A Computational Efficient Algorithm for Protein Sequence Classification

Yiming Li and Hsiao-Mei Lu[1]

National Nano Device Laboratories
Microelectronics and Information Systems Research Center, National Chiao Tung University
P.O. Box 25-178, Hsinchu 300, Taiwan, ymli@faculty.nctu.edu.tw
[1]Department of Bioengineering, University of Illinois at Chicago, Chicago, USA, hlu7@uic.edu

ABSTRACT

In this paper we present statistical algorithms to classify the stability of proteins by their sequence. A protein sequence consists of successive amino acid codes and can be considered as multivariate categorical data. Based on the statistical variance analysis for data set in each group (stable or unstable protein), the weights are calculated and become an important clue for the effects of the combination of amino acids codes on protein stability. Once the weights for every combination of amino acid codes have been decided, we can assign each protein a score presenting its stability. The distribution of the score for a stable protein is different from the score of an unstable protein. Our algorithm is well suit in the protein stability analysis by its sequence. We propose weighting algorithms and compare them as the results of protein stability classification. It provides an alternative for the protein stability classification and a predictable result as the reference before the protein mutation.

Keywords: Protein stability; Classification of protein sequence; Prediction model; Statistical analysis; Computational statistics.

1 INTRODUCTION

The protein sequence analysis has been of great interests recently [1-5]. Proteins are large, organic molecules and are among the most important components in the cells of living organisms. They are more diverse in structure and function than any other kind of molecule. Proteins are known to degrade rapidly when conformations are altered due to abnormality in the sequences. Normal cellular proteins also display a wide range of half-life - turnover rates of individual proteins can differ as much as 1000-fold. Sequence specific properties, global features and the location of a protein in the cell are found to be important in deciding in the intracellular stability of a protein. In order to identify sequence dependent properties, it is necessary to analyze the stable and less stable protein sequences. In addition, the observation of the dipeptide composition in stable proteins is also an important thing in developing a theoretical method to predict protein stability from its amino acid sequence information. Through many on-line protein data banks [6-8], a protein sequence is easy obtained and ready analyzed for various studies and applications. Conventional computational intelligence algorithms, such as neural network and genetic algorithm have been proposed and applied for the problem of protein classification; unfortunately, these methods need many empirical parameters and cannot have reasonable predictions for various cases [9,10].

In this paper, we present efficient statistical algorithms to classify the stability of proteins based on their sequence. A protein sequence consists of successive amino acid codes and can be considered as multivariate categorical data. Based on the statistical variance analysis for data set in each group, stable or unstable proteins, the weights are calculated and become an important clue for the effects of the combination of amino acids codes on protein stability. Once the weights for every combination of two successive amino acid residues have been decided, we can assign each protein a score which is from a function of the sequence presenting its stability. The distribution of the score for a stable protein is different from the score of an unstable protein. In testing various protein problems, the proposed approach is well suit in large scale protein stability analysis by its sequence. The stability of a protein is defined by its half life time, or the rate of turnover, the time it takes a protein to degenerate as a random shape in vivo. Generally, the definition of a stable protein is when its half life time is greater than 5 hours, and for the unstable protein that is less than 2 hours [9,10].

There are many factors may affect a protein's stability [9-12]. For example, the simulation features of proteins proposed to increase rates of turnover include: (1) global properties: size, charge, hydrophobicity, thermal instability, flexibility, proteolytic susceptibility; (2) sequence-specific parameters: PEST sequence, RNase pentapeptide, a-Amino terminus, Asn, His, Cys, and Met oxidation; (3) location: Assembled or unassembled, bound/diffusible. This problem was studied by the PEST and Instability Index (*II*) method [12,13]. Our method is based on the *II* score, which has a good reliability, and also considers the effect of the α-Amino terminus which is one of the most important clues of protein's instability from the sequence [11]. The prediction results show good improvement by using our weighted algorithm with the α-Amino terminus consideration.

This paper is organized as follows. Section 2 states the simulation methodology. Section 3 shows the result and discussion. Conclusion and suggestion for future works are in the section 4.

2 SIMULATION METHODOLOGY

Our computation methodology is mainly based *II* score. As shown in Fig. 1, the statistical related computational procedure for the classification model problem includes some steps. First of all, we need to generate the *II* score look up table for each different protein dipeptide, and then we can calculate the instability index for any protein. Then we have to verify if it needs to do more adjustment according to its n-terminal residue. The details of the computational algorithm are discussed as follows.

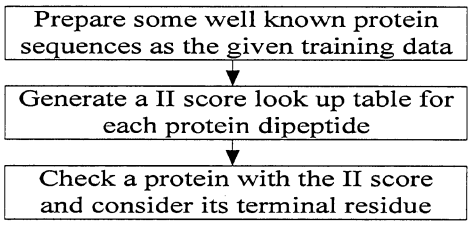

Figure 1: An illustration flow chart for the computational procedure.

In the first step, we need to classify protein sequences as two sets, the stable and unstable set. The expected value for every dipeptide in each set is calculated with:

$$N^e(a_i,a_{i+1}) = [N^0(a_i)N^0(a_{i+1})/T^2]\sum_{a_i}\sum_{a_{i+1}} N^0(a_i,a_{i+1}), \quad (1)$$

where $N^0(a_i)$ is the number of amino acid, a_i, occurred in the set and $N^0(a_i,a_{i+1})$ is the number of dipeptide, $a_i - a_{i+1}$, occurred in the set. T is total number of amino acid in the set.

Secondly, by using the equation 2 we calculate the chi-square value for every dipeptide in the set of the stable proteins and also in the set of unstable proteins. The another chi-square value represents the difference of these two sets as the 3rd set, and the average chi-square values, $\underline{\chi}^2$, in each set are computed according to Eqs. 3 and 4, respectively.

$$\chi^2(a_i,a_{i+1}) = [N^0(a_i,a_{i+1}) - N^e(a_i,a_{i+1})]^2 / N^0(a_i,a_{i+1}), \quad (2)$$

$$\chi^2_{s-u}(a_i,a_{i+1}) = [N^0_s(a_i,a_{i+1}) - N^e_s(a_i,a_{i+1})]^2 / N^0_s(a_i,a_{i+1}), \quad (3)$$

$$\underline{\chi}^2(a_i,a_{i+1}) = [\sum_{a_i}\sum_{a_{i+1}} \chi^2(a_i,a_{i+1})]/400. \quad (4)$$

From these average chi-square values, some qualified dipeptides are selected if their chi-square value is greater than the average chi-square value.

$$\chi^2(a_i,a_{i+1}) \geq \underline{\chi}^2(a_i,a_{i+1}) \quad (5)$$

In the third step, the occurrence probability P for each selected dipeptide in each set is calculated respectively by the following equation:

$$P(a_i,a_{i+1}) = N^0(a_i,a_{i+1})/N^e(a_i,a_{i+1}). \quad (6)$$

According to P, a dipeptide may satisfy only one or some of the 7 conditions shown in Tab. 1. The Impact Factor (*IF*) and its adjusted Impact Factor (*IF**) 1 to 7 for each condition in Tab. 1 can be calculated by the following two equations 7 and 8, respectively:

$$IF = N^0_s(a_i,a_{i+1})/N^0_{us}(a_i,a_{i+1}), \quad (7)$$

$$IF^* = 2 + \frac{IF}{|\min\{IF\}|}. \quad (8)$$

Subsets	Conditions
1	$P_u(a_i,a_{i+1}) \geq 1.5$ and $P_s(a_i,a_{i+1}) < 1.5$
2	$P_s(a_i,a_{i+1}) \geq 1.5$ and $P_u(a_i,a_{i+1}) < 1.5$
3	$P_u(a_i,a_{i+1}) \leq 0.64$ and $P_s(a_i,a_{i+1}) > 0.64$
4	$P_s(a_i,a_{i+1}) \leq 0.64$ and $P_u(a_i,a_{i+1}) > 0.64$
5	$P_{s-u}(a_i,a_{i+1}) \geq 1.5$ and
6	$P_{s-u}(a_i,a_{i+1}) < 0.64$
7	otherwise

Table 1: The conditions of seven subsets for P probability in each amino acid dipeptide.

Therefore, the Dipeptide Impact Factor (*DIF*) are calculated by summing the adjusted Impact Factor (*IF**) from 1 to 7 for each didpeptide with Eq. 9:

$$DIF(a_i,a_{i+1}) = \sum_{x=1\sim7} IF^*_x(a_i,a_{i+1}). \quad (9)$$

We have finished the calculation of the *II* score look up table for each dipeptide for the further calculation of *II* score of any protein. As shown in Fig. 2, it is the summary of the procedure of the computation for the II score look up table. The outline of the algorithm is: (1) Calculate chi-square values for the set of stable proteins, unstable proteins, and the difference between stable and unstable proteins; (2) Calculate average chi-square values for these 3 sets respectively; (3) Select peptides which qualified in their sets by the average chi-square values respectively; (4) Calculate the occurrence probability of each set and use it

to classify every peptide into 7 groups; (5) Calculate the impact factors for the seven groups and its adjustment; and (6) Calculate the dipeptide impact factor for the computation of the instability index score of any protein.

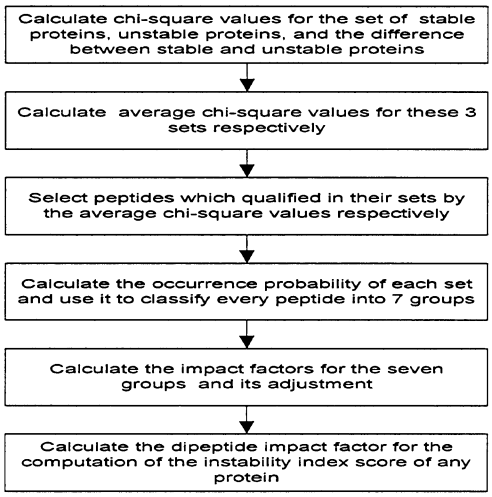

Figure 2: The flow chart of the computational procedure of the II score look up table.

Finally, we can obtain the *II* score as the summation of the *DIF*s of a protein sequence and adjust it as the adjusted Instability Index (*II**) by:

$$II = \sum_{i=1}^{L-1} DIF\,(a_i, a_{i+1}), \tag{10}$$

$$AII = \omega II \delta(a_n) + II, \tag{11}$$

where L is the length of the protein sequence, n is the length of the protein; ω and $\delta(a_n)$ are given as follows:

$$\omega = \frac{1}{\mathrm{var}(IF(a_i, a_{i+1}))}, \tag{12}$$

$$\delta(a_n) = \begin{cases} 0, \text{if } a_n \text{ is Met/Ser/Ala/Thr/Val/Gly} \\ 1, \text{otherwise} \end{cases}. \tag{13}$$

3 RESULTS AND DISCUSSION

We have presented efficient statistical algorithms above to classify the stability of proteins based on their sequence. As shown in Fig. 3, it is the result of the Instability Index (II) by testing 42 proteins without considering the effect of the α-Amino terminus contributed to proteins' stability (*algorithm* 1) [3]. As we can see, the classification result is acceptable; however, there are still some known stable and unstable proteins mixed near the decision line.

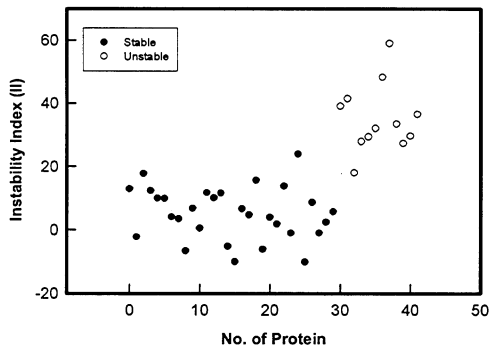

Figure 3: The II calculated without considering the effect of α-amino acid terminus.

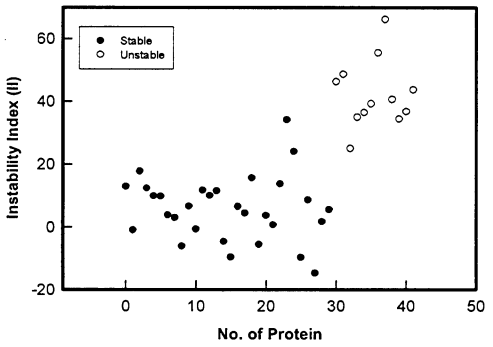

Figure 4: The II* calculated with adaptive method.

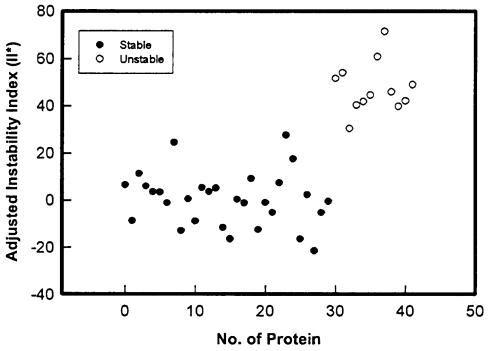

Figure 5: The II* calculated with considering the effect of α-amino acid terminus.

Figs. 4 and 5 are the results by using the adaptive method (*algorithm* 2) and considering the factor of α-amino acid terminus (*algorithm* 3), respectively. The stability of two classes which is calculated by II* is much easier to

identify than the one by the *algorithm* 1. In order to compare the reliability among these three algorithms, we calculate the average prediction rate using the leave-one-out method, leave one out of our testing data and then predict its stability. The comparison is shown in Table 2 below. The average prediction rate shows that our proposed *algorithms* 2 and 3 have better classification than that of the *algorithm* 1. Furthermore, to demonstrate the efficiency of the proposed *algorithms* 2 and 3, we briefly present, as shown in Figs. 6 and 7, some results of more testing protein sequences by using the adaptive method and considering the factor of α-amino acid terminus, respectively.

Algorithm	Average prediction rate (%)
1	74
2	78
3	89

Table 2: The comparison of the average prediction rate between the algorithms (1) without considering the effect of α-amino acid terminus, (2) with adaptive weighted, and (3) with considering the effect of α-amino acid terminus.

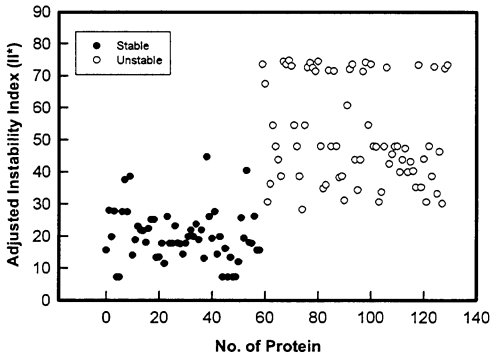

Figure 6: The II* calculated by the adaptive method.

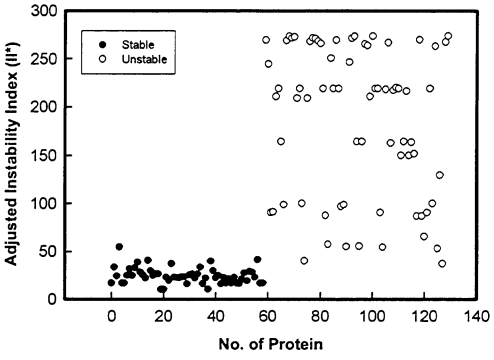

Figure 7: The result of II* calculation with considering the α–amino terminus.

4 CONCLUSIONS

In conclusion, we have proposed a statistical approach to study the stability of proteins classification problem. An adaptive approach has been developed to calculate the adjusted Instability Index by considering the effect of α-amino terminus for each protein. Our computational result has good reliability and significant difference between stable and unstable protein, and hence the problem is easier to identify. The adjusted instability index (II*) can be used as a prediction of a protein's stability before some experiments when it may be altered due to mutations. It provides an alternative for the protein stability classification and a predictable result as the reference before the protein mutation.

There are some future works which can be studied to improve the prediction of protein stability, such as add more realistic term related to this problem, build an on-line complete data and software for public search and feedback, and apply this algorithm to other classification problem's study.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Council of Taiwan under contract number NSC 91-2112-M-317-001 and the 2002 Research Fellowship Award of the Pan Wen-Yuan Foundation in Taiwan.

REFERENCES

[1] Kazufumi Takano and Katsuhide Yutani, Protein Eng. 14, 525-528, 2001.

[2] Francesca Trejo, Josep Ll. Gelpi, Albert Ferrer, Albert Boronat, Montserrat Busquets and Antoni Cortes, Protein Eng. 14, 911-917, 2001.

[3] K. Guruprasad, BV Reddy, and MW PAndit, Protein Eng. 4, 155-161, 1990.

[4] Yoshihiro Sambongi, Susunu Vchiyama, Yuji Kobayashi, Yasuo Igarashi, and Yun Hasegawa, Euro. J. Biochem. 269, 3355-3361, 2002.

[5] Jannic Boehm, Yansheng He, Axel Greiner, Louis Staudt, and Thomas Wirth, EMBO J. 20, 4153-4162, 2001.

[6] http://www.rcsb.org/pdb/

[7] http://www.ncbi.nlm.nih.gov/

[8] http://pir.georgetown.edu/

[9] Andreas Bachmair, Daniel Finley, and Alexander Varshavsky, Science 234, 179, 1986.

[10] Martin Rechsteiner, Scott Rogers, and Kavin Rote, TIBS **12**, 390, 1987.

[11] Li Xiao and Barry Honing, JMB 289, 5, 1435-1444, 1999.

[12] Scott Rogers, Rodeny Wells, and Martin Rechsteiner, Science 234, 364, 1986.

[13] J. Fred Dice, and Alfred L. Goldberg, Archives Biochem. Biophys. 170, 213-219, 1975.