國 立 交 通 大 學

生物資訊研究所

碩 士 論 文

蛋白質主鏈結構預測

The Prediction of the Backbone Conformation from Protein Sequences

研 究 生：施建華

指導教授：黃鎮剛 教授

中 華 民 國 九 十 四 年 七 月

蛋白質主鏈結構預測

The Prediction of the Backbone Conformation from Protein Sequences

研 究 生：施建華　　Student：Chien-Hua Shih

指導教授：黃鎮剛　　Advisor：Jenn-Kang Hwang

國 立 交 通 大 學

生 物 資 訊 研 究 所

碩 士 論 文

A Thesis

Submitted to Department of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Bioinformatics

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

蛋白質主鏈結構預測

學生：施建華　　指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

摘　要

建立蛋白質在空間上的結構有助於瞭解蛋白質的功能與機制。現今可藉由 X-ray 與 NMR 的方法來建構蛋白質的三級結構。但這些方法仍需耗費金錢、人力與時間。目前在 PDB 資料庫中已有三萬多個蛋白質三級結構。若可以利用此資料庫作為知識背景，利用機器學習的方法來預測未知結構的蛋白質序列的結構與功能。這將使得我們可以更容易瞭解蛋白質的功能與機制。

本研究是尋找適合描述蛋白質區域結構的定義，探討由序列預測主鏈結構的可能性。我們採用了兩種不同的定義 Ramachandran Plot 及 Protein Blocks 來定義一個蛋白質殘基的區域結構。並以支持向量機來預測，得到相當不錯的結果。

最後我們將研究蛋白質區域結構的保守性與區域結構熵的關係。發現在經由不同的蛋白質區域結構定義所的到的結構熵有相當的一致性。

The Prediction of the Backbone Conformation from Protein Sequences

Student : Chien-Hua Shih Advisor : Jenn-Kang Hwang

Institute of Bioinformatics

National Chiao Tung University

ABSTRACT

The knowledge of protein structure conformation is useful in understanding the functions and mechanism of proteins. Nowadays, we can use the X-ray or NMR technique to construct the 3-dimensional structure of proteins. But these methods cost lots of time and money. Fortunately, we now have an easier alternative to the knowledge of protein structures. That is, from the increasing amount of data in the protein structure database, such as PDB database. Up to the present, there are 30 thousand records stored in the PDB database. If we use the PDB database as our knowledge and use the machine learning technique to predict the 3-dimensional structure from the sequence of an unknown protein structure. That can let us easier understand the protein's function and mechanism.

This research is to find a suitable definition to describe protein backbone conformation, and to predict backbone conformation from sequence information. We adapt two different definitions of Ramachandran Plot and Protein Blocks. Our results by Support Vector Machines (SVM) get a relatively good performance.

Finally, we will study on the relationship between the conservation of protein backbone conformation and structural entropy. We've discovered that there is a consistency in structural entropy by different definitions of protein backbone conformation.

誌　　謝

因為在論文完成之前已經說過要寫兩頁的的誌謝，所以由此開始。

當學生是最快樂了的，尤其是在我們指導老師黃鎮剛教授的實驗室下當學生。不管是在實驗設備上或是研究環境都是一流的。因此非常感謝我們指導老師黃教授給我們有這麼樣的環境以及獨立研究的自由，並適時的給於鼓勵與啟發。兩年內從老師那學到的不是去完成指派的任務而是自己去找出可以完成什麼任務。這些我相信是在做科學研究所應有的熱情與態度。也特別感謝楊進木教授、盧錦隆教授以及黃憲達教授的教導，讓我有更多專業上的技能得以應用在研究的主題上。

接下來我要感謝所有再實驗室一同研究的所有的夥伴。林勇欣學長在幫我增長了不少生物養殖的知識，我不會忘記青蛙是會淹死的。我第一個接觸的生物資訊程式就是詹鎮熊學長的，在這裡得到不少知識與想法。在研究設施的維護方面特別感謝尤禎祥學長，游景盛學長以及黃存操同學，有你們費心的維護我才得以順利完成此論文。也特別感謝羅淵仁學長與游景盛學長時常爬山讓我得到不少登山常識，雖然我一次都沒去。還有養魚的知識與情報的交換要感謝羅淵仁學長、陸志豪學長、林志鵬學長以及隔壁楊老師實驗室的同學們，讓我在研究之餘還有其他的休閒活動。梁涵堃學長提供了不少汽車資訊。劉力彰學長感謝你要介紹女生給我認識，雖然在我寫這篇誌謝的時候我還沒認識到，不過白紙黑字你跑不掉了。在我剛

到實驗室的時候陳志杰學長與王志淵學長給我不少的建議與協助，尤其在我還沒有地方住的時候幫我安排地方居住在此特別感謝。在研究的過程中黃少偉、黃存操、陳啟德、林肇基、徐蔚倫及陳孟琪同學們互相砥礪與研究，讓我感到非常的快樂。並感謝黃少偉同學提供研究的數據資料，在此特別感謝因為有你的幫助我才得以完成此論文。本論文的英文校正要感謝尤禎祥學長、黃少偉、簡思樸與盧慧同學，有你們的幫助讓我的論文不會出現一個句子兩個動詞以及一堆(and)的贅字，還有尤禎祥學長指導我投影片的製作，讓我能夠在論文口試的過程順利報告。葉書瑋同學提供統計上的建議以及張世瑜同學常常幫忙買飲料。以上皆使得研究生活變的更豐富更快樂，在此感謝你們。

還有呢?要感謝所有的不是女朋友的人。這樣我就不用在寫上感謝你的包容或是惹你我生氣的事。這是團員的福利。

當然還要感謝鼓勵我敢研究所的老師與同學，沒有你們的支持與協助我也不可能有機會寫這一頁誌謝。

最後要感謝我的父母，從高中以前我一直都是一個問題人物。總是要上不上要下不下的狀況。感謝你們沒有放棄我一直提供資源給我我才有機會完成今天這個學業。謹以此論文作為碩士生涯的總結，並獻給所有幫助過我的人。

Contents

FIGUE CAPTIONS

INTRODUCTION

The conformation and structure of the polypeptide backbone in a folded protein are influenced by local amino acid sequence.[1,2] Since the first descriptions of protein structures by Pauling and Corey,[3] the researches in this field have been studied into the two principal point of view, assignment and prediction. The secondary structure of a protein is a summary of the general conformation and hydrogen bonding pattern of the amino acid backbone.[4] And the general definition of secondary structure is DSSP.[5] Using this definition to predict secondary structure can be achieved with the success rate of 80%.[6] But the description of the secondary structure of a protein is accurate enough to characterize the complete structure of proteins.[7] For the reason given above, many researches focus on characterizing protein peptide conformations.[8-10] Therefore, we now focus on the fundamental descriptors of a residue's conformation, i.e., the backbone torsional angles and protein blocks[11]. In the protein, the backbone conformation of each residue can be specified by the ϕ , ψ angles (Figure 1) used in the Ramachandran Plot.[12,13] Previous researchers have provide various classification schemes of the distribution of the backbone torsional angles. [3,14-19] We chose the definition of Ross as our classifying method for the Ramachandran Plot.[16] Because it can assign secondary structure that using a simple algorithm. In the Yang and Baker's research[18,19], they got the similar results of prediction backbone torsional angles using Ramachandran Plot. But they only use three kinds of classes to define the conformation for one residue in a protein. In their results, it raise the question of that why a polypeptide is predicted into the wrong conformation in the hypothesis of conformation depend on the amino acid sequence. It is reasonable for the coil structure because the numbers of the coil conformation's types are more than the helix and β-strand conformation's types. But the uniform conformation of helix and β-strand are also predicted in wrong conformation. Other research also mentioned this problem and tried to explain the reason. Meiler and Baker used two kind folding pathways to explain that a certain

segment first might adopt a secondary structure preferred by the local sequence and later be transformed to another secondary structure because of non-local interactions with a segment distant along the sequence in the folding process.[20] Minor and Kim discovered that the same local 11-residue segment can adopt helix or β-strand structure[21] , and Sudarsanam found this kind of residue segment from the Protein Data Bank.[22] In these problems, we need to use the local structure entropy to find what kind of amino acid sequence will not follow the rule of conformation depend on the amino acid sequence. The local structure entropy proved to be successful in discovering the relation between structural conservation and local structural stability that used directly statistics on the database or machine learning method.[11,23] But those method are only use secondary structural to descript local structural conformation, and we have discussed earlier that the secondary structural is not enough to descript local structural conformation. And using the backbone torsional angles and protein blocks will be suitable and enough to descript the local backbone conformation in the protein. In this reason, we think our research can find the relation between sequence and the fields about folding process, disorder sequence and protein design.

METHODS

Datasets

We use three kinds of datasets in our experimentation. The characteristic of each dataset are above.

RS126 : A nonhomologous data set that includes 126 globular and 4 membrane protein chains with less than 25% pair wise similarity for length > 80 residues. There are 23,706 cases have complete ϕ and ψ values in this data set.[24]

Sequence Unique Database : The data set is generated by the algorithm that is similar to the PDB select algorithm, but rather than focusing on maximizing the size of the subset. The algorithm focuses on getting representative structures of highest available quality. The dataset's pair wise sequence identity is less than 30%, and the resolution and R-factor criteria are 2.0 and 0.21. The total cases are 509,316, and the total chains are 2,295 in this data set. (version Nov/2004) (http://swift.cmbi.kun.nl/whatif/select/).

Protein Blocks Databanks : It is based on the PAPIA server, which allow to select protein with a maximum sequence identity and a RMSD value. The parameters are here : a. Sequence identity < 30 %, b. RMSD > 10 Å, c. Only X-ray structures, d. R-factor < 0.2 and e. resolution < 2.0Å . Total has 675 protein chains corresponding to 146,253 amino acids. In order to use the five-fold cross-validation. In this dataset includes some redundant chains (ie. duplicate chains), and we filter out the redundant chains. Finally this dataset has 651 protein chains for our works.

(http://www.ebgm.jussieu.fr/~debrevern/DOWN/DB/papia_databank_22082001.txt)

Definition of Backbone Torsional Angle Class

We use 36 classes to define different ϕ , ψ torsional angle pairs. And each class will be represented by a letter code (Figure 2), and we named this definition *Grid1*. Why we use this definition, because it can use three continuous letter codes to convert secondary structure for the central letter. The ϕ , ψ -values for each residues are mapped into the closest letter code. Assignment of residues to the secondary structure category is then straightforward. Progressing along sequence, conformation codes for each triple consecutive residues, $< i, i+1, i+2 >$, are used to classify the central residue, $i+1$, into (a) helix if all three residues are in {O,P}, (b) β-strand if all three residues are in {A,F,G,L,M,R}, (c) β-turn if $< i, i+1 >$ or $< i+1, i+2 >$

match a combination in this set { OO, OP, OJ, PO, PP, PJ, JO, JP, JJ, Mo, Mp, Ro, Rp, Rj, oo, op, oj, pp, pj, jo, jp, jj, mO, mP, mJ, rO, rP, rJ}, (d) Pii-helix if the residues have not already been classified as above classes and in {M,R}, or (e) coil in all other cases.[16] A class is assigned to each position along the sequence except for the one residue on each end of a sequence.

And we merge some classes into a new class (TABLE 1). This can make the different classes definition with *Grid1*. And *Grid2*, *Grid3* and *Grid4* (Figures 3-5) are used in our method.

Definition of Protein Blocks Class [25-27]

We use 16 kinds of *Protein Blocks (PBs)* (Figure 6), and each class will be assigned to one letter code $(a \sim p)$. The *PBs* classes are overlapping fragments, $M = 5$ residue in length, position i at the central of the fragment, encoded as sequence windows of $2(M-1)$ consecutive torsional angles vector $(\psi_{i-2}, \phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \phi_{i+1}, \psi_{i+1}, \phi_{i+2})$. They were obtained by an unsupervised classifier similar to Kohonen maps [28] and hidden Markov models [29]. *PBs* m and d correspond to the prototypes for the central α-helix and the central β-strand, respectively. *PBs* a through c represent β-strand N-caps and e and f, C-caps. *PBs* g through j are specific to coils, k and l to α-helix N-caps, and n through p to α-helix C-caps. Protein structures are encoded as sequences of ϕ-ψ torsional angles, and cut into consecutive overlapping fragments, each M amino acids in length. The fragment is defined by a vector of $2(M-1)$ torsional angular values. The fragment vector is compared with each *PBs* vector (Table 2) with the Root Mean Square Deviations on Angular(RMSDA) values measure.

$$RMSDA(V_1,V_2)=\sqrt{\frac{\sum_{i=1}^{i=M-1}\left[\psi_i(V_1)-\psi_i(V_2)\right]^2+\left[\phi_{i+1}(V_1)-\phi_{i+1}(V_2)\right]^2}{2(M-1)}} \quad (1)$$

The lowest RMSDA value for $2(M-1)$ angles determines the PB assignment. A PB is assigned to each position along the sequence except for the two residues on each end of a sequence. (Figure 7)

PSI-BLAST Profile

The inputs of SVM are parsed from PSSM (Position-Specific Scoring Matrix).The PSSM is generated by Position-Specific Iterated BLAST (PSI-BLAST) program. This program is an iterative program to search a database for proteins and it can generate a profile contained probabilities of residues, taking into account the significance of each sequence and distant homologues.[30] The PSSM log-odd score profile was obtained after five iterations with E-value threshold 0.001 against the non-redundant protein database. This profile has M×20 elements, where M is the target sequence and each element represents the log-likelihood of the length of the particular residue substitution at the template (based on a weighted average of BLOSUM62 matrix scores for given alignment position).

The Support Vector Machines (SVM)

The SVM are a family of algorithms for classification problems. [31] It's successfully applied to secondary structure prediction.[32], subcellular localization[33] prediction and other biological pattern classification problems. We use the publicly available LIBSVM [34] package

in our experiments, and kernel type is radial basis function (RBF).(Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

The basic idea of SVM is the use of a linear separating hyperplane, which can maximize the distance between two classes. If the problem can't be linearly separated in the original input space, then SVM employs two techniques to deal with it. Firstly SVM uses a soft margin hyperplane, which adds a penalty function to allow training errors. Secondly SVM non-linearly transforms the original input space into a higher dimension feature space. Hence in this new feature space it is more possible to find a linear optimal separating hyperplane to classify the data. SVM training always seeks a global optimized solution and avoids over fitting, so it has ability to deal with a large number of features.

Given input vectors x_i, $i = 1,...,l$ and a vector y defined as: $y_i = 1$ if x_i is in one class, and $y_i = -1$ if x_i is in the other class, SVM tried to find a separating hyperplane to separate these two classes of data. Here x_i can be a local sequence window centered on the interested residue. The variable y_i is used to denote the local conformation class of the residue, and the local conformation classes are defined at previous section. The support vector technique tries to find the separating hyperplanes $w^T x_i + b = 0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane. This requirement is equivalent to minimization of $\frac{1}{2} w^T w$ with respect to w and b under the constraint that $y_i(w^T x_i + b = 0) \geq 1$. However, in practice, these data that is to be classified may not be linearly separable. To overcome this difficulty, SVN non-linearly transforms the original input space into higher dimensional feature space by $\Phi(x) = (\phi_1(x), \phi_2(x), ...)$ and tries to minimize with respect to w, b and ξ, $\frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$, where $C > 0$ is the penalty parameter, under the constraint that $y_i[w^T \Phi(x_i) + b] \geq 1 - \xi_i$ where $\xi_i \geq 0$. This procedure has the advantage of allowing training errors. It should be noted that only some of

the x_i's are used to construct w and b, and these data are called support vectors. $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function, and we use the RBF kernel ($K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \quad \gamma > 0$) to calculate all inner product. All SVM calculations are performed using LIBSVM[34], a general library for support vector classification and regression.

In the SVM training procedure, we need use cross-validation to find the best parameter C and γ for RBF kernel. and the LIBSVM can use grid-search to find the best parameters. Using the wrong parameters will generate worst prediction result.

SVM-Inputs

Our SVM inputs utilize W×21-bits for each residue which is central in the W-length fragment, and each residue in this fragment will represented by 21 bits(1-bit indicates the null amino acid and the other 20-bits indicate the score for each amino acid from the PSI-BLAST profile). And the score are normalized to the range [0,1] by the following scaling function.[32]

$$f(x) = \begin{cases} 0.0 & if \quad x \le -5 \\ 0.5 + 0.1x & if \quad -5 < x \le 5 \\ 1.0 & if \quad x > 5 \end{cases} \qquad (2)$$

Each fragment will be represented by W×21-bits vector. W is the length of the sliding window size on the sequence. The window size W is chosen to be an odd numbers so that the target residue is always centered in the sliding window (Figure 8). In this work, we used 7, 9, 11 and 15 for the sliding window size. The SVM output is one classes for different definition *Grid1*, *Grid2*, *Grid3*, *Grid4* or *PBs*.

Performance Measure

In our experiments, we use five-fold cross validation for the dataset, and the accuracy to judge our results.

$$\text{Accuracy} = \frac{the \quad number \quad of \quad correct \quad predictions}{total \quad nunber \quad cases} \tag{3}$$

Structure Entropy Calculation [11]

The LIBSVM package has Support Vector Regression(SVR) model. Use this model can output the probability estimates for each classes using the –b option. And summing of each output probability is one.

When a protein sequence Γ of length L is denoted by $\Gamma = \delta_1\delta_2\cdots\delta_L$, where δ_i is the amino acid at i-th position. And the SVR output for the δ_i can be represented by a vector P_i, that is,

$$P_i = (p_{i1}, p_{i2}, ..., p_{iN}), \tag{4}$$

where the p_{ij} is the probability of the j-th class at δ_i. Using Shannon's information theory, [35] we compute the conformational entropy of δ_i by

$$Entropy_i = -\sum_{j}^{N} p_{ij} \log_b p_{ij}, \tag{5}$$

where the b is the number of the classes. (Figure 9)

RESULTS AND DISCUSSION

Predict *Grid1* using *RS126* dataset

We test different window sizes on *RS126* dataset. The results are consisting of each class for the *RS126* dataset when the window sizes are 7, 9, and 11, respectively. (Table 3) The accuracy of certain classes is 0%, and the test cases of those classes are under 100, implying that the classes they belong to are forbidden for this kind of torsional angle pairs of protein residues. The accuracy of the classes L and O are around 70% and 85%, respectively. And the total test cases of L and O classes are 4013 and 6853. It means classes L and O are more favorable for a residue in a protein sequence. In some classes P and R, they all have about two thousand test cases but their accuracy is only around 20% and 0%, respectively. We think this result can explain that the P class has more relationship with protein sequence than R class. In j, p and q classes, the number of test cases are around 300 but the accuracy are around 20%. It suggests some protein sequence fragments have a tendency to reside in these classes.

Comparison between *RS126* and *Sequence Unique Database* using *Grid1* definition

Comparing the results of *RS126* dataset (Table3) and *Sequence Unique Database* of window size 9 (Table 4) (Figure 10), *Sequence Unique Database* is better than *RS126* dataset in each class, especially in class R and j. For the case of R class, the difference has been explained in the previous section. Utilizing the *RS126* dataset, 0% of prediction accuracy is resulted from the small size of *RS126* dataset and hence insufficient Protein sequence information.

The total accuracy of *Sequence Unique Database* using *Grid1* definition can reach

54.81%. Numbers of the test case in the *Sequence Unique Database* is 20 times larger than that of *RS126* dataset, and therefore the *Sequence Unique Database* is too time-consuming in the SVM training process. Furthermore, the *RS126* dataset is too small in size to represent the scale of *PDB database*. Thus, it is required to develop a new database suitable to describe the whole space of PDB database while the size for training in SVM.

The results of *Grid2* definition

In *Grid2* definition, we merge *Grid1*'s 36 classes into eight classes. The total accuracy of *RS126* dataset (Table 5) and *Sequence Unique Database* (Table 6) of window size 9 are 54.7% and 64%, respectively. The prediction of *Grid2* is better than *Grid1* definition, and each class in *Grid2* is better than the merged class from *Grid1*. Only class E' of the *Grid2* definition is lower than it's mapping class O of the *Grid1* definition.

The results of *Grid3* and *Grid4* definition

The *Grid3* and *Grid4* definitions(Table 7-8) are similar to Yang's[18], including the results. The progress in accuracy is contributed by merging the classes into a larger one. Although progressive in accuracy, these definitions may lead to the problem of unable to get the reasonable class consisting true relationship between sequence and local conformation.

Comparison the test case in each datasets using *PBs* definition

The occurrence frequency of test cases in this definition is quite consistent within each dataset. (Figure 11) This suggests *PBs* definition is a comprehensive and relatively stable descriptor that the occurrence frequency would almost make no difference between each

dataset.

Comparison the result between de Brevern's[27] and our method using *PBs* definition

Our result is predicted by SVM and using the same dataset and window size with de Brevern's method. It suggests that using SVM and PSSM profile as input can get a better accuracy than de Brevern's method (Figure 12). But in the classes g, h, i and j, it is lower than de Brevern's method. The reason of this result may be because of the limit of SVM. The occurrence frequency of test cases in class g, h, i, and j are under 3%. If the SVM input of one class cannot be distinguished from others, the accuracy would drop to a low level. This may somehow conclude that the class g, h, i, and j have weak relationship between the local conformation and the sequence information.

Comparison different window size in *RS126* and *Protein Blocks Databanks* using *PBs* definition

The comparison result shows consistency between window size 9 and 15. This also reveals that enough sequence information is obtained to predict PBs when window size is set to 9. (Figure 13-14)

Comparison different dataset using *PBs* definition

We compared the accuracy of *RS126*, *Sequence Unique Database* and *Protein Blocks Databanks* using window size 9. The result is in Figure 15, and it shows the large dataset can get better accuracy than smaller one. The number of test case in *Sequence Unique Database* is 5 times bigger than *Protein Blocks Databanks*, while the number of test case in *Sequence*

Unique Database is 6 times bigger than *RS126*. This is certainly because the larger dataset covers almost the entire sequence space than the smaller dataset.

The relation between structure conservation and structural entropy

We compare 4 kinds of structural entropy calculated from different local structural definitions: (a) *PBs*, (b) *Grid1*, (c) *Grid2*, and (d) SSE. There are fifteen results shown in Figure 15-30. In our result, the structural entropy definition from *PBs*, *Grid1* and *Grid2* are very consistent. Because the definition of PBs and Grid2 are correlated with Grid1, this result seems quite reliable. In these figures (e) labeling the secondary structure by different colors: red for helix, blue for β strand and green for coil. In Huang's research, the Secondary Structural Entropy (*SSE*) has the relationship with structure conservation. And the lower entropy can map to the conservation region using *SSE*. We discovered that the inconsistent regions are almost coil regions. This suggests *SSE* is relatively insensitive due to its insufficient description power to coil conformations, which only containing three kinds of local conformation, helix, β strand, and coil from secondary structure. Using our definitions can make the structural entropy more sensitive in coil conformation, so it can improve *SSE*'s method.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181(96):223-230.
2. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 1999;24(1):26-33.
3. Pauling L, Corey RB. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc Natl Acad Sci U S A 1951;37(5):235-240.
4. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23(4):566-579.
5. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577-2637.
6. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47(2):228-235.
7. Camproux AC, Gautier R, Tuffery P. A hidden markov model derived structural alphabet for proteins. J Mol Biol 2004;339(3):591-605.
8. Estrada E. A protein folding degree measure and its dependence on crystal packing, protein size, secondary structure, and domain structural class. J Chem Inf Comput Sci 2004;44(4):1238-1250.
9. Sims GE, Choi IG, Kim SH. Protein conformational space in higher order phi-Psi maps. Proc Natl Acad Sci U S A 2005;102(3):618-621.
10. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins 2002;48(3):463-486.
11. Huang SW, Hwang JK. Computation of conformational entropy from protein sequences using the machine-learning method--application to the study of the relationship between structural conservation and local structural stability. Proteins 2005;59(4):802-809.
12. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95-99.
13. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283-438.
14. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10(3):599-612.
15. Kang HS, Kurochkina NA, Lee B. Estimation and use of protein backbone angle probabilities. J Mol Biol 1993;229(2):448-460.
16. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content

predicts folding rates for simple, two-state proteins. J Mol Biol 2003;327(5):1149-1154.
17. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40(1):135-144.
18. Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20(10):1612-1621.
19. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301(1):173-190.
20. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci U S A 2003;100(21):12105-12110.
21. Minor DL, Jr., Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature 1996;380(6576):730-734.
22. Sudarsanam S. Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. Proteins 1998;30(3):228-231.
23. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostabilty. Proteins 2004;57(4):684-691.
24. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232(2):584-599.
25. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41(3):271-287.
26. de Brevern AG, Valadie H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 2002;11(12):2871-2886.
27. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: improved prediction methods. Proteins 2005;59(4):810-827.
28. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59-69.
29. Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc IEEE 1989;77:257-285.
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-3402.
31. Vapnik V. The Nature of statistical learning theory. New York; 1995.
32. Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 2003;16(8):553-560.
33. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for

Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 2004;13(5):1402-1406.

34. Hsu CW, Chung CC, Lin CJ. A Practical Guide to Support Vector Classification. Taipei.

35. Shannon CE. A mathematical theory of communication. The Bell System Technical Journal 1948;27:379-423, 623-656.

TABLES

TABLE 1. *Grid2, Grid3, Grid4* Definition

Grid2		*Grid3*		*Grid4*	
Type	Types merge from *Grid1*	Type	Types merge from *Grid1*	Type	Types merge from *Grid1*
A’	A, F	A’’	A, F, G, L	A’’’	A, F, G, L, M, R
B’	G, L				
C’	M, R	B’’	M, R		
D’	J, P	C’’	J, P, O	B’’’	J, P, O
E’	O				
F’	m, g, r				
G’	j, o , p	D’’	j, o, p	C’’’	j, o, p
H’	B, C, D ,E ,H, I, K, N, Q , h, i, k, l, n, q	E’’	B, C, D ,E ,H, I, K, N, Q , g, h, i, k, l, m, n, q, r	D’’’	B, C, D ,E ,H, I, K, N, Q , g, h, i, k, l, m, n, q, r
non-class	S, T, U, V ,W, X	non-class	S, T, U, V ,W, X	non-class	S, T, U, V ,W, X

Table 2. The angles on the central residue *i* for the 5-residues *Protein Blocks.*

PB type	ψ_{i-2}	ϕ_{i-1}	ψ_{i-1}	ϕ_i	ψ_i	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}
a	41.14	75.53	13.92	-99.8	131.88	-96.27	122.08	-99.68
b	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.9
c	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
d	141.98	-112.79	132.2	-114.79	140.11	-111.05	139.54	-103.16
e	133.25	-112.37	137.64	-108.13	133	-87.3	120.54	77.4
f	116.4	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
g	0.4	-81.83	4.91	-100.59	85.5	-71.65	130.78	84.98
h	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-90.88
i	130.68	-56.92	119.26	77.85	10.42	-99.43	141.4	-98.01
j	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
k	117.16	-95.41	140.4	-59.35	-29.23	-72.39	-25.08	-76.16
l	139.2	-55.96	-32.7	-68.51	-26.09	-74.44	-22.6	-71.74
m	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
n	-35.34	-65.03	-38.12	-66.34	-29.51	-89.1	-2.91	77.9
o	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
p	-27.09	-86.14	0.3	59.85	21.51	-96.3	132.67	-92.91

TABLE 3. Prediction Results for *RS126* using *Grid1* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio(%)		
			Window Size		
			7	9	11
A	604	2.55	6.95	1.82	5.13
B	34	0.14	0	0	0
C	32	0.13	0	0	0
D	42	0.18	0	0	0
E	56	0.24	0	0	0
F	298	1.26	0	0	0
G	1791	7.56	13.18	5.63	10.22
H	88	0.37	0	0	0
I	367	1.55	1.63	0	1.36
J	1303	5.50	16.42	10.05	15.73
K	435	1.83	4.14	0.46	2.76
L	4013	16.93	67.73	74.13	69.05
M	991	4.18	12.11	3.63	9.38
N	50	0.21	0	0	0
O	6853	28.91	83.07	86.3	84.15
P	2546	10.74	22.31	19.56	22.94
Q	231	0.97	0.87	0.87	1.73
R	2236	9.43	30.81	28.18	28.53
S	11	0.05	0	0	0
T	13	0.05	0	0	0
U	77	0.32	0	0	0
V	15	0.06	0	0	0
W	28	0.12	0	0	0
X	33	0.14	0	0	0
g	119	0.50	8.4	2.52	6.72
h	57	0.24	0	0	0
i	31	0.13	0	0	0
j	249	1.05	24.9	12.85	21.69
k	35	0.15	5.71	0	0
l	26	0.11	0	0	0
m	110	0.46	8.18	0	4.55
n	45	0.19	0	0	0
o	336	1.42	10.71	5.36	11.01
p	391	1.65	25.32	22.25	27.37
q	52	0.22	3.85	0	1.92
r	108	0.46	1.85	0	1.85
Total	23706		44.41	43.57	44.32

TABLE 4. Prediction Results for *Sequence Unique database* using *Grid1* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio (%)
			Window Size
			9
A	13258	2.60	19.44
B	305	0.06	6.56
C	141	0.03	0.71
D	183	0.04	1.64
E	571	0.11	2.28
F	5218	1.02	3.53
G	34613	6.80	18.51
H	966	0.19	4.87
I	5716	1.12	6.74
J	32101	6.30	31.80
K	7209	1.42	12.00
L	85634	16.81	75.02
M	22077	4.33	17.48
N	157	0.03	0
O	157494	30.92	87.26
P	54652	10.73	33.70
Q	3311	0.65	7.52
R	53418	10.49	45.6
S	15	0.00	0
T	24	0.00	0
U	80	0.02	0
V	7	0.00	0
W	59	0.01	0
X	74	0.01	0
g	2200	0.43	17.72
h	517	0.10	12.38
i	137	0.03	0
j	4413	0.87	20.82
k	276	0.05	0
l	298	0.06	6.38
m	1977	0.39	12.49
n	185	0.04	5.95
o	8514	1.67	27.58
p	11041	2.17	49.75
q	397	0.08	6.30
r	2078	0.41	18.77
Total	509316		54.81

TABLE 5. Prediction Results for *RS126* using *Grid2* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio (%)
			Window Size
			9
A’	936	3.89	0.53
B’	5971	24.84	76.99
C’	3332	13.86	32.89
D’	3922	16.32	39.09
E’	6915	28.77	78.50
F’	347	1.44	3.17
G’	995	4.14	45.63
H’	1620	6.74	1.42
Total	24038		54.69

TABLE 6. Prediction Results for *Sequence Unique database* using *Grid2* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio (%)
			Window Size
			9
A’	18487	3.63	14.70
B’	120320	23.62	76.22
C’	75550	14.83	48.85
D’	86787	17.04	53.20
E’	157554	30.93	82.86
F’	6260	1.23	27.84
G’	23982	4.71	58.37
H’	20379	4.00	10.41
Total	509319		63.99

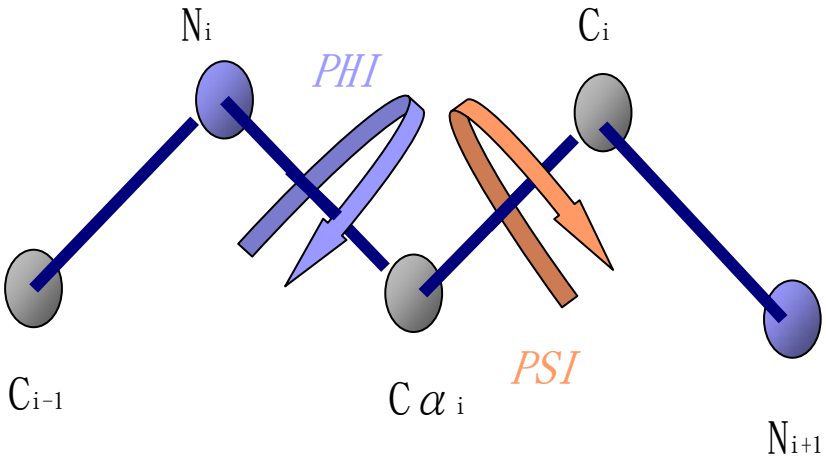
TABLE 7. Prediction Results for *RS126* using *Grid3* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio (%)
			Window Size
			9
A''	6907	28.73	72.88
B''	3332	13.86	24.58
C''	10837	45.08	87.59
D''	995	4.14	33.57
E''	1967	8.18	4.52
Total	24038		65.60

TABLE 8. Prediction Results for *RS126* using *Grid4* Definition

Type	Number of cases	Occurrence frequency ratio (%)	Correctly predicted ratio (%)
			Window Size
			9
A'''	10239	42.60	83.86
B'''	10837	45.08	80.86
C'''	995	4.14	31.46
D'''	1967	8.18	1.12
Total	24038		73.57

Figures



The range of PHI、PSI angles：-180~180

Figure 1. The definition of PHI and PSI (φ - ψ) angles in a amino acid residue i.

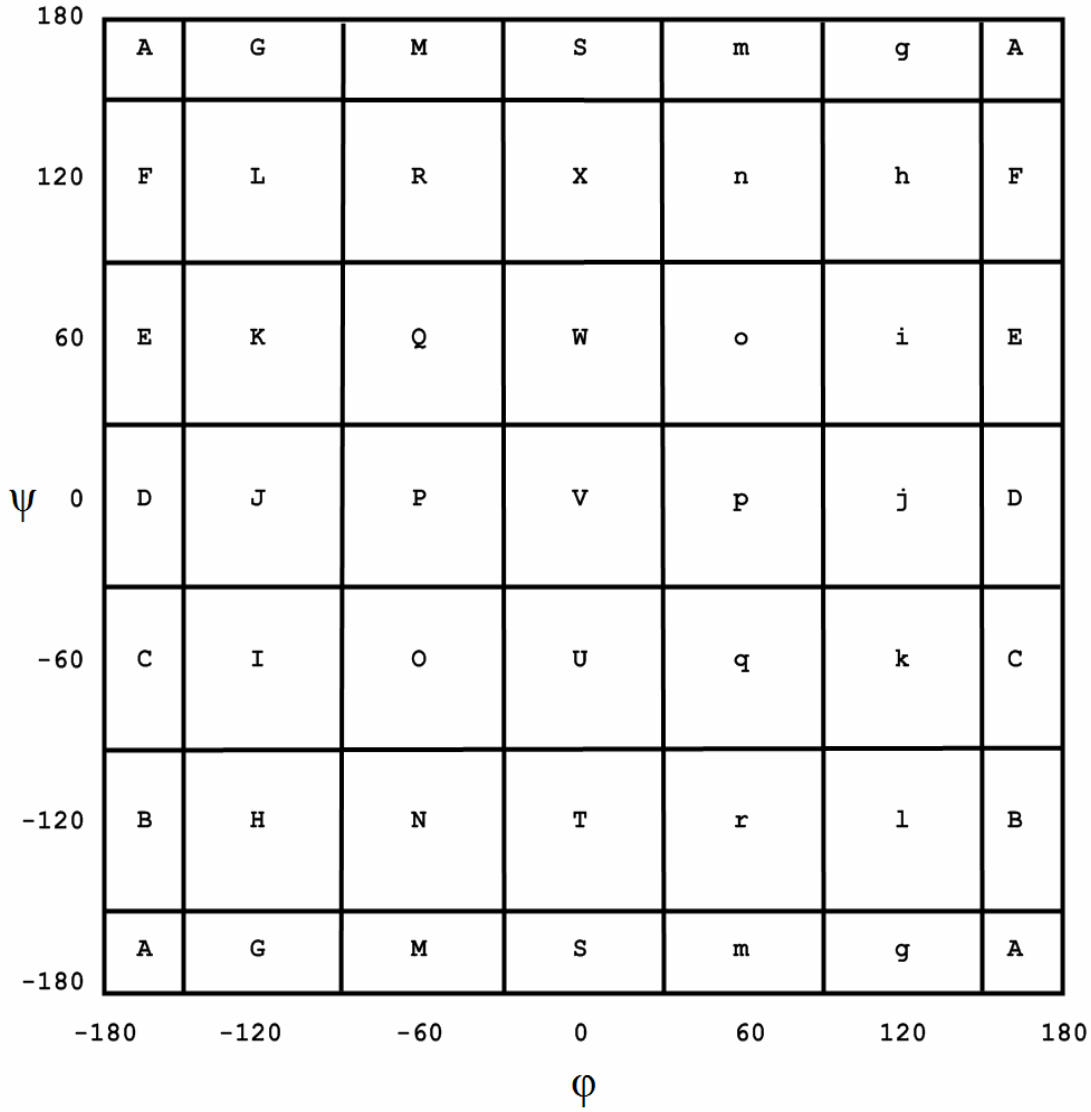


Figure 2. *Grid1* : The torsional angle was been classified according this φ - ψ map. This map was divided into 36 classes, and each class has be assigned a letter code. [16]

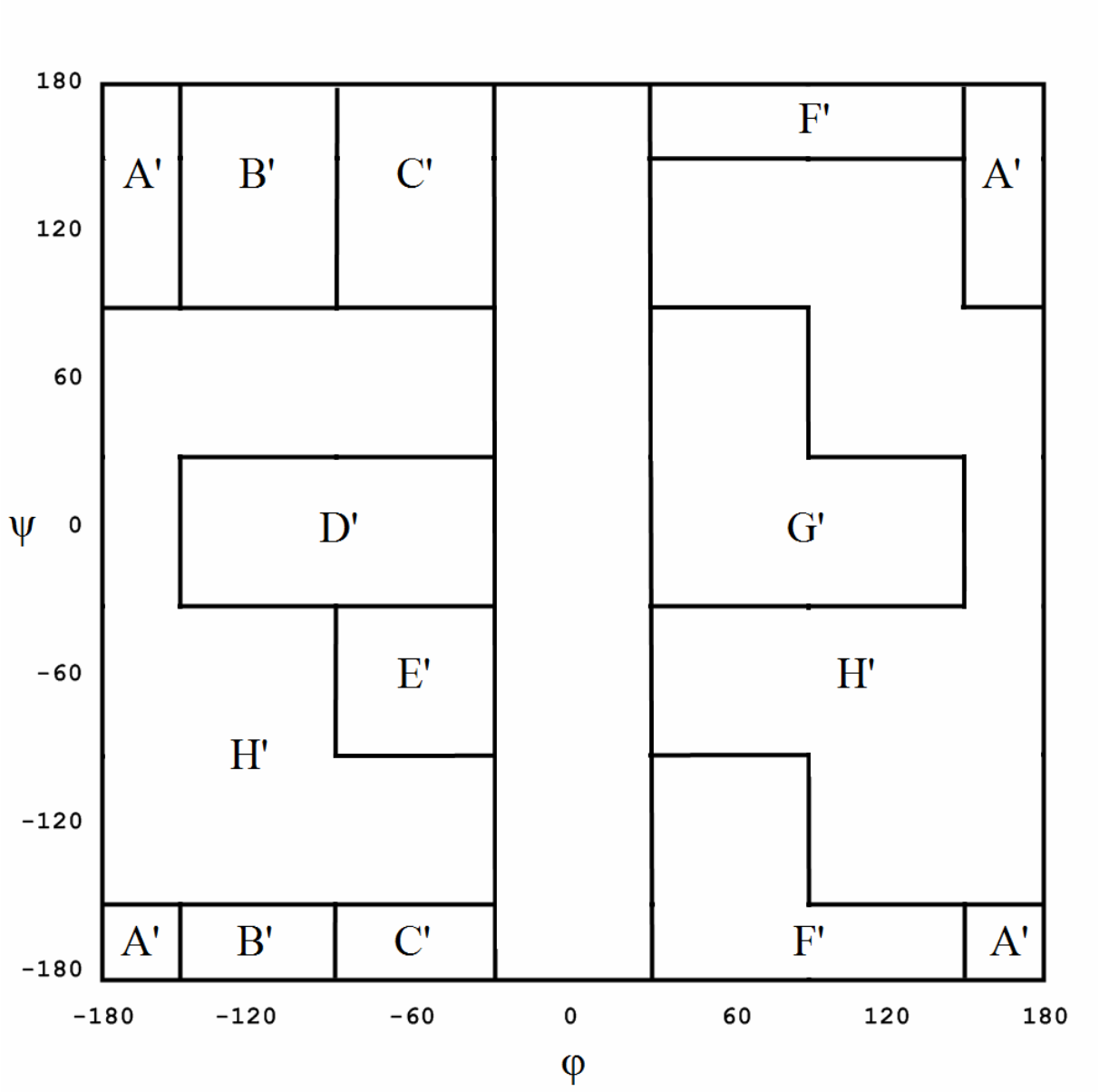


Figure 3. *Grid2* : The torsional angle was been classified according this φ - ψ map. This map was divided into 8 classes, and each class has be assigned a letter code.

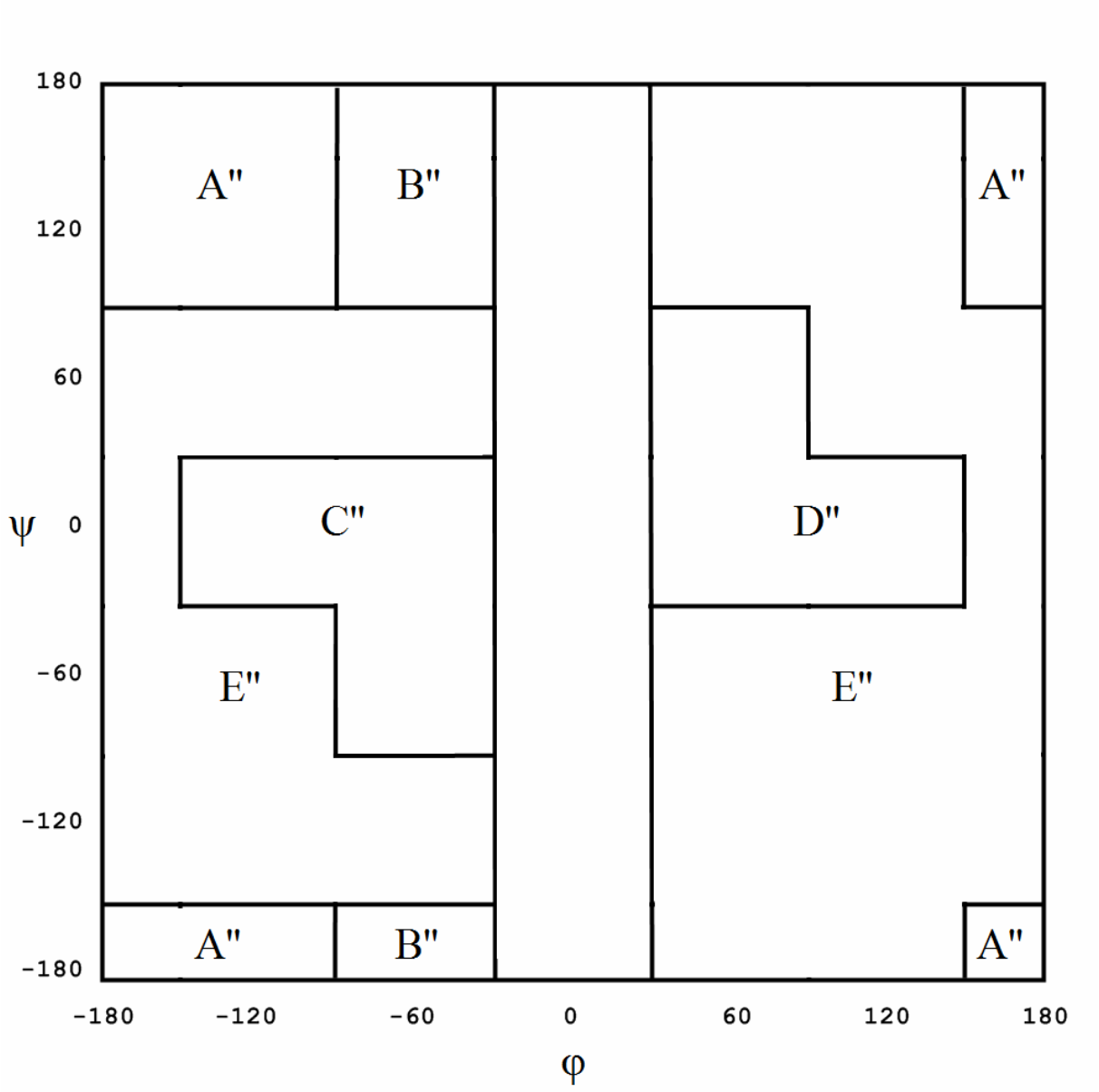


Figure 4. *Grid3* : The torsional angle was been classified according this φ - ψ map. This map was divided into 5 classes, and each class has be assigned a letter code.

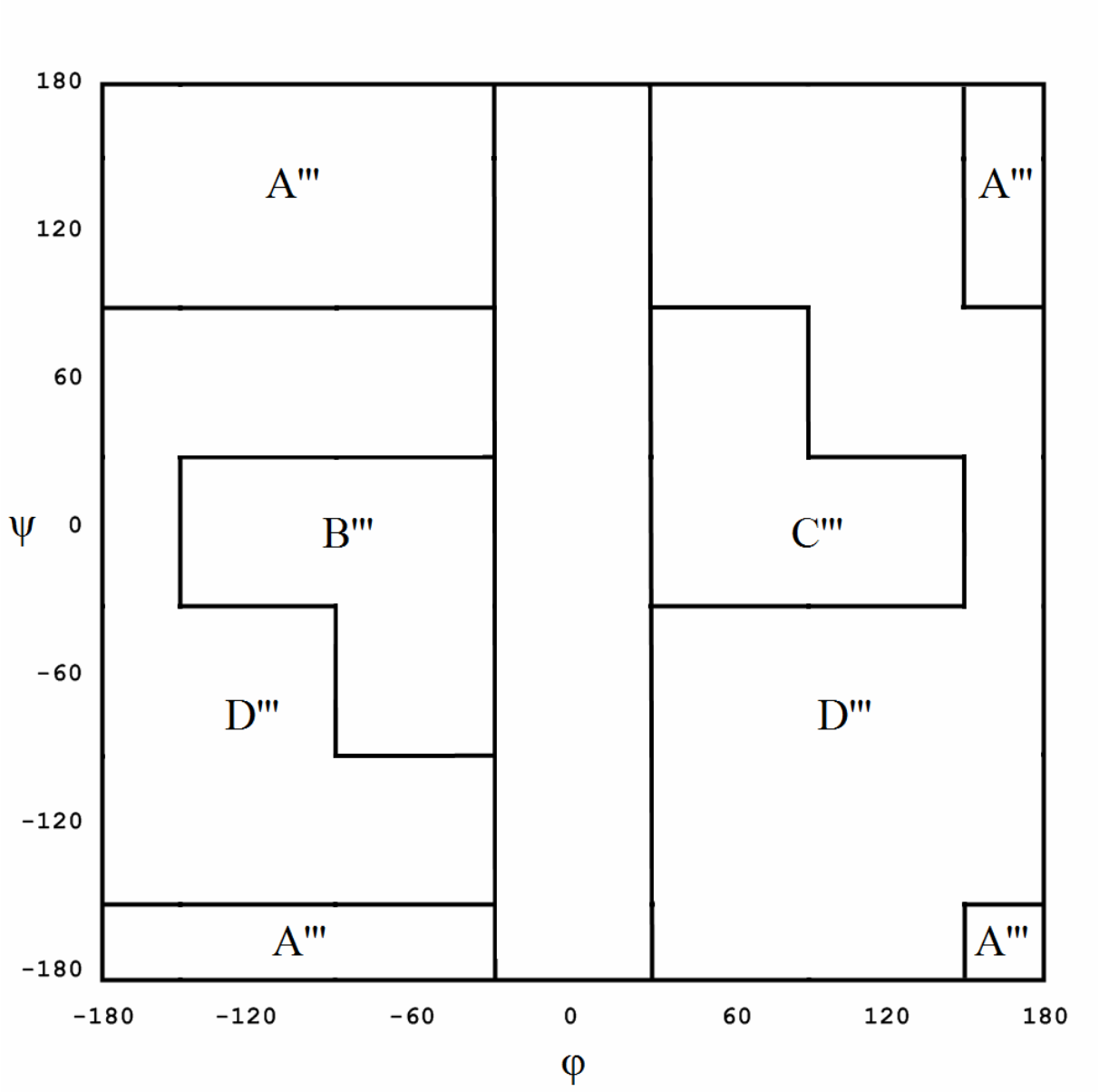


Figure 5. *Grid4* : The torsional angle was been classified according this φ - ψ map. This map was divided into 5 classes, and each class has be assigned a letter code.

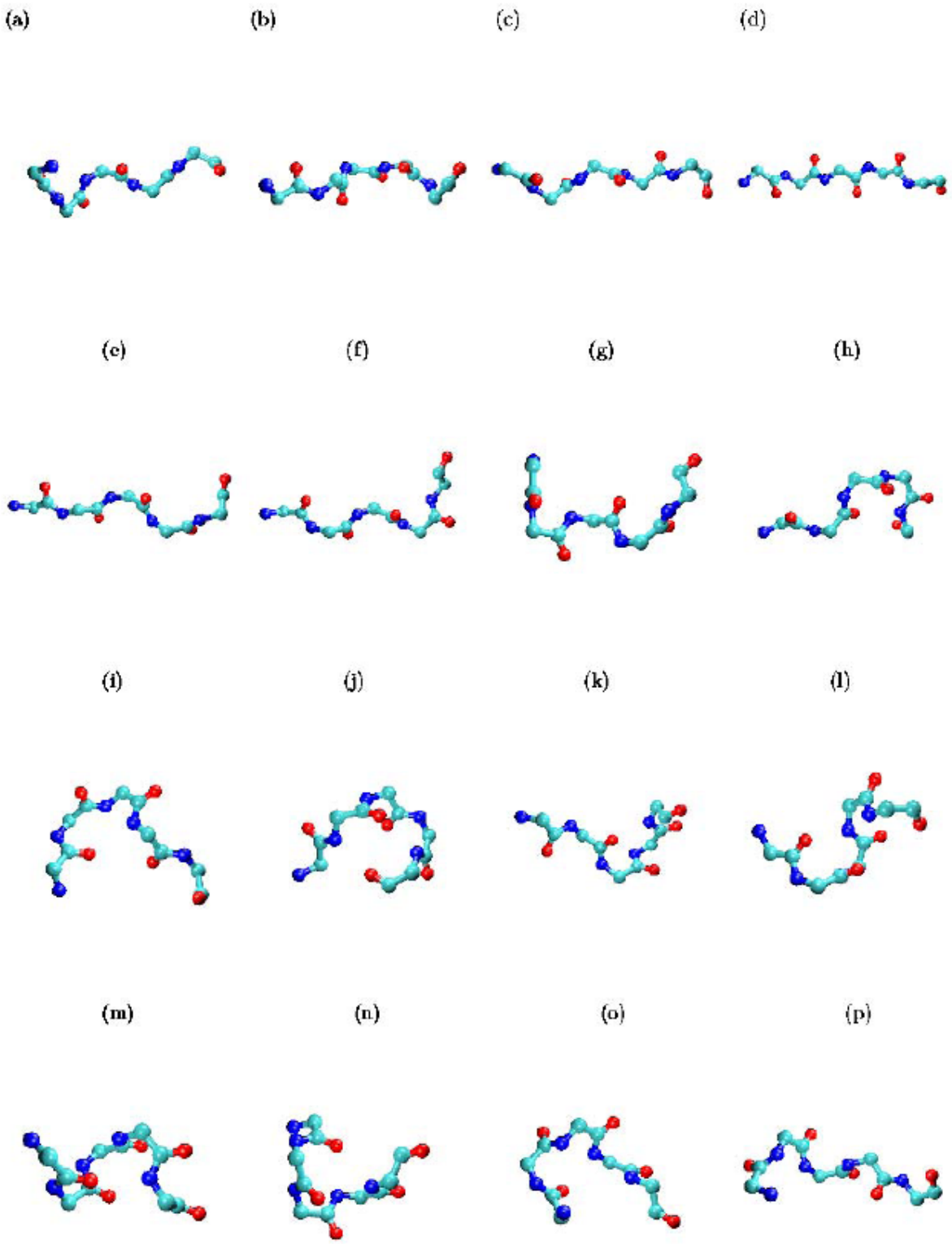


Figure 6. *Protein Blocks*. From left to right and top to bottom the 16 Protein Blocks (labeled from *a* to *p*). For each PB, the N-cap is on the left and the C-cap is on the right.

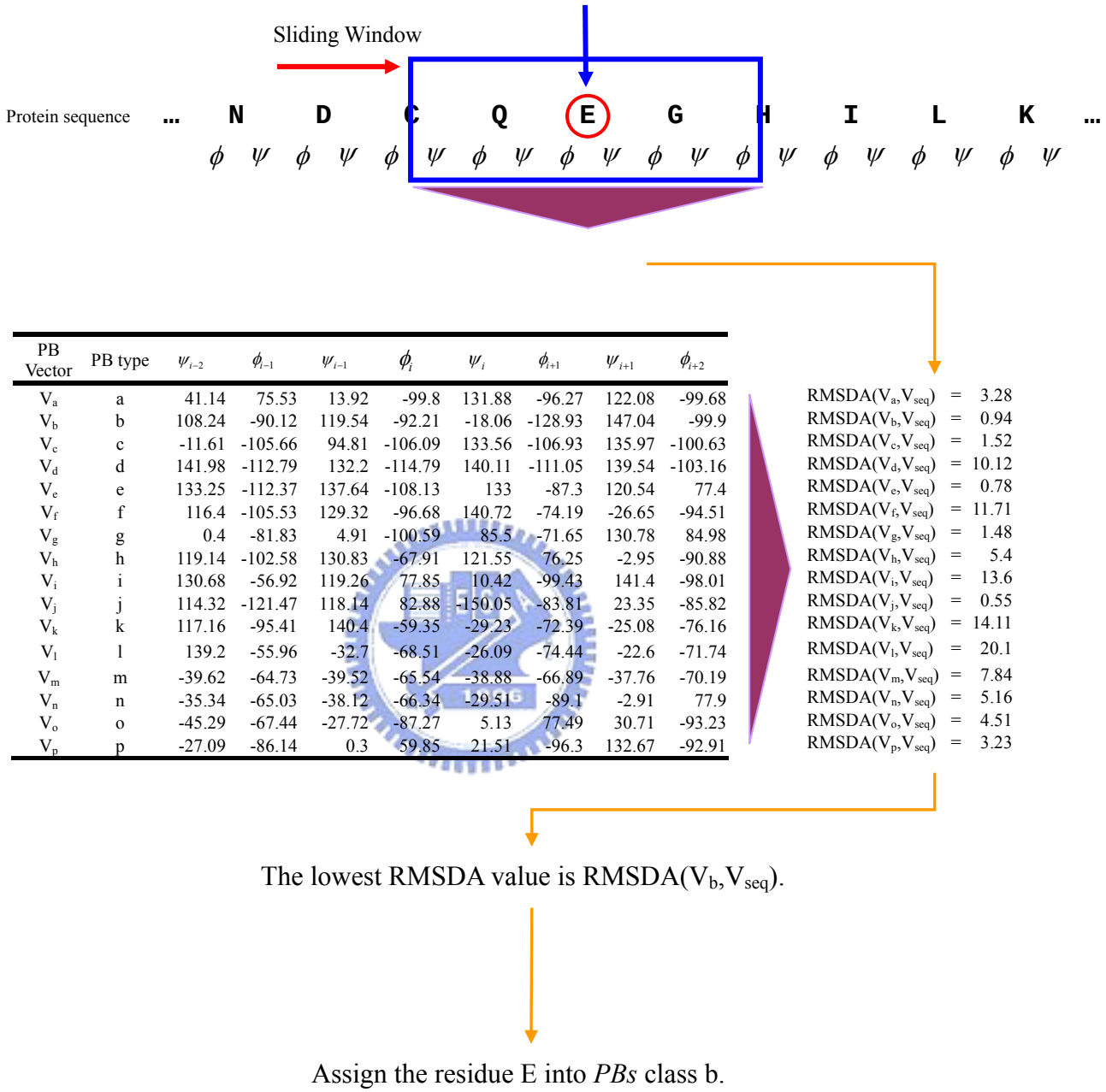


Figure 7. The flow chart of assignment PBs class.

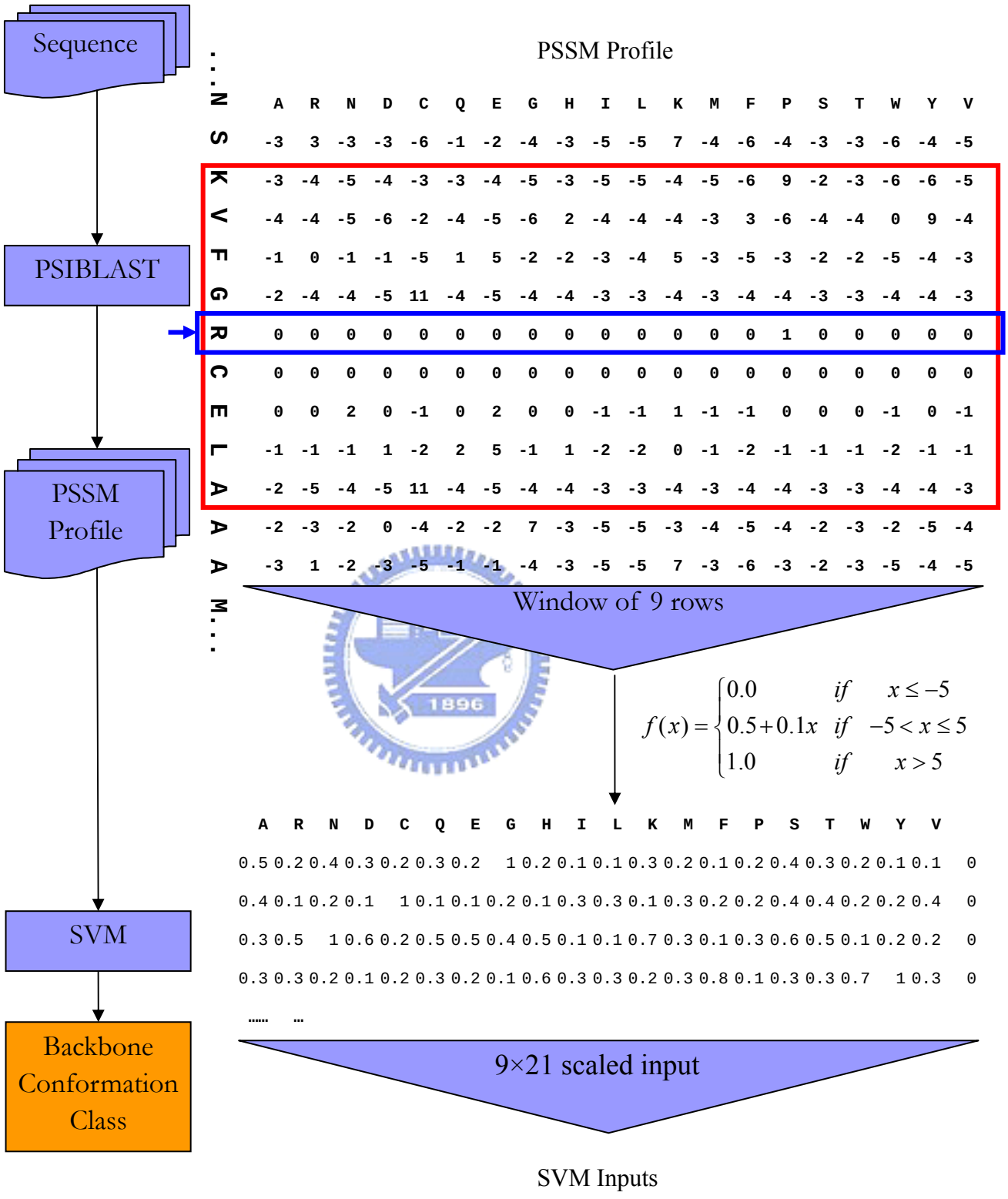


Figure 8. Encoding and prediction flow chart.

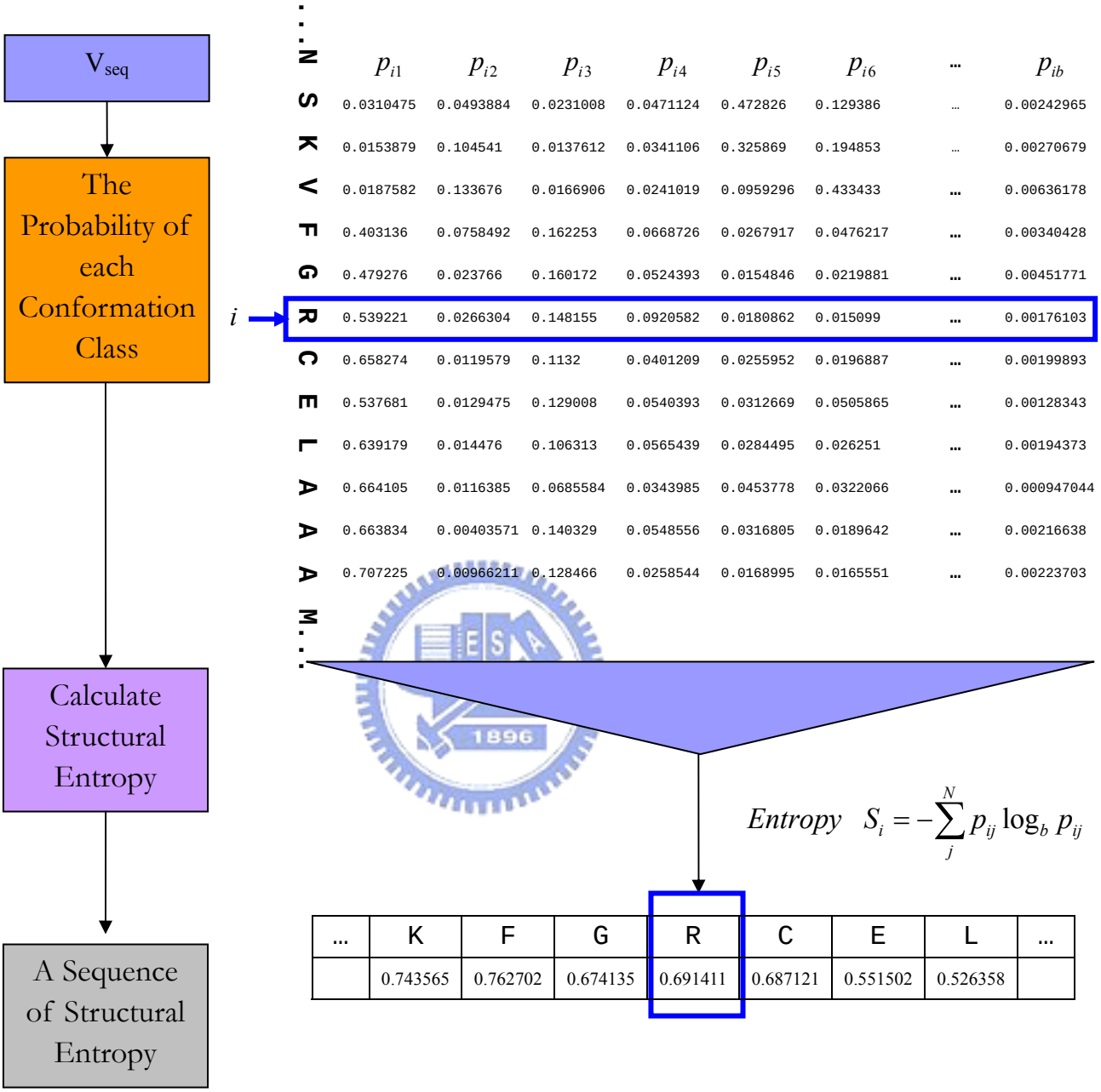


Figure 9. Calculate the structural entropy using SVR output.

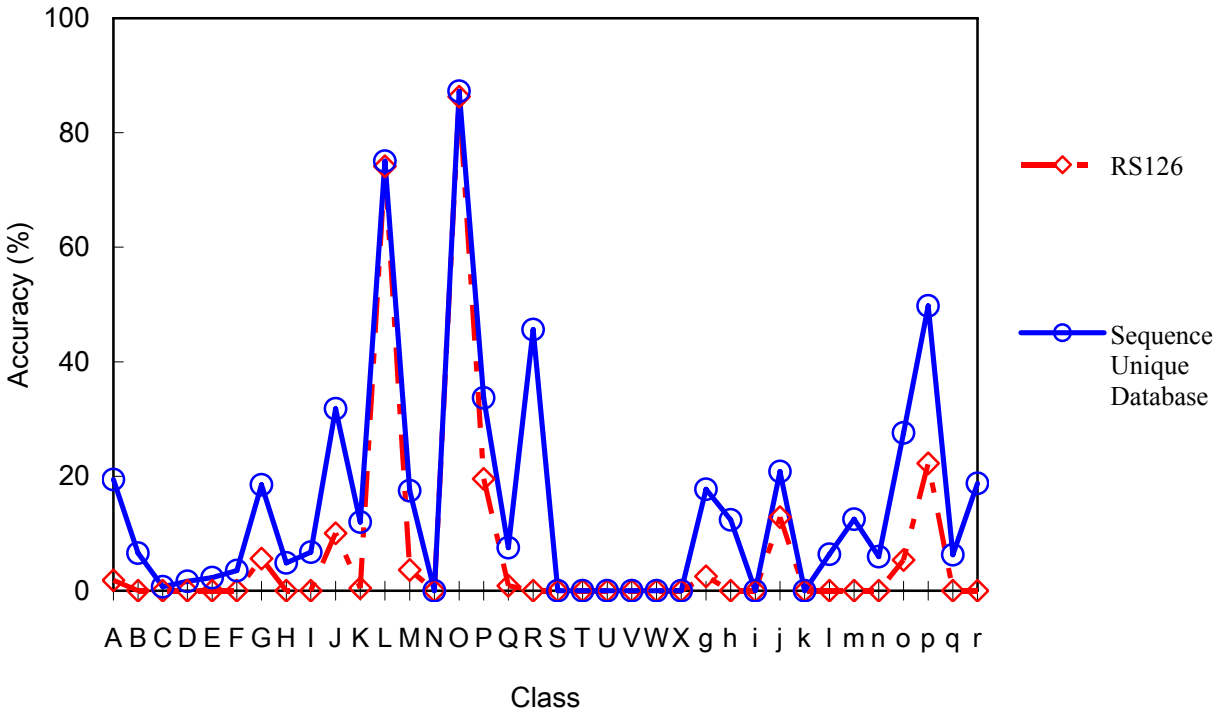


Figure 10. It shows that the accuracy of using *Sequence Unique Database* is better than using *RS126 database* with window size 9.

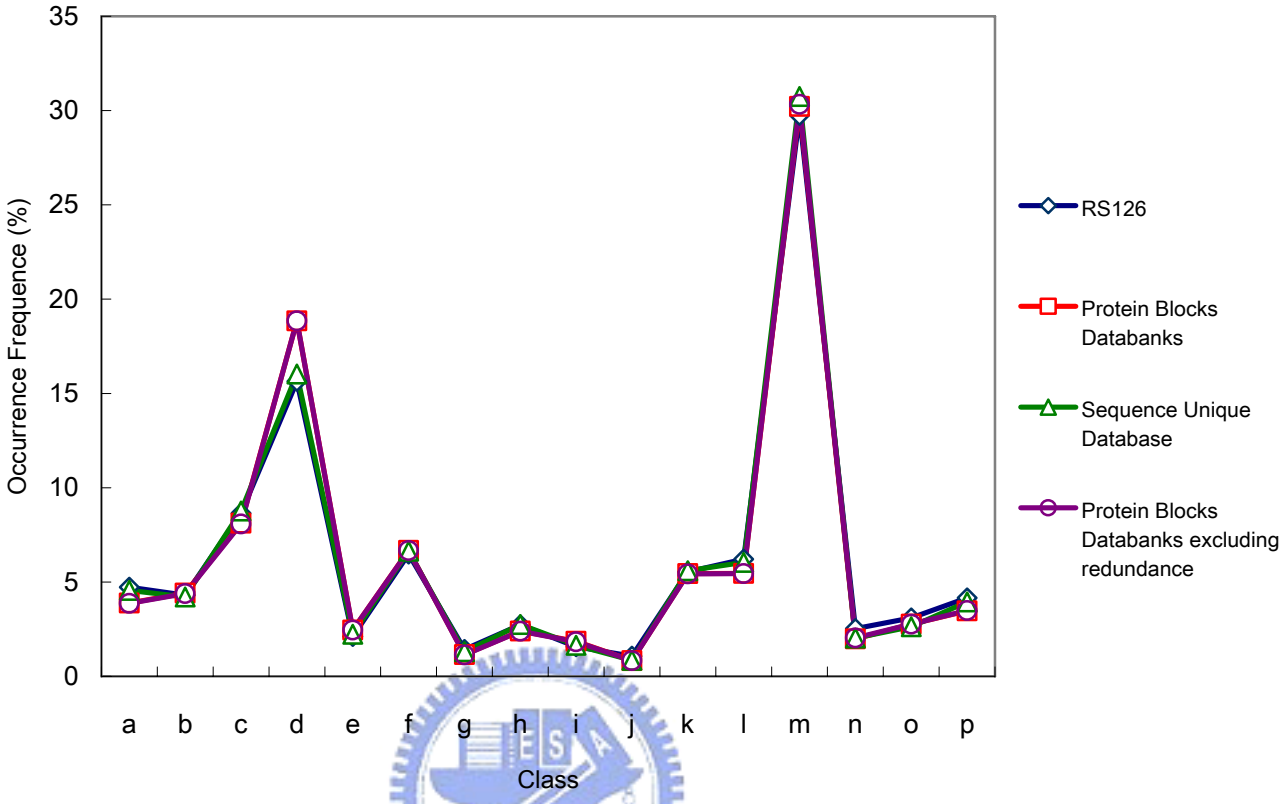


Figure 11. Compare the test case in each datasets using *PBs* definition.

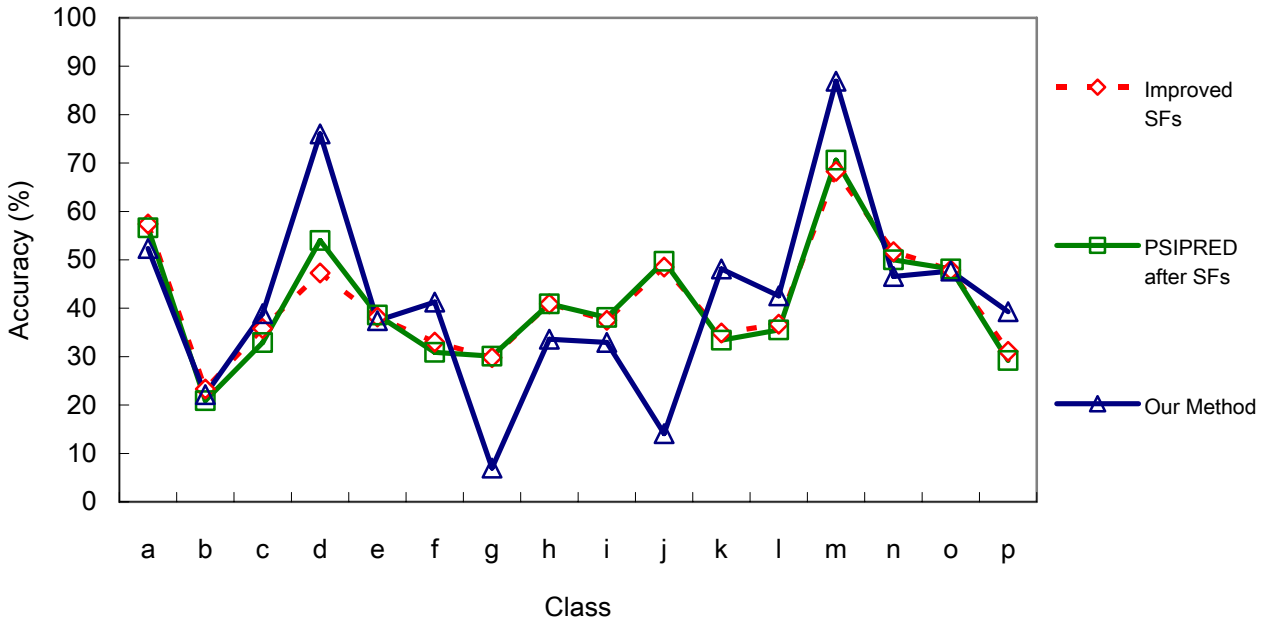


Figure 12. Compare the result between de Brevern's and ours using *PBs* definition, window size 9 and the same dataset.

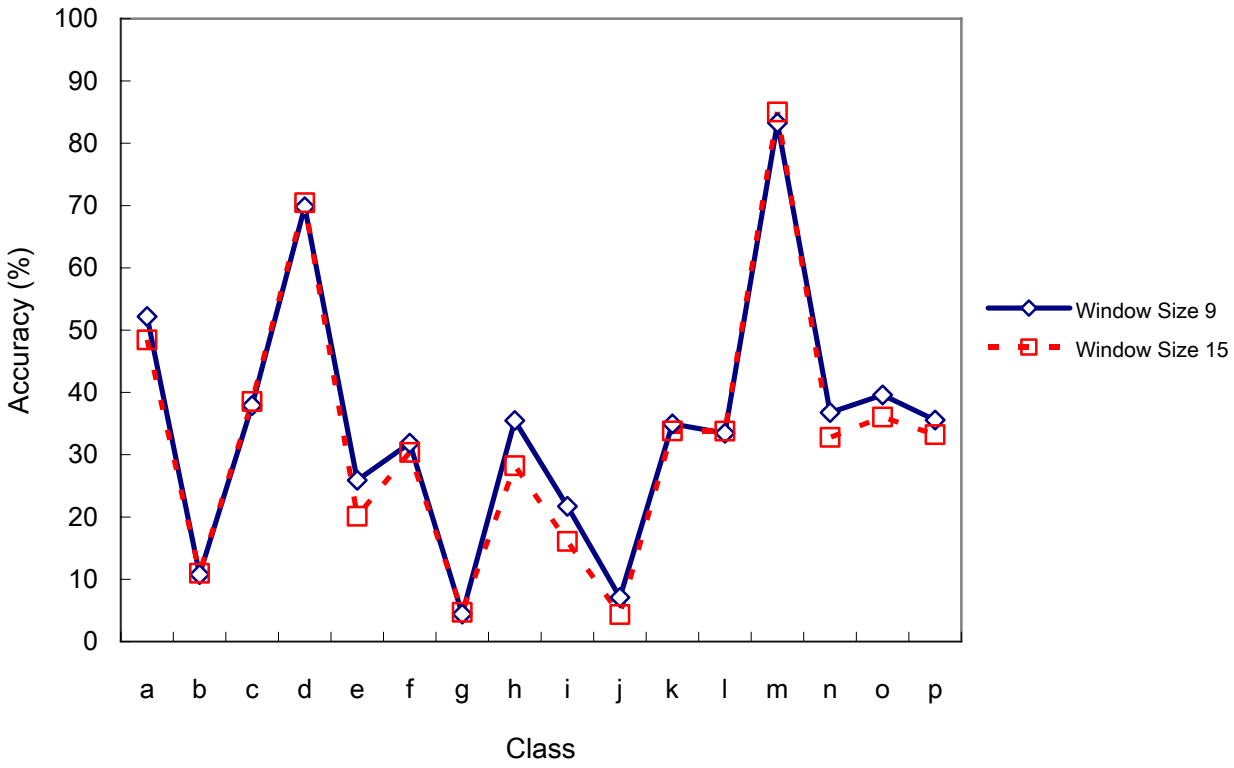


Figure 13. Compare different window size in *RS126* using *PBs* definition.

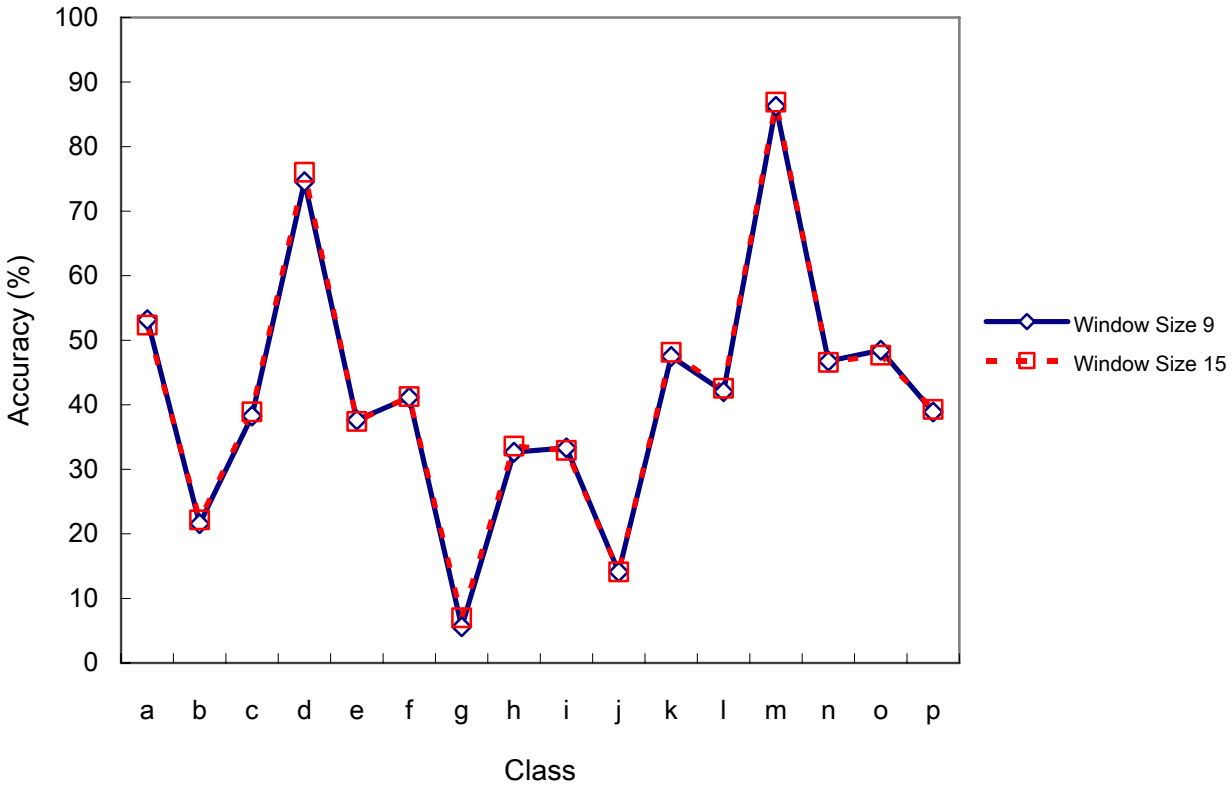


Figure 14. Compare different window size in *Protein Blocks Databanks* using *PBs* definition.

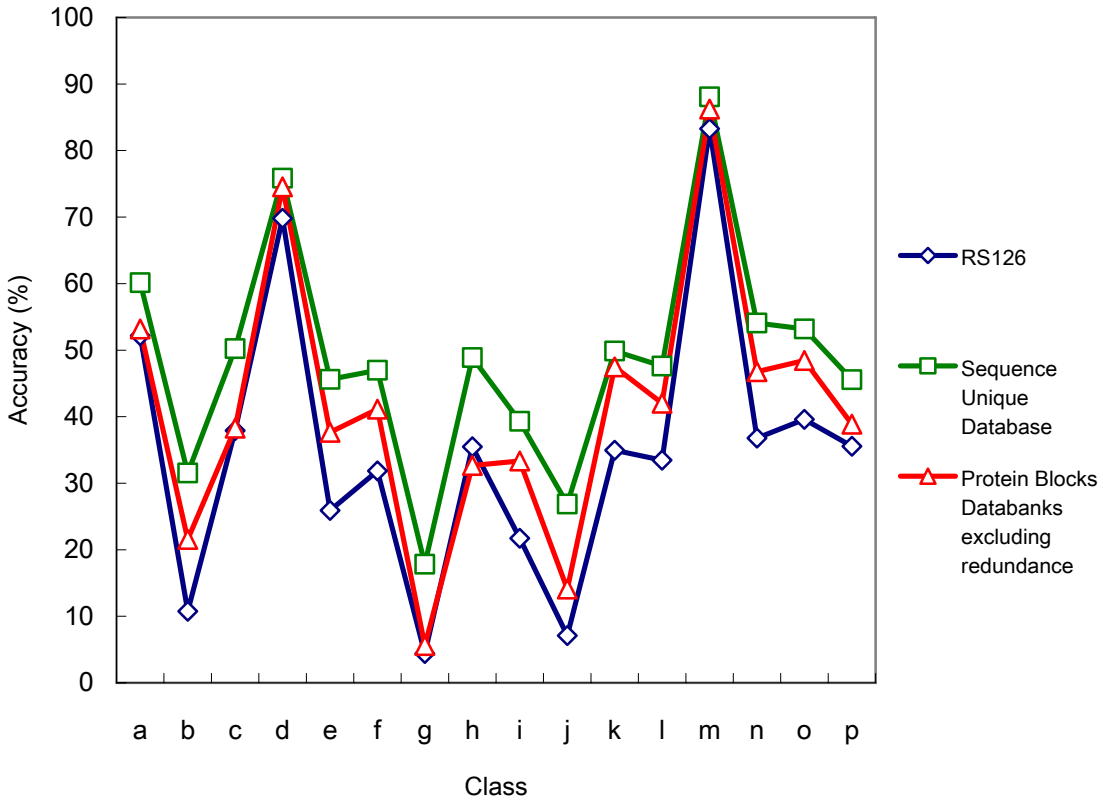


Figure 15. Compare different dataset using *PBs* definition.

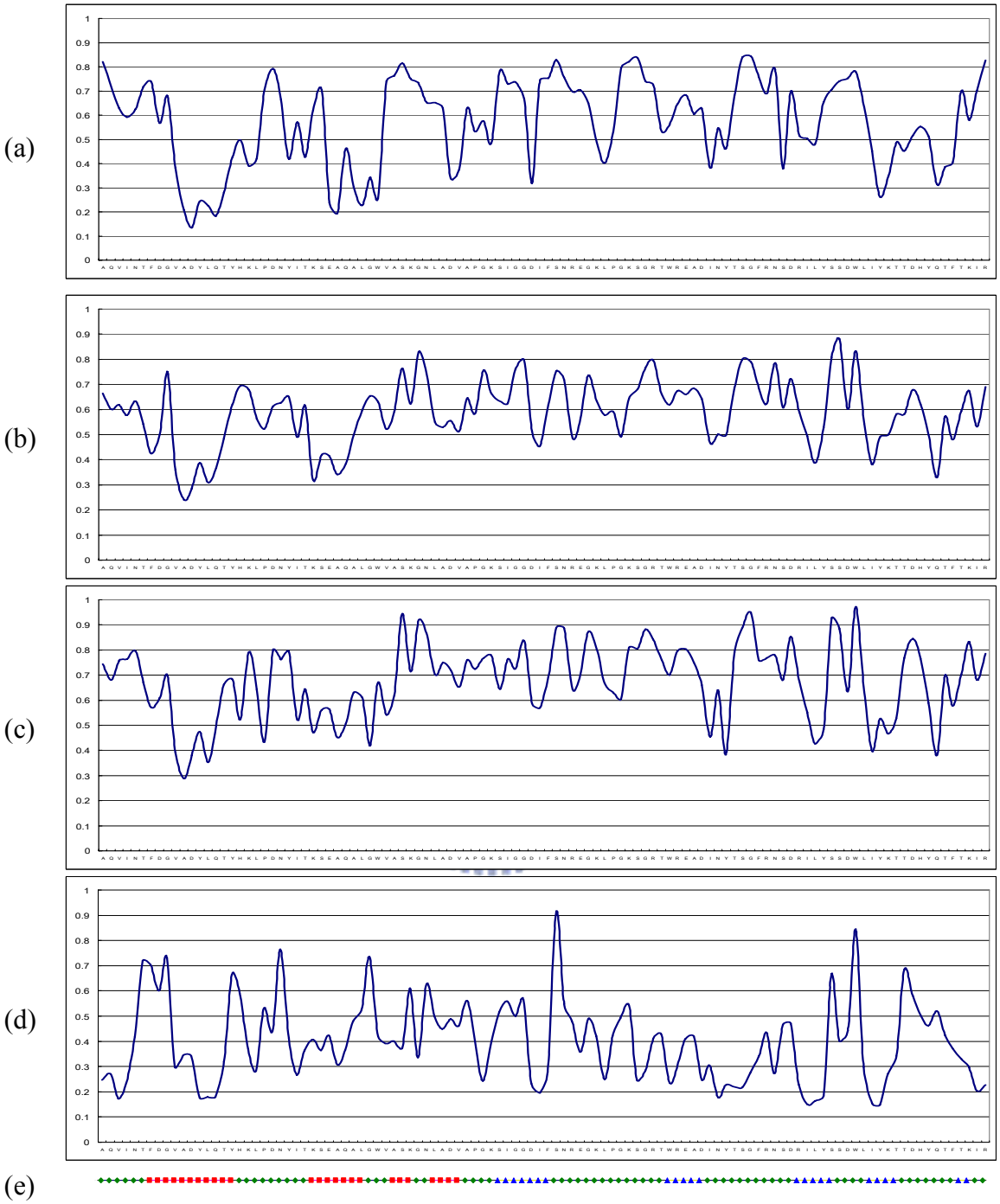


Figure 16.

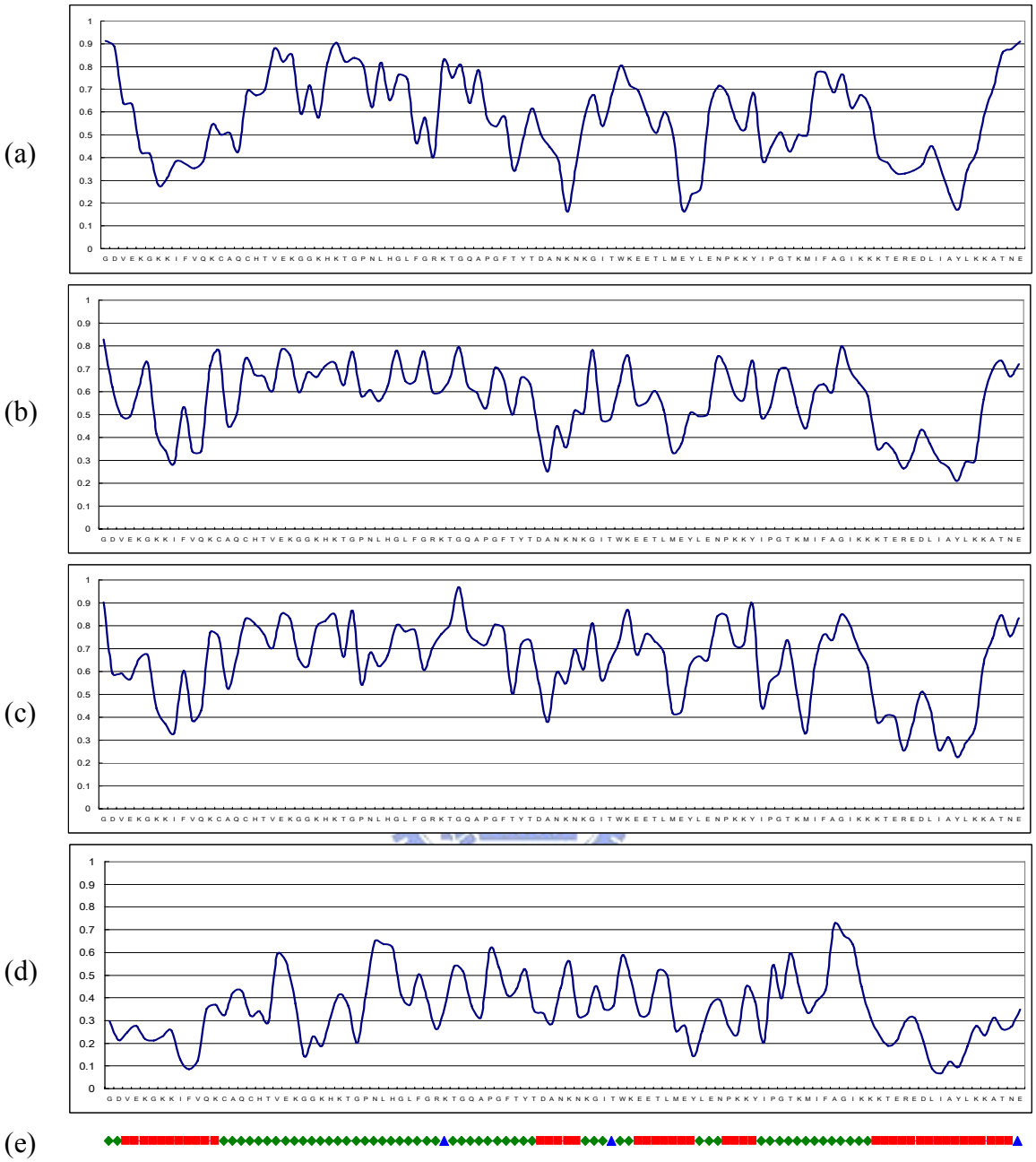
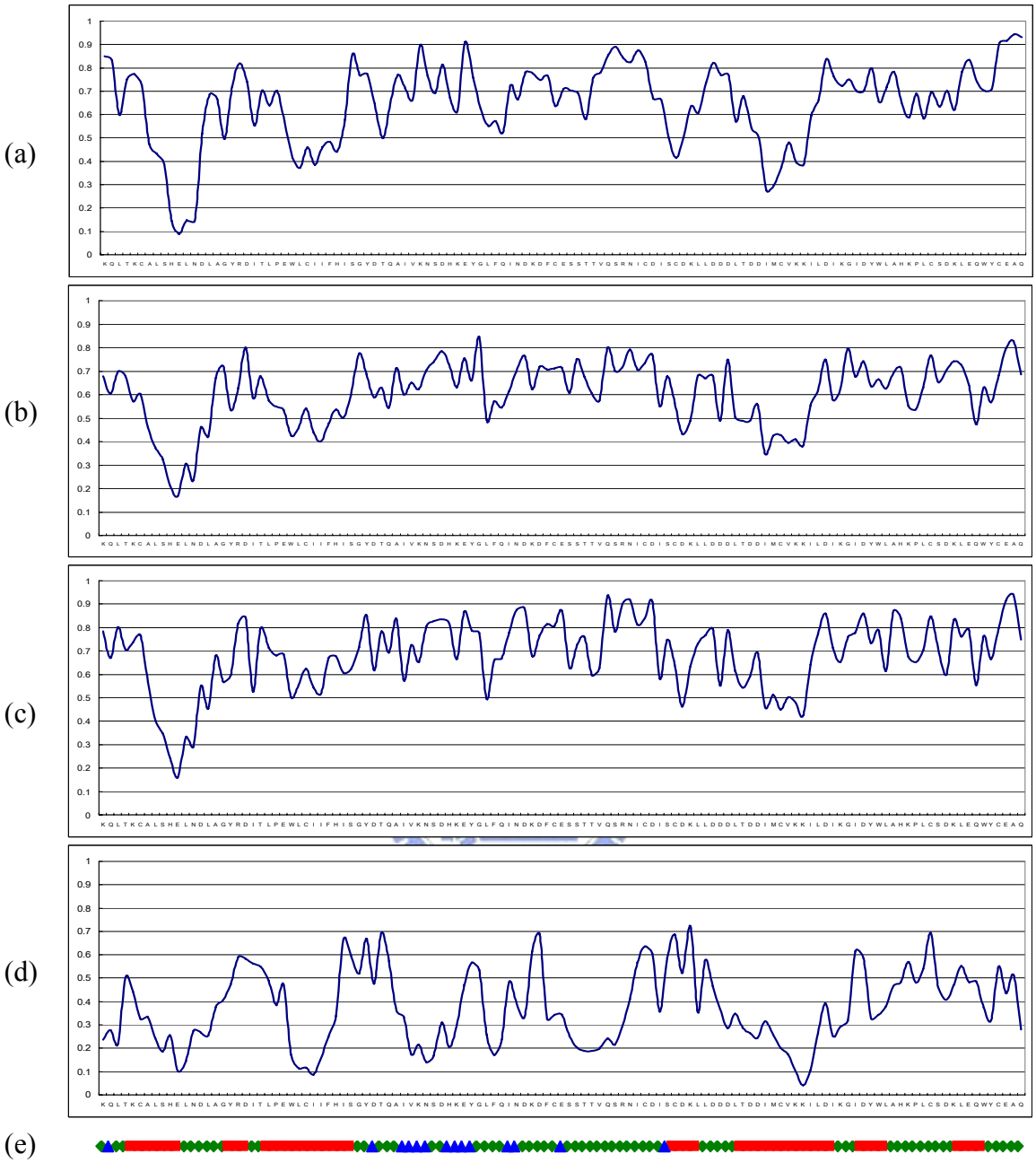
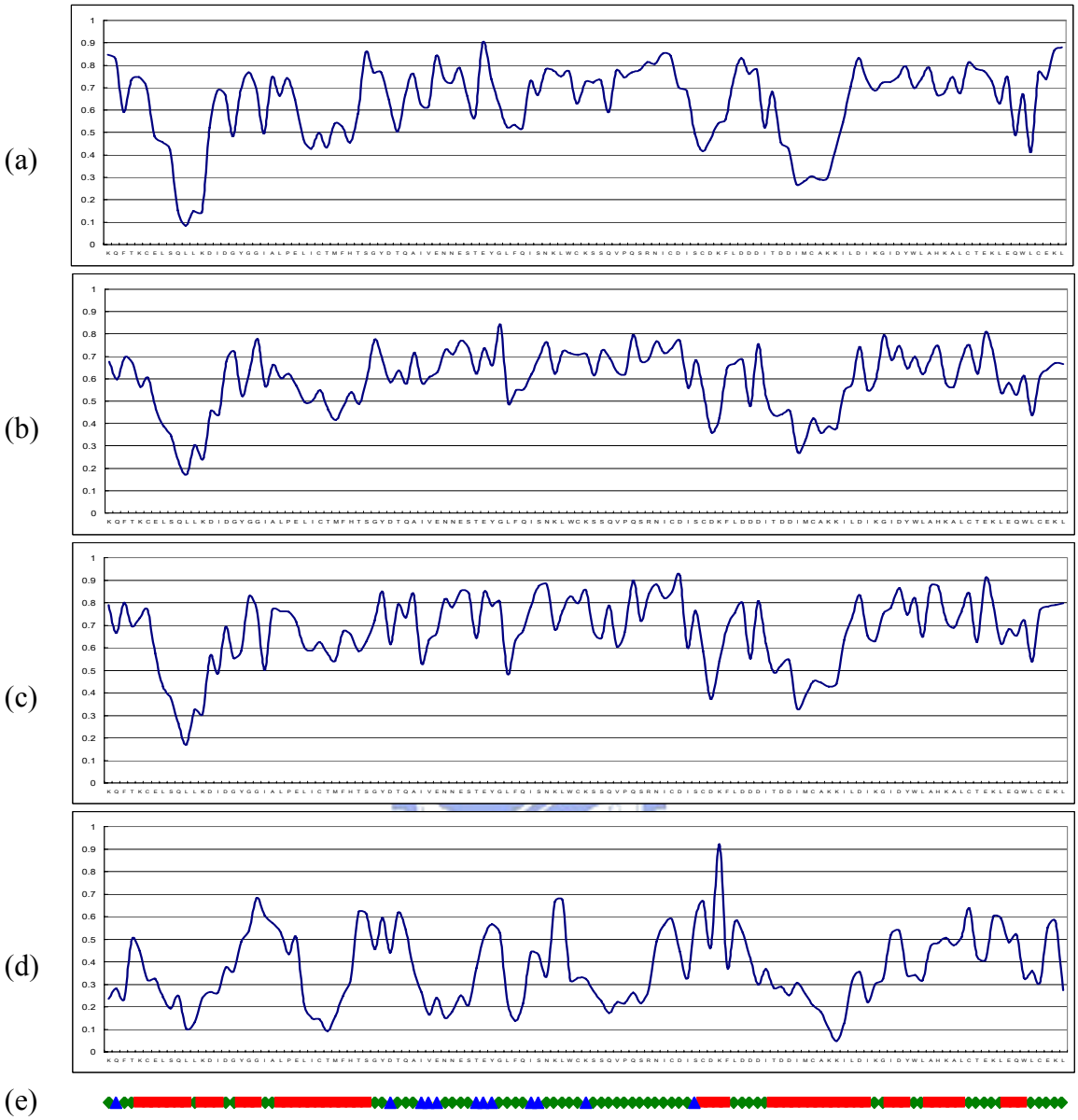


Figure 17.



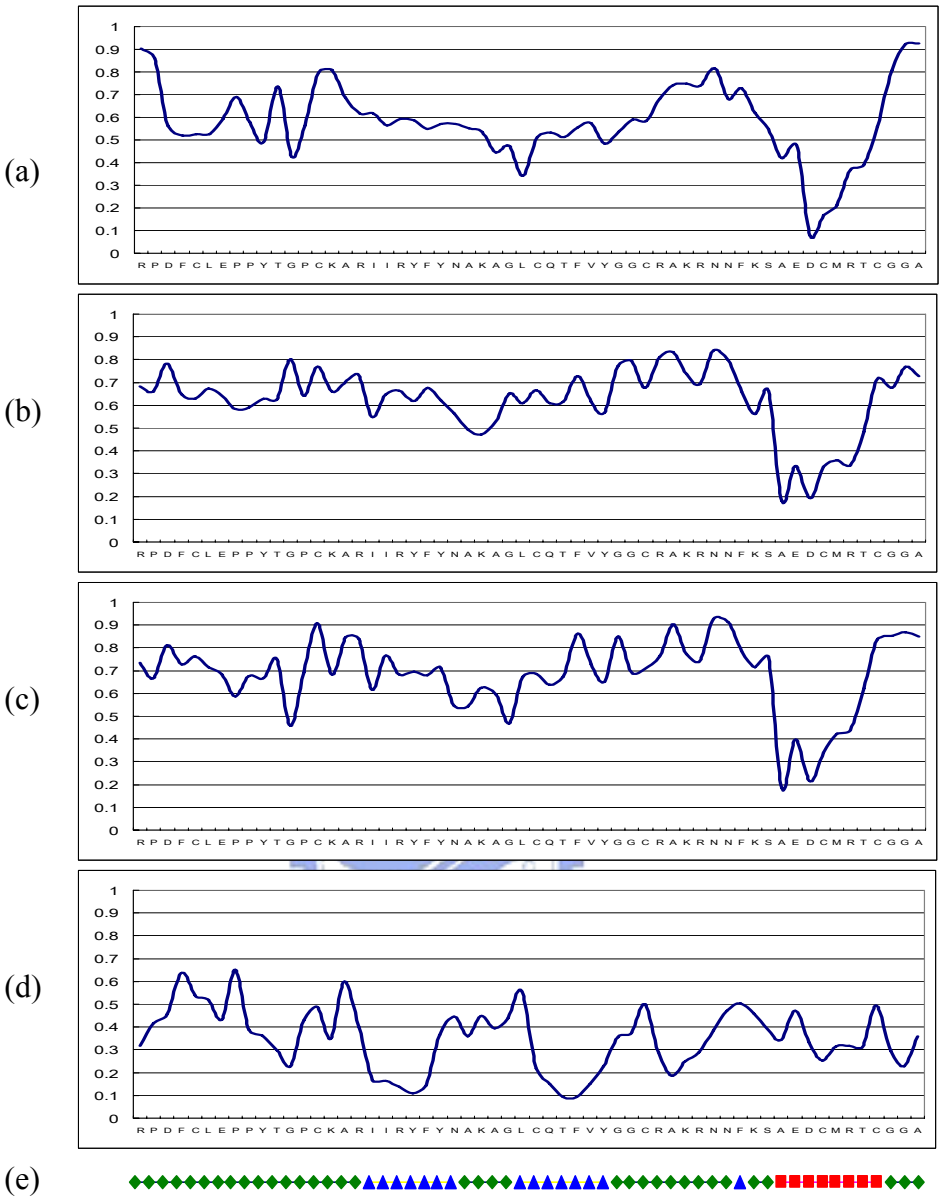
1HFX

Figure 18.



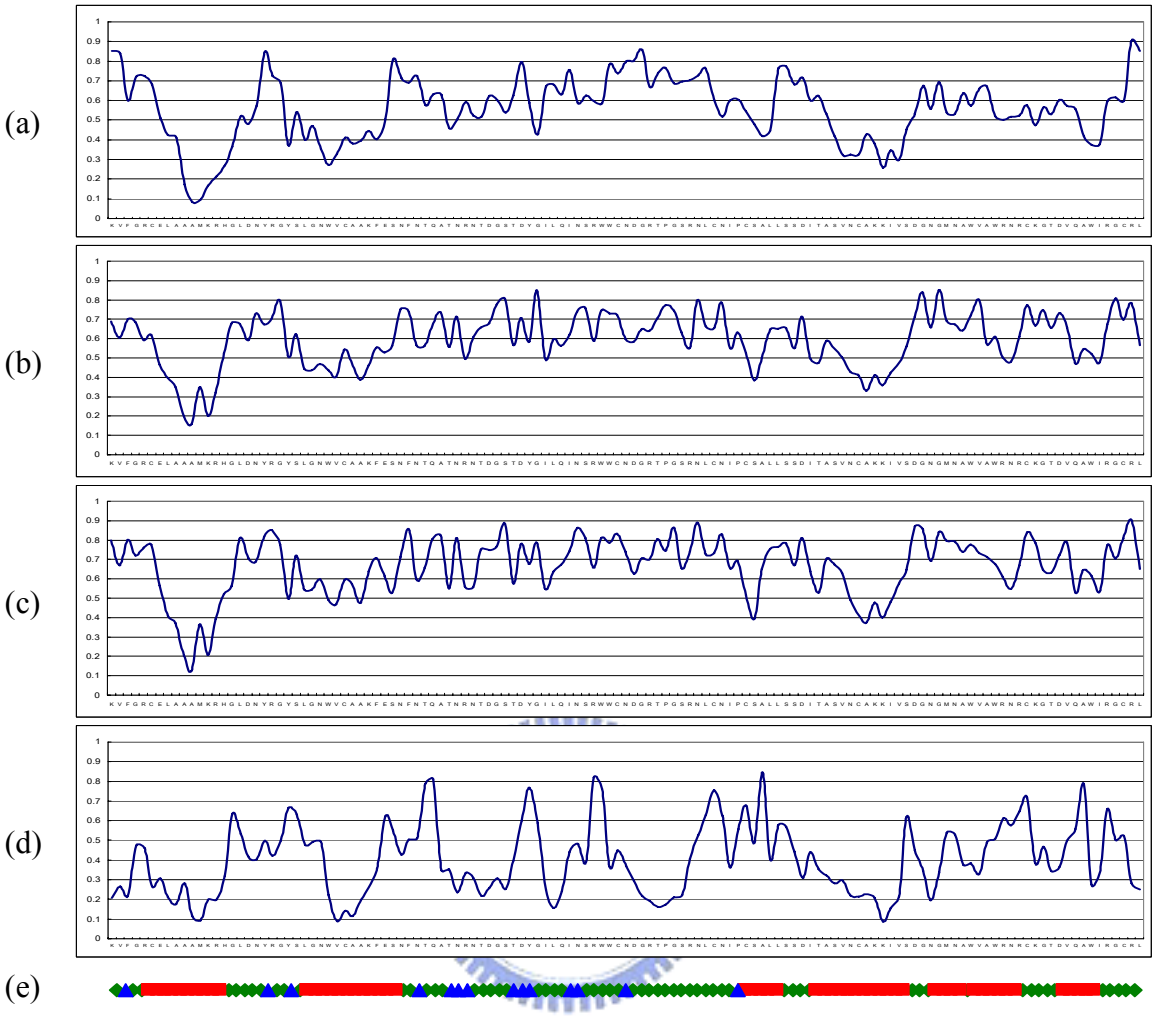
1HML

Figure 19.



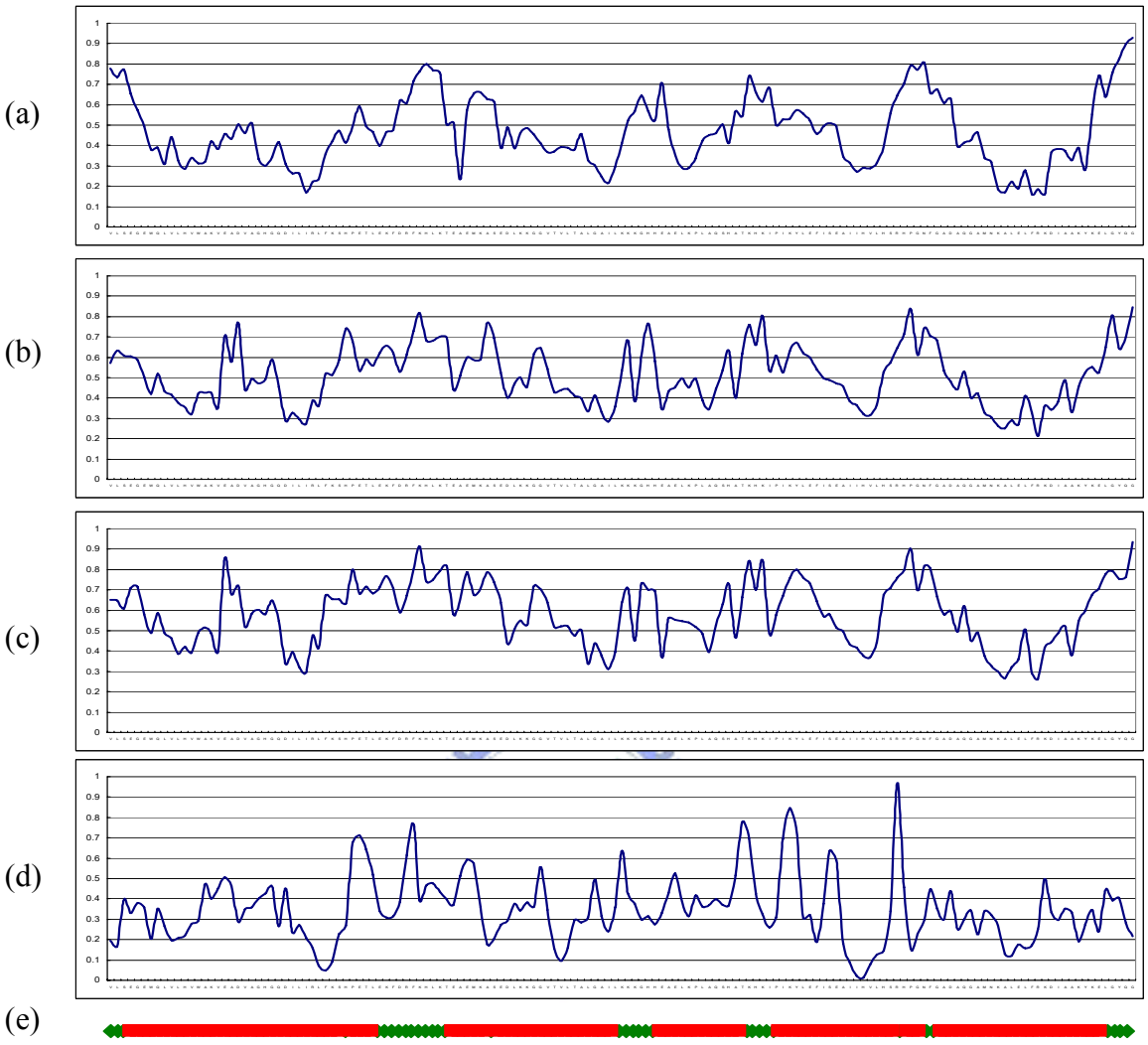
5PTI

Figure 20.



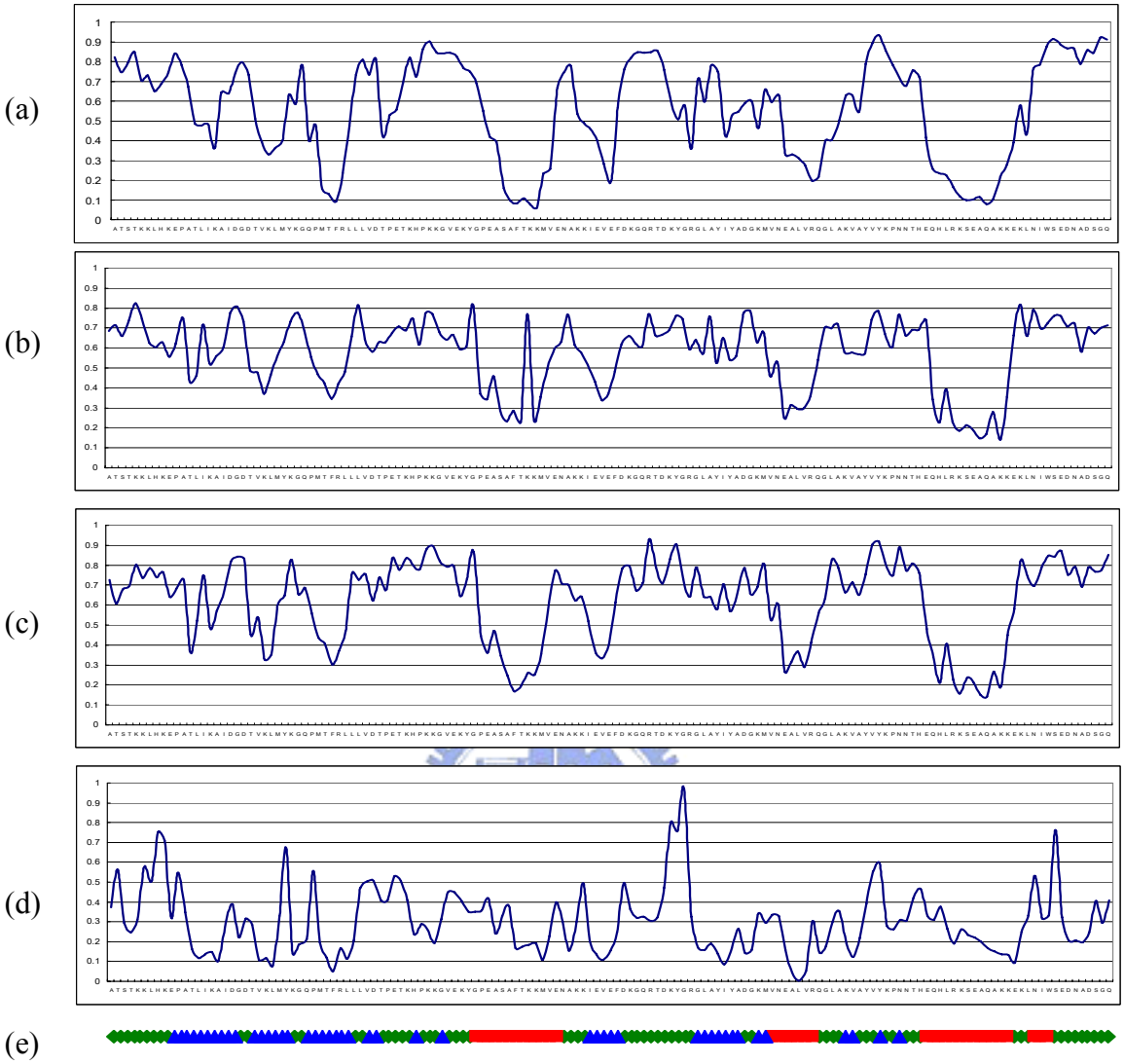
1HEL

Figure 21



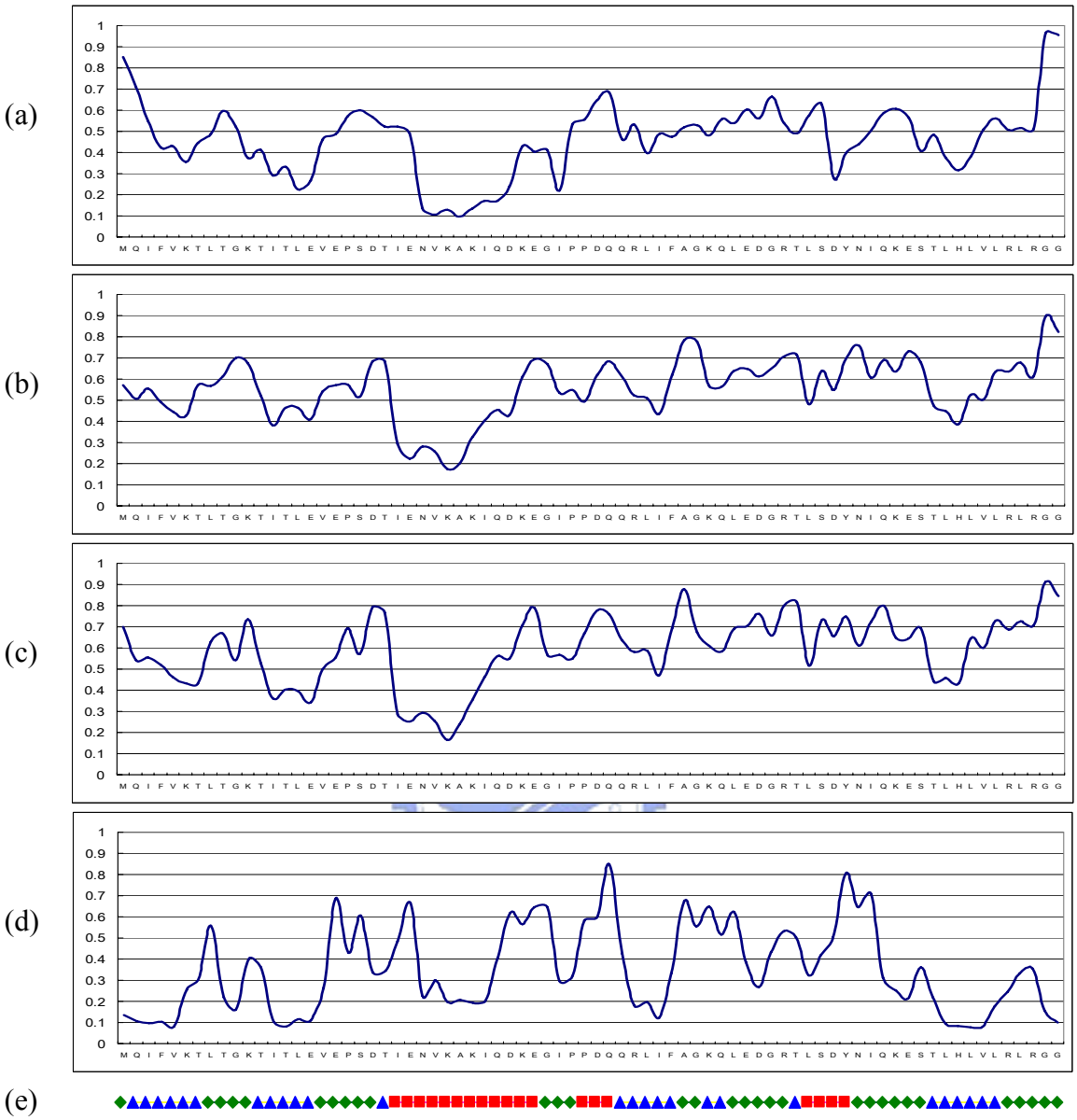
1MBO

Figure 22



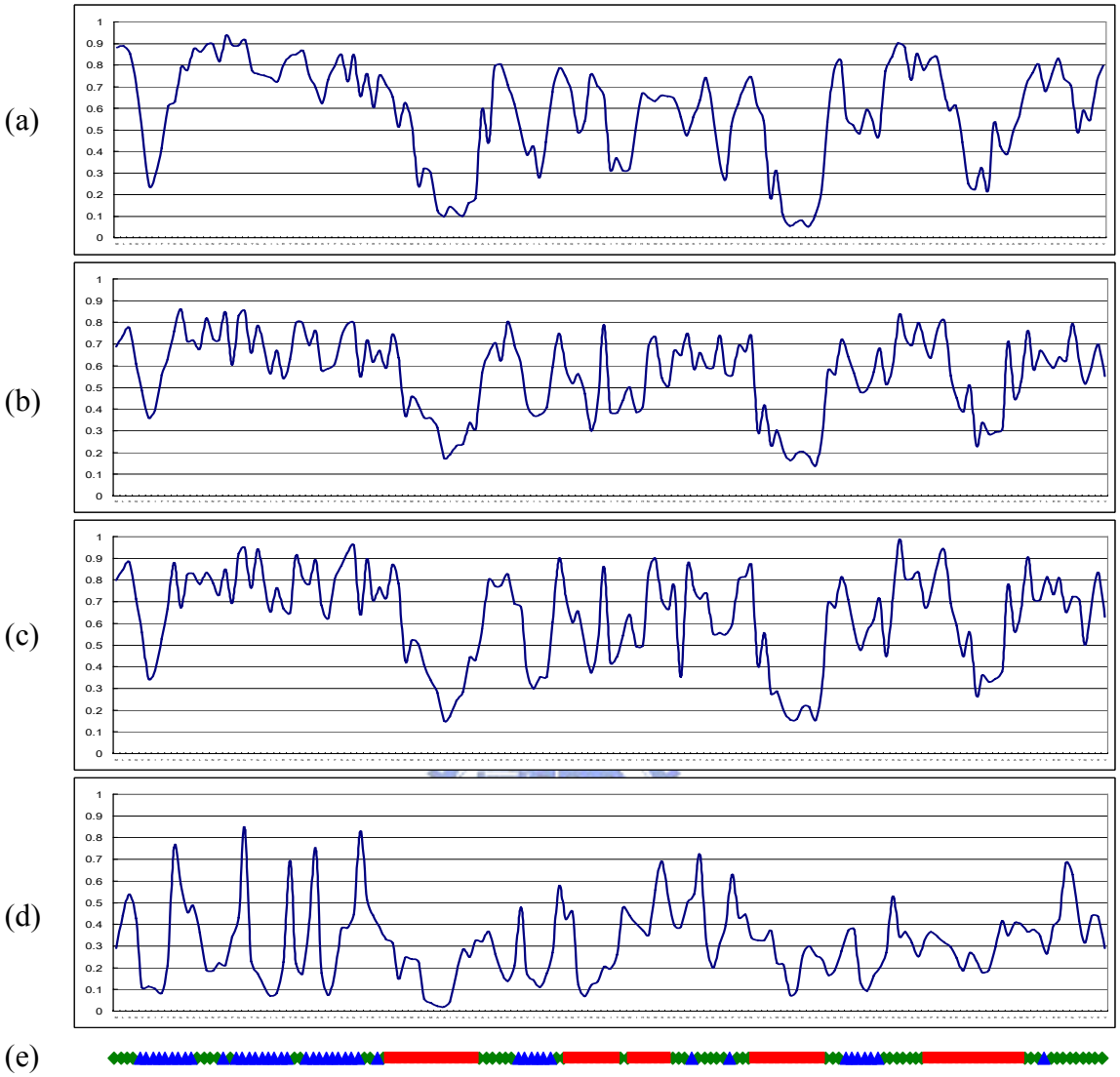
1STN

Figure 23



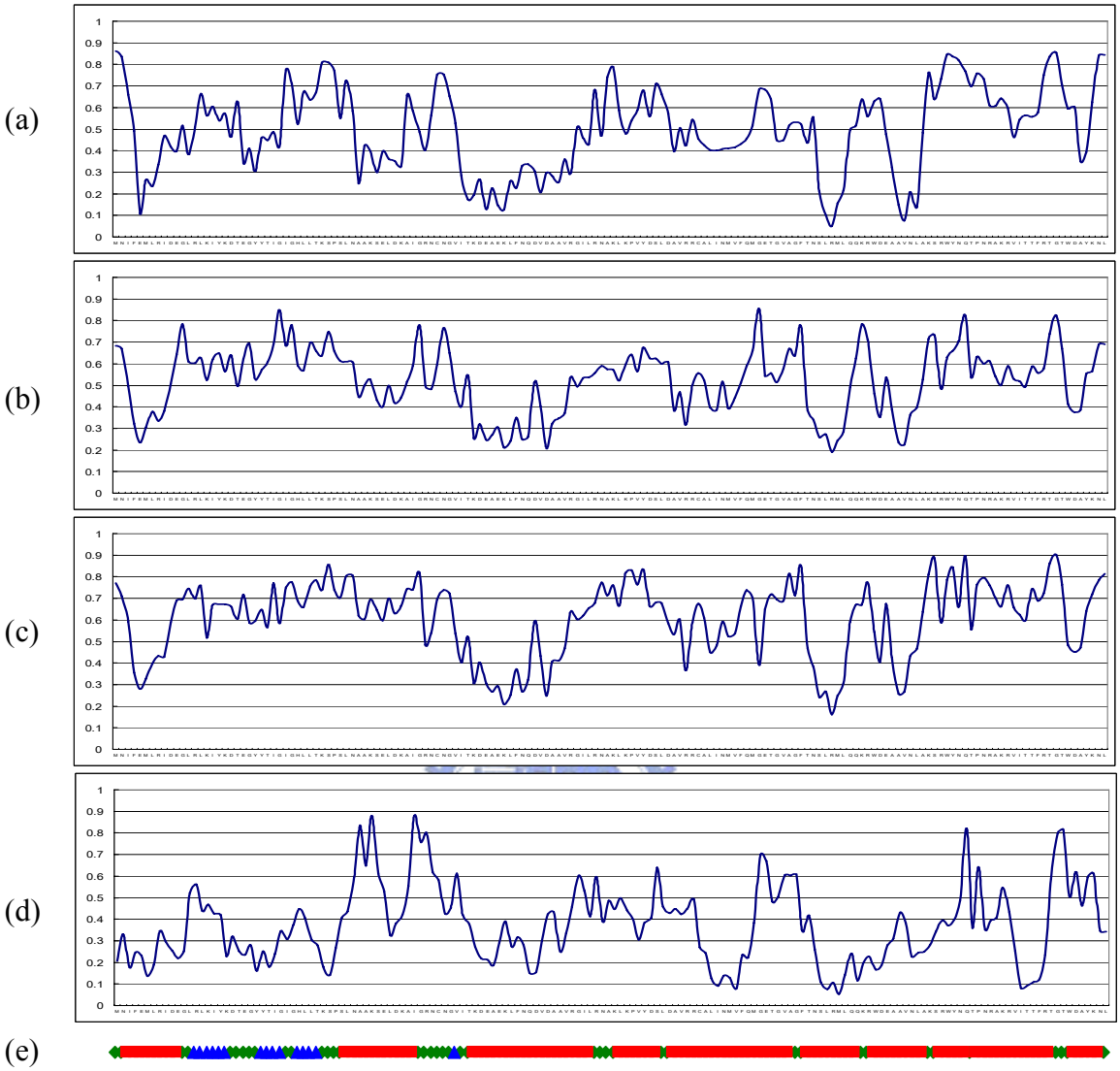
1UBI

Figure 24



1F21_A

Figure 25.



2LZM

Figure 26.

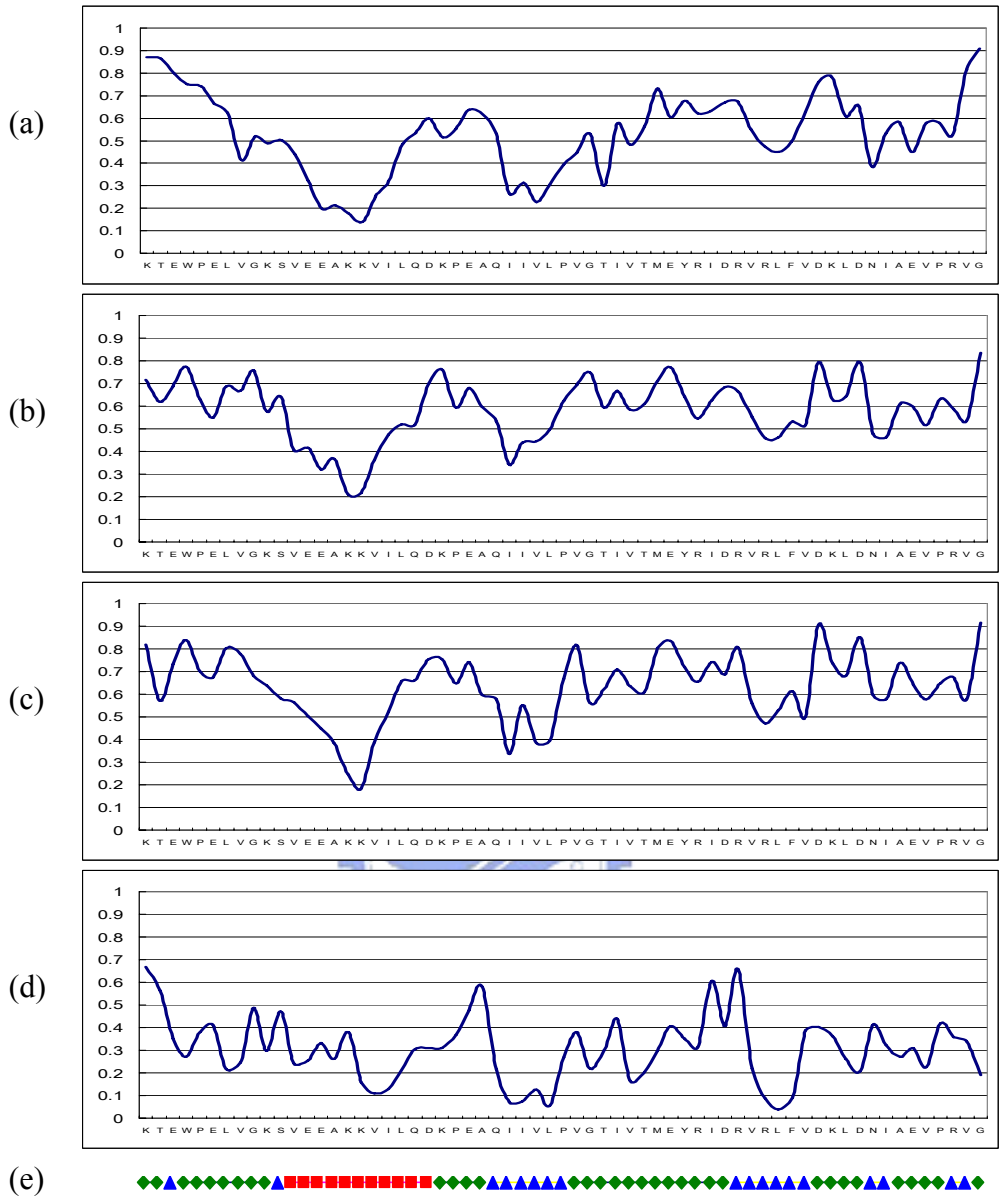

(a)
(b)
(c)
(d)
(e)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
K T E W P E L V G K S V E E A K K V I L Q D K P E A Q I I V L P V G T I V T M E Y R I D R V R L F V D K L D N I A E V P R V G
2CI2_I

Figure 27.

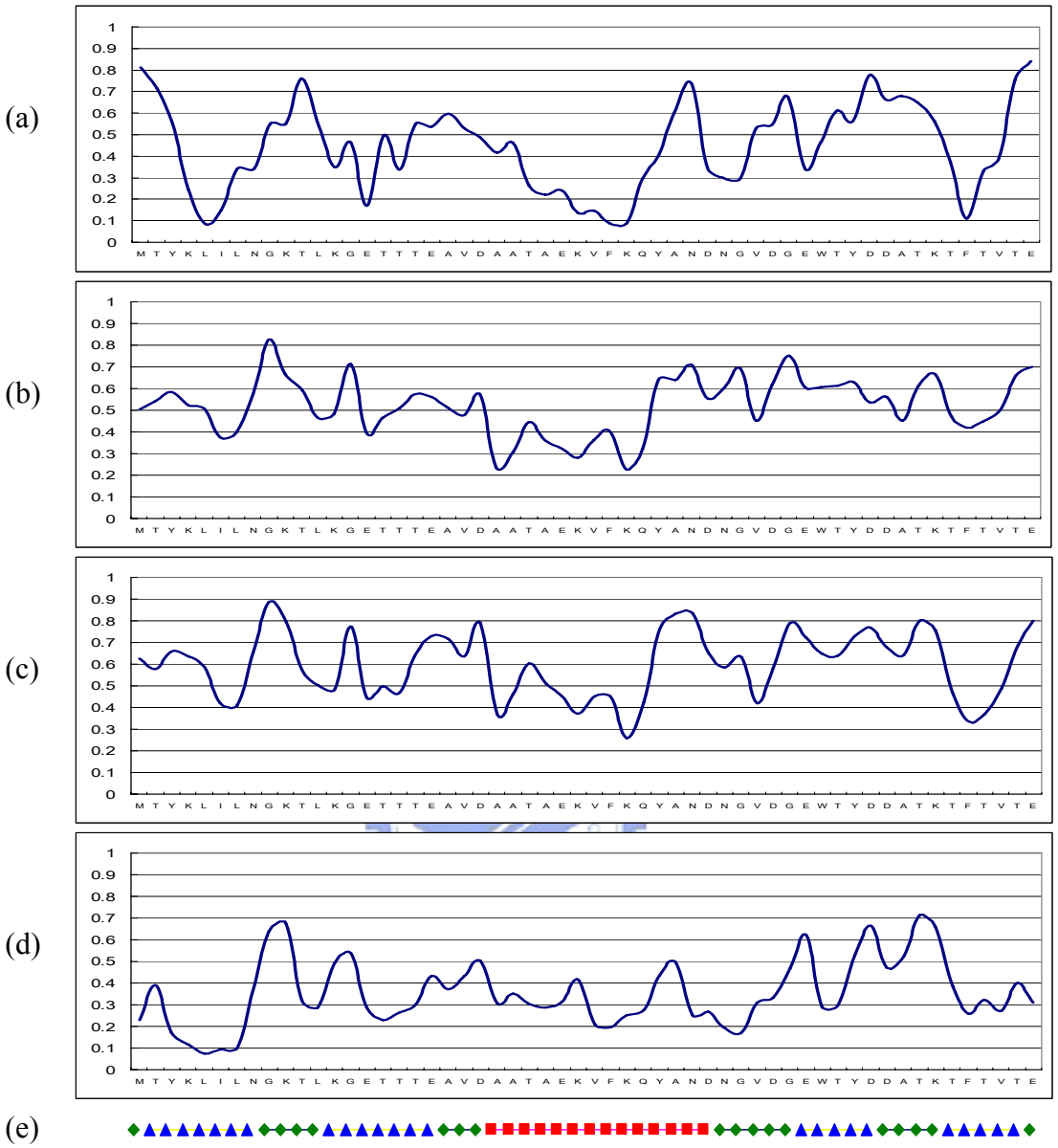

(a)
(b)
(c)
(d)
(e)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
M T Y K L I L N G K T L K G E T T T E A V D A A T A E K V F K Q Y A N D N G V D G E W T Y D D A T K T F T V T E
1PGA

Figure 28.

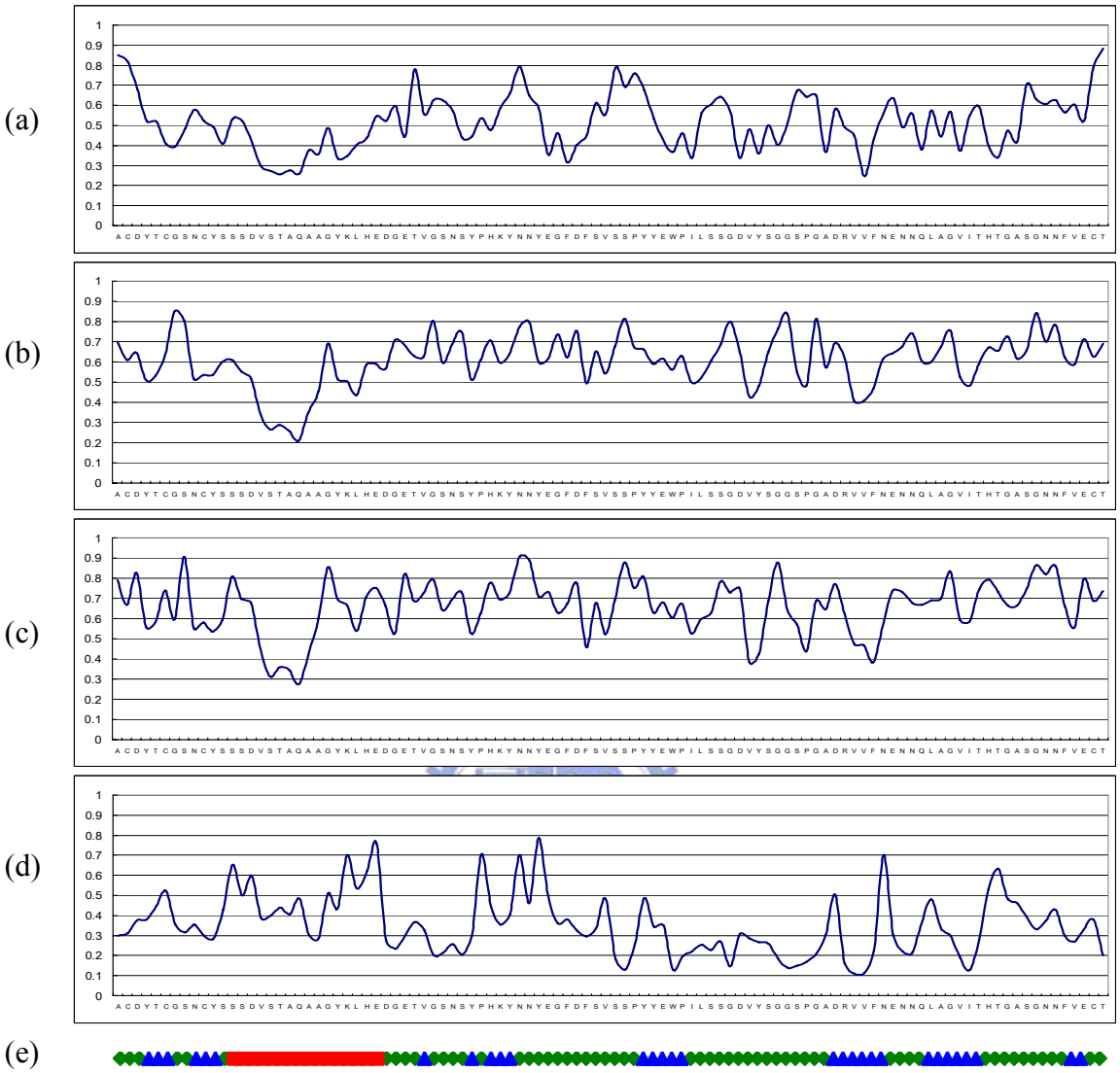


Figure 29.

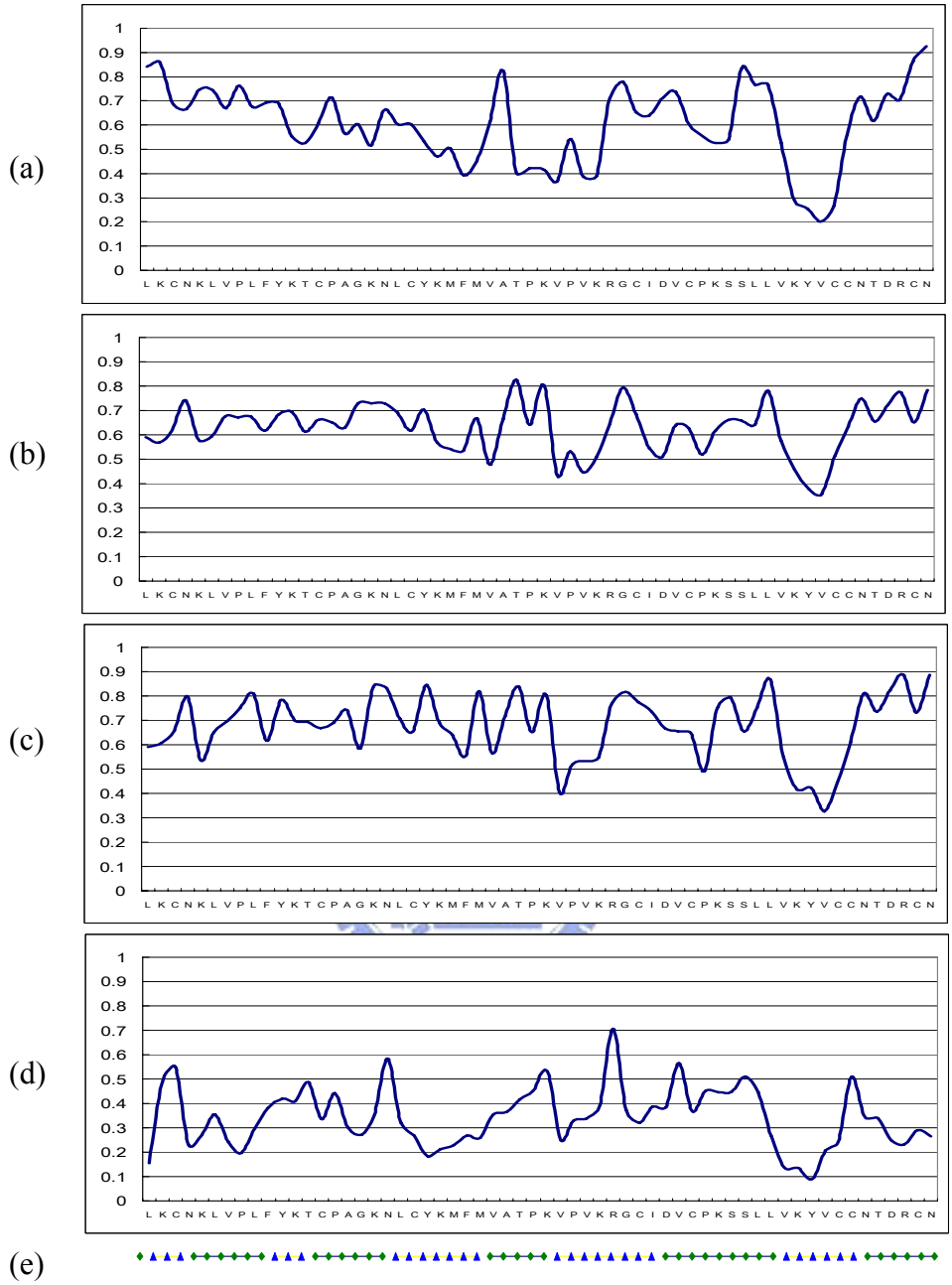

(a)
(b)
(c)
(d)
(e)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
L K C N K L V P L F Y K T C P A G K N L C Y K M F M V A T P K V P V K R G C I D V C P K S S L L V K Y V C C N T D R C N
2CRT

Figure 30.

Appendix

Appendix 1. *RS126* Dataset

256b_A	5er2_E	1ovo_A	1BKS_A	6cts_	2ltn_A	4ts1_A	4cpa_I
9api_B	1G6N_A	2mhu_	1acx_	1DUR_A	2pcy_	1prc_H	6dfr_
7cat_A	3hmg_A	2rsp_A	1bds_	2gls_A	1rhd_	2alp_	1fxi_A
6cpa_	1lap_	1tgs_I	1CDH_	2i1b_	4sgb_I	4bp2_	1hip_
3ebx_	2or1_L	2wrp_R	1cse_I	1LMB_3	1tnf_A	4cms_	9ins_B
2FOX_	1r09_2	6acn_	1fdl_H	9pap_	1prc_C	1eca_	1mcp_L
6hir_	7rsa_	1bbp_A	1gd1_O	1rbp_	2AK3_A	1IQZ_A	2phh_
1l58_	2tgp_I	2ccy_A	5hvp_A	3SDH_A	3blm_	4gr1_	4rhv_4
2mev_4	9wga_A	1crn_	2lhb_	2tmv_P	3cln_	1il8_A	2sod_B
1pyp_	8abp_	1fc2_C	1paz_	4xia_A	5cyt_R	5lyz_	1ubq_
3rnt_	1CYO_	1A45_	1ppt_	3ait_	1FND_	3pgm_	1prc_M
2stv_	1cc5_	2hmz_A	1s01_	1bmv_2	1gp1_A	4rhv_3	
2utg_A	4cpv_	1GDJ_	6tmn_E	3cla_	7icd_	2sns_	
2aat_	1etu_	2pab_A	1BKS_B	2cyp_	2ltn_B	2tsc_A	
1azu_	2gbp_	1mrt_	8adh_	1fkf_	4pfk_	1prc_L	
1cbh_	3hmg_B	4rxn_	1bmv_1	2gn5_	4rhv_1	9api_A	
6cpp_	5ldh_	3tim_A	1cdt_A	3icb_	1sh1_	2cab_	

Appendix 2. *Protein Blocks Databanks* filter out the redundant chains

153l__	19gsA_	1a12A_	1a1x__	1a2pA_	1a2zA_	1a3aA_	1a3h__	1a44__	1a4iB_
1a4uA_	1a6q__	1a6q__1	1a76__	1a7uA_	1a8e__	1a8h__	1a8l__	1a8p__	1aew__
1af7__	1afwB_	1agjA_	1ah7__	1ahc__	1ai3__	1ai9A_	1aj8A_	1ajz__	1ako__
1amm__	1amp__	1aocA_	1aop__	1aop__3	1apyA_	1apyB_	1aq0A_	1aqb__	1arb__
1aru__	1atlA_	1110	1axn__	1ayoA_	1ayx__	1az9__	1azo__	1azo__1	1b00A_
1b1cA_	1b2pA_	1b3mA_	1b4kA_	1b5eA_	1b5qB_	1b71A_	1b8aA_	1b8pA_	1b94A_
1b9hA_	1bbpA_	1bd8__	1bea__	1bf2__	1bf6B_	1bfd__	1bg6__	1bg6__1	1bgf__
1bgvA_	1bhe__	1bhtA_	1bj7__	1bjwA_	1bkpB_	1bkzA_	1bn8A_	1bolA_	1bqk__
1bslB_	1bsmA_	1btl__	1btn__	1bu7A_	1buoA_	1bxaA_	1byfB_	1byqA_	1bz0A_
1bzyA_	1c02A_	1c1fA_	1c1kA_	1c2pA_	1c3jA_	1c3jA_1	1c3jA_2	1c3kA_	1c3qA_
1c44A_	1c8kA_	1c8uA_	1ca1__	1cczA_	1celA_	1cem__	1ceqA_	1ceqA_1	1cewI_
1cfb__	1chd__	1chmA_	1cjcA_	1ckeA_	1ckeA_1	1cmbA_	1cnzA_	1cozA_	1cp2A_
1cpn__	1cq3A_	1cqxA_	1crzA_	1cs6A_	1css__	1cv8__	1cvl__	1cvl__1	1cvrA_
1cy9A_	1cy9A_1	1cy9A_2	1cz1A_	1czfA_	1cznA_	1cztA_	1d0bA_	1d0qA_	1104
1d2mA_	1d2mA_1	1d2mA_2	1d2oA_	1d2tA_	1d2tA_1	1d2vA_	1d2vC_	1d3sA_	1d4oA_
1d6oA_	1d8wA_	1d9cA_	1dbfA_	1dbxA_1	1dciA_	1dd3A_	1dd9A_	1dd9A_1	1dd9A_2
1ddt__	1ddt__1	1ddvA_	1dfx__	1dg3A_	1dg3A_1	1dg3A_3	1dg3A_4	1dgwA_	1dgyA_
1dgyA_1	1dhn__	1dixA_	1dj0A_	1dk0A_	1dk8A_	1dlwA_	1dmgA_	1dmgA_1	1dmhA_
1dmr__	1doi__	1dorA_	1dowA_	1dozA_	1dp4A_	1dqeA_	1dqgA_	1dqiA_	1dqtA_
1dqzA_	1ds0A_	1dsbA_	1dtoA_	1dtoA_1	1dts__	1dugA_	1dupA_	1dusA_	1duwA_
1dxeA_	1dxy__	1dysA_	1dytA_	1dz3A_	1dzfA_	1e0cA_	1e15A_	1e19A_	1e29A_
1e2hB_	1e2hB_1	1e2hB_2	1e2hB_3	1e2uA_	1e3aA_	1e3uB_	1e4fT_	1e4fT_1	1e5mA_
1e6oL_	1e6qM_	1e6uA_	1270	1e87A_	1ecsA_	1edg__	1edqA_	1edt__	1ee8A_
1eehA_	1eehA_1	1eejA_	1ef8C_	1ef8C_1	1eg9A_	1eg9B_	1eguA_	1eif__	1eif__1
1eimA_	1ej2A_	1ejbA_	1ejdA_	1ejjA_	1ek0A_	1ekgA_	1el4A_	1el6A_	1emvB_
1eo6B_	1eo9A_	1eo9B_	1eokA_	1ep0A_	1eq6A_	1erzA_	1esfA_	1esfA_1	1esgB_
1esl__	1eswA_	1eu3A_	1eu8A_	1euaA_	1euhA_	1eur__	1evxA_	1ew0A_	1ew4A_
1ew6A_	1ex2A_	1extA_	1ey0A_	1eyhA_	1eyqA_	1eyvB_	1ez3A_	1f00I_	1f08A_
1f0kA_	1f0xA_	1f0xA_1	1f2dA_	1f2tB_	1f2uA_	1f32A_	1f39A_	1f3uA_	1f3uF_2
1f5mB_	1f5vA_	1f5wA_	1f6kA_	1f7sA_	1f8aB_	1f8aB_1	1f8mA_	1f9zA_	1fc3A_
1fc9A_	1fd7D_	1ff9A_	1ff9A_1	1fgyA_	1fi2A_	1fit__	1fj2A_	1fjhA_	1fjhA_1
1fkmA_	1fl2A_	1flmA_	1flp__	1fm0E_	1fm0E_1	1fmtA_	1fmtA_1	1fn9A_	1fnc__
1fnnA_	1fnnA_1	1fp1D_	1fp1D_1	1fp2A_	1fp3A_	1fs7A_	1ft5A_	1ftrA_	1fua__
1fupA_	1fus__	1fvaB_	1fvuB_	1fvuB_1	1fx7B_	1fx7B_1	1fxlA_	1fyeA_	1fyeA_1
1fzqA_	1g0sB_	1g12A_	1g13A_	1g1bA_	1g1kA_	1g291_	1g3qA_	1g57B_1	1g5tA_
2227	1g6sA_	1g72A_	1g73A_	1g73D_	1g8iA_	1g8iA_1	1g8lA_	1gakA_	1gceA_

Continued

1gceA_1	1gcuA_	1gcyA_	1gcyA_1	1gd0A_	1gd1O_	1gefA_	1geqB_	1geqB_1	1gg6B_
1ggxA_	1gia__	1gnd__	1gof__	1gp1A_	1gpeA_	1gpr__	1gsh__	1gsh__1	1gsh__2
1h2rL_	1h2rS_	1hcl__	1hcl__1	1hcz__	1hd5A_	1hd5A_1	1hd7A_	1hd7A_2	1he7A_
1hf8A_	1hfc__	1hhsA_	1ho1B_	1ho1B_1	1hruA_	1hsbA_	1htrB_	1hxqA_	1hxqA_1
1i0dA_	1i0rB_	1i1qB_	1i1qB_1	1i1qB_2	1i39A_	1i6pA_	1i9yA_	1i9yA_1	1i9yA_2
1iab__	1iakA_	1iakB_	1iazA_	1149	1icjA_	1ido__	1ie8A_	1ie8A_1	1igs__
1ihgA_	1io7A_	1jbc__	1jfrA_	1jkmB_	1kdj__	1koe__	1kpf__	1lam__	1lenC_
1lib__	1lki__	1lml__	1lml__1	1lml__2	1ltm__	1ltm__1	1ltsA_	1mba__	1mgtA_
1mkaA_	1mla__	1mml__	1227	1mugA_	1muyA_	1nah__	1nar__	1nbaA_	1nbcA_
1nfn__	1nfn__1	1nkr__	1nlr__	1nox__	1npk__	1nseA_	1nsf__	1nsj__	1nsyA_
1nwpA_	1nzyA_	1obwA_	1oen__	1oen__1	1oen__2	1onrA_	1pamA_	1pbn__	1pbv__
1pbwA_	1pdo__	1phm__	1phm__1	1phnA_	1php__	1pmi__	1pnkB_	1poa__	1ppn__
1pprM_	1prn__	1puc__	1pud__	1qadA_	1qadA_1	1qazA_	1qb0A_	1qb8A_	1qbiB_
1qbiB_1	1qbiB_2	1qcxA_	2326	1qd9A_	1qdeA_	1qdeA_1	1qe3A_	1qe3A_1	1qe3A_2
1qftA_	1qg8A_	1qg8A_1	1qgiA_	1qh4A_	1qh5A_	1qhqA_	1qhvA_	1qi7A_	1qj5A_
1qj5A_1	1qjdA_	1qk8A_	1qksA_	1qnrA_	1qnxA_	1qqjA_	1qsaA_	1qstA_	1qtoA_
1qtsA_	1qu1F_	1regX_	1rhs__	1rl6A_	1rmg__	1rom__	1rpjA_	1rro__	1sacA_
1seiA_	1sftA_	1skf__	1sll__	1smlA_	1sra__	1srvA_	1stmA_	1sur__	1svb__
1svy__	1tca__	1tf4A_	1tfe__	1thfD_	1thm__	1thv__	1tib__	1tkiA_	1tl2A_
1tolA_	1tolA_1	1tpfA_	1trkA_	1ttqB_	1tyu__	1tyu__1	1tyu__2	1uch__	1uch__1
1udh__	1uok__	1up1__	1uroA_	1vcaA_	1vfrA_	1vid__	1vjs__	1vjs__1	1vls__
1vpnB_	1vsd__	1vsrA_	1wab__	1wdcC_	1041	1wgtA_	1whi__	1xer__	1xgsA_
1xib__	1xnb__	1xsoA_	1xwl__	1yacA_	1yge__	1zin__	1zpdA_	256bA_	2abh__
2abk__	2baa__	2bbkH_	2bbkL_	2bc2A_	2bc2A_1	2cba__	2cpl__	2ctb__	2cuaA_1
2e2c__	2end__	2fcbA_	2gdm__	2hft__1	2hft__2	2hmzA_1	2hrvA_	2hvm__	2i1b__
2lisA_	2mcm__	2mnr__	2nacA_	2pgd__	2pia__	2pii__	2plc__	2por__	2pth__
2rn2__	2rspA_	2rspA_1	2scpA_	2sil__	2spcA_	2tgi__	2tlxA_	2tnfA_	2vhbA_
2vhbA_1	3chy__	3cyr__	3daaA_	3dni__	3dni__1	3grs__	3hsc__	3lzm__	3mbp__
3pah__	3pte__	3sdhA_	3stdA_	3thiA_	3vub__	3wrp__	4fgf__	4pgaA_	5nll__
7nn9__									

Appendix 3. *Sequence Unique Database*

1a2p_A	1m5s_A	2abk_	1m3u_A	3c2c_	1ly2_A	2sli_	1m7y_A	2fcr_	1m3s_A
1a2z_A	1m6k_A	2acy_	1m48_A	3ebx_	1lzl_A	2spc_A	1mb3_A	2hbg_	1m4v_A
1a4m_A	1mc2_A	2ahj_A	1m4l_A	3fap_B	1m0k_A	2tgi_	1mhn_A	2hmz_A	1m55_A
1a58_	1me4_A	2ahj_B	1m6y_A	5pal_	1m1n_B	3csu_C	1mix_A	2lis_A	1m65_A
1a73_A	1mhw_A	2ayh_	1m7s_A	5tmp_A	1m22_A	3ezm_A	1mju_H	2nlr_A	1m6j_A
1a8b_	1mi3_A	2cbp_	1mai_	7aat_A	1m2g_A	3psg_	1mlw_A	2oat_A	1m6s_A
1a8e_	1mjf_A	2cy3_	1mfg_A	9wga_A	1m33_A	3sdh_A	1mml_	2psp_A	1m7j_A
1a8q_	1mkz_A	2eng_	1mgr_A	1a1i_A	1m4i_A	3vub_	1mmq_	2pvi_A	1m8z_A
1abe_	1ml4_A	2fcb_A	1mid_A	1a3a_A	1m8a_A	4ubp_A	1mo1_A	2sga_	1mba_
1agi_	1mla_	2ilk_	1mj4_A	1a6m_	1mdc_	5hpg_A	1mo9_A	2wea_	1md6_A
1ail_	1mng_A	2ltn_B	1mj5_A	1a88_A	1mhw_C	5nll_	1moq_	3cla_	1mdx_A
1aj8_A	1mog_A	2mcm_	1mqk_H	1a8l_	1mpg_A	7atj_A	1mqq_A	3dfr_	1mg4_A
1aky_	1mol_A	2mhr_	1mqk_L	1a8u_A	1mqi_A	8tln_E	1mro_B	3grs_	1mgt_A
1al3_	1mpx_A	2pii_	1mr3_F	1ajj_	1mr8_A	9rnt_	1mro_C	3lzm_	1mju_L
1aoe_A	1mr7_A	2pvb_A	1mrj_	1amf_	1mtp_A	1a4i_A	1muc_A	3nul_	1mn8_A
1aoz_A	1mtz_A	2rn2_	1mro_A	1amk_	1mtp_B	1aap_A	1mwq_A	3pcc_A	1mrg_
1atg_	1muw_A	2sn3_	1msk_	1atl_A	1mug_A	1ah7_	1mxg_A	3std_A	1n40_A
1ayf_A	1myt_	2trx_A	1mv8_A	1atz_A	1mve_A	1aoh_A	1mzn_A	4fgf_	1n57_A
1b0u_A	1n1j_A	2wrp_R	1mvo_A	1axn_	1mxr_A	1arb_	1n0q_A	4ubp_B	1n7s_A
1b4p_A	1n4w_A	3cox_	1mwp_A	1b1c_A	1my5_A	1arv_	1n13_B	4ubp_C	1n83_A
1b65_A	1n7s_B	3lzt_	1mwv_A	1b25_A	1my6_A	1avb_A	1n1t_A	7ahl_A	1na3_A
1b67_A	1n7s_C	3mbp_	1mxi_A	1b3a_A	1mz4_A	1ayx_	1n67_A	7fd1_A	1naq_A
1b8o_A	1n9b_A	3seb_	1n3l_A	1b5f_B	1n0w_B	1azq_A	1n71_A	1a12_A	1nc5_A
1b8p_A	1nar_	4pga_A	1n45_A	1b6t_A	1n13_A	1b0b_	1n7f_A	1a7s_	1ne2_B
1bbz_A	1nc7_A	5cyt_R	1n6a_A	1b8d_A	1n1f_A	1b0n_A	1n93_X	1aac_	1nf9_A
1bfd_	1ncx_	5tim_A	1n7h_A	1b9o_A	1n1j_B	1b4k_A	1na5_A	1ag9_A	1nki_A
1bi5_A	1nd1_A	7a3h_A	1n7s_D	1bgp_	1n2a_A	1b5q_A	1nb9_A	1agj_A	1no5_A
1bio_	1ne7_A	153l_	1n82_A	1bkj_A	1n62_A	1bb1_A	1nbc_A	1aho_	1nps_A
1bm8_	1nep_A	1a34_A	1n8v_A	1bkr_A	1n62_C	1bbh_A	1nh2_D	1ajs_A	1nqj_A
1bqb_A	1nhk_L	1a44_	1n97_A	1bqc_A	1n8n_A	1bdo_	1nk0_A	1ake_A	1nqu_A
1brf_A	1nkg_A	1a8s_	1ne9_A	1bs0_A	1nh2_B	1bea_	1nls_	1alv_A	1nrw_A
1bte_A	1nkx_A	1ado_A	1nfp_	1bsm_A	1nkr_	1beb_A	1nof_A	1aun_	1ns5_A
1bud_A	1nln_A	1aew_	1ng2_A	1bxy_A	1noa_	1bf6_A	1nox_	1auo_A	1nsc_A
1byi_	1nme_A	1afw_A	1nh0_A	1c02_A	1npk_	1bg6_	1nq7_A	1ava_C	1ntn_
1bz4_A	1nme_B	1agq_A	1nh2_C	1c1k_A	1nq6_A	1bhp_	1nty_A	1awd_	1nuu_A
1c0p_A	1nng_A	1ak0_	1nh8_A	1c2a_A	1nrj_A	1bj7_	1nvm_B	1ay7_B	1nxm_A

Continued

1c1d_A	1nnh_A	1ako_	1nhc_A	1c5y_B	1nuy_A	1bkp_A	1nwp_A	1ayl_	1ny1_A
1c1f_A	1nns_A	1ayo_A	1nlf_A	1c7k_A	1nvm_A	1bsg_	1nww_A	1b16_A	1o13_A
1c5e_A	1nnx_A	1b5e_A	1nm8_A	1cgh_A	1nwz_A	1bu8_A	1nxq_A	1b2p_A	1o1y_A
1cb8_A	1np7_A	1b5f_A	1nn6_A	1chd_	1nxc_A	1c26_A	1nyk_A	1b8a_A	1o6a_A
1ccr_	1npi_A	1b6a_	1nog_A	1cjw_A	1o1z_A	1c3c_A	1nyt_A	1b8j_A	1o6i_A
1cka_A	1npy_A	1b93_A	1nqc_A	1cmb_A	1o4r_A	1c52_	1o08_A	1bd8_	1o7e_A
1clc_	1nr0_A	1b94_A	1nvr_A	1cns_A	1o4v_A	1cbs_	1o0e_A	1bgv_A	1o7i_A
1csh_	1nr4_A	1bgf_	1nwa_A	1co6_A	1o4w_A	1cc3_A	1o2d_A	1bhe_	1o81_A
1ctj_	1nrg_A	1blx_A	1nyc_A	1cp2_A	1o54_A	1cc8_A	1o3u_A	1bkz_A	1o8v_A
1cuk_	1ntf_A	1bn7_A	1nzj_A	1cqx_A	1o58_A	1cd0_A	1o4s_A	1bmd_A	1obb_A
1cv8_	1nth_A	1c3p_A	1o02_A	1cru_A	1o5u_A	1cg5_A	1o4t_A	1bn8_A	1ock_A
1cvl_	1nu2_A	1c44_A	1o0x_A	1ctf_	1o7q_A	1ci9_A	1o4y_A	1bqk_	1oe1_A
1czn_A	1nu5_A	1c75_A	1o0y_A	1d3h_A	1o82_A	1cjc_A	1o60_A	1bue_A	1oej_A
1d2i_A	1nul_A	1c7n_A	1o1x_A	1d4a_A	1oaq_H	1cnv_	1o6v_B	1bx4_A	1of8_A
1d7d_A	1nxj_A	1ccz_A	1o5k_A	1d4x_G	1oaq_L	1cnz_A	1o7j_A	1bx7_	1oko_A
1d7o_A	1nxp_A	1cex_	1o6s_B	1d5n_A	1ob3_A	1cs1_A	1o91_A	1bxv_A	1olr_A
1dc1_A	1nxu_A	1ci3_M	1o7n_B	1db1_A	1obd_A	1cse_I	1oaa_	1byb_	1on3_A
1dci_A	1nzy_A	1coj_A	1obf_O	1dbx_A	1ocy_A	1cy5_A	1oaf_A	1c5c_L	1onw_A
1dde_A	1o0i_A	1cot_	1oc2_A	1df4_A	1od3_A	1cyo_	1oao_C	1c9o_A	1oot_A
1dgw_Y	1o26_A	1cqy_A	1ocb_A	1df7_A	1od6_A	1d2n_A	1odm_A	1ccw_A	1os8_A
1dl2_A	1o50_A	1cuo_A	1oi6_A	1dfu_P	1odk_A	1d4t_A	1ofz_A	1ccw_B	1ova_A
1dnu_C	1o69_A	1czf_A	1ok7_A	1din_	1odo_A	1d8w_A	1oi2_A	1cel_A	1oxs_C
1dos_A	1o7n_A	1d0d_A	1okh_A	1dj0_A	1ofl_A	1dbf_A	1ojh_A	1cg5_B	1p2f_A
1dow_B	1o97_C	1d0q_A	1oll_A	1dl5_A	1ofw_A	1deu_A	1ojx_A	1cke_A	1p5d_X
1dry_A	1o97_D	1d3v_A	1ols_B	1dlf_H	1oh4_A	1dgw_A	1olm_A	1cl1_A	1p5f_A
1dug_A	1o98_A	1d4o_A	1onj_A	1dlw_A	1ohl_A	1dhn_	1onr_A	1cpo_	1pby_B
1dus_A	1o9r_A	1d7u_A	1ooe_A	1dly_A	1ok0_A	1djt_A	1oo2_A	1cpq_	1php_
1dwk_A	1oa3_A	1dfx_	1oqc_A	1doi_	1oki_A	1dpj_A	1orr_A	1ctq_A	1pjx_A
1e58_A	1oai_A	1dgf_A	1os6_A	1dqs_A	1onc_	1dqz_A	1osy_A	1cxy_A	1pk6_C
1e5p_A	1oap_A	1dk0_A	1ouw_A	1dsz_A	1one_A	1duv_G	1otf_A	1cyd_A	1pko_A
1e85_A	1ob8_A	1dnu_A	1oyg_A	1dtd_B	1onh_A	1dw0_A	1ou8_A	1cyj_	1plc_
1ea7_A	1ogd_A	1dow_A	1p0z_A	1dys_A	1ooy_A	1dxe_A	1ow4_A	1czp_A	1puc_
1eex_B	1ogq_A	1doz_A	1p1h_A	1e29_A	1oqf_A	1dxj_A	1owl_A	1czt_A	1pvm_A
1eex_G	1ojq_A	1dqi_A	1p1m_A	1e5k_A	1oqj_A	1dxy_	1p0b_A	1dgw_X	1pyo_A
1egw_A	1oks_A	1e19_A	1p1x_A	1e6i_A	1oqv_A	1dz4_A	1p7g_A	1dk8_A	1pyo_B
1eif_	1ol0_A	1e43_A	1p5v_B	1e6y_C	1oru_A	1e0c_A	1p90_A	1dlf_L	1q08_A
1ejb_A	1ooh_A	1e6y_B	1p77_A	1e8a_A	1ox0_A	1e4m_M	1pa2_A	1dmg_A	1q7e_A

Continued

1enf_A	1opb_A	1ea5_A	1p7w_A	1eax_A	1ozn_A	1e5m_A	1pbe_	1dpt_A	1q9u_A
1ep0_A	1oq1_A	1ec7_A	1p9h_A	1ecd_	1p3c_A	1eaj_A	1pby_C	1dqp_A	1qah_A
1epf_A	1oth_A	1edq_A	1pby_A	1ed5_A	1p4c_A	1eaz_A	1pk6_A	1dsz_B	1qaz_A
1esg_A	1oxf_A	1ek0_A	1pe1_A	1ede_	1p5z_B	1eb6_A	1pl8_A	1dun_	1qd9_A
1et1_A	1oyj_A	1ekg_A	1pe9_A	1edm_B	1p9g_A	1edg_	1pmh_X	1dxg_A	1qgi_A
1euv_A	1oz2_A	1en2_A	1pm4_A	1edt_	1pa7_A	1ei5_A	1pmi_	1dyr_	1qgj_A
1ex2_A	1oz9_A	1eur_	1poa_	1eex_A	1paz_	1ej0_A	1pmy_	1dzp_A	1qgw_C
1f0x_A	1p3d_A	1euv_B	1ppo_	1ef8_A	1pbj_A	1ej8_A	1pn0_A	1e42_A	1qmg_A
1f20_A	1p6o_A	1ew0_A	1pq4_A	1egu_A	1pee_A	1eok_A	1pt6_A	1e6b_A	1qnn_A
1f4g_A	1p99_A	1ew2_A	1ptf_	1ejd_A	1pfb_A	1erz_A	1pzg_A	1e6w_A	1qpc_A
1f5j_A	1pb7_A	1ew4_A	1pv5_A	1ek6_A	1pfv_A	1esw_A	1q16_B	1e9g_A	1qqf_A
1f5v_A	1pcf_A	1ex7_A	1pwa_A	1elj_A	1pgs_	1evy_A	1q1c_A	1ea2_A	1qqp_2
1f60_A	1pdo_	1exm_A	1pym_A	1elr_A	1pkh_A	1exr_A	1q33_A	1ecs_A	1qqp_3
1f60_B	1pgv_A	1ext_A	1pz4_A	1elt_	1ppf_E	1eye_A	1q52_A	1ejx_A	1qre_A
1f74_A	1pk6_B	1f1e_A	1pzs_A	1elw_A	1pq7_A	1eyv_A	1q6h_A	1ejx_B	1qsa_A
1fe0_A	1pot_	1f1g_A	1q2w_A	1epx_A	1prn_	1f1u_A	1q8r_A	1el1_A	1qtn_B
1ff3_A	1psr_A	1f46_A	1q7t_A	1eq9_A	1puo_A	1f2t_A	1q9b_A	1ept_A	1qtw_A
1fhv_A	1pu6_A	1f8e_A	1q92_A	1erm_A	1pva_A	1f41_A	1q9r_B	1eqc_A	1qu9_A
1fi2_A	1pwb_A	1fg7_A	1qad_A	1eu3_A	1pyf_A	1fc9_A	1qav_A	1esf_A	1qv1_A
1fjj_A	1pwg_A	1fjs_A	1qb7_A	1eu8_A	1pyq_A	1fec_A	1qcz_A	1eug_A	1qve_A
1fl2_A	1px0_A	1fjs_L	1qd1_A	1eua_A	1pz3_A	1feh_A	1qh4_A	1euw_A	1qwg_A
1flt_V	1pxf_A	1fkn_A	1qft_A	1euj_A	1q0e_A	1ff4_A	1qhf_A	1ezw_A	1qwy_A
1fm0_D	1pyl_A	1flj_A	1qgd_A	1ew6_A	1q0g_A	1fh9_A	1qho_A	1f2t_B	1qz9_A
1fm4_A	1q0q_A	1fm0_E	1qgw_B	1eyh_A	1q0r_A	1fit_	1qhv_A	1f3z_	1qzn_A
1fmc_A	1q35_A	1fn9_A	1qj5_A	1f0l_B	1q0u_A	1fiu_A	1qjd_A	1f7d_A	1r2d_A
1fnc_	1q3f_A	1fo8_A	1qj8_A	1f94_A	1q16_C	1fjh_A	1ql3_A	1f8m_A	1r89_A
1fnl_A	1q40_C	1fsf_A	1qlw_A	1f9m_A	1q1a_A	1fkm_A	1qnr_A	1f9y_A	1rb9_
1fr3_A	1q4r_A	1fsg_A	1qq5_A	1fa8_A	1q6o_A	1frd_	1qop_A	1fas_	1rcf_
1ft5_A	1q4u_A	1ftr_A	1qqp_1	1fcq_A	1q7r_A	1frr_A	1qq9_A	1fj2_A	1rdg_
1fus_	1q5y_A	1fvk_A	1qv9_A	1fcy_A	1q8u_A	1g0c_A	1qsg_A	1fle_I	1rm8_A
1fux_A	1q5z_A	1fxl_A	1qwd_A	1fd3_A	1qb5_D	1g1s_A	1qwl_A	1fon_A	1rqb_A
1fvg_A	1q7f_A	1fxo_A	1qy5_A	1fe6_A	1qfl_A	1g2o_A	1qwr_A	1fqt_A	1rro_
1fvu_A	1q8b_A	1fye_A	1qy6_A	1fk5_A	1qgw_A	1g3k_A	1r0m_A	1fvu_B	1rv9_A
1fx2_A	1q8d_A	1fz1_A	1r03_A	1fkj_	1qh5_A	1g3m_A	1r1q_A	1fzy_A	1ry6_A
1fz1_E	1q8f_A	1fz1_C	1r0t_B	1flm_A	1qhd_A	1g4i_A	1r26_A	1g0s_A	1ryq_A
1g1b_A	1qav_B	1g2q_A	1r0u_A	1flt_X	1qip_A	1g5a_A	1r2q_A	1g1t_A	1s5e_B
1g2a_A	1qb0_A	1g2r_A	1r12_A	1fny_A	1qj4_A	1g6u_A	1r2r_A	1g3p_	1s5u_A

Continued

1g2b_A	1qcx_A	1g68_A	1r3q_A	1fs7_A	1qkr_A	1g72_B	1r4p_A	1g6s_A	1s69_A
1g57_A	1qf8_A	1g7f_A	1r4p_B	1fsj_B	1qnt_A	1g8q_A	1r4v_A	1g87_A	1sbx_A
1g60_A	1qfm_A	1ga6_A	1r55_A	1fua_	1qnx_A	1gca_	1r7a_A	1g8s_A	1sgv_A
1g61_A	1qhq_A	1gd1_O	1r5l_A	1fxd_	1qo2_A	1gco_A	1r8o_A	1g9o_A	1slt_A
1g72_A	1qk8_A	1ggx_A	1r66_A	1g5t_A	1qoz_A	1gg6_C	1ra9_	1gar_A	1snc_
1g8k_B	1qks_A	1ghe_A	1r75_A	1g66_A	1qqp_4	1gl2_C	1rcy_	1gde_A	1spg_B
1g9z_A	1ql0_A	1gk9_A	1r7o_A	1g8a_A	1qw2_A	1gmx_A	1rh4_	1geg_A	1sqe_A
1gai_	1qmv_A	1gkk_A	1r88_A	1g9g_A	1qwz_A	1gnd_	1rj9_A	1gg6_B	1sqs_A
1gci_	1qop_B	1gkp_A	1r8o_B	1gad_O	1qxy_A	1gnu_A	1rqj_A	1gk6_A	1srv_A
1gcq_A	1qtn_A	1gl2_D	1r9l_A	1gl2_A	1qzm_A	1gp0_A	1rsy_	1gk7_A	1ss4_A
1gcq_C	1que_	1glq_A	1rec_	1gl2_B	1r0k_A	1gpe_A	1ru4_A	1gk9_B	1svi_A
1gd0_A	1qwk_A	1gp6_A	1rfs_	1gmy_A	1r2m_A	1gu2_A	1rvk_A	1gny_A	1sxr_A
1gdv_A	1qwo_A	1gqi_A	1rj1_A	1gnl_A	1r4u_A	1gvj_A	1rw7_A	1gpq_A	1t0a_A
1ghx_L	1qy1_A	1gqz_A	1rkq_A	1gq6_A	1r5r_A	1gvz_A	1rwi_A	1gqa_A	1t2d_A
1gk8_A	1qz5_A	1gs5_A	1rku_A	1gq8_A	1r6j_A	1gxy_A	1rwj_A	1gqn_A	1t3q_A
1gk8_I	1r0r_I	1gsi_A	1rli_A	1gu1_A	1r6x_A	1gz2_A	1rxq_A	1gt1_A	1t92_A
1gpr_	1r4x_A	1gsk_A	1rop_A	1guu_A	1r85_A	1gz7_A	1ry9_A	1gtk_A	1ta3_B
1gqv_A	1r8h_A	1gt9_1	1rqw_A	1gvk_B	1r9c_A	1h03_P	1s0p_A	1gu7_A	1tc5_A
1gud_A	1rdo_1	1gtf_A	1rrm_A	1gwe_A	1rcq_A	1h0b_A	1sff_A	1gve_A	1tgx_A
1gut_A	1rds_	1gui_A	1rwr_A	1gxm_A	1rew_A	1h1a_A	1sfp_	1gvf_A	1tif_
1gv5_A	1reg_X	1gv9_A	1rxj_A	1gyu_A	1rg8_A	1h1y_A	1sgw_A	1gx5_A	1tjy_A
1gvn_A	1rew_C	1gwm_A	1ryh_A	1gyx_A	1rgx_A	1h2c_A	1sh8_A	1gzc_A	1tlu_A
1gvn_B	1rhc_A	1gyg_A	1s3e_A	1gz8_A	1roc_A	1h46_X	1sj1_A	1h0h_L	1tmy_
1gvp_	1rhs_	1gzt_A	1s9u_A	1h12_A	1rqp_A	1h6t_A	1sjd_A	1h2b_A	1tqj_A
1gvt_A	1rie_	1h05_A	1sf9_A	1h32_A	1rtq_A	1h72_C	1sml_A	1h4x_A	1tu6_A
1gwi_A	1rjd_A	1h1n_A	1sfs_A	1h4g_A	1rw1_A	1hc9_A	1snr_A	1h8d_H	1tw9_A
1gx1_A	1rk4_A	1h32_B	1sfx_A	1h4r_A	1rwy_A	1hdi_A	1sr7_A	1h9m_A	1u00_A
1gxu_A	1rl9_A	1hd2_A	1sjw_A	1h5q_A	1rxy_A	1hdo_A	1su7_A	1hcz_	1u11_A
1gyo_A	1rlh_A	1hdh_A	1sk4_A	1h6f_A	1s1d_A	1hfc_	1swx_A	1hfx_	1u4g_A
1h75_A	1rq2_A	1hfe_L	1sl5_A	1h6h_A	1s1p_A	1hfe_S	1sy7_A	1hh5_A	1u69_A
1h8p_A	1rtu_	1hj8_A	1spg_A	1h6l_A	1s4k_A	1hg8_A	1t0b_A	1hkk_A	1ual_A
1heu_A	1rx0_A	1hlq_A	1sq2_N	1h7e_A	1s5a_A	1hgx_A	1t3i_A	1hpg_A	1ub3_A
1hk0_X	1rya_A	1hm9_A	1ssx_A	1h7m_A	1s8i_A	1hh8_A	1t56_A	1hqs_A	1ubk_S
1hnj_A	1rz3_A	1hmt_	1svy_	1h97_A	1s8n_A	1hle_A	1t5o_A	1htw_A	1ucd_A
1hoz_A	1rzf_L	1hq1_A	1szn_A	1h98_A	1s95_A	1hpi_	1t61_C	1hx0_A	1ucs_A
1hs6_A	1s68_A	1hsl_A	1t06_A	1hcb_	1s9r_A	1hpm_	1t64_A	1hye_A	1ud9_A
1ht6_A	1s7z_A	1htr_B	1t1v_A	1hdk_A	1sbp_	1htr_P	1t6c_A	1i0d_A	1udc_

Continued

1hx6_A	1sdi_A	1hvy_A	1t46_A	1hfu_A	1shk_A	1hxh_A	1t6g_C	1i1n_A	1udz_A
1hyo_A	1sds_A	1hw1_A	1t4b_A	1hj9_A	1sii_A	1hy7_A	1t8h_A	1i2s_A	1ugp_A
1hz4_A	1sdw_A	1hxn_	1t5b_A	1hjx_A	1smo_A	1i24_A	1ta3_A	1i3c_A	1ukf_A
1i07_A	1sen_A	1hxp_A	1t82_A	1hp1_A	1sox_A	1i2t_A	1tc1_A	1i40_A	1ukz_
1i19_A	1sft_A	1hyp_	1t8k_A	1hw6_A	1st9_A	1i4u_A	1te2_A	1i8a_A	1umk_A
1i6l_A	1shu_X	1hzt_A	1t9h_A	1hxi_A	1su2_A	1i6n_A	1tgs_I	1i8o_A	1uoy_A
1i71_A	1six_A	1i0r_A	1tbb_A	1i1j_A	1svb_	1i76_A	1thf_D	1i9g_A	1upi_A
1i7q_A	1sjv_A	1i2m_A	1thm_	1i1w_A	1syy_A	1iae_	1tiq_A	1iaz_A	1usp_A
1iej_A	1smb_A	1id0_A	1tib_	1i2k_A	1t07_A	1iat_A	1tks_A	1idr_A	1uuf_A
1ikq_A	1smx_A	1ida_A	1tjo_A	1i7b_B	1t0t_V	1icx_A	1to4_A	1ijv_A	1uv7_A
1iq6_A	1snb_	1ie0_A	1toa_A	1i7q_B	1t15_A	1igq_A	1tr0_A	1im5_A	1uvq_A
1iro_	1sr8_A	1ifc_	1tu1_A	1i9s_A	1t2a_A	1ikt_A	1tua_A	1ioo_A	1uw4_A
1is3_A	1t0h_A	1ig5_A	1tu9_A	1iby_A	1t3q_C	1iom_A	1tzb_A	1iqq_A	1v02_A
1isp_A	1t1j_A	1ihg_A	1tuw_A	1icr_A	1t9m_A	1ird_A	1tzv_A	1iqz_A	1v7z_A
1iue_A	1t61_A	1ihr_A	1tzp_A	1ihb_A	1tbf_A	1isu_A	1tzy_C	1is9_A	1v9f_A
1iwd_A	1t6t_1	1inl_A	1u02_A	1iit_A	1td4_A	1iua_A	1u14_A	1iu8_A	1vca_A
1iy8_A	1t7r_A	1ist_A	1u0k_A	1ijq_A	1tdz_A	1iuz_	1u60_A	1iuq_A	1vdw_A
1iyn_A	1t8t_A	1iye_A	1u9c_A	1io0_A	1ten_	1ix1_A	1uai_A	1ixk_A	1vfa_B
1j0o_A	1tad_A	1iyh_A	1u9d_A	1ir1_S	1tfe_	1ix2_A	1ueh_A	1iys_A	1vg8_A
1j1u_A	1tca_	1j1q_A	1uar_A	1ird_B	1thg_	1ixl_A	1ufb_A	1j09_A	1vh5_A
1j31_A	1ton_	1j3v_A	1uas_A	1isc_A	1thx_	1j05_A	1ufo_A	1j1y_A	1vhf_A
1j34_A	1tqg_A	1j58_A	1ufy_A	1itw_A	1tk4_A	1j05_B	1ugp_B	1j2l_A	1vhu_A
1j8b_A	1tqh_A	1j5w_A	1ug6_A	1itx_A	1tml_	1j1n_A	1ui0_A	1jay_A	1vi6_A
1j8u_A	1tvd_A	1j7g_A	1uhe_A	1iv8_A	1tp6_A	1j27_A	1ujn_A	1jd0_A	1vi9_A
1j9l_A	1tx4_B	1jd1_A	1umd_B	1iw0_A	1ts9_A	1j30_A	1ukk_A	1jdh_B	1via_A
1ja9_A	1tzy_B	1jev_A	1umg_A	1j0h_A	1tuh_A	1j34_B	1uku_A	1jfb_A	1vio_A
1jb9_A	1tzy_D	1jjf_A	1umn_A	1j3a_A	1twd_A	1j34_C	1ukv_G	1jiw_I	1vje_A
1jer_	1u1w_A	1jjt_A	1umz_A	1j54_A	1tx2_A	1j3w_A	1ukv_Y	1jln_A	1vk1_A
1jg1_A	1u6d_X	1jkm_A	1unn_C	1j71_A	1tx4_A	1j48_A	1ulr_A	1jlt_A	1vkb_A
1jh6_A	1u7i_A	1jm0_A	1urd_A	1j79_A	1tzy_A	1j6o_A	1umv_X	1jo0_A	1vla_A
1jif_A	1ua4_A	1jmx_G	1us5_A	1j97_A	1u7p_A	1j7x_A	1unq_A	1jpz_A	1vlj_A
1jkv_A	1uay_A	1jov_A	1usg_A	1j9b_A	1uc7_A	1j83_A	1uqr_A	1jtv_A	1vps_A
1jmx_A	1ubi_	1jqe_A	1uv0_A	1jak_A	1ucr_A	1j8q_A	1uqx_A	1juh_A	1vsd_
1jni_A	1ubk_L	1jr8_A	1uvq_B	1jb3_A	1uek_A	1jae_	1urr_A	1jwq_A	1vsr_A
1jp4_A	1ugx_A	1ju3_A	1uxz_A	1jbo_B	1uj2_A	1jat_A	1uw1_A	1jy3_N	1w2p_A
1jr2_A	1uha_A	1jue_A	1uy4_A	1jbw_A	1uow_A	1jat_B	1uwv_A	1k12_A	1w2w_B
1jug_	1uj6_A	1jw9_B	1v0l_A	1jfu_A	1upg_A	1jbe_A	1uww_A	1k27_A	1w3i_A

Continued

1jy3_O	1ukj_A	1jy3_P	1v1h_A	1jfx_A	1upq_A	1jbm_A	1uyl_A	1k2x_A	1w3o_A
1jyh_A	1umd_A	1k1b_A	1v3e_A	1jhd_A	1usc_A	1jbo_A	1uyp_A	1k3i_A	1wfa_A
1k5c_A	1unk_A	1k20_A	1v6h_A	1jig_A	1uti_A	1jdl_A	1uzi_A	1k3y_A	1wkr_A
1k5n_B	1up9_A	1k3x_A	1v6s_A	1jix_A	1uu3_A	1jdw_	1v5v_A	1k4m_A	1wmu_A
1k6d_A	1us0_A	1k4n_A	1v84_A	1jk7_A	1uux_A	1jfl_A	1v8f_A	1k55_A	1wtl_A
1k77_A	1usl_A	1k5n_A	1va4_A	1jke_A	1uw4_B	1jfr_A	1v9m_A	1kdi_	1xc1_A
1k9u_A	1utx_A	1k6u_A	1vap_A	1jlj_A	1v4v_A	1ji1_A	1vc4_A	1kex_A	1xfo_A
1kap_P	1uuj_A	1k6w_A	1vcl_A	1jlt_B	1v58_A	1ji7_A	1vfr_A	1kf3_A	1xfp_A
1kl9_A	1uuq_A	1k6x_A	1vhe_A	1jmx_B	1v8a_A	1jks_A	1vfs_A	1kfw_A	1xm8_A
1klu_B	1uwc_A	1ka1_A	1vhn_A	1jnr_A	1vd5_A	1jm1_A	1vhh_	1kgd_A	1xnb_
1klx_A	1uwu_B	1kao_	1vhw_A	1jo8_A	1vgg_A	1jnd_A	1vim_A	1kgn_A	1xuu_A
1kol_A	1uxx_X	1kcq_A	1vjo_A	1jr0_D	1vie_	1jnr_B	1vju_A	1kko_A	1yai_A
1kpf_	1uzb_A	1kdk_A	1vk4_A	1jsf_	1vjf_A	1jps_T	1vjw_	1kl1_A	1yal_
1kpt_A	1v2d_A	1kew_A	1vk5_A	1ju2_A	1vk2_A	1jvb_A	1vkf_A	1kli_H	1yna_
1kq3_A	1v3w_A	1kgs_A	1vk8_A	1jv1_A	1vki_A	1jzt_A	1vkn_A	1klu_A	2a0b_
1kq6_A	1v4x_A	1kjq_A	1vke_A	1jvw_A	1vl4_A	1k07_A	1vl1_A	1kmv_A	2abh_
1kqw_A	1v4x_B	1kli_L	1vkh_A	1jyr_A	1vlp_A	1k1e_A	1vlr_B	1ko3_A	2arc_A
1ks8_A	1v7r_A	1knm_A	1vkm_A	1k2x_B	1vma_A	1k38_A	1vlt_A	1koe_	2bbk_H
1ktg_A	1v93_A	1kok_A	1vl5_A	1k4i_A	1vmb_A	1k7c_A	1vp8_A	1kp6_A	2end_
1kth_A	1v9y_A	1kqp_A	1vl7_A	1k75_A	1vmf_A	1k7h_A	1vyb_A	1kqf_B	2fbj_H
1kve_A	1viy_A	1kw3_B	1vly_A	1k7j_A	1vmh_A	1k8u_A	1vyi_A	1kr7_A	2fdn_
1kve_B	1viz_A	1kwg_A	1vm0_A	1k94_A	1vmj_A	1kcz_A	1vyr_A	1krn_	2hlc_A
1kxv_C	1vjl_A	1kyf_A	1vme_A	1khc_A	1vp6_A	1kdg_A	1w53_A	1kt7_A	2igd_
1ky2_A	1vjp_A	1kzf_A	1vmg_A	1khq_A	1w15_A	1kg2_A	1wab_	1kv0_A	2mbr_
1l1d_A	1vkk_A	1kzq_A	1vns_	1kmz_A	1w2w_A	1khi_A	1whi_	1kwm_A	2nll_B
1l6x_A	1vkp_A	1l3k_A	1vp2_A	1kng_A	1w8m_A	1kqr_A	1wmz_A	1l3p_A	2por_
1l9x_A	1vlc_A	1l6p_A	1vyf_A	1kop_A	1wd3_A	1l6x_B	1wri_A	1l6w_A	2pth_
1lid_	1vp4_A	1l9f_A	1wdn_A	1kqf_C	1wp5_A	1l9l_A	1x82_A	1l7a_A	2sic_I
1lj5_A	1vpd_A	1lam_	1wmd_A	1kso_A	1wr8_A	1lit_	1x8q_A	1lau_E	3cao_A
1lkp_A	1w0h_A	1lb6_A	1woq_A	1kug_A	1x7v_A	1lk3_L	1xa1_A	1ld8_B	3daa_A
1lm6_A	1w0n_A	1lbu_	1xau_A	1kv7_A	1xi3_A	1lk5_A	1xff_A	1le6_A	3hts_B
1lni_A	1w2f_A	1ld8_A	1xc2_A	1kv9_A	1xsz_A	1lkf_A	1xhd_A	1lg7_A	3pcc_M
1lop_A	1w2i_A	1ldt_L	1xdz_A	1l2i_A	1ycc_	1lm5_A	1xmt_A	1lkk_A	3rp2_A
1lqv_A	1w2y_A	1lf2_A	1xfj_A	1l5o_A	1ygh_A	1lo6_A	1xsv_A	1lm4_A	3tgl_
1lqy_A	1wad_	1lfa_A	1xfk_A	1l6r_A	1ytb_A	1lq9_A	1xyp_A	1lm8_C	3ukd_
1lr0_A	1wba_	1lj8_A	1xyz_A	1l6s_A	2bc2_A	1lqt_B	1yac_A	1lmb_3	451c_
1lri_A	1who_	1lk9_A	256b_A	1lc5_A	2bop_A	1lqv_C	1zin_	1lo7_A	4fiv_

Continued

1lst_	1wkq_A	1lkd_A	2aza_A	1lfw_A	2ccy_A	1ls6_A	1zpd_A	1lp8_A	6rxn_
1lua_A	1wlg_A	1ll2_A	2bbk_L	1lk2_B	2cxb_A	1ls9_A	2act_	1lqa_A	7taa_
1lug_A	1wms_A	1llp_	2gep_	1lln_A	2gdm_	1ltm_	2ae2_A	1lr7_A	7tim_A
1lwb_A	1wmu_B	1lm8_V	2ltn_A	1lm8_B	2hvm_	1lvw_A	2ak3_A	1lsh_B	830c_A
1m07_A	1x6o_A	1lmq_	2sil_	1lmi_A	2mnr_	1lyx_A	2apr_	1lts_C	
1m0s_A	1xg5_A	1ltz_A	2tnf_A	1loq_A	2msb_A	1lzj_A	2cmd_	1luc_A	
1m1f_A	1xgs_A	1luc_B	2tps_A	1ls1_A	2plt_	1m5w_A	2cpg_A	1lv7_A	
1m1q_A	1xso_A	1m0w_A	2utg_A	1lu4_A	2rhe_	1m70_A	2dri_	1m2r_A	
1m2a_A	1znb_A	1m1n_A	2vhb_A	1lwd_A	2rmc_A	1m7g_A	2erl_	1m2x_A