WL	F*T***SNE	IA***E*CL	PA***A**H	NM***TK*H	C*K***A*H
T****TIKL	Y****KIKE	GPL*****H	NM***TK*H	FR***MR*E	PA***A**H
*TS***KMH	DH***AVWL	C*K***A*H	FR***MR*E	IMR***EWE	NM***TK*H
E*RGH	IMM****SL	PA***A**H	IMR***EWE	*AA***TVE	FR***MR*E
FC***VP*H	VA***TEHL	NM***TK*H	*AA***TVE	*D***RRNE	IMR***EWE
YH***IKYH	EGT****WL	FR***MR*E	*KE****NL	*KE****NL	*AA***TVE
PSV***VDH	T****TIKL	*KE****NL	FV***NFYL	FV***NFYL	*D***RRNE
IRS***Q*H		IA***E*CL	IA***E*CL	IA***E*CL	*KE****NL
VER****DE	R 232	GPL*****H	LF***FEHL	LF***FEHL	FV***NFYL
F*T***SNE	GPL*****H	C*K***A*H	GPL*****H	GPL*****H	Q*H***LWL
Y****KIKE	PA***A**H	PA***A**H	L*Y***MHH	L*Y***MHH	IA***E*CL
DH***AVWL	*KE****NL	NM***TK*H	C*K***A*H	C*K***A*H	LF***FEHL
IMM****SL	IA***E*CL	FR***MR*E	PA***A**H	PA***A**H	GPL*****H
VA***TEHL	GPL*****H	IMR***EWE	NM***TK*H	NM***TK*H	L*Y***MHH
EGT****WL	PA***A**H	*KE****NL	FR***MR*E	FR***MR*E	C*K***A*H

PA***A**H	Q*H***LWL	QI***TANE	**M***CSH	DWQ**QG*E	AMA**VCIH
NM***TK*H	IA***E*CL	G****NGWE	DWQ**QG*E	*FM**I*CE	**M***CSH
FR***MR*E	LF***FEHL	W*N***LSL	*FM**I*CE	QI***TANE	SMQ***LIE
IMR***EWE	GPL*****H	S****TAKL	QI***TANE	G****NGWE	NVK**NCCE
*AA***TVE	L*Y***MHH	KEI****VL	G****NGWE	W*N***LSL	DWQ**QG*E
*D***RRNE	C*K***A*H	V****Y*FL	W*N***LSL	S****TAKL	*FM**I*CE
*KE****NL	PA***A**H	QI***TANE	S****TAKL	M****ANPL	QI***TANE
FV***NFYL	NM***TK*H	G****NGWE	KEI****VL	KEI****VL	G****NGWE
Q*H***LWL	FR***MR*E	W*N***LSL	*Q***QKTL	*Q***QKTL	W*N***LSL
IA***E*CL	*R***VASE	S****TAKL	V****Y*FL	V****Y*FL	S****TAKL
LF***FEHL	IMR***EWE	KEI****VL	AMA**VCIH	K*F***PHL	M****ANPL
GPL*****H	*AA***TVE	V****Y*FL	**M***CSH	AMA**VCIH	KEI****VL
L*Y***MHH	*D***RRNE	*FM**I*CE	DWQ**QG*E	**M***CSH	*Q***QKTL
C*K***A*H	*KE****NL	QI***TANE	*FM**I*CE	DWQ**QG*E	V****Y*FL
PA***A**H	FV***NFYL	G****NGWE	QI***TANE	*FM**I*CE	K*F***PHL
NM***TK*H	Q*H***LWL	W*N***LSL	G****NGWE	QI***TANE	AMA**VCIH
FR***MR*E	Y****IT*L	S****TAKL	W*N***LSL	G****NGWE	**M***CSH
*R***VASE	IA***E*CL	KEI****VL	S****TAKL	W*N***LSL	SMQ***LIE
IMR***EWE	LF***FEHL	*Q***QKTL	M****ANPL	S****TAKL	NVK**NCCE
*AA***TVE	GPL*****H	V****Y*FL	KEI****VL	M****ANPL	DWQ**QG*E
*D***RRNE	L*Y***MHH	DWQ**QG*E	*Q***QKTL	KEI****VL	*FM**I*CE
*KE****NL	C*K***A*H	*FM**I*CE	V****Y*FL	*Q***QKTL	QI***TANE
FV***NFYL	PA***A**H	QI***TANE	K*F***PHL	V****Y*FL	G****NGWE
Q*H***LWL	H*F***MAH	G****NGWE	AMA**VCIH	K*F***PHL	W*N***LSL
IA***E*CL	NM***TK*H	W*N***LSL	**M***CSH	AMA**VCIH	S****TAKL
LF***FEHL	FR***MR*E	S****TAKL	DWQ**QG*E	**M***CSH	M****ANPL
GPL*****H	*R***VASE	KEI****VL	*FM**I*CE	SMQ***LIE	KEI****VL
L*Y***MHH	IMR***EWE	*Q***QKTL	QI***TANE	DWQ**QG*E	*Q***QKTL
C*K***A*H	*AA***TVE	V****Y*FL	G****NGWE	*FM**I*CE	V****Y*FL
PA***A**H	*D***RRNE	DWQ**QG*E	W*N***LSL	QI***TANE	RF***I**L
NM***TK*H	*KE****NL	*FM**I*CE	S****TAKL	G****NGWE	K*F***PHL
FR***MR*E	FV***NFYL	QI***TANE	M****ANPL	W*N***LSL	AMA**VCIH
*R***VASE	Q*H***LWL	G****NGWE	KEI****VL	S****TAKL	**M***CSH
IMR***EWE	Y****IT*L	W*N***LSL	*Q***QKTL	M****ANPL	SMQ***LIE
*AA***TVE	IA***E*CL	S****TAKL	V****Y*FL	KEI****VL	NVK**NCCE
*D***RRNE	LF***FEHL	KEI****VL	K*F***PHL	*Q***QKTL	DWQ**QG*E
*KE****NL		*Q***QKTL	AMA**VCIH	V****Y*FL	*FM**I*CE
FV***NFYL	N 283	V****Y*FL	**M***CSH	K*F***PHL	QI***TANE

G****NGWE	*Q***QKTL	AMA**VCIH	*IS***YGH	A*F****TE	GSA***FIL
W*N***LSL	V****Y*FL	**M***CSH	SMQ***LIE	*LE****PE	RDL*****H
S****TAKL	RF***I**L	SMQ***LIE	NVK**NCCE	*L***G*WE	F****VRNH
M****ANPL	K*F***PHL	NVK**NCCE	DWQ**QG*E	ERQ****GE	HE***I**E
KEI****VL	AMA**VCIH	DWQ**QG*E	Y*N****GE	RRR**DN*L	A*F****TE
*Q***QKTL	**M***CSH	Y*N****GE	*FM**I*CE	GSA***FIL	*LE****PE
V****Y*FL	SMQ***LIE	*FM**I*CE	QI***TANE	RDL*****H	*M***VS*E
RF***I**L	NVK**NCCE	QI***TANE	G****NGWE	F****VRNH	*L***G*WE
K*F***PHL	DWQ**QG*E	G****NGWE	W*N***LSL	HE***I**E	ERQ****GE
AMA**VCIH	*FM**I*CE	W*N***LSL	S****TAKL	A*F****TE	RRR**DN*L
M*CSH	QI***TANE	S****TAKL	P*P***QRL	*LE****PE	QDQ***ENL
SMQ***HDE	G****NGWE	P*P***QRL	M****ANPL	*L***G*WE	GG***R*EL
SMQ***LIE	W*N***LSL	M****ANPL	KEI****VL	ERQ****GE	GSA***FIL
NVK**NCCE	S****TAKL	KEI****VL	*Q***QKTL	RRR**DN*L	RDL*****H
DWQ**QG*E	M****ANPL	*Q***QKTL	V****Y*FL	GSA***FIL	F****VRNH
*FM**I*CE	KEI****VL	V****Y*FL	RF***I**L	RDL*****H	HE***I**E
QI***TANE	*Q***QKTL	RF***I**L	K*F***PHL	F****VRNH	A*F****TE
G****NGWE	V****Y*FL	K*F***PHL		HE***I**E	*LE****PE
W*N***LSL	RF***I**L	AMA**VCIH	D 298	A*F****TE	*M***VS*E
S****TAKL	K*F***PHL	**M***CSH	RDL*****H	*LE****PE	*L***G*WE
M****ANPL	AMA**VCIH	SMQ***LIE	F****VRNH	*M***VS*E	*T***VLQE
KEI****VL	**M***CSH	NVK**NCCE	HE***I**E	*L***G*WE	ERQ****GE
*Q***QKTL	SMQ***LIE	DWQ**QG*E	*LE****PE	ERQ****GE	RRR**DN*L
V****Y*FL	NVK**NCCE	Y*N****GE	*L***G*WE	RRR**DN*L	QDQ***ENL
RF***I**L	DWQ**QG*E	*FM**I*CE	ERQ****GE	QDQ***ENL	GG***R*EL
K*F***PHL	Y*N****GE	QI***TANE	RRR**DN*L	GG***R*EL	GSA***FIL
AMA**VCIH	*FM**I*CE	G****NGWE	GSA***FIL	GSA***FIL	RDL*****H
M*CSH	QI***TANE	W*N***LSL	RDL*****H	RDL*****H	F****VRNH
SMQ***LIE	G****NGWE	S****TAKL	F****VRNH	F****VRNH	HE***I**E
NVK**NCCE	W*N***LSL	P*P***QRL	HE***I**E	HE***I**E	A*F****TE
DWQ**QG*E	S****TAKL	M****ANPL	*LE****PE	A*F****TE	*LE****PE
*FM**I*CE	P*P***QRL	KEI****VL	*L***G*WE	*LE****PE	*M***VS*E
QI***TANE	M****ANPL	*Q***QKTL	ERQ****GE	*M***VS*E	*L***G*WE
G****NGWE	KEI****VL	V****Y*FL	RRR**DN*L	*L***G*WE	QHV***M*E
W*N***LSL	*Q***QKTL	RF***I**L	GSA***FIL	ERQ****GE	*T***VLQE
S****TAKL	V****Y*FL	K*F***PHL	RDL*****H	RRR**DN*L	ERQ****GE
M****ANPL	RF***I**L	AMA**VCIH	F****VRNH	QDQ***ENL	RRR**DN*L
KEI****VL	K*F***PHL	**M***CSH	HE***I**E	GG***R*EL	QDQ***ENL

GG***R*EL	*M***VS*E	*T***VLQE	QDQ***ENL	RDL*****H	*LE****PE
GSA***FIL	*L***G*WE	ERQ****GE	GG***R*EL	F****VRNH	*M***VS*E
RDL*****H	QDT***PHE	RRR**DN*L	GSA***FIL	HE***I**E	*L***G*WE
F****VRNH	QHV***M*E	*H***M*QL	P*G****HL	A*F****TE	QDT***PHE
HE***I**E	*T***VLQE	QDQ***ENL	RDL*****H	*LE****PE	QHV***M*E
A*F****TE	ERQ****GE	GG***R*EL	F****VRNH	*M***VS*E	*T***VLQE
*LE****PE	RRR**DN*L	GSA***FIL	HE***I**E	*L***G*WE	ERQ****GE
*M***VS*E	*H***M*QL	P*G****HL	A*F****TE	QDT***PHE	RRR**DN*L
*L***G*WE	QDQ***ENL	RDL*****H	*LE****PE	QHV***M*E	*H***M*QL
QHV***M*E	GG***R*EL	F****VRNH	*M***VS*E	*T***VLQE	QDQ***ENL
*T***VLQE	GSA***FIL	HE***I**E	*L***G*WE	ERQ****GE	GG***R*EL
ERQ****GE	P*G****HL	A*F****TE	QDT***PHE	RRR**DN*L	GSA***FIL
RRR**DN*L	RDL*****H	*LE****PE	QHV***M*E	*H***M*QL	P*G****HL
QDQ***ENL	F****VRNH	*M***VS*E	*T***VLQE	QDQ***ENL	
GG***R*EL	HE***I**E	*L***G*WE	ERQ****GE	GG***R*EL	C 445
GSA***FIL	A*F****TE	QDT***PHE	RRR**DN*L	GSA***FIL	*****QRAH
RDL*****H	*LE****PE	QHV***M*E	*H***M*QL	P*G****HL	KC***Q*QH
F****VRNH	*M***VS*E	*T***VLQE	QDQ***ENL	RDL*****H	*****KLRH
HE***I**E	*L***G*WE	ERQ****GE	GG***R*EL	F****VRNH	*I***E*RH
A*F****TE	QDT***PHE	RRR**DN*L	GSA***FIL	HE***I**E	F*F***RWH
*LE****PE	QHV***M*E	*H***M*QL	P*G****HL	A*F****TE	KL***L*DE
*M***VS*E	*T***VLQE	QDQ***ENL	RDL*****H	*LE****PE	GY***FPNE
*L***G*WE	ERQ****GE	GG***R*EL	F****VRNH	*M***VS*E	D*V***WEE
QDT***PHE	RRR**DN*L	GSA***FIL	HE***I**E	*L***G*WE	KCT***LLL
QHV***M*E	*H***M*QL	P*G****HL	A*F****TE	QDT***PHE	*IA***NIL
*T***VLQE	QDQ***ENL	RDL*****H	*LE****PE	QHV***M*E	LMR***SEL
ERQ****GE	GG***R*EL	F****VRNH	*M***VS*E	*T***VLQE	*****QRAH
RRR**DN*L	GSA***FIL	HE***I**E	*L***G*WE	ERQ****GE	KC***Q*QH
*H***M*QL	P*G****HL	A*F****TE	QDT***PHE	RRR**DN*L	*****KLRH
QDQ***ENL	RDL*****H	*LE****PE	QHV***M*E	*H***M*QL	*I***E*RH
GG***R*EL	F****VRNH	*M***VS*E	*T***VLQE	QDQ***ENL	F*F***RWH
GSA***FIL	HE***I**E	*L***G*WE	ERQ****GE	GG***R*EL	TWE***AGH
P*G****HL	A*F****TE	QDT***PHE	RRR**DN*L	GSA***FIL	KL***L*DE
RDL*****H	*LE****PE	QHV***M*E	*H***M*QL	P*G****HL	GY***FPNE
F****VRNH	*M***VS*E	*T***VLQE	QDQ***ENL	RDL*****H	D*V***WEE
HE***I**E	*L***G*WE	ERQ****GE	GG***R*EL	F****VRNH	KCT***LLL
A*F****TE	QDT***PHE	RRR**DN*L	GSA***FIL	HE***I**E	*IA***NIL
*LE****PE	QHV***M*E	*H***M*QL	P*G****HL	A*F****TE	LMR***SEL

AYI***Q*L	KL***L*DE	LMR***SEL	FRD**ICWL	KCT***LLL	GY***FPNE
*****QRAH	GY***FPNE	AYI***Q*L	*****DYWH	*IA***NIL	D*V***WEE
KC***Q*QH	D*V***WEE	*****QRAH	*****QRAH	LMR***SEL	LN***LWGE
*****KLRH	KCT***LLL	KC***Q*QH	KC***Q*QH	AYI***Q*L	VGE***A*L
*I***E*RH	*IA***NIL	*****KLRH	*****KLRH	FRD**ICWL	KCT***LLL
F*F***RWH	LMR***SEL	*F***MQCH	*F***MQCH	*****DYWH	*IA***NIL
TWE***AGH	AYI***Q*L	*I***E*RH	*I***E*RH	*****QRAH	LMR***SEL
KL***L*DE	*****QRAH	EQA**TCNH	EQA**TCNH	KC***Q*QH	AYI***Q*L
GY***FPNE	KC***Q*QH	F*M**V*YH	F*M**V*YH	*****KLRH	FRD**ICWL
D*V***WEE	*****KLRH	F*F***RWH	F*F***RWH	*F***MQCH	*****DYWH
KCT***LLL	*F***MQCH	TWE***AGH	HYE****CH	*I***E*RH	*****QRAH
*IA***NIL	*I***E*RH	KL***L*DE	TWE***AGH	EQA**TCNH	KC***Q*QH
LMR***SEL	EQA**TCNH	GY***FPNE	KL***L*DE	F*M**V*YH	*****KLRH
AYI***Q*L	F*M**V*YH	D*V***WEE	GY***FPNE	F*F***RWH	*F***MQCH
*****QRAH	F*F***RWH	LN***LWGE	D*V***WEE	HYE****CH	*I***E*RH
KC***Q*QH	TWE***AGH	VGE***A*L	LN***LWGE	TWE***AGH	EQA**TCNH
*****KLRH	KL***L*DE	KCT***LLL	VGE***A*L	KL***L*DE	F*M**V*YH
*F***MQCH	GY***FPNE	*IA***NIL	KCT***LLL	GY***FPNE	F*F***RWH
*I***E*RH	D*V***WEE	LMR***SEL	*IA***NIL	D*V***WEE	HYE****CH
EQA**TCNH	KCT***LLL	AYI***Q*L	LMR***SEL	LN***LWGE	TWE***AGH
F*F***RWH	*IA***NIL	*****QRAH	AYI***Q*L	VGE***A*L	PT***N**E
TWE***AGH	LMR***SEL	KC***Q*QH	FRD**ICWL	KCT***LLL	KL***L*DE
KL***L*DE	AYI***Q*L	*****KLRH	*****DYWH	*IA***NIL	GY***FPNE
GY***FPNE	*****QRAH	*F***MQCH	*****QRAH	LMR***SEL	D*V***WEE
D*V***WEE	KC***Q*QH	*I***E*RH	KC***Q*QH	AYI***Q*L	LN***LWGE
KCT***LLL	*****KLRH	EQA**TCNH	*****KLRH	FRD**ICWL	VGE***A*L
*IA***NIL	*F***MQCH	F*M**V*YH	*F***MQCH	*****DYWH	KCT***LLL
LMR***SEL	*I***E*RH	F*F***RWH	*I***E*RH	*****QRAH	*IA***NIL
AYI***Q*L	EQA**TCNH	TWE***AGH	EQA**TCNH	KC***Q*QH	LMR***SEL
*****QRAH	F*M**V*YH	KL***L*DE	F*M**V*YH	*****KLRH	AYI***Q*L
KC***Q*QH	F*F***RWH	GY***FPNE	F*F***RWH	*F***MQCH	FRD**ICWL
*****KLRH	TWE***AGH	D*V***WEE	HYE****CH	*I***E*RH	CW***H*IL
*F***MQCH	KL***L*DE	LN***LWGE	TWE***AGH	EQA**TCNH	*****DYWH
*I***E*RH	GY***FPNE	VGE***A*L	KL***L*DE	F*M**V*YH	*****QRAH
EQA**TCNH	D*V***WEE	KCT***LLL	GY***FPNE	F*F***RWH	V****SPWH
F*M**V*YH	LN***LWGE	*IA***NIL	D*V***WEE	HYE****CH	KC***Q*QH
F*F***RWH	KCT***LLL	LMR***SEL	LN***LWGE	TWE***AGH	*****KLRH
TWE***AGH	*IA***NIL	AYI***Q*L	VGE***A*L	KL***L*DE	*F***MQCH

*I***E*RH	KCT***LLL	V****SPWH	F*F***RWH	NN***GKVE	LCC**YTWL
EQA**TCNH	*IA***NIL	KC***Q*QH	HYE****CH	VGE***A*L	FRD**ICWL
F*M**V*YH	LMR***SEL	*****KLRH	TWE***AGH	P*P***WWL	CW***H*IL
F*F***RWH	AYI***Q*L	*F***MQCH	PT***N**E	KCT***LLL	N*V**AIDL
HYE****CH	LCC**YTWL	*I***E*RH	KL***L*DE	*IA***NIL	
TWE***AGH	FRD**ICWL	*L***KLKH	GY***FPNE	LMR***SEL	Q 287
PT***N**E	CW***H*IL	EQA**TCNH	D*V***WEE	AYI***Q*L	R****LRRH
KL***L*DE	N*V**AIDL	F*M**V*YH	LN***LWGE	*A***IR*L	**D***AHH
GY***FPNE	*****DYWH	F*F***RWH	NN***GKVE	LCC**YTWL	**S***GRE
D*V***WEE	*****QRAH	HYE****CH	VGE***A*L	FRD**ICWL	K****LF*E
LN***LWGE	V****SPWH	TWE***AGH	P*P***WWL	CW***H*IL	LS***A*EL
VGE***A*L	KC***Q*QH	PT***N**E	KCT***LLL	N*V**AIDL	LF***SLVL
KCT***LLL	*****KLRH	KL***L*DE	*IA***NIL	*****DYWH	*Q***SKVL
*IA***NIL	*F***MQCH	GY***FPNE	LMR***SEL	*****QRAH	R****LRRH
LMR***SEL	*I***E*RH	D*V***WEE	AYI***Q*L	V****SPWH	**D***AHH
AYI***Q*L	EQA**TCNH	LN***LWGE	*A***IR*L	KC***Q*QH	**S***GRE
FRD**ICWL	F*M**V*YH	NN***GKVE	LCC**YTWL	*****KLRH	K****LF*E
CW***H*IL	F*F***RWH	VGE***A*L	FRD**ICWL	*F***MQCH	LS***A*EL
N*V**AIDL	HYE****CH	P*P***WWL	CW***H*IL	*I***E*RH	LF***SLVL
*****DYWH	TWE***AGH	KCT***LLL	N*V**AIDL	*L***KLKH	*Q***SKVL
*****QRAH	PT***N**E	*IA***NIL	*****DYWH	EQA**TCNH	*****LKAL
V****SPWH	KL***L*DE	LMR***SEL	*****QRAH	F*M**V*YH	R****LRRH
KC***Q*QH	GY***FPNE	AYI***Q*L	V****SPWH	F*F***RWH	**D***AHH
*****KLRH	D*V***WEE	*A***IR*L	KC***Q*QH	HYE****CH	**S***GRE
*F***MQCH	LN***LWGE	LCC**YTWL	*****KLRH	TWE***AGH	K****LF*E
*I***E*RH	NN***GKVE	FRD**ICWL	*F***MQCH	PT***N**E	LS***A*EL
EQA**TCNH	VGE***A*L	CW***H*IL	*I***E*RH	KL***L*DE	LF***SLVL
F*M**V*YH	P*P***WWL	N*V**AIDL	*L***KLKH	GY***FPNE	*Q***SKVL
F*F***RWH	KCT***LLL	*****DYWH	EQA**TCNH	D*V***WEE	*****LKAL
HYE****CH	*IA***NIL	*****QRAH	F*M**V*YH	LN***LWGE	R****LRRH
TWE***AGH	LMR***SEL	V****SPWH	F*F***RWH	NN***GKVE	**D***AHH
PT***N**E	AYI***Q*L	KC***Q*QH	HYE****CH	VGE***A*L	**S***GRE
KL***L*DE	LCC**YTWL	*****KLRH	TWE***AGH	P*P***WWL	K****LF*E
GY***FPNE	FRD**ICWL	*F***MQCH	PT***N**E	KCT***LLL	LS***A*EL
D*V***WEE	CW***H*IL	*I***E*RH	KL***L*DE	*IA***NIL	LF***SLVL
LN***LWGE	N*V**AIDL	*L***KLKH	GY***FPNE	LMR***SEL	*Q***SKVL
NN***GKVE	*****DYWH	EQA**TCNH	D*V***WEE	AYI***Q*L	*****LKAL
VGE***A*L	*****QRAH	F*M**V*YH	LN***LWGE	*A***IR*L	R****LRRH

D*AHH	LEM**KQHH	**S***GRE	HQ***KI*L	LF***SLVL	LF***SLVL
S*GRE	**S***GRE	G*Q***RIE	LF***SLVL	YA***DSPL	YA***DSPL
K****LF*E	G*Q***RIE	K****LF*E	YA***DSPL	*Q***SKVL	*Q***SKVL
LS***A*EL	K****LF*E	LS***A*EL	*Q***SKVL	YTR*****L	YTR*****L
LF***SLVL	LS***A*EL	HQ***KI*L	YTR*****L	L*L***RGL	L*L***RGL
*Q***SKVL	HQ***KI*L	LF***SLVL	L*L***RGL	AL***SFVL	AL***SFVL
*****LKAL	LF***SLVL	YA***DSPL	AL***SFVL	*****LKAL	*****LKAL
R****LRRH	*Q***SKVL	*Q***SKVL	*****LKAL	*EA****KL	*EA****KL
D*AHH	*****LKAL	YTR*****L	R****LRRH	R****LRRH	R****LRRH
S*GRE	R****LRRH	AL***SFVL	**D***AHH	**D***AHH	**D***AHH
K****LF*E	**D***AHH	*****LKAL	LEM**KQHH	LEM**KQHH	LEM**KQHH
LS***A*EL	LEM**KQHH	R****LRRH	QQF***AEH	QQF***AEH	QQF***AEH
LF***SLVL	**S***GRE	**D***AHH	*DM***M*H	*DM***M*H	*DM***M*H
*Q***SKVL	G*Q***RIE	LEM**KQHH	**S***GRE	**S***GRE	**S***GRE
*****LKAL	K****LF*E	QQF***AEH	RPG***PWE	RPG***PWE	RPG***PWE
R****LRRH	LS***A*EL	*DM***M*H	G*Q***RIE	G*Q***RIE	G*Q***RIE
D*AHH	HQ***KI*L	**S***GRE	K****LF*E	K****LF*E	K****LF*E
LEM**KQHH	LF***SLVL	G*Q***RIE	LS***A*EL	LS***A*EL	LS***A*EL
S*GRE	*Q***SKVL	K****LF*E	HQ***KI*L	HQ***KI*L	HQ***KI*L
G*Q***RIE	YTR*****L	LS***A*EL	LF***SLVL	LF***SLVL	LF***SLVL
K****LF*E	*****LKAL	HQ***KI*L	YA***DSPL	YA***DSPL	YA***DSPL
LS***A*EL	R****LRRH	LF***SLVL	*Q***SKVL	*Q***SKVL	*Q***SKVL
LF***SLVL	**D***AHH	YA***DSPL	YTR*****L	YTR*****L	YTR*****L
*Q***SKVL	LEM**KQHH	*Q***SKVL	L*L***RGL	L*L***RGL	L*L***RGL
*****LKAL	QQF***AEH	YTR*****L	AL***SFVL	AL***SFVL	AL***SFVL
R****LRRH	**S***GRE	L*L***RGL	*****LKAL	*****LKAL	*****LKAL
D*AHH	G*Q***RIE	AL***SFVL	*EA****KL	*EA****KL	*EA****KL
LEM**KQHH	K****LF*E	*****LKAL	R****LRRH	R****LRRH	R****LRRH
S*GRE	LS***A*EL	R****LRRH	**D***AHH	**D***AHH	**D***AHH
G*Q***RIE	HQ***KI*L	**D***AHH	LEM**KQHH	LEM**KQHH	LEM**KQHH
K****LF*E	LF***SLVL	LEM**KQHH	QQF***AEH	QQF***AEH	QQF***AEH
LS***A*EL	*Q***SKVL	QQF***AEH	*DM***M*H	*DM***M*H	*DM***M*H
HQ***KI*L	YTR*****L	*DM***M*H	**S***GRE	**S***GRE	**S***GRE
LF***SLVL	*****LKAL	**S***GRE	RPG***PWE	RPG***PWE	RPG***PWE
*Q***SKVL	R****LRRH	RPG***PWE	G*Q***RIE	G*Q***RIE	G*Q***RIE
*****LKAL	**D***AHH	G*Q***RIE	K****LF*E	K****LF*E	K****LF*E
R****LRRH	LEM**KQHH	K****LF*E	LS***A*EL	LS***A*EL	LS***A*EL
D*AHH	QQF***AEH	LS***A*EL	HQ***KI*L	HQ***KI*L	HQ***KI*L

LF***SLVL	*IL***SVL	GA***MSSH	GA***MSSH	*IL***SVL	GA***MSSH
YA***DSPL	ATG***HFH	*PL***F*H	*PL***F*H	GGV***HLL	*PL***F*H
*Q***SKVL	REG***QFH	REG***QFH	REG***QFH	ATG***HFH	*****DAYH
YTR*****L	N****D*PE	N****D*PE	R****KFDH	GA***MSSH	REG***QFH
L*L***RGL	SC****DHL	D*G***FVE	N****D*PE	*PL***F*H	R****KFDH
AL***SFVL	*IL***SVL	SC****DHL	D*G***FVE	*****DAYH	E****IT*H
*****LKAL	ATG***HFH	*IL***SVL	SC****DHL	REG***QFH	N****D*PE
*EA****KL	GA***MSSH	GGV***HLL	*IL***SVL	R****KFDH	D*G***FVE
R****LRRH	*PL***F*H	ATG***HFH	GGV***HLL	N****D*PE	SC****DHL
D*AHH	REG***QFH	GA***MSSH	ATG***HFH	D*G***FVE	*IL***SVL
LEM**KQHH	N****D*PE	*PL***F*H	GA***MSSH	SC****DHL	GGV***HLL
QQF***AEH	SC****DHL	REG***QFH	*PL***F*H	*IL***SVL	ATG***HFH
*DM***M*H	*IL***SVL	N****D*PE	REG***QFH	GGV***HLL	RFA***P*H
S*GRE	ATG***HFH	D*G***FVE	R****KFDH	ATG***HFH	GA***MSSH
RPG***PWE	GA***MSSH	SC****DHL	N****D*PE	GA***MSSH	*PL***F*H
G*Q***RIE	*PL***F*H	*IL***SVL	D*G***FVE	*PL***F*H	*****DAYH
K****LF*E	REG***QFH	GGV***HLL	SC****DHL	*****DAYH	REG***QFH
LS***A*EL	N****D*PE	ATG***HFH	*IL***SVL	REG***QFH	R****KFDH
HQ***KI*L	SC****DHL	GA***MSSH	GGV***HLL	R****KFDH	E****IT*H
LF***SLVL	*IL***SVL	*PL***F*H	ATG***HFH	N****D*PE	N****D*PE
YA***DSPL	ATG***HFH	REG***QFH	GA***MSSH	D*G***FVE	KC***GFVE
*Q***SKVL	GA***MSSH	R****KFDH	*PL***F*H	SC****DHL	D*G***FVE
YTR*****L	*PL***F*H	N****D*PE	REG***QFH	*IL***SVL	SC****DHL
L*L***RGL	REG***QFH	D*G***FVE	R****KFDH	GGV***HLL	*IL***SVL
AL***SFVL	N****D*PE	SC****DHL	N****D*PE	ATG***HFH	GGV***HLL
*****LKAL	D*G***FVE	*IL***SVL	D*G***FVE	RFA***P*H	IH***A**L
*EA****KL	SC****DHL	GGV***HLL	SC****DHL	GA***MSSH	
	*IL***SVL	ATG***HFH	*IL***SVL	*PL***F*H	G 125
E 187	ATG***HFH	GA***MSSH	GGV***HLL	*****DAYH	NYL*****H
SC****DHL	GA***MSSH	*PL***F*H	ATG***HFH	REG***QFH	NYL*****H
*IL***SVL	*PL***F*H	REG***QFH	GA***MSSH	R****KFDH	NYL*****H
N****D*PE	REG***QFH	R****KFDH	*PL***F*H	N****D*PE	L****KQ*E
SC****DHL	N****D*PE	N****D*PE	*****DAYH	D*G***FVE	*K***S*AL
*IL***SVL	D*G***FVE	D*G***FVE	REG***QFH	SC****DHL	NYL*****H
ATG***HFH	SC****DHL	SC****DHL	R****KFDH	*IL***SVL	L****KQ*E
REG***QFH	*IL***SVL	*IL***SVL	N****D*PE	GGV***HLL	*K***S*AL
N****D*PE	GGV***HLL	GGV***HLL	D*G***FVE	ATG***HFH	NYL*****H
SC****DHL	ATG***HFH	ATG***HFH	SC****DHL	RFA***P*H	VYV***GPE

L****KQ*E	NYL*****H	L****KQ*E	*K***S*AL	GEG****LE	TAQ***PYL
*K***S*AL	VYV***GPE	G****D*EL		F****D*IL	K*S***VEH
NYL*****H	L****KQ*E	Y*F****LL	H 340	TAQ***PYL	HD***L**H
VYV***GPE	Y*F****LL	*K***S*AL	K*S***VEH	K*S***VEH	NR***K**H
L****KQ*E	*K***S*AL	PQ***DDMH	HD***L**H	HD***L**H	PWE****MH
*K***S*AL	PQ***DDMH	*RD****WH	NR***K**H	NR***K**H	*****SQLH
NYL*****H	*RD****WH	NYL*****H	C****ER*H	C****ER*H	C****ER*H
VYV***GPE	NYL*****H	VYV***GPE	*G***KT*E	*G***KT*E	*G***KT*E
L****KQ*E	VYV***GPE	L****KQ*E	E****GVCE	E****GVCE	GRL*****E
Y*F****LL	L****KQ*E	G****D*EL	F****D*IL	I*G***HCE	E****GVCE
*K***S*AL	Y*F****LL	Y*F****LL	K*S***VEH	*NL**STCE	I*G***HCE
PQ***DDMH	*K***S*AL	*K***S*AL	HD***L**H	GEG****LE	*NL**STCE
NYL*****H	PQ***DDMH	PQ***DDMH	NR***K**H	F****D*IL	GEG****LE
VYV***GPE	*RD****WH	*RD****WH	C****ER*H	TAQ***PYL	F****D*IL
L****KQ*E	NYL*****H	WWR**V**H	*G***KT*E	K*S***VEH	*W***HRDL
Y*F****LL	VYV***GPE	NYL*****H	E****GVCE	HD***L**H	TAQ***PYL
*K***S*AL	L****KQ*E	VYV***GPE	I*G***HCE	NR***K**H	K*S***VEH
PQ***DDMH	Y*F****LL	L****KQ*E	F****D*IL	C****ER*H	HD***L**H
NYL*****H	*K***S*AL	G****D*EL	TAQ***PYL	*G***KT*E	NR***K**H
VYV***GPE	PQ***DDMH	Y*F****LL	K*S***VEH	GRL*****E	PWE****MH
L****KQ*E	*RD****WH	*K***S*AL	HD***L**H	E****GVCE	*****SQLH
Y*F****LL	NYL*****H	PQ***DDMH	NR***K**H	I*G***HCE	C****ER*H
*K***S*AL	VYV***GPE	*RD****WH	C****ER*H	*NL**STCE	*G***KT*E
PQ***DDMH	L****KQ*E	WWR**V**H	*G***KT*E	GEG****LE	GRL*****E
NYL*****H	Y*F****LL	NYL*****H	E****GVCE	F****D*IL	E****GVCE
VYV***GPE	*K***S*AL	VYV***GPE	I*G***HCE	TAQ***PYL	I*G***HCE
L****KQ*E	PQ***DDMH	L****KQ*E	*NL**STCE	K*S***VEH	*NL**STCE
Y*F****LL	*RD****WH	G****D*EL	GEG****LE	HD***L**H	GEG****LE
*K***S*AL	NYL*****H	Y*F****LL	F****D*IL	NR***K**H	F****D*IL
PQ***DDMH	VYV***GPE	*K***S*AL	TAQ***PYL	*****SQLH	*W***HRDL
*RD****WH	L****KQ*E	PQ***DDMH	K*S***VEH	C****ER*H	TAQ***PYL
NYL*****H	G****D*EL	*RD****WH	HD***L**H	*G***KT*E	K*S***VEH
VYV***GPE	Y*F****LL	WWR**V**H	NR***K**H	GRL*****E	HD***L**H
L****KQ*E	*K***S*AL	NYL*****H	C****ER*H	E****GVCE	NR***K**H
Y*F****LL	PQ***DDMH	VYV***GPE	*G***KT*E	I*G***HCE	PWE****MH
*K***S*AL	*RD****WH	L****KQ*E	E****GVCE	*NL**STCE	*****SQLH
PQ***DDMH	NYL*****H	G****D*EL	I*G***HCE	GEG****LE	C****ER*H
*RD****WH	VYV***GPE	Y*F****LL	*NL**STCE	F****D*IL	*G***KT*E

GRL*****E	*NL**STCE	F****D*IL	GEG****LE	*CS***MFE	*G***KT*E
E****GVCE	GEG****LE	*F***LD*L	F****D*IL	*NL**STCE	GRL*****E
I*G***HCE	F****D*IL	*W***HRDL	*F***LD*L	GEG****LE	CHI*****E
*CS***MFE	*F***LD*L	TAQ***PYL	*W***HRDL	F****D*IL	HH***GP*E
*NL**STCE	*W***HRDL	K*S***VEH	TAQ***PYL	*F***LD*L	YI***DQFE
GEG****LE	TAQ***PYL	HD***L**H	*K***DPVL	*W***HRDL	E****GVCE
F****D*IL	K*S***VEH	NR***K**H	K*S***VEH	TAQ***PYL	I*G***HCE
*W***HRDL	HD***L**H	PWE****MH	HD***L**H	*K***DPVL	*CS***MFE
TAQ***PYL	NR***K**H	*****SQLH	NR***K**H	K*S***VEH	*NL**STCE
K*S***VEH	PWE****MH	C****ER*H	PWE****MH	HD***L**H	GEG****LE
HD***L**H	*****SQLH	*G***KT*E	*****SQLH	CWK**I*EH	F****D*IL
NR***K**H	C****ER*H	GRL*****E	C****ER*H	NR***K**H	*F***LD*L
PWE****MH	*G***KT*E	HH***GP*E	*G***KT*E	PWE****MH	*W***HRDL
*****SQLH	GRL*****E	YI***DQFE	GRL*****E	*****SQLH	TAQ***PYL
C****ER*H	HH***GP*E	E****GVCE	HH***GP*E	C****ER*H	DP***NHEL
*G***KT*E	E****GVCE	I*G***HCE	YI***DQFE	*G***KT*E	*K***DPVL
GRL*****E	I*G***HCE	*CS***MFE	E****GVCE	GRL*****E	K*S***VEH
HH***GP*E	*CS***MFE	*NL**STCE	I*G***HCE	CHI*****E	HD***L**H
E****GVCE	*NL**STCE	GEG****LE	*CS***MFE	HH***GP*E	CWK**I*EH
I*G***HCE	GEG****LE	F****D*IL	*NL**STCE	YI***DQFE	NR***K**H
*CS***MFE	F****D*IL	*F***LD*L	GEG****LE	E****GVCE	PWE****MH
*NL**STCE	*F***LD*L	*W***HRDL	F****D*IL	I*G***HCE	*****SQLH
GEG****LE	*W***HRDL	TAQ***PYL	*F***LD*L	*CS***MFE	C****ER*H
F****D*IL	TAQ***PYL	*K***DPVL	*W***HRDL	*NL**STCE	*G***KT*E
*W***HRDL	K*S***VEH	K*S***VEH	TAQ***PYL	GEG****LE	GRL*****E
TAQ***PYL	HD***L**H	HD***L**H	*K***DPVL	F****D*IL	CHI*****E
K*S***VEH	NR***K**H	NR***K**H	K*S***VEH	*F***LD*L	HH***GP*E
HD***L**H	PWE****MH	PWE****MH	HD***L**H	*W***HRDL	YI***DQFE
NR***K**H	*****SQLH	*****SQLH	NR***K**H	TAQ***PYL	E****GVCE
PWE****MH	C****ER*H	C****ER*H	PWE****MH	DP***NHEL	I*G***HCE
*****SQLH	*G***KT*E	*G***KT*E	*****SQLH	*K***DPVL	*CS***MFE
C****ER*H	GRL*****E	GRL*****E	C****ER*H	K*S***VEH	*NL**STCE
*G***KT*E	HH***GP*E	HH***GP*E	*G***KT*E	HD***L**H	GEG****LE
GRL*****E	E****GVCE	YI***DQFE	GRL*****E	CWK**I*EH	F****D*IL
HH***GP*E	I*G***HCE	E****GVCE	HH***GP*E	NR***K**H	*F***LD*L
E****GVCE	*CS***MFE	I*G***HCE	YI***DQFE	PWE****MH	*W***HRDL
I*G***HCE	*NL**STCE	*CS***MFE	E****GVCE	*****SQLH	TAQ***PYL
*CS***MFE	GEG****LE	*NL**STCE	I*G***HCE	C****ER*H	DP***NHEL

*K***DPVL	*L***M*NH	*L***M*NH	G****EF*H	NEQ***AVH	*IL****YH
	NEQ***AVH	NEQ***AVH	ISI****GE	G****EF*H	*DA***N*H
I 237	ISI****GE	G****EF*H	W*N***CLE	ISI****GE	*****AQYH
*****AQYH	W*N***CLE	ISI****GE	LPM**IYWE	W*N***CLE	**V***FCH
ISI****GE	ETQ****SE	W*N***CLE	ETQ****SE	LPM**IYWE	*L***M*NH
ETQ****SE	ALG***SHL	LPM**IYWE	ALG***SHL	ETQ****SE	NEQ***AVH
ALG***SHL	M****F*QL	ETQ****SE	M****F*QL	ALG***SHL	G****EF*H
*****AQYH	L*D****WL	ALG***SHL	L*D****WL	M****F*QL	ISI****GE
ISI****GE	*****AQYH	M****F*QL	*IL****YH	L*D****WL	GC***P**E
W*N***CLE	*L***M*NH	L*D****WL	*****AQYH	*IL****YH	W*N***CLE
ETQ****SE	NEQ***AVH	*IL****YH	*L***M*NH	*DA***N*H	LPM**IYWE
ALG***SHL	ISI****GE	*****AQYH	NEQ***AVH	*****AQYH	ETQ****SE
*****AQYH	W*N***CLE	*L***M*NH	G****EF*H	**V***FCH	EHV**MAWL
ISI****GE	ETQ****SE	NEQ***AVH	ISI****GE	*L***M*NH	ALG***SHL
W*N***CLE	ALG***SHL	G****EF*H	W*N***CLE	NEQ***AVH	M****F*QL
ETQ****SE	M****F*QL	ISI****GE	LPM**IYWE	G****EF*H	L*D****WL
ALG***SHL	L*D****WL	W*N***CLE	ETQ****SE	ISI****GE	*IL****YH
M****F*QL	*****AQYH	LPM**IYWE	ALG***SHL	W*N***CLE	*DA***N*H
L*D****WL	*L***M*NH	ETQ****SE	M****F*QL	LPM**IYWE	*****AQYH
*****AQYH	NEQ***AVH	ALG***SHL	L*D****WL	ETQ****SE	**V***FCH
*L***M*NH	ISI****GE	M****F*QL	*IL****YH	ALG***SHL	*L***M*NH
NEQ***AVH	W*N***CLE	L*D****WL	*DA***N*H	M****F*QL	NEQ***AVH
ISI****GE	ETQ****SE	*IL****YH	*****AQYH	L*D****WL	G****EF*H
W*N***CLE	ALG***SHL	*****AQYH	*L***M*NH	*IL****YH	ISI****GE
ETQ****SE	M****F*QL	*L***M*NH	NEQ***AVH	*DA***N*H	GC***P**E
ALG***SHL	L*D****WL	NEQ***AVH	G****EF*H	*****AQYH	HDS***IDE
M****F*QL	*****AQYH	G****EF*H	ISI****GE	**V***FCH	W*N***CLE
L*D****WL	*L***M*NH	ISI****GE	W*N***CLE	*L***M*NH	LPM**IYWE
*****AQYH	NEQ***AVH	W*N***CLE	LPM**IYWE	NEQ***AVH	ETQ****SE
*L***M*NH	G****EF*H	LPM**IYWE	ETQ****SE	G****EF*H	EHV**MAWL
NEQ***AVH	ISI****GE	ETQ****SE	ALG***SHL	ISI****GE	ALG***SHL
ISI****GE	W*N***CLE	ALG***SHL	M****F*QL	GC***P**E	M****F*QL
W*N***CLE	ETQ****SE	M****F*QL	L*D****WL	W*N***CLE	L*D****WL
ETQ****SE	ALG***SHL	L*D****WL	*IL****YH	LPM**IYWE	*IL****YH
ALG***SHL	M****F*QL	*IL****YH	*DA***N*H	ETQ****SE	*DA***N*H
M****F*QL	L*D****WL	*****AQYH	*****AQYH	ALG***SHL	*****AQYH
L*D****WL	*IL****YH	*L***M*NH	**V***FCH	M****F*QL	**V***FCH
*****AQYH	*****AQYH	NEQ***AVH	*L***M*NH	L*D****WL	*L***M*NH

NEQ***AVH
G****EF*H
ISI****GE
GC***P**E
HDS***IDE
W*N***CLE
LPM**IYWE
ETQ****SE
EHV**MAWL
ALG***SHL
M****F*QL
L*D****WL

L 103
HVL***Q*H
I****F*EL
HVL***Q*H
I****F*EL
HVL***Q*H
I****F*EL
HVL***Q*H
I****F*EL
HVL***Q*H
I****F*EL
HVL***Q*H
V*I***K*E
I****F*EL
HVL***Q*H
V*I***K*E
I****F*EL
HVL***Q*H
V*I***K*E

I****F*EL
LQL*****L
HVL***Q*H
V*I***K*E
N****ML*L
I****F*EL
LQL*****L
HVL***Q*H
V*I***K*E
N****ML*L
I****F*EL
LQL*****L
HVL***Q*H
V*I***K*E
N****RILE
N****ML*L
Q****APIL
I****F*EL
LQL*****L
HVL***Q*H
V*I***K*E
N****RILE
QD***K*GL
N****ML*L
HVL***Q*H
V*I***K*E
N****RILE

QD***K*GL
N****ML*L
Q****APIL
I****F*EL
LQL*****L
HVL***Q*H
V*I***K*E
N****RILE
QD***K*GL
N****ML*L
Q****APIL
I****F*EL
LQL*****L
HVL***Q*H
YAR***FLE
V*I***K*E
N****RILE
QD***K*GL
N****ML*L
Q****APIL
I****F*EL
LQL*****L
FTA***LKH
HVL***Q*H
YAR***FLE
V*I***K*E
*G***LYTE
N****RILE
QD***K*GL
N****ML*L

Q****APIL
I****F*EL
LQL*****L

K 139
QI***AR*H
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H
D****QHAE
DC***K*TL
LSM***YAL
DDS****PL
QI***AR*H

D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
DDS****PL
GII**LDWH
QI***AR*H

NFM**I*HL
LHQ***PPL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
*NA***DCL
DDS****PL
NFM**I*HL
WCI****IL
LHQ***PPL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
*NA***DCL
DDS****PL
NFM**I*HL
WCI****IL
LHQ***PPL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
*NA***DCL
DDS****PL
NFM**I*HL
WCI****IL
LHQ***PPL
GII**LDWH
QI***AR*H
D****QHAE
DC***K*TL
*NA***DCL
DDS****PL
NFM**I*HL
WCI****IL
LHQ***PPL

GII**LDWH	*L***LM*E	IWC***TPH	MVE***HYH	Q*E***QHH	GC***LMAL
QI***AR*H	*****VGFL	Q*E***QHH	GLR*****H	S*V**NPHE	*****VGFL
D****QHAE	PS***G**L	S*V**NPHE	P****CCGH	EYS****IE	PS***G**L
DC***K*TL	MVE***HYH	*L***LM*E	*H***NKNH	*L***LM*E	K*V**SDVL
*NA***DCL	Q*E***QHH	GC***LMAL	IWC***TPH	GC***LMAL	MCV**MHGL
DDS****PL	*L***LM*E	*****VGFL	QFE***QRH	*****VGFL	IPF***V*L
NFM**I*HL	*****VGFL	PS***G**L	Q*E***QHH	PS***G**L	SQS***D*L
WCI****IL	PS***G**L	K*V**SDVL	S*V**NPHE	K*V**SDVL	MVE***HYH
LHQ***PPL	IPF***V*L	IPF***V*L	EYS****IE	MCV**MHGL	GLR*****H
GII**LDWH	MVE***HYH	SQS***D*L	*L***LM*E	IPF***V*L	P****CCGH
CQS*****H	Q*E***QHH	MVE***HYH	GC***LMAL	SQS***D*L	*H***NKNH
QI***AR*H	*L***LM*E	GLR*****H	*****VGFL	MVE***HYH	IWC***TPH
D****QHAE	*****VGFL	P****CCGH	PS***G**L	GLR*****H	QFE***QRH
DC***K*TL	PS***G**L	IWC***TPH	K*V**SDVL	P****CCGH	Q*E***QHH
*NA***DCL	IPF***V*L	Q*E***QHH	IPF***V*L	*H***NKNH	S*V**NPHE
DDS****PL	SQS***D*L	S*V**NPHE	SQS***D*L	IWC***TPH	EYS****IE
NFM**I*HL	MVE***HYH	*L***LM*E	MVE***HYH	QFE***QRH	*L***LM*E
WCI****IL	GLR*****H	GC***LMAL	GLR*****H	Q*E***QHH	QWD**VC*L
LHQ***PPL	Q*E***QHH	*****VGFL	P****CCGH	S*V**NPHE	GC***LMAL
*****MW*H	S*V**NPHE	PS***G**L	*H***NKNH	EYS****IE	*****VGFL
GII**LDWH	*L***LM*E	K*V**SDVL	IWC***TPH	*L***LM*E	PS***G**L
CQS*****H	*****VGFL	IPF***V*L	QFE***QRH	GC***LMAL	K*V**SDVL
QI***AR*H	PS***G**L	SQS***D*L	Q*E***QHH	*****VGFL	MCV**MHGL
D****QHAE	IPF***V*L	MVE***HYH	S*V**NPHE	PS***G**L	IPF***V*L
DC***K*TL	SQS***D*L	GLR*****H	EYS****IE	K*V**SDVL	SQS***D*L
*NA***DCL	MVE***HYH	P****CCGH	*L***LM*E	MCV**MHGL	MVE***HYH
DDS****PL	GLR*****H	IWC***TPH	GC***LMAL	IPF***V*L	GLR*****H
NFM**I*HL	Q*E***QHH	QFE***QRH	*****VGFL	SQS***D*L	P****CCGH
WCI****IL	S*V**NPHE	Q*E***QHH	PS***G**L	MVE***HYH	*H***NKNH
LHQ***PPL	*L***LM*E	S*V**NPHE	K*V**SDVL	GLR*****H	IWC***TPH
	GC***LMAL	EYS****IE	IPF***V*L	P****CCGH	QFE***QRH
M 297	*****VGFL	*L***LM*E	SQS***D*L	*H***NKNH	Q*E***QHH
Q*E***QHH	PS***G**L	GC***LMAL	MVE***HYH	IWC***TPH	S*V**NPHE
*L***LM*E	K*V**SDVL	*****VGFL	GLR*****H	QFE***QRH	EYS****IE
*****VGFL	IPF***V*L	PS***G**L	P****CCGH	Q*E***QHH	*L***LM*E
PS***G**L	SQS***D*L	K*V**SDVL	*H***NKNH	S*V**NPHE	QWD**VC*L
MVE***HYH	MVE***HYH	IPF***V*L	IWC***TPH	EYS****IE	GC***LMAL
Q*E***QHH	GLR*****H	SQS***D*L	QFE***QRH	*L***LM*E	*****VGFL

PS***G**L	MCV**MHGL	K*V**SDVL	G****VQAE	*AF****PL	L****IH*H
K*V**SDVL	IPF***V*L	MCV**MHGL	GEV***A*L	L****IH*H	*KD***LWH
MCV**MHGL	SQS***D*L	IPF***V*L	*S***A*PL	*KD***LWH	F*D***NPE
IPF***V*L	MVE***HYH	SQS***D*L	*AF****PL	GKF***TYH	G****VQAE
SQS***D*L	GLR*****H	CWM**LPNH	GKF***TYH	F*D***NPE	PS***H*GL
MVE***HYH	P****CCGH	WWH**PN*H	G****VQAE	G****VQAE	GEV***A*L
GLR*****H	YD***Q*SH	MVE***HYH	GEV***A*L	GEV***A*L	*S***A*PL
P****CCGH	*H***NKNH	GLR*****H	*S***A*PL	*S***A*PL	E****YW*L
*H***NKNH	IWC***TPH	P****CCGH	*AF****PL	E****YW*L	*AF****PL
IWC***TPH	QFE***QRH	YD***Q*SH	GKF***TYH	*AF****PL	L****IH*H
QFE***QRH	Q*E***QHH	*H***NKNH	G****VQAE	L****IH*H	*KD***LWH
Q*E***QHH	S*V**NPHE	IWC***TPH	GEV***A*L	*KD***LWH	F*D***NPE
S*V**NPHE	EYS****IE	QFE***QRH	*S***A*PL	GKF***TYH	G****VQAE
EYS****IE	*L***LM*E	Q*E***QHH	*AF****PL	F*D***NPE	PS***H*GL
*L***LM*E	QWD**VC*L	S*V**NPHE	GKF***TYH	G****VQAE	GEV***A*L
QWD**VC*L	GC***LMAL	EYS****IE	G****VQAE	GEV***A*L	*S***A*PL
GC***LMAL	*****VGFL	*L***LM*E	GEV***A*L	*S***A*PL	E****YW*L
*****VGFL	PS***G**L	QWD**VC*L	*S***A*PL	E****YW*L	*AF****PL
PS***G**L	K*V**SDVL	GC***LMAL	*AF****PL	*AF****PL	L****IH*H
K*V**SDVL	MCV**MHGL	*****VGFL	GKF***TYH	L****IH*H	*KD***LWH
MCV**MHGL	IPF***V*L	PS***G**L	F*D***NPE	*KD***LWH	F*D***NPE
IPF***V*L	SQS***D*L	K*V**SDVL	G****VQAE	GKF***TYH	G****VQAE
SQS***D*L	WWH**PN*H	MCV**MHGL	GEV***A*L	F*D***NPE	PS***H*GL
MVE***HYH	MVE***HYH	IPF***V*L	*S***A*PL	G****VQAE	GEV***A*L
GLR*****H	GLR*****H	SQS***D*L	*AF****PL	GEV***A*L	*S***A*PL
P****CCGH	P****CCGH		GKF***TYH	*S***A*PL	E****YW*L
*H***NKNH	YD***Q*SH	F 159	F*D***NPE	E****YW*L	*AF****PL
IWC***TPH	*H***NKNH	GKF***TYH	G****VQAE	*AF****PL	L****IH*H
QFE***QRH	IWC***TPH	G****VQAE	GEV***A*L	L****IH*H	*KD***LWH
Q*E***QHH	QFE***QRH	GEV***A*L	*S***A*PL	*KD***LWH	F*D***NPE
S*V**NPHE	Q*E***QHH	*S***A*PL	E****YW*L	GKF***TYH	G****VQAE
EYS****IE	S*V**NPHE	*AF****PL	*AF****PL	F*D***NPE	PS***H*GL
*L***LM*E	EYS****IE	GKF***TYH	GKF***TYH	G****VQAE	GEV***A*L
QWD**VC*L	*L***LM*E	G****VQAE	F*D***NPE	PS***H*GL	*S***A*PL
GC***LMAL	QWD**VC*L	GEV***A*L	G****VQAE	GEV***A*L	E****YW*L
*****VGFL	GC***LMAL	*S***A*PL	GEV***A*L	*S***A*PL	*AF****PL
PS***G**L	*****VGFL	*AF****PL	*S***A*PL	E****YW*L	L****IH*H
K*V**SDVL	PS***G**L	GKF***TYH	E****YW*L	*AF****PL	*KD***LWH

F*D***NPE	EW***M**L	E*E***VME	II***S*PH	*QL****RE	E*E***VME
G****VQAE	**T***TIL	*K***A*ME	FYV***D*E	MYI*****L	*K***A*ME
PS***H*GL	KTK**QWRL	*QL****RE	**N***DKE	LWA***C*L	N****MW*E
GEV***A*L	ALK***YKL	MYI*****L	E*E***VME	EW***M**L	*QL****RE
*S***A*PL	*QL****RE	EW***M**L	*K***A*ME	**T***TIL	MYI*****L
E****YW*L	MYI*****L	**T***TIL	*QL****RE	KTK**QWRL	LWA***C*L
*AF****PL	EW***M**L	KTK**QWRL	MYI*****L	ALK***YKL	EW***M**L
L****IH*H	**T***TIL	ALK***YKL	EW***M**L	DRV***WVL	**T***TIL
*KD***LWH	KTK**QWRL	KPV**T*YH	**T***TIL	WY***RIDL	KTK**QWRL
F*D***NPE	ALK***YKL	C*F***D*H	KTK**QWRL	KPV**T*YH	ALK***YKL
G****VQAE	E*E***VME	II***S*PH	ALK***YKL	WE***VR*H	DRV***WVL
PS***H*GL	*K***A*ME	**N***DKE	DRV***WVL	*A***IF*H	WY***RIDL
GEV***A*L	*QL****RE	E*E***VME	KPV**T*YH	C*F***D*H	KPV**T*YH
*S***A*PL	MYI*****L	*K***A*ME	WE***VR*H	II***S*PH	MWP****YH
E****YW*L	EW***M**L	*QL****RE	C*F***D*H	SM***TY*E	WE***VR*H
*AF****PL	**T***TIL	MYI*****L	II***S*PH	FYV***D*E	*A***IF*H
L****IH*H	KTK**QWRL	EW***M**L	FYV***D*E	**N***DKE	EW***FIWH
*KD***LWH	ALK***YKL	**T***TIL	**N***DKE	E*E***VME	PVG***M*H
F*D***NPE	C*F***D*H	KTK**QWRL	E*E***VME	*K***A*ME	C*F***D*H
G****VQAE	E*E***VME	ALK***YKL	*K***A*ME	N****MW*E	II***S*PH
PS***H*GL	*K***A*ME	DRV***WVL	N****MW*E	*QL****RE	SM***TY*E
GEV***A*L	*QL****RE	KPV**T*YH	*QL****RE	MYI*****L	FYV***D*E
*S***A*PL	MYI*****L	C*F***D*H	MYI*****L	LWA***C*L	**N***DKE
E****YW*L	EW***M**L	II***S*PH	EW***M**L	EW***M**L	E*E***VME
*AF****PL	**T***TIL	FYV***D*E	**T***TIL	**T***TIL	*K***A*ME
L****IH*H	KTK**QWRL	**N***DKE	KTK**QWRL	KTK**QWRL	N****MW*E
*KD***LWH	ALK***YKL	E*E***VME	ALK***YKL	ALK***YKL	*QL****RE
F*D***NPE	C*F***D*H	*K***A*ME	DRV***WVL	DRV***WVL	SID***PRE
G****VQAE	E*E***VME	*QL****RE	KPV**T*YH	WY***RIDL	MYI*****L
PS***H*GL	*K***A*ME	MYI*****L	WE***VR*H	KPV**T*YH	LWA***C*L
GEV***A*L	*QL****RE	EW***M**L	*A***IF*H	MWP****YH	EW***M**L
*S***A*PL	MYI*****L	**T***TIL	C*F***D*H	WE***VR*H	**T***TIL
E****YW*L	EW***M**L	KTK**QWRL	II***S*PH	*A***IF*H	KTK**QWRL
*AF****PL	**T***TIL	ALK***YKL	FYV***D*E	C*F***D*H	ALK***YKL
	KTK**QWRL	DRV***WVL	**N***DKE	II***S*PH	DRV***WVL
P 365	ALK***YKL	KPV**T*YH	E*E***VME	SM***TY*E	WY***RIDL
*QL****RE	C*F***D*H	WE***VR*H	*K***A*ME	FYV***D*E	KPV**T*YH
MYI*****L	**N***DKE	C*F***D*H	N****MW*E	**N***DKE	MWP****YH

WE***VR*H	*WK**HIDE	PVG***M*H	MYI*****L	EW***FIWH	LW****G*H
*A***IF*H	MYI*****L	C*F***D*H	LWA***C*L	PVG***M*H	*VS***RSH
EW***FIWH	LWA***C*L	II***S*PH	EW***M**L	C*F***D*H	DQM****YH
PVG***M*H	EW***M**L	SM***TY*E	**T***TIL	II***S*PH	M****VELL
C*F***D*H	**T***TIL	FYV***D*E	ALK***YKL	SM***TY*E	LW****G*H
II***S*PH	KTK**QWRL	**N***DKE	DRV***WVL	FYV***D*E	*VS***RSH
SM***TY*E	ALK***YKL	E*E***VME	WY***RIDL	**N***DKE	DQM****YH
FYV***D*E	DRV***WVL	*K***A*ME	IC***G*CH	E*E***VME	M****VELL
N*DKE	WY***RIDL	N****MW*E	KPV**T*YH	*K***A*ME	T****LGRL
E*E***VME	KPV**T*YH	*QL****RE	MWP****YH	N****MW*E	LW****G*H
*K***A*ME	MWP****YH	V*Q***NDE	WE***VR*H	*QL****RE	*VS***RSH
N****MW*E	WE***VR*H	SID***PRE	*A***IF*H	V*Q***NDE	DQM****YH
*QL****RE	*A***IF*H	*WK**HIDE	EW***FIWH	SID***PRE	F*A**L*WL
SID***PRE	EW***FIWH	MYI*****L	PVG***M*H	*WK**HIDE	M****VELL
MYI*****L	PVG***M*H	LWA***C*L	C*F***D*H	MYI*****L	T****LGRL
LWA***C*L	C*F***D*H	EW***M**L	II***S*PH	LWA***C*L	LW****G*H
EW***M**L	II***S*PH	**T***TIL	SM***TY*E	EW***M**L	*VS***RSH
T*TIL	SM***TY*E	ALK***YKL	FYV***D*E	**T***TIL	DQM****YH
KTK**QWRL	FYV***D*E	DRV***WVL	**N***DKE	ALK***YKL	F*A**L*WL
ALK***YKL	**N***DKE	WY***RIDL	E*E***VME	DRV***WVL	M****VELL
DRV***WVL	E*E***VME	KPV**T*YH	*K***A*ME	WY***RIDL	T****LGRL
WY***RIDL	*K***A*ME	MWP****YH	N****MW*E		LW****G*H
KPV**T*YH	N****MW*E	WE***VR*H	*QL****RE	F 162	*VS***RSH
MWP****YH	*QL****RE	*A***IF*H	V*Q***NDE	LW****G*H	DQM****YH
WE***VR*H	SID***PRE	EW***FIWH	SID***PRE	DQM****YH	**H***CME
*A***IF*H	*WK**HIDE	PVG***M*H	*WK**HIDE	M****VELL	F*A**L*WL
EW***FIWH	MYI*****L	C*F***D*H	MYI*****L	LW****G*H	M****VELL
PVG***M*H	LWA***C*L	II***S*PH	LWA***C*L	*VS***RSH	T****LGRL
C*F***D*H	EW***M**L	SM***TY*E	EW***M**L	DQM****YH	LW****G*H
II***S*PH	**T***TIL	FYV***D*E	**T***TIL	M****VELL	*VS***RSH
SM***TY*E	ALK***YKL	**N***DKE	ALK***YKL	LW****G*H	DQM****YH
FYV***D*E	DRV***WVL	E*E***VME	DRV***WVL	*VS***RSH	**H***CME
N*DKE	WY***RIDL	*K***A*ME	WY***RIDL	DQM****YH	F*A**L*WL
E*E***VME	KPV**T*YH	N****MW*E	KPV**T*YH	M****VELL	M****VELL
*K***A*ME	MWP****YH	*QL****RE	MWP****YH	LW****G*H	T****LGRL
N****MW*E	WE***VR*H	V*Q***NDE	*ER****RH	*VS***RSH	LW****G*H
*QL****RE	*A***IF*H	SID***PRE	WE***VR*H	DQM****YH	*VS***RSH
SID***PRE	EW***FIWH	*WK**HIDE	*A***IF*H	M****VELL	DQM****YH

H*CME	*VS***RSH	F*A**L*WL	PVA*****E	PVA*****E	MK***VR*H
F*A**L*WL	DQM****YH	GQE***QAL	GK***L*LH	QLI***V*L	**R***KLH
M****VELL	*M***V*CE	M****VELL	LD***LLRH	IP***MMHL	*CS***VDE
T****LGRL	**H***CME	T****LGRL	*YR***MHE	GK***L*LH	*YR***MHE
LW****G*H	F*A**L*WL	G*V***Y*L	PVA*****E	LD***LLRH	F*M***G*E
*VS***RSH	GQE***QAL	*L***SE*L	IP***MMHL	T*M***CVH	PVA*****E
DQM****YH	M****VELL	S****VRVH	GK***L*LH	**R***KLH	*IM****GE
H*CME	T****LGRL	LW****G*H	LD***LLRH	*CS***VDE	QLI***V*L
F*A**L*WL	*L***SE*L	*VS***RSH	*YR***MHE	*YR***MHE	IP***MMHL
M****VELL	LW****G*H	*C***LPLH	PVA*****E	F*M***G*E	GK***L*LH
T****LGRL	*VS***RSH	DQM****YH	IP***MMHL	PVA*****E	LD***LLRH
LW****G*H	DQM****YH	*M***V*CE	GK***L*LH	QLI***V*L	T*M***CVH
*VS***RSH	*M***V*CE	**H***CME	LD***LLRH	IP***MMHL	MK***VR*H
DQM****YH	**H***CME	F*A**L*WL	*CS***VDE	GK***L*LH	**R***KLH
H*CME	F*A**L*WL	GQE***QAL	*YR***MHE	LD***LLRH	*CS***VDE
F*A**L*WL	GQE***QAL	M****VELL	PVA*****E	T*M***CVH	*YR***MHE
M****VELL	M****VELL	T****LGRL	QLI***V*L	MK***VR*H	F*M***G*E
T****LGRL	T****LGRL	G*V***Y*L	IP***MMHL	**R***KLH	PVA*****E
LW****G*H	G*V***Y*L	*L***SE*L	GK***L*LH	*CS***VDE	*IM****GE
*VS***RSH	*L***SE*L	S****VRVH	LD***LLRH	*YR***MHE	QLI***V*L
DQM****YH	LW****G*H	LW****G*H	*CS***VDE	F*M***G*E	IP***MMHL
*M***V*CE	*VS***RSH	*VS***RSH	*YR***MHE	PVA*****E	GK***L*LH
H*CME	*C***LPLH	*C***LPLH	PVA*****E	QLI***V*L	LD***LLRH
F*A**L*WL	DQM****YH	*QT***NDH	QLI***V*L	IP***MMHL	T*M***CVH
M****VELL	*M***V*CE	DQM****YH	IP***MMHL	GK***L*LH	MK***VR*H
T****LGRL	**H***CME	*M***V*CE	GK***L*LH	LD***LLRH	**R***KLH
*L***SE*L	F*A**L*WL	**H***CME	LD***LLRH	T*M***CVH	*CS***VDE
LW****G*H	GQE***QAL	F*A**L*WL	*CS***VDE	MK***VR*H	*YR***MHE
*VS***RSH	M****VELL	GQE***QAL	*YR***MHE	**R***KLH	F*M***G*E
DQM****YH	T****LGRL	M****VELL	PVA*****E	*CS***VDE	PVA*****E
*M***V*CE	G*V***Y*L	T****LGRL	QLI***V*L	*YR***MHE	*IM****GE
H*CME	*L***SE*L	G*V***Y*L	IP***MMHL	F*M***G*E	QLI***V*L
F*A**L*WL	LW****G*H	*L***SE*L	GK***L*LH	PVA*****E	IP***MMHL
GQE***QAL	*VS***RSH		LD***LLRH	QLI***V*L	GK***L*LH
M****VELL	*C***LPLH	T 236	T*M***CVH	IP***MMHL	LD***LLRH
T****LGRL	DQM****YH	GK***L*LH	*CS***VDE	GK***L*LH	T*M***CVH
*L***SE*L	*M***V*CE	LD***LLRH	*YR***MHE	LD***LLRH	MK***VR*H
LW****G*H	**H***CME	*YR***MHE	F*M***G*E	T*M***CVH	**R***KLH

*CS***VDE	MK***VR*H	W*S**THEE	RDA****VL	FS***R*LH	T****Y*IE
*YR***MHE	**R***KLH	*IM****GE	QLI***V*L	*D***SA*H	*H***WCYE
YD***ML*E	*CS***VDE	QLI***V*L	NRL****IL	T****Y*IE	D*K***IKE
F*M***G*E	*YR***MHE	NRL****IL	IP***MMHL	*H***WCYE	*GK***GGE
PVA*****E	SWP****ME	IP***MMHL	FTF***HEL	D*K***IKE	*AM**VYPL
*IM****GE	YD***ML*E	FTF***HEL		*GK***GGE	PDL***QQL
QLI***V*L	F*M***G*E	GK***L*LH	W 368	*AM**VYPL	G****MY*L
IP***MMHL	PVA*****E	LD***LLRH	FS***R*LH	PDL***QQL	K*M***L*L
GK***L*LH	*IM****GE	T*M***CVH	*D***SA*H	G****MY*L	SNQ***K*L
LD***LLRH	QLI***V*L	MK***VR*H	T****Y*IE	K*M***L*L	*NI**MHKL
T*M***CVH	IP***MMHL	**R***KLH	D*K***IKE	SNQ***K*L	VD***T**H
MK***VR*H	GK***L*LH	*CS***VDE	*GK***GGE	VD***T**H	FS***R*LH
R*KLH	LD***LLRH	*YR***MHE	*AM**VYPL	FS***R*LH	*D***SA*H
*CS***VDE	T*M***CVH	SWP****ME	PDL***QQL	*D***SA*H	T****Y*IE
*YR***MHE	MK***VR*H	YD***ML*E	G****MY*L	T****Y*IE	*H***WCYE
YD***ML*E	**R***KLH	F*M***G*E	K*M***L*L	*H***WCYE	D*K***IKE
F*M***G*E	*CS***VDE	PVA*****E	FS***R*LH	D*K***IKE	*GK***GGE
PVA*****E	*YR***MHE	W*S**THEE	*D***SA*H	*GK***GGE	*AM**VYPL
*IM****GE	SWP****ME	*IM****GE	T****Y*IE	*AM**VYPL	PDL***QQL
QLI***V*L	YD***ML*E	RDA****VL	D*K***IKE	PDL***QQL	G****MY*L
IP***MMHL	F*M***G*E	QLI***V*L	*GK***GGE	G****MY*L	S*E***HLL
GK***L*LH	PVA*****E	NRL****IL	*AM**VYPL	K*M***L*L	K*M***L*L
LD***LLRH	*IM****GE	IP***MMHL	PDL***QQL	SNQ***K*L	SNQ***K*L
T*M***CVH	QLI***V*L	FTF***HEL	G****MY*L	VD***T**H	VSA***T*L
MK***VR*H	NRL****IL	GK***L*LH	K*M***L*L	FS***R*LH	*NI**MHKL
R*KLH	IP***MMHL	LD***LLRH	VD***T**H	*D***SA*H	VD***T**H
*CS***VDE	FTF***HEL	T*M***CVH	FS***R*LH	T****Y*IE	ALQ***SDH
*YR***MHE	GK***L*LH	MK***VR*H	*D***SA*H	*H***WCYE	FS***R*LH
SWP****ME	LD***LLRH	H****FMMH	T****Y*IE	D*K***IKE	*D***SA*H
YD***ML*E	T*M***CVH	**R***KLH	*H***WCYE	*GK***GGE	T****Y*IE
F*M***G*E	MK***VR*H	*CS***VDE	D*K***IKE	*AM**VYPL	*H***WCYE
PVA*****E	**R***KLH	*YR***MHE	*GK***GGE	PDL***QQL	D*K***IKE
*IM****GE	*CS***VDE	SWP****ME	*AM**VYPL	G****MY*L	*GK***GGE
QLI***V*L	*YR***MHE	YD***ML*E	PDL***QQL	K*M***L*L	*AM**VYPL
IP***MMHL	SWP****ME	F*M***G*E	G****MY*L	SNQ***K*L	PDL***QQL
GK***L*LH	YD***ML*E	PVA*****E	K*M***L*L	VD***T**H	G****MY*L
LD***LLRH	F*M***G*E	W*S**THEE	SNQ***K*L	FS***R*LH	S*E***HLL
T*M***CVH	PVA*****E	*IM****GE	VD***T**H	*D***SA*H	K*M***L*L

SNQ***K*L	VD***T**H	*NI**MHKL	VSA***T*L	*H***R*HE	ALQ***SDH
VSA***T*L	ALQ***SDH	W*K**K*HH	AWK*****L	*AM**VYPL	VYE****YH
*NI**MHKL	FS***R*LH	VD***T**H	*NI**MHKL	PDL***QQL	QKI**RKPH
VD***T**H	*D***SA*H	ALQ***SDH	W*K**K*HH	G****MY*L	FS***R*LH
ALQ***SDH	T****Y*IE	FS***R*LH	VD***T**H	E****DEKL	*D***SA*H
FS***R*LH	*H***WCYE	*D***SA*H	F*V**RCWH	S*E***HLL	T****Y*IE
*D***SA*H	D*K***IKE	T****Y*IE	ALQ***SDH	K*M***L*L	D*K***IKE
T****Y*IE	*GK***GGE	*H***WCYE	VYE****YH	P*S***HTL	*GK***GGE
*H***WCYE	*H***R*HE	D*K***IKE	QKI**RKPH	SNQ***K*L	*H***R*HE
D*K***IKE	*AM**VYPL	*GK***GGE	FS***R*LH	VSA***T*L	GGT**C*NL
*GK***GGE	PDL***QQL	*H***R*HE	*D***SA*H	AWK*****L	*AM**VYPL
*AM**VYPL	G****MY*L	*AM**VYPL	T****Y*IE	*NI**MHKL	PDL***QQL
PDL***QQL	S*E***HLL	PDL***QQL	*H***WCYE	W*K**K*HH	G****MY*L
G****MY*L	K*M***L*L	G****MY*L	D*K***IKE	VD***T**H	E****DEKL
S*E***HLL	P*S***HTL	S*E***HLL	*GK***GGE	F*V**RCWH	S*E***HLL
K*M***L*L	SNQ***K*L	K*M***L*L	*H***R*HE	ALQ***SDH	K*M***L*L
SNQ***K*L	VSA***T*L	P*S***HTL	*AM**VYPL	VYE****YH	P*S***HTL
VSA***T*L	AWK*****L	SNQ***K*L	PDL***QQL	QKI**RKPH	SNQ***K*L
*NI**MHKL	*NI**MHKL	VSA***T*L	G****MY*L	FS***R*LH	VSA***T*L
VD***T**H	W*K**K*HH	AWK*****L	S*E***HLL	*D***SA*H	AWK*****L
ALQ***SDH	VD***T**H	*NI**MHKL	K*M***L*L	T****Y*IE	*NI**MHKL
FS***R*LH	ALQ***SDH	W*K**K*HH	P*S***HTL	D*K***IKE	W*K**K*HH
*D***SA*H	FS***R*LH	VD***T**H	SNQ***K*L	*GK***GGE	VD***T**H
T****Y*IE	*D***SA*H	ALQ***SDH	VSA***T*L	*H***R*HE	F*V**RCWH
*H***WCYE	T****Y*IE	FS***R*LH	AWK*****L	*AM**VYPL	ALQ***SDH
D*K***IKE	*H***WCYE	*D***SA*H	*NI**MHKL	PDL***QQL	VYE****YH
*GK***GGE	D*K***IKE	T****Y*IE	W*K**K*HH	G****MY*L	QKI**RKPH
*AM**VYPL	*GK***GGE	*H***WCYE	VD***T**H	E****DEKL	FS***R*LH
PDL***QQL	*H***R*HE	D*K***IKE	F*V**RCWH	S*E***HLL	*D***SA*H
G****MY*L	*AM**VYPL	*GK***GGE	ALQ***SDH	K*M***L*L	T****Y*IE
S*E***HLL	PDL***QQL	*H***R*HE	VYE****YH	P*S***HTL	D*K***IKE
K*M***L*L	G****MY*L	*AM**VYPL	QKI**RKPH	SNQ***K*L	WDH**F*AE
P*S***HTL	S*E***HLL	PDL***QQL	FS***R*LH	VSA***T*L	*GK***GGE
TDC***H*L	K*M***L*L	G****MY*L	*D***SA*H	AWK*****L	*H***R*HE
SNQ***K*L	P*S***HTL	S*E***HLL	T****Y*IE	*NI**MHKL	GGT**C*NL
VSA***T*L	SNQ***K*L	K*M***L*L	*H***WCYE	W*K**K*HH	*AM**VYPL
AWK*****L	VSA***T*L	P*S***HTL	D*K***IKE	VD***T**H	PDL***QQL
*NI**MHKL	AWK*****L	SNQ***K*L	*GK***GGE	F*V**RCWH	G****MY*L

E****DEKL	PWD**S**E	*M***VTHL	PWD**S**E	DN***I*NE	DN***I*NE
S*E***HLL	DN***I*NE	*PQ***P*L	I****ECPE	TFQ***VHE	TFQ***VHE
K*M***L*L	*PQ***P*L	M****IK*L	DN***I*NE	VEM**M*EL	VEM**M*EL
P*S***HTL	M****IK*L	MTK*****H	TFQ***VHE	*M***VTHL	*M***VTHL
SNQ***K*L	MTK*****H	*P***QRTH	VEM**M*EL	*PQ***P*L	*PQ***P*L
VSA***T*L	*P***QRTH	T*V***YYH	*M***VTHL	M****IK*L	M****IK*L
AWK*****L	T*V***YYH	G*M***QGH	*PQ***P*L	MTK*****H	MTK*****H
*NI**MHKL	G*M***QGH	PWD**S**E	M****IK*L	*P***QRTH	*P***QRTH
W*K**K*HH	PWD**S**E	DN***I*NE	MTK*****H	T*V***YYH	T*V***YYH
VD***T**H	DN***I*NE	TFQ***VHE	*P***QRTH	G*M***QGH	G*M***QGH
F*V**RCWH	*PQ***P*L	*M***VTHL	T*V***YYH	PWD**S**E	PWD**S**E
ALQ***SDH	M****IK*L	*PQ***P*L	G*M***QGH	I****ECPE	I****ECPE
VYE****YH	MTK*****H	M****IK*L	PWD**S**E	DN***I*NE	MQ***RAFE
QKI**RKPH	*P***QRTH	MTK*****H	I****ECPE	TFQ***VHE	DN***I*NE
FS***R*LH	T*V***YYH	*P***QRTH	DN***I*NE	VEM**M*EL	TFQ***VHE
*D***SA*H	G*M***QGH	T*V***YYH	TFQ***VHE	*M***VTHL	VEM**M*EL
T****Y*IE	PWD**S**E	G*M***QGH	VEM**M*EL	*PQ***P*L	*M***VTHL
D*K***IKE	DN***I*NE	PWD**S**E	*M***VTHL	M****IK*L	*PQ***P*L
WDH**F*AE	TFQ***VHE	DN***I*NE	*PQ***P*L	MTK*****H	M****IK*L
*GK***GGE	*PQ***P*L	TFQ***VHE	M****IK*L	*P***QRTH	MTK*****H
*H***R*HE	M****IK*L	*M***VTHL	MTK*****H	T*V***YYH	*P***QRTH
GGT**C*NL	MTK*****H	*PQ***P*L	*P***QRTH	G*M***QGH	T*V***YYH
*AM**VYPL	*P***QRTH	M****IK*L	T*V***YYH	PWD**S**E	G*M***QGH
PDL***QQL	T*V***YYH	MTK*****H	G*M***QGH	I****ECPE	PWD**S**E
G****MY*L	G*M***QGH	*P***QRTH	PWD**S**E	MQ***RAFE	I****ECPE
E****DEKL	PWD**S**E	T*V***YYH	I****ECPE	DN***I*NE	MQ***RAFE
S*E***HLL	DN***I*NE	G*M***QGH	DN***I*NE	TFQ***VHE	DN***I*NE
K*M***L*L	TFQ***VHE	PWD**S**E	TFQ***VHE	VEM**M*EL	TFQ***VHE
P*S***HTL	*M***VTHL	I****ECPE	VEM**M*EL	*M***VTHL	VEM**M*EL
SNQ***K*L	*PQ***P*L	DN***I*NE	*M***VTHL	*PQ***P*L	*M***VTHL
VSA***T*L	M****IK*L	TFQ***VHE	*PQ***P*L	M****IK*L	*PQ***P*L
AWK*****L	MTK*****H	*M***VTHL	M****IK*L	MTK*****H	M****IK*L
*NI**MHKL	*P***QRTH	*PQ***P*L	MTK*****H	*P***QRTH	MTK*****H
	T*V***YYH	M****IK*L	*P***QRTH	T*V***YYH	H****G*SH
Y 247	G*M***QGH	MTK*****H	T*V***YYH	G*M***QGH	*P***QRTH
*P***QRTH	PWD**S**E	*P***QRTH	G*M***QGH	PWD**S**E	T*V***YYH
T*V***YYH	DN***I*NE	T*V***YYH	PWD**S**E	I****ECPE	G*M***QGH
G*M***QGH	TFQ***VHE	G*M***QGH	I****ECPE	MQ***RAFE	PWD**S**E

I****ECPE	MTK*****H	W****KLKE	*NS***HYL	T****MG*E	Q*I***CLE
MQ***RAFE	H****G*SH	*DI***RRE	AN***TKYL	VQP***CWL	W****KLKE
DN***I*NE	*FK***C*H	AN***TKYL	*II***KTL	QHF***GCL	*DI***RRE
TFQ***VHE	*P***QRTH	*RK****DH	*RK****DH	*NS***HYL	F****LYKE
VEM**M*EL	T*V***YYH	TK***A*GH	TK***A*GH	AN***TKYL	T****MG*E
*M***VTHL	G*M***QGH	Q*I***CLE	Q*I***CLE	*II***KTL	*IV***LNE
*PQ***P*L	PWD**S**E	W****KLKE	W****KLKE	*RK****DH	VQP***CWL
M****IK*L	I****ECPE	*DI***RRE	*DI***RRE	ECR***VYH	QHF***GCL
MTK*****H	MQ***RAFE	AN***TKYL	T****MG*E	TK***A*GH	*NS***HYL
H****G*SH	DN***I*NE	*RK****DH	VQP***CWL	*NA***GCH	AN***TKYL
*P***QRTH	H*K***K*E	TK***A*GH	*NS***HYL	Q*I***CLE	*II***KTL
T*V***YYH	TFQ***VHE	Q*I***CLE	AN***TKYL	W****KLKE	*RK****DH
G*M***QGH	VEM**M*EL	W****KLKE	*II***KTL	*DI***RRE	ECR***VYH
PWD**S**E	*M***VTHL	*DI***RRE	*RK****DH	F****LYKE	TK***A*GH
I****ECPE	*PQ***P*L	AN***TKYL	TK***A*GH	T****MG*E	*NA***GCH
MQ***RAFE	M****IK*L	*RK****DH	Q*I***CLE	VQP***CWL	Q*I***CLE
DN***I*NE		TK***A*GH	W****KLKE	QHF***GCL	W****KLKE
TFQ***VHE	V 205	Q*I***CLE	*DI***RRE	*NS***HYL	*DI***RRE
VEM**M*EL	TK***A*GH	W****KLKE	T****MG*E	AN***TKYL	F****LYKE
*M***VTHL	*DI***RRE	*DI***RRE	VQP***CWL	*II***KTL	T****MG*E
*PQ***P*L	AN***TKYL	VQP***CWL	*NS***HYL	*RK****DH	*IV***LNE
M****IK*L	TK***A*GH	AN***TKYL	AN***TKYL	ECR***VYH	VQP***CWL
MTK*****H	Q*I***CLE	*RK****DH	*II***KTL	TK***A*GH	QHF***GCL
H****G*SH	*DI***RRE	TK***A*GH	*RK****DH	*NA***GCH	*NS***HYL
*FK***C*H	AN***TKYL	Q*I***CLE	TK***A*GH	Q*I***CLE	AN***TKYL
*P***QRTH	TK***A*GH	W****KLKE	Q*I***CLE	W****KLKE	*II***KTL
T*V***YYH	Q*I***CLE	*DI***RRE	W****KLKE	*DI***RRE	*RK****DH
G*M***QGH	W****KLKE	T****MG*E	*DI***RRE	F****LYKE	ECR***VYH
PWD**S**E	*DI***RRE	VQP***CWL	T****MG*E	T****MG*E	TK***A*GH
I****ECPE	AN***TKYL	AN***TKYL	VQP***CWL	VQP***CWL	*NA***GCH
MQ***RAFE	TK***A*GH	*II***KTL	*NS***HYL	QHF***GCL	Q*I***CLE
DN***I*NE	Q*I***CLE	*RK****DH	AN***TKYL	*NS***HYL	W****KLKE
H*K***K*E	W****KLKE	TK***A*GH	*II***KTL	AN***TKYL	*DI***RRE
TFQ***VHE	*DI***RRE	Q*I***CLE	*RK****DH	*II***KTL	F****LYKE
VEM**M*EL	AN***TKYL	W****KLKE	TK***A*GH	*RK****DH	T****MG*E
*M***VTHL	*RK****DH	*DI***RRE	Q*I***CLE	ECR***VYH	*IV***LNE
*PQ***P*L	TK***A*GH	T****MG*E	W****KLKE	TK***A*GH	VQP***CWL
M****IK*L	Q*I***CLE	VQP***CWL	*DI***RRE	*NA***GCH	QHF***GCL

*NS***HYL	*NA***GCH	*IV***LNE	*RK****DH	*DI***RRE	*NS***HYL
AN***TKYL	Q*I***CLE	VQP***CWL	ECR***VYH	F****LYKE	AN***TKYL
*II***KTL	W****KLKE	QHF***GCL	TK***A*GH	T****MG*E	*II***KTL
*RK****DH	*DI***RRE	*NS***HYL	*NA***GCH	*IV***LNE	
ECR***VYH	F****LYKE	AN***TKYL	Q*I***CLE	VQP***CWL	
TK***A*GH	T****MG*E	*II***KTL	W****KLKE	QHF***GCL	

Vita

Yen-Wei Chu was born on July. 26, 1969 in Keelung, Taiwan, Republic of China. He received the B. D degree in the Department of Computer and Information Science from National Chiao Tung University, in 1996. Since Sep. 1996, he was a graduate student in the Institute of Computer and Information Science at National Chiao Tung University.

Since Sep. 1998, he is working for the Ph.D. degree in the Institute of Computer Science at National Chiao Tung University. His current research interests include Bioinformatics, Evolutionary Computation, Data Mining and Medical Informatics.