行政院國家科學委員會專題研究計畫　成果報告

用序列與結構排列法，預測蛋白質功能與交互作用。

計畫類別：個別型計畫
計畫編號：NSC94-2218-E-216-007-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：中華大學生物資訊學系

計畫主持人：許文龍

計畫參與人員：許文龍，吳信宏，劉智偉。

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 31 日

行政院國家科學委員會補助專題研究計畫 **■ 成 果 報 告**
□期中進度報告

用序列與結構排列法，預測蛋白質功能與交互作用

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC-94-2218-E-216-007-
執行期間：2005 年 8 月 1 日 至 2006 年 7 月 31 日

計畫主持人：許文龍
共同主持人：
計畫參與人員：許文龍，吳信宏，劉智偉。

成果報告類型(依經費核定清單規定繳交):■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：中華大學

中　華　民　國　95　年　10　月　31　日

計畫中文摘要：

關鍵詞：蛋白質功能與交互作用，HMM 模型，序列與結構排列。

蛋白質體學被認爲是後基因時代非常重要的領域，然而在生物實驗室中，蛋白質比基因更難處理，生物資訊研究者的任務在分析已知蛋白質資料，用演算法及智識庫的技術，來預測未知蛋白質的特性，這個研究成果可節省生物實驗室巨大的時間、人力、金錢。

在中華大學生物資訊實驗室中，我們已成功地完成 2D 及 3D 結構預測，在這個計劃，我們利用自創的預測 2D 結構的 HMM 模型，來預測蛋白質的功能及交互作用〈有關 4D 結構〉。這個模型可以產生整體的機率，此機率結合氨基酸碼出現在 2D 元件機率，也結合下一氨基酸碼之轉換機率。這種方法稱爲序列及結構排列法。

蛋白質被分成十二種功能，我們的序列及結構排列法，可用來預測蛋白質功能，也可預測未知蛋白質位於表面的子序列，是否會與交互蛋白質智識庫中之子序列產生交互作用，然後再用智識庫統計之能量，找出所有交互作用的子序列對，和子序列對的交互排列。

計畫英文摘要：

Keywords: Protein function and interaction, HMM, sequence-structure alignment.

Proteomics is considered to be a very important field in post-genome era. However protein is much difficult to handle than gene in biological laboratory. Bioinformatics researcher's responsibility is to analyze known protein data, and use algorithms and knowledge base technology to predict unknown protein's characteristics. This research result can save tremendous time, manpower and money for biological laboratory.

In the bioinformatics laboratory of Chung-Hua University, we have successfully developed our approaches to predict 2D and 3D protein structure. In this project, we are trying to predict protein function and interaction (relating 4D structure) basing on our HMM which is created to predict protein 2D structure. This model will generate an overall probability which combines not only the probability of amino acid code appearing in a particular 2D component, also their transition probability to next code. This method is called sequence-structure alignment approach.

All proteins are categorized as 12 functions and sequence-structure alignment approach is used to predict protein function. This approach can also be applied to predict if query protein's surface subsequence can be interactive with any subsequences in protein interacting library. Then knowledge base statistical energy is used to identify all interacting subsequence pairs and the alignment of interacting subsequence pair.

(三)報告內容：

1. Preface

Protein-protein interactions play an important role in predicting protein function. Identification of protein-protein interaction sites and detection of specific amino acid residues that participate in protein interactions is an important problem ranging from rational drug design to analysis of metabolic and signal transduction networks. Experimental proteomics projects have already resulted in complete 'interactomes'. While such efforts yield a catalog of interacting proteins, experimental detection of residues in protein-protein interaction surfaces must come from determination of the structure of protein-protein complexes. However, determination of protein-complex structures using X-ray and NMR methods lags far behind the number of known protein sequences. Hence, there is a need for the development of reliable computational methods for identifying protein-protein interaction residues.

Many cellular events involve the formation of protein–protein complexes. Elucidation of the structural details of these complexes will undoubtedly contribute to our understanding of their functional properties, and thus is a major goal of structural biology. However, only a small fraction of experimentally determined structures are of protein–protein complexes. Therefore, it is of substantial interest to develop computational docking methods that, given the structures of the individual component proteins, are able to assemble them into the complex in an accurate and reliable way.

Today, new genomes are constantly increasing. It is only possible to assign function to 40% of all protein sequences based on sequence similarity. There is a great need for protein function prediction methods. If a protein sequence is not similar to any other known protein, it is reasonable to expect that proteins with related function will have similar properties.

2. Research purpose

The main purpose of this research is to predict the interactions and functions of proteins. A HMM model which is related to 1D sequence and secondary states is built to predict protein interaction and protein function. The result of this research can be used to extend gene network databases. These databases can be further applied for medical diagnosis, drug discovery and biological researches.

3. Previous Efforts

Based on different characteristics of known protein-protein interaction sites, several methods have been proposed for predicting interface residues using a

combination of protein sequence and structural information. For example, based on their observation that proline residues occur frequently near interfaces, Kini and Evans [1] predicted potential protein-protein interaction sites by detecting the presence of ‘‘proline brackets.’’ Jones and Thornton [2][3] successfully predicted interfaces in a set of 59 structures using a scoring function based on six parameters: solvation potential, residues interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area. Gallet et al. [4] identified interacting residues using an analysis of sequence hydrophobicity based on a method previously developed by Eisenberg et al. [5] for detecting membrane and surface segments of proteins. Lu et al. [6] have developed statistical potentials for interfaces and used them in a structure-based multimeric threading algorithm to assign quaternary structures and predict protein interaction partners for proteins in the yeast genome.

Several groups have used neural networks to predict protein-protein interaction sites. Zhou and Shan [7] and Fariselli et al. [8] have independently used neural network algorithms to predict whether or not a residue is located in an interaction site using the spatial neighbors of the target residues as input, and achieved accuracy of 70% and 73%, respectively. Ofran and Rost [9] have successfully predicted protein-protein interaction sites using a neural network method based on their observations that the majority of protein-protein interaction residues are clustered on a sequence and that the protein-protein interfaces differ from the rest of the protein surface in residue composition.

There are a number of methods to predict protein function, such as: 1. Pairwise sequence similarity methods [10, 11, 12, 13, 14]. 2. Iterative search methods among multiple sequences [15,16]. 3. Super family database methods [17, 18]. 4. Alignment based methods using functional links [19, 20]. 5. Predicting protein function via structure [21, 22]. 6. Neural network using protein feature [23]. In [24], Jenson had developed 17 protein features from sequence or secondary structure. Many physical, chemical or biological characteristics are involved to compute these features. Then neural network is trained to recognize specific property of these features. Our HMM statistic model is relatively simple to gene ontology [25]. This model has considered the alignment corresponding relationships between sequence and secondary structure. For orphan protein which has similar property and dissimilar sequence, the propagation possibilities of HMM can recognize such property and provide better prediction.

4. Research Methods

Our research includes two parts: interaction and docking. We acquired all protein data files from Protein Data Bank (http://www.rcsb.org/pdb/Welcome.do). All

the files are about protein primary structures and second structures. And we download protein interaction dataset from Yan's website (http:// www.cs.iastate.edu /~yan330 /p-p /p- p.htm). We create a database to store these data records.

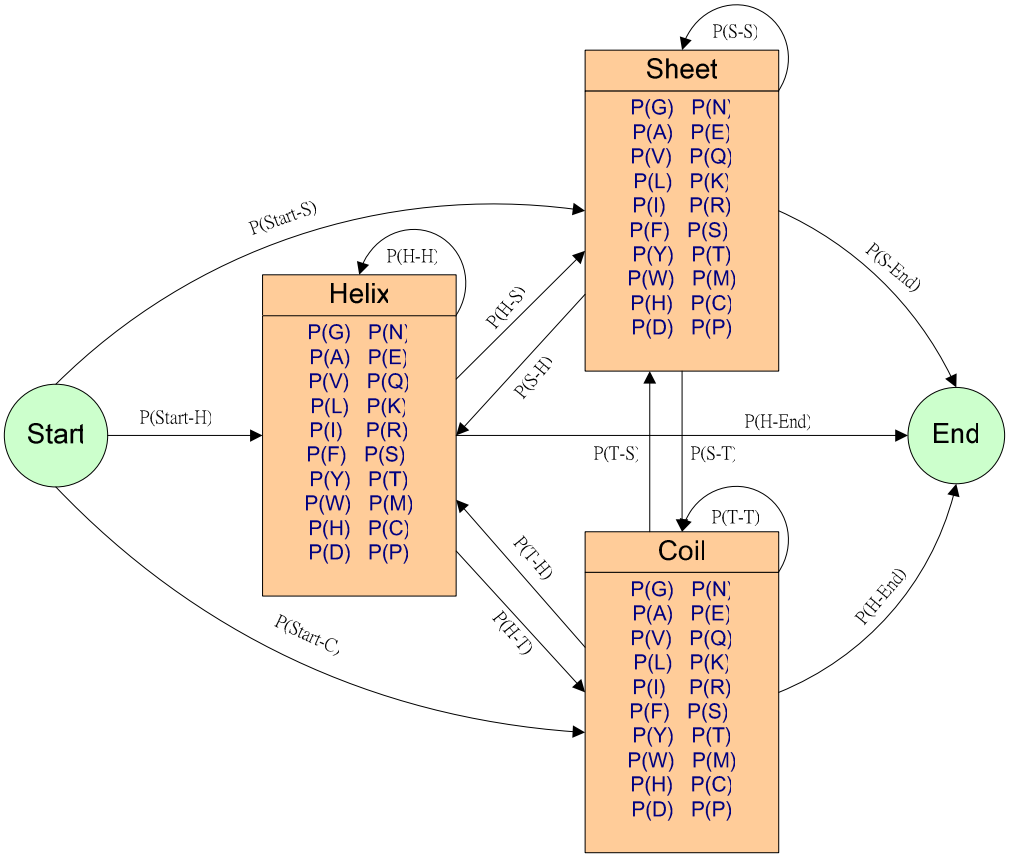

Figure 1. HMM Model of Our Approach.

We train our statistic model in Figure1 using the database. After training, we also test the statistic model. The statistic model achieved 71% accuracy (complicated model) and 80% accuracy (easy model). The statistic model can predict which residues participate in protein-protein interaction when you input a protein sequence.

The second step is about docking. After predicting protein interaction site, interactions between proteins become critical in biology. With the advent of genome and proteome projects, there is much interest in predicting the structures of protein-protein complexes. We use genetic algorithm to modify FTDock. Then we displace MultiDock by our statistic model and cooperate RPDock to be fitness function [26].

The Gene Ontology (GO) Annotation project at European Bioinformatics Institute has mapped identifiers of Protein Data Bank (PDB) into GO terms. Three organizing principles of GO are cellular component, biological process and molecular function. Table 1 collects 6362 proteins from Protein Data Bank. These proteins are

classified according to six categories of GO cellular component. These six categories are envelope, extracellular matrix, extracellular region, organelle, protein complex and virion. The protein quantity for each classification is also showing in the table.

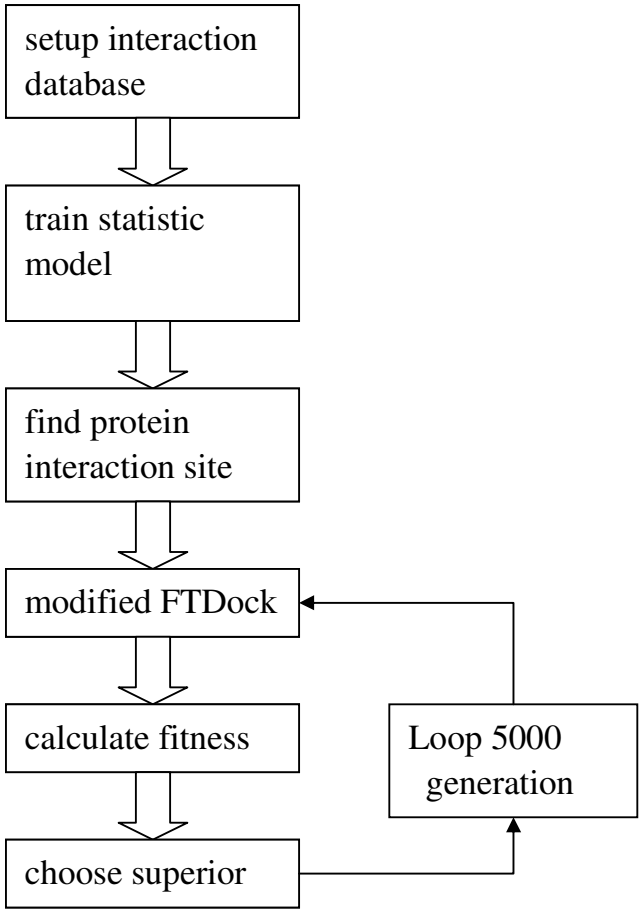

Figure 2. Genetic Algorithm for FTDock.

In this research [27, 28], a HMM predictor have been trained for a subset of Gene Ontology classes. Since protein function is closely related to both protein sequence and structure. To record a large amount of protein properties, the input data for this HMM are amino acid sequences aligning with the corresponding secondary structure sequences. This mathematical approach is practical and easy to compute. The accuracy of this method can reach to 62.97%.

5. Results and Conclusions:

There are 10 protein datasets (10% of 77 protein datasets) are used to test our statistic model. After training our model, we have all the logistic probability of every route and amino acid in a block. We input our test sets to test if our statistic model is good or not. First we load test datasets to our database like training datasets in the

same way. Every division of test sets is input our interaction model and non-interaction model. All kind of logistic probability (all route and amino acid in a block) that test sets match are summed. So we have input interaction test sets in our interaction model and non-interaction model. The statistic model achieved 71% accuracy (complicated model) and 80% accuracy (easy model).

We have developed a method to predict protein function. Our algorithm, a Hidden Markov Model, has been built to solve our problem with statistical concept. At present, many researchers are devoted to the study of this field. But several methods are too complex because many factors of the protein are considered. We utilize advantage of the computer to deal with a large amount of biological data quickly. The mathematical model is trained to get the probability. We can use our method to predict protein function after training. The relation between the protein function and its molecular structure is close, so the secondary structure is considered in our method. Finally, the average accuracy of our method is 62.97%. In the future, this method can be applied to the prediction of other functions of biochemistry such as Enzyme and Non-enzyme.

參考文獻

[1] Kini RM, Evans HJ (1996) Prediction of potential protein-protein interaction sites from amino acid sequence identification of a fibrin polymerization site. FEBS Lett 385:81-86

[2] Jones S, Thornton JM (1997a) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272:121-132

[3] Jones S, Thornton JM (1997b) Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272:133-143

[4] Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) A fast method to predict protein interaction sites from sequences. J Mol Biol 302:917-926

[5] Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 179:125-142

[6] Lu L, Lu H, Skolnick J (2003) Development of Unified Statistical Potentials describing Protein-protein interactions. Biophy J 84:1895–1901

[7] Zhou H, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44:336-343

[8] Fariselli P, Pazos F, Valencia A, Casadia R (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269:1356-1361

[9] Ofran Y, Rost B (2003b) Predicted protein-protein interaction sites from local sequence information. FEBS Lett 544:236-239

[10] Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. Prot. Eng., 7:1323-1332.

[11] Mott, R. (2001). Maximum likelihood estimation of the statistical distribution of smith

waterman local sequence similarity scores.

[12] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. Nature, 402:83-86, Bull. Math. Biol., 54:59-75.

[13] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. Proc. Natl. Acad. U.S.A., 85:2444-2448.

[14] Smith, T. F. and Waterman, M. S. (1981). Identi_cation of common molecular subsequences. J. Mol. Biol., 147:195-197.

[15] Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. (1998). EUCLID: automatic classi_cation of proteins in functional classes by their database annotations. Bioinformatics, 14:542-543.

[16] Bolten, E., Schliep, A., Schneckener, S., Schomburg, D., and Schrader, R. (2001). Clustering protein sequences structure prediction by transitive homology. Bioinformatics, 17:935-941.

[17] Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and A, B. (2002). The PROSITE database, its status in 2002. Nucleic Acids Res., 30:235-238.

[18] [13] Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res., 30:268-272.

[19] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. Science, 285:751-753.

[20] Yanai, I., Mellor, J. C., and DeLisi, C. (2001). Identifying functional links between genes using conserved chromosomal proximity. Trends in Biotechnology, 19:S61-S66.

[21] Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the

behavioural sciences. PhD thesis, Harvard University.

[22] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol., 307:1113-1143.

[23] Norin, M. and Sundstr□om, M. (2002). Structural proteomics: development in structure-to-function predictions. Trends in Biotechnology, 20:79-84.

[24] Lars Juhl Jensen (2002)，"Prediction of Protein Function from Sequence Derived Protein Features"，Center for Biological Sequence Analysis，BioCentrum-DTU，Technical University of Denmark，Lyngby 2002.

[25] Gene Ontology http://www.geneontology.org/

[26] Jyh-Wei Liou, "Predict Protein Interactions and Quauternar Structure", Master Dissertation (to appear).

[27] Wen-Lung Shu, Hsin-Hung Wu and Jyh-Wei Liou, "Predict Protein Functions Using Sequence-Structure Alignment Method." 民生電子暨信號處理研討會論文, Dec 16~17 2006.

[28] Hsin-Hung Wu, "Predict Gene Ontology Functions Using Sequence-Structure Alignment Method", Master Dissertation, CSIE Department, July 2006.

計畫成果自評：

蛋白質相互作用之預測及功能預料，達成之目標與原計畫相符，超出部份爲：產生基因演算法之 FFT 3D Dock 軟體，及未來可發展一個精確度八成以上的 gene ontology 預測軟體。基因演算法之 FFT 3D Dock 有學術或應用價值，是否申請專利應評估其是否能有效產生經濟價值而定。此種 docking 方法應研究如何應用在藥物設計才能發揮經濟價值，尤其是中草藥之研究。Gene ontology 預測軟體則有待進一步發展，用以預測 gene pathways。

可供推廣之研發成果資料表

□ 可申請專利　■ 可技術移轉　日期:95 年 10 月 31 日

國科會補助計畫	計畫名稱：用序列與結構排列法，預測蛋白質功能與交互作用。 計畫主持人： 計畫編號：94-2218-E-216-007-　學門領域：量子計算
技術/創作名稱	基因演算法之 FFT 3D Dock 軟體
發明人/創作人	許文龍
技術說明	中文：此一軟體可用來預測蛋白質四極結構，此軟體即 FTDock 是由英國癌症研究基金會提供之免費公開軟體修改而成，其精確度可由 12 度旋轉角度提昇至 1 度，基因演算法可有效降低大量增加之計算量，此研究所發展之蛋白質交互作用之數學模型，可當作基因演算法之 fitness function。
	英文：The technique can be used to predict protein quaternary Structure. This program “FTDock” is modified from the cancer foundation institute of united kingdom. The accurate of rotating angles can be increased from 12 degrees to 1 degree. The computing time can be greatly reduced by using genetic algorithm. The mathematical model developed from the protein interacting research can be used as fitness function of genetic algorithm.
可利用之產業 及 可開發之產品	用於中草藥研究，可應用來開發藥物。
技術特點	精確度可由 12 度旋轉角度提昇至 1 度，基因演算法可有效降低大量增加之計算量
推廣及運用的價值	藥物設計極需快速且精準之 3D dock 軟體，電腦輔助藥物設計的成功機會才可提昇。

附錄

(民生電子暨信號處理研討會論文, Dec 16~17 2006)

Predict Protein Functions Using Sequence-Structure Alignment Method

Wen-Lung Shu[1], Hsin-Hung Wu[2] and Jyh-Wei Liou[2]
Bioinformatics Department[1]
Department of Computer Science and Information Engineering[2]
Chung Hua University, Hsinchu 300, Taiwan
wlshu@chu.edu.tw

Abstract

Today, new genomes are constantly increasing. It is only possible to assign function to 40% of all protein sequences based on sequence similarity. There is a great need for protein function prediction methods. If a protein sequence is not similar to any other known protein, it is reasonable to expect that proteins with related function will have similar properties.

In this paper, a HMM predictor have been trained for a subset of Gene Ontology classes. Since protein function is closely related to both protein sequence and structure. To record a large amount of protein properties, the input data for this HMM are amino acid sequences aligning with the corresponding secondary structure sequences. This mathematical approach is practical and easy to compute. The accuracy of this method can reach to 62.97%.

1 Introduction

New genomes are increasing rapidly. When analyzing the protein coding genes, it is typically only possible to assign function to 40% of all protein sequences based on sequence similarity [1, 2, 3, 4]. There is a great need for protein function prediction methods. There are a number of methods to predict protein function, such as: 1. Pairwise sequence similarity methods [5, 6, 7, 8, 9]. 2. Iterative search methods among multiple sequences [10, 11]. 3. Super family database methods [12, 13]. 4. Alignment based methods using functional links [14, 15]. 5. Predicting protein function via structure [16,17,18]. 6. Neural network using protein features.[19]

The Gene Ontology project [20] provides a controlled vocabulary to describe gene and gene product in any organism. The GO consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes. The three organizing principles of Gene Ontology are cellular component, biological process and molecular function. A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions.

If a protein sequence is not similar to any other known protein, it is reasonable to expect that proteins with related function will have similar properties, even if they are not evolutionarily related. In this paper, a HMM predictor have been trained for a subset (cellular component) of Gene Ontology classes. Similar technique can be applied for biological process and molecular function of GO classes. Since protein function is closely related to both protein sequence and structure. To record a large amount of protein properties, the input data for this HMM model are amino acid sequences aligning with the corresponding secondary structure sequences. This mathematical approach is practical and easy to compute. The accuracy of this method can reach to 62.97%.

In [19], Jenson had developed 17 protein features from sequence or secondary structure. Many physical, chemical or biological characteristics are involved to compute these features. Then neural network is trained to recognize specific property of these features. Our HMM statistic model is relatively simple to predict function. This model has considered the alignment corresponding relationships between sequence and secondary structure. For orphan protein which has similar property and dissimilar sequence, the propagation possibilities of HMM can recognize such property and provide better prediction.

A data set and HMM graph is introduced in Section 2. This HMM is trained in Section 3. Then a prediction method is applied in Section 4. Finally, the performance of our approach is evaluated in Section 5.

2 Data Set and Hidden Markov Model

The Gene Ontology (GO) Annotation project at European Bioinformatics Institute has mapped identifiers of Protein Data Bank (PDB) into GO terms. Three organizing principles of GO are cellular component, biological process and molecular function. Table 1 collects 6362 proteins from Protein Data Bank. These proteins are classified according to six categories of GO cellular component. These six categories are envelope, extracellular matrix, extracellular region, organelle, protein complex and virion. The protein quantity for each classification is also showing in the table.

Table 1. Classify PDB proteins into 6 GO terms.

Classification name	quantity	test set#
envelope	22	8
extracellular_region	2292	639
extracellular_matrix	184	49
organelle	1830	444
protein complex	1574	416
virion	460	129

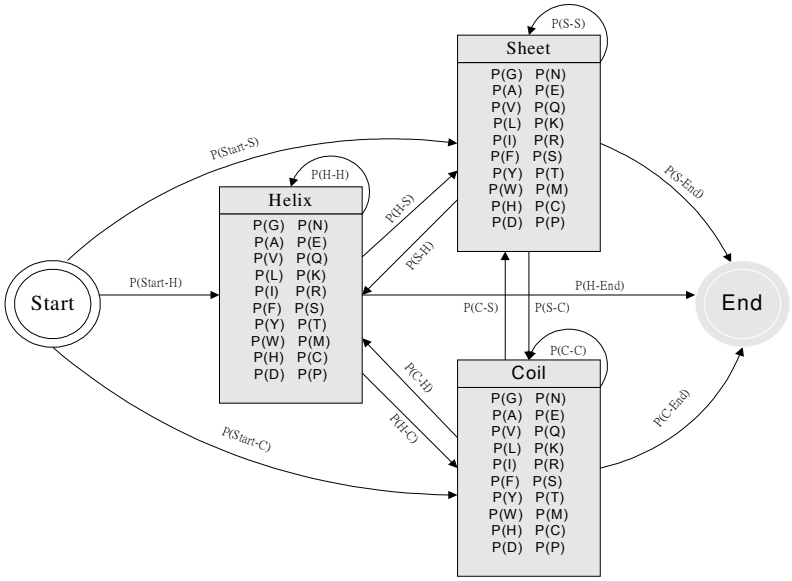

Figure 1. HMM graph using for sequence - structure alignment method

The HMM graph of our sequence-structure method is given in Figure 1. Three states are helix, sheet and coil. Each state includes 20 probabilities that represent the probabilities of 20 amino acids appearing in this state. The propagation probabilities are used to represent all the probabilities of transiting from one state to another.

3 Training Hidden Markov Model

Three arrays are created: h[residue], s[residue] and c[residue]. The index "residue" represents 20 different amino acids. Every parameter represents individual probability. $h_s means the frequency of transition from helix to sheet. $h[0] means the frequency of amino acid G appearing in α-helix.

Example 1:

Assume that protein sequence is defined as array aa[0..4] = "AQGQM", and the corresponding 2D structure sequence is array ss[0..4] = "THHHE".

When the symbol of the 2D structure is "E", it represents β-sheet state. The symbol of the 2D structure "H" represents α-helix state. All the symbols including the blank differ from these two represents the state is in the coil,. The following 3 steps are used to train HMM:

Step1: calculate the frequency in each path:

At index=0, ss[index]="T" and aa[index]="A" in Example 1. Since ss[index] = "T", the state is transiting from start to coil. Therefore amino acid "A" is in state of coil. $start_c and $c[1] are incremented.

At index=1, ss[index]="H" and aa[index]="Q". because ss[index] is equal to "H", the state is from coil to helix. $c_h and h[12] are incremented by 1.

At index=2, ss[index]="H" and aa[index]="G". $h_h and $h[0] are incremented.

At index=3, ss[index]="H" and aa[index]="Q". $h_h and $h[12] are incremented.
At index=4, ss[index]="E" and aa[index]="M". $h_s and $s[17] areincremented.
After all, it's the end of the string, the state is from sheet to end. Then $s_end=$s_end+1.

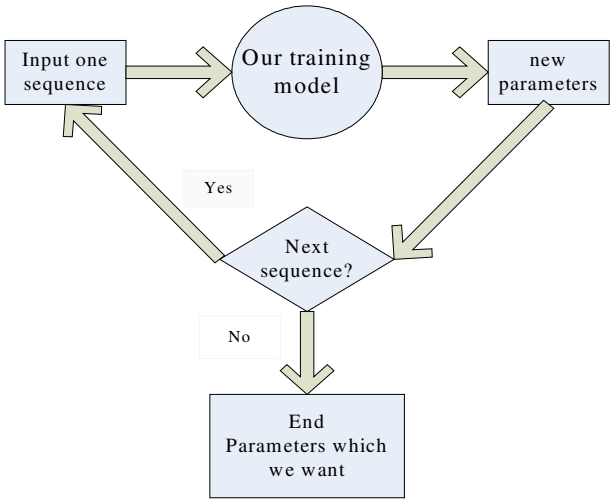

Figure 2. Training the HMM model.

Step2: calculate the probability of every parameter

We have to know the probability of each parameter. Utilizing the data counted, we can get the expected value of each parameter. In helix, sheet or coil, the total sum of probabilities of 20 amino acids equals 1. The probability of amino acid G is the frequency of amino acid G divided by total frequency of 20 amino acids, as:

$$P(G)=\frac{N(G)}{N(G)+N(A)+N(V)+...+N(C)}$$

It is noted that the starting and ending protein sequences are usually corresponding to coil at most of time.

Step3: The calculated probability translates into log value, and the base of logarithm is e (= 2.71828). There is every parameter's log value about this example. Because we can not take log of 0, represent it by §.

4 Predicting Method

We got the protein data from protein data bank and had said we focus on the cellular component. It has eight categories at present, but we only fetch six categories. After running our training model, we can get envelope model by training set of envelope category. Extracellular matrix, extracellular region, organelle, protein complex and virion are the same. After training, we can begin to test. See the figure 3-7.

The predicting method is to input an unknown protein into the model, and we multiply every parameter of passing through. Then we can finally get the expected value. But the last value may be very difficult to calculate because the value is too small, like 3.95E-70. It is all the multiplication of decimals, so the result is too small to show and the performance is not very good on procedure execution. Because of this shortcoming, we utilize logarithmic function. Not only the result shows easily, the procedure is relatively high efficiency on addition operation. We assume there are two sequences, protein sequence and its 2D structure sequence, and the length of them is m. The sequence imports to our training model and calculate the expected value. The array "aa" denotes this protein sequence. P(aa) denotes its expected value, but it's probably in helix, sheet or coil. The array "A" denotes a series of the transition of state. It represents probability when the state is changed.

$$A=h\_h||h\_s||h\_c||s\_h||s\_s||s\_c||c\_h||c\_s||c\_c$$

$$\log_T(XY)=\log_T X+\log_T Y$$

There is a formula about logarithmic function.

$$\log_e P(aa)=\log_e P(start)+\sum_{i=0}^{m-1}[\log_e P(aa[0])$$

$$+\log_e P(A_0)]+\log_e P(end)$$

Assume the base of logarithm is e and rewrite the formula above.

$$P(aa)=P(start)*\prod_{i=0}^{m-1}[P(aa[i])*P(A_i)]$$

$$*P(end)$$

Then we get a simple prediction utilizing above training data. Given an unknown protein sequence and it's 2D structure.

protein sequence: PSGQM

2D structure sequence: GGHHH

Log value of Expected = P(start_c)+P(P in coil)+P(c_c)+P(S in coil)+P(c_h)+P(G in helix)
+P(h_h)+P(Q in helix)+P(h_h)+P(M in helix)+P(h_end)
= -2.3552

Expected value = 0.0948365875081179

H(G)	0.0462	S(G)	0.0426	C(G)	0.09
H(A)	0.106	S(A)	0.0838	C(A)	0.0856
H(V)	0.0836	S(V)	0.1276	C(V)	0.0846
H(L)	0.0999	S(L)	0.0799	C(L)	0.0615
H(I)	0.0704	S(I)	0.0732	C(I)	0.0419
H(F)	0.03	S(F)	0.0472	C(F)	0.0292
H(Y)	0.0351	S(Y)	0.0529	C(Y)	0.0328
H(W)	0.0167	S(W)	0.0259	C(W)	0.0132
H(H)	0.0084	S(H)	0.0159	C(H)	0.0158
H(D)	0.0493	S(D)	0.0239	C(D)	0.0518
H(N)	0.0514	S(N)	0.0415	C(N)	0.0686
H(E)	0.0601	S(E)	0.0364	C(E)	0.0412
H(Q)	0.0401	S(Q)	0.0304	C(Q)	0.0341
H(K)	0.0598	S(K)	0.0488	C(K)	0.0507
H(R)	0.0507	S(R)	0.0409	C(R)	0.0432
H(S)	0.0818	S(S)	0.0817	C(S)	0.089
H(T)	0.0622	S(T)	0.082	C(T)	0.0806
H(M)	0.0137	S(M)	0.0156	C(M)	0.0123
H(C)	0.0085	S(C)	0.0152	C(C)	0.0117
H(P)	0.025	S(P)	0.0334	C(P)	0.0613
Start_S	0	Start_H	0	Start_C	1
S_S	0.7896	S_H	0.0023	S_C	0.208
H_S	0.0006	H_H	0.9022	H_C	0.097
C_S	0.0826	C_H	0.0212	C_C	0.8883
S_End	0	H_End	0	C_End	0.0079

5 Performance of Our Approach

1 Data source and our database

We get data from protein data bank, and put them into our database with Perl. Perl is often applied to the bioinformation because it's easy to handle string. Capacity of the file is usually very big and the string is mostly too long, like DNA sequence. Putting the data into our database, the program fetches the data from the database directly. Our systematic environment is as follows.

Apache Web Server Version 1.3.33
PHP Script Language Version 4.3.10
MySQL Database Version 4.1.8
Zend Optimizer Version 2.5.7
phpMyAdmin Database Manager Version 2.6.1-rc2

2 Training

The program fetches the data from database and put them into our mathematical model. The result of virion category is shown as follows. It's every variable's frequency of passing through of virion.
After knowing this, we can calculate every variable's probability..
The probability translates into log value through the logarithmic function and the base of logarithm is e.

3 Performance

Table 4-6 Rank table of test dataset

Function\rank	1	2	3	4	5	6
envelope	7	0	0	0	0	1
extracellular_matrix	26	8	0	6	5	4
extracellular_region	427	83	80	37	12	0
organelle	275	85	49	14	12	9
protein complex	266	106	33	9	2	0
virion	60	14	13	26	12	4
total	1061	296	175	92	43	18

We know protein sequence and its secondary structure of every protein in our dataset, then utilizing our approach to predict. We see an example of secondary structure in figure 4-2.
After testing, we see the output of envelope category. It's shown in table 4-5. In the table 4-5, each score is log value and every protein has data of six different categories. We assume one certain protein belongs to the category which score is the biggest. According to the output of envelope category, there are seven proteins belong to the envelope category. We see the result of all and it's shown in table 4-6. The accuracy: 1061/1685=62.97%. Sensitivity: TP / (TP+FN) = 1061 / (1061+624) = 0.6297
Accuracy: 63%

4 Conclusion

We have developed a method to predict protein function. Our algorithm, a Hidden Markov Model, has been built to solve our problem with statistical concept. At present, many researchers are devoted to the study of this field. But several methods are too complex because many factors of the protein are considered. We utilize advantage of the computer to deal with a large amount of biological data quickly. The mathematical model is trained to get the

probability. We can use our method to predict protein function after training. The relation between the protein function and its molecular structure is close, so the secondary structure is considered in our method. Finally, the average accuracy of our method is above 60%. In the future, this method can be applied to the prediction of other functions of biochemistry such as Enzyme and Nonenzyme.

6 Reference

[1] Altschul, S. F., Madden, T. L., Scha_er, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25:3389-3402.

[2] A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res., 25:3389-3402.

[3] Li, W., Pio, F., Pawlowski, K., and Godzik, A. (2000). Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. Bioinformatics, 16:1105-1110.

[4] Spang, R. and Vingron, M. (2001). Limits of homology detection by pairwise sequence comparison. Bioinformatics, 17:338-342.

[5] Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. Prot. Eng., 7:1323-1332.

[6] Mott, R. (2001). Maximum likelihood estimation of the statistical distribution of smith waterman local sequence similarity scores.

[7] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. Nature, 402:83-86.

Bull. Math. Biol., 54:59-75.

[8] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. Proc. Natl. Acad. U.S.A., 85:2444-2448.

[9] Smith, T. F. and Waterman, M. S. (1981). Identi_cation of common molecular subsequences. J. Mol. Biol., 147:195-197.

[10] Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. (1998). EUCLID: automatic classi_cation of proteins in functional classes by their database annotations. Bioinformatics, 14:542-543.

[11] Bolten, E., Schliep, A., Schneckener, S., Schomburg, D., and Schrader, R. (2001). Clustering protein sequences structure prediction by transitive homology. Bioinformatics, 17:935-941.

[12] Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and A, B. (2002). The PROSITE database, its status in 2002. Nucleic Acids Res., 30:235-238.

[13] Gough, J. and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res., 30:268-272.

[14] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. Science, 285:751-753.

[15] Yanai, I., Mellor, J. C., and DeLisi, C. (2001). Identifying functional links between genes using conserved chromosomal proximity. Trends in Biotechnology, 19:S61-S66.

[16] Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioural sciences. PhD thesis, Harvard University.

[17] Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol., 307:1113{1143.

[18] Norin, M. and Sundstr□om, M. (2002). Structural proteomics: development in structure-to-function predictions. Trends in Biotechnology, 20:79{84.

[19] Lars Juhl Jensen (2002)，”Prediction of Protein Function from Sequence Derived Protein Features”，Center for Biological Sequence Analysis，BioCentrum-DTU，Technical University of Denmark，Lyngby 2002

[20] Gene Ontology http://www.geneontology.org/