行政院國家科學委員會補助專題研究計畫 
成果報告

利用密碼取代法預測低相似序列的功能

計畫類別：■ 個別型計畫  □ 整合型計畫
計畫編號：NSC 97－2621－B－006－002－MY2
執行期間：97年03月01日 至 98年07月31日

計畫主持人：吳煥麟
共同主持人：張素瓊
計畫參與人員：賴建志

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：生物多樣性研究所

中華民國98年10月26日
摘要

如果一種新發現的蛋白質序列與已知的序列相似時，經常有可能利用同源比對的方法來推斷其結構與功能。然而當序列歧異度太大時，序列/結構/功能的關係是非常難以判斷的。我們認為經過適當的氨基酸互換，同功但差異度大的序列仍可產生可偵測的序列樣式(patterns)。其實這種字母(alphabet)互換規則長久以來已成功地被利用在訊息的加密(encryption)與解密(decryption)中。因此我們設計了一個機率模型來判別同功蛋白質之間的互換規則。鑑於二十個氨基酸字母的所有互換規則過於龐大，我們亦設計了演化演算法(evolutionary algorithm)來完成所須的計算。訊號序列(signal peptide)具有短且歧異度大的特性，所以非常適合用來驗證我們的想法。我們分別測試了病毒蛋白質紅血球凝集素(hemagglutinin)及哺乳類蛋白質 HLA class I compatible antigens 的訊號序列，發現這個方法確實是顯著而可行的。本研究成果可提供一種新方法去尋找歧異度大的序列的結構或功能，而這是傳統的方法較難以做到的。

關鍵字：序列比對、密碼學、演化演算法
Abstract

Once a new found amino acid sequence shows similarity to known proteins, one can infer its possible structure and/or function by conventional alignment methods. However these homology methods cannot do much when the similarity of the sequences is really low. A substitution rule is a man-made regulation to determine whether amino acids are replaceable or not. Under some substitution rules, several short stretches of matched sequence could be found, suggesting that a specific substitution rule enables us to identify regions in protein sequences that share the similar functional characteristics. We established a mathematical method to detect the possible substitution rules of protein sequences, and designed an evolutionary algorithm in practice. Signal peptides are extremely variable, both in their length and in their amino acid composition. We think that the signal peptides are an ideal target to validate our methods. We have tested the signal peptides of viral proteins, hemagglutinin, and mammalian proteins, HLA class I compatible antigens respectively. The results showed our methods are practicable and significant. Our model may offer an alternative to investigate those diverse sequences that show no significant sequence similarity using conventional alignment methods.

Keywords: sequence alignment, cryptography, evolutionary algorithm
# Table of contents

Chinese abstract  

English abstract  

1. Introduction  

2. Methods  
   2.1 Cryptography  
   2.2 The formula of the model  
   2.3 Evolutionary algorithm  

3. Results and Discussions  

References  

計畫成果自評部份
1. Introduction

Proteins are the main catalysts, structural elements, signaling messengers and molecular machines of biological tissues (Branden 2000). Until recently, there have been two principal ways to learn more about the functions of protein molecules. All primary knowledge of function has come from some biochemical, genetic or structural experiment on an individual protein. Besides, once a function has been assigned to an individual protein, one can search for other proteins with related functions by seeking proteins whose amino acid sequences are similar to the original protein. This ‘homology method’ is used widely to extend knowledge of protein function from one protein to its cousins. It is estimated that these annotated sequences is only a half of ORFs from 600 known genomes and the roles of many ORFs have been incorrectly assigned (annotated) in databases (Stubbe 2007). These errors lead to additional misassignments as new sequences are deposited. Determining the function of a protein is thus a huge challenge that requires creative, multidisciplinary approaches.

Prediction of protein function from sequence and structure is still a difficult problem, because homologous proteins often have different functions (Whisstock 2003). Take the eye lens protein as an example. This protein in the duck is identical in sequence to active lactate dehydrogenase and enolase in other tissues, although they do not encounter the substrates in the eye (Wistow 1987). Conversely, non-homologous proteins may have similar functions. Chymotrypsin and subtilisin, two proteinases are not homologous and show entirely different folding patterns (Wistow 1987). They are a standard example of convergent evolution. This and other examples show that it is not possible to reason that if two proteins have different sequences and/or folding patterns they must have different functions. Therefore we can summarize these coherent and incoherent phenomena of protein prediction here: similar sequences normally produce similar protein structures and functions, but exceptions abound (Ponting 2001); conversely, similar structures and/or functions are often found with very different sequences.

Different approaches have been proposed to prediction of protein structure/function from amino acid sequence alone, which focus on three folds, overall sequence similarity, signature patterns of active sites, and profile methods. These varied thought and methods for data mining mainly rely on the sequence alignment. Moreover the existing alignment methods all base on a few fixed substitution matrices, shah as BLOSUM or PAM (Dayhoff 1981). Since these methods are sample-dependent, Aravind and Koonin argued that current sequence search methods do not pick up even all the proteins known to be genuine homologues (Aravind 1999). We aim to develop a method to find those dissimilar sequences whose function could not be detected using conventional alignment methods. This method combines the concepts form cryptography and formal language theory. To our knowledge, it is the first time that these mixed concepts are applied to bioinformatical study.

2. Methods

Identification and analysis of conserved sequences is an important part of current bio-scientific research. Substitution matrices are the core of sequence alignment. The existing scoring schemes of substitution matrices for sequence alignment are based on observed
substitutions of amino acids. To extend this idea of substitution, we first review an analogue case from cryptography.

2.1 Cryptography

The substitution cipher has long been used to keep secret messages. The original message is called the plaintext and the encrypted message is called a ciphertext (Singh 1999). The substitution cipher is a substitution rule that transforms each letter of the plaintext into some other symbol(s). For example, the Caesar cipher has a rule as below. Note here that the correspondence between letters is not necessary to be one-to-one in other more complex cases. The substitution rule can be regarded as a partition mathematically. Here the Caesar cipher is denoted as the partition \([\{A, N\}, \{B, O\}, \{C, P\}, \ldots, \{M, Z\}\]).

![Caesar Cipher Diagram]

Hence the plaintext ‘BIOSCIENCE’ is transformed into ciphertext ‘OVBFVPVRAPR’ according to the Caesar cipher.

If we get a ciphered message ‘OVBFVPVRAPR’, how do we get its meaning? A basic method is to try all possible substitution rules and to see whether these plaintexts are ‘understandable’. It may happen that several ‘understandable’ plaintexts exist. However the number of ‘understandable’ plaintexts will decrease when the length of text increases. When the right substitution rule is found, then the other messages ciphered by the same substitution cipher will be obtained.

Following this logical thought, the procedure can be applied to bio-sequences which have similar structure and/or function (as ‘BIOSCIENCE’ and ‘OVBFVPVRAPR’ with the same meaning) but have no significant sequence similarity (as ‘BIOSCIENCE’ and ‘OVBFVPVRAPR’ with totally different texts).

However bio-sequences, comparatively, are more complicated since the ciphertext is not obtained by one transformation but several times of point, insert and deletion mutation. And basically there is no naturally ‘understandable’ bio-sequence to see whether the substitution rule is ‘better’. Therefore we design a mathematical method to judge which substitution rule is more possible a ‘better’ one. We first give a description of this method and explain why this method may work. For convenience we use some terminology of formal language theory, such as sequence (bio-sequence or amino acid sequence in biology) and letter (nucleotide or amino acid in biology) in general (Moll 1988) and use biological terminology when we want to emphasize its biological characteristic.

Given an alphabet set, various substitution rules can be envisaged. For example, given an alphabet \([A, B, C, D]\), then \([A, B, C, D]\), \([\{A, D\}, C, B]\) and \([A, [B, C, D]]\) are three different substitution rules (or partitions), where symbols inside the square brackets are replaceable to one
another. Take \([B, C, D]\) as an example. B, C and D are regarded as the replaceable letters, briefly B=C=D. The question then arises as to how we assess which substitution rules are better than others.

2.2. The formula of the model

For simplicity we take only two sequences under consideration. According to each substitution rule, a Dotter-type (Sonnhammer 1995) analysis will be used to identify local similarities in these given sequences (Fig. 1). Note here that the order of local-similar segments is not necessary to be the same. Then the probability that these local-similar segments occur by chance will be calculated. Those substitution rules that indicate local-similar segments with lower probabilities of occurring by chance will be retained for the further rounds of analysis and refinement.

![Figure 1](image-url) According to the substitution rules, two sequences (pink and sky-blue lines) have local-similar segments (square boxes). The black arrows indicate which subsequences are the same.

We now present the mathematical formula used in above probability calculation. Let \(X\) denote an alphabet (amino acids in biology) and \(|X|\) the number of \(X\). Assume \(S=x_1 \ldots x_m\) is a sequence, and there are \(e_i\) letters replaceable with \(x_i\) under a give substitution rule. We define the volume \(V\) of \(S\) as \(e_1 \times e_2 \times \ldots \times e_m\). For example, assume \({A, [B, C, D]}\) is a substitution rule and \(S=ABC\). Then the volume of \(S\) is \(1 \times 3 \times 3 = 9\). That is, the volume of \(S\) is the number of all sequences that is the same as \(S\) under the given substitution rule. Firstly, we deal with a simple case (Fig. 2). Given one sequence \(S\) with length \(L\) and one local-similar segment \(U_1\) with length \(n_1\) and volume \(v_1\), then the probability \(P_1\) of \(U_1\) occurring in \(S\) by chance is equal to or smaller than

\[
(i) \quad v_1 \times \{ t \times |X|^{t-1} - [(t-n_1) \times (t-n_1+1)/2] \times |X|^{t-1-n_1} + [ (t-2n_1) \times (t-2n_1+1) \times (t-2n_1+2)/6] \times |X|^{t-1-2n_1} \} / |X|^L,
\]

where \(t=L-n_1+1\).

![Figure 2](image-url) A sequence \(S\) (pink line) with a local-similar segment \(U_1\) (light orange square box).

This formula has three terms:

\[
\begin{align*}
&v_1 \times \{ t \times |X|^{t-1} \} / |X|^L; \\
&- v_1 \times \{ [(t-n_1) \times (t-n_1+1)/2] \times |X|^{t-1-n_1} \} / |X|^L \text{ and}
\end{align*}
\]
\[ v_1 \times \{(t-2n_1) \times (t-2n_1+1) \times (t-2n_1+2)/6 \} \times |X|^{(t-1-2n_1)/|X|L}. \]

More terms can be added, if required, for more accurate estimate.

Now consider a more realistic example (Fig. 3). Given one sequence \( S_1 \) with length \( L_1 \) and \( m \) disjoint local-similar segments \( U_1, U_2, \ldots, U_m \) in \( S_1 \) with length \( n_i \), where \( i=1, \ldots, m \), then the probability that the \( m \) local-similar segments occur by chance in \( S_1 \) is equal to or smaller than:

(ii) \[ P_1 \times P_2 \times \cdots \times P_m \]

where \( P_i, i=2, \ldots, m \), are from equation (i) but \( L \) are replaced by \( L-\Sigma n_j \).

**Figure 3.** A sequence \( S_1 \) (pink line) with local-similar segments \( U_i \) (square boxes).

If \( P \) is the probability of local-similar segments \( U_i \) occurring by chance in \( S_1 \) and \( Q \) is the probability of the local-similar segments occurring by chance in a second sequence \( S_2 \), then the probability of the local-similar segments occurring by chance in \( S_1 \) and \( S_2 \) is equal to or smaller than

(iii) \[ P \times Q. \] (Fig. 4)

The formula can be extended to more than two sequences by mathematical induction.

**Figure 4.** Two sequence \( S_1 \) and \( S_2 \) ((pink and sky-blue lines) with local-similar segments square boxes).

Now we explain what meaning this formula has. Assume two or more sequences with the similar function are collected. Numbers of probabilities are obtained from these sequences according to different substitution rules. These probabilities tell us how high these events occur by chance. The smaller probability means the event is less likely to occur by chance. Therefore, the corresponding substitution rule could carry some information about the function of those sequences.

2.3. Evolutionary algorithm

Time complexity is about \( 10^{15} \) if we test all substitution rules of twenty amino acids. It is inefficient and infeasible to examine all of them. We, therefore, designed an evolutionary algorithm and try to find a heuristic solution.

Figure 5 is the flow chart of our algorithm where italic red letters represents parameters. The
algorithm firstly generates a population of m partitions, and then uses our proposed formula to calculate probability with given sequences. After selecting n best partitions with lower probabilities, we disconnect these partitions into pairs. Since every partition is composed with pairs, we filter our partition with theses pairs with k higher appearance. Finally the pairs are combined together and test if it can achieve lower probability. If we can not get a better partition in this iteration, we then decrease k by 1 and do the same test again. The whole flow will be continued until the stop condition has been satisfied.
Create $m$ partitions of amino acids

Use our formula to get $n$ best partitions with lower probability

Get pair sections from these partitions and select $k$ pairs with higher appearance.

$k = k - 1$

If $k > 1$

Test if these $k$ pairs can assemble better partitions?

yes

no

Test if there exists better partition from any combination of currently partition?

yes

no

Replace selected partition to new one

no

Stop condition

yes

Finally selected partition

Figure 5. Evolutionary algorithm for finding partitions
3. Results and Discussions

Signal peptides (SP) are extremely variable, both in their length and in their amino acid composition (Hegde 2006). Hence we think that the signal peptides are an ideal target to validate our method. These signal peptides are obtained from Signal Peptide Website: An Information Platform for Signal Sequences and Signal Peptides (http://www.signalpeptide.de/).

We initially tested the hemagglutinin SPs in viruses. The SPs are divided into three statuses, the confirmed and the potential hemagglutinin SPs, and the non-hemagglutinin SPs in the database.

**Result 1.** Partition training with SPs in hemagglutinin

**Signal peptides picked**

Seventy confirmed SPs in hemagglutinin were obtained from the database. Fifty of these SPs were picked randomly for training to obtain a better partition. The twenty remaining SPs, all of the potential SPs and all viral SPs (hemagglutinin and non-hemagglutinin) were used to validate the efficiency of this method.

**Parameters** (Fig. 5)

- Iteration number: 15
- Total partitions: 10000
- The number of better partitions chosen each time: 1000
- The minimal length of the same sequence for probability calculation: 5

**The better peptide model and partition obtained:**

- Accession number: Q6DQ22 (http://www.signalpeptide.de/)
- Sequences: KIVLLLAIVSLVKS
- Partition: [AFILTV], [C], [D], [E], [G], [H], [K], [M], [N], [P], [Q], [R], [S], [W], [Y]

**The probability distribution analyses:** the probability P is converted to log (P).

These data shows (Fig. 6, 7, 8):

1. Applying the clustering analyses to these data (Fig. 6, 8), our method can discriminate hemagglutinin and non-hemagglutinin SPs significantly.
2. The method re-proves that the potential hemagglutinin SPs are valid (Fig. 6, 7).
Figure 6.

Probability distribution of SPs in verifying cipher training with virus hemagglutinin (testing)

Figure 7.

Probability distribution of SPs in verifying cipher training with virus hemagglutinin (potential)

Figure 8
Moreover we tested SPs in mammalian HLA class I compatible antigen. The result is more significant (Fig. 9, 10).

Result 2: Partition training with SPs in mammalian HLA class I compatible antigen
The methods and parameters are similar to those in Result 1. Here 21 HLA SPs are used for training and 70 SPs are randomly picked from all mammalian SPs in the database.

The better model peptide obtained:
Accession number: P30457 ([http://www.signalpeptide.de/](http://www.signalpeptide.de/))
Sequences: MAVMAPRTLVLSSGALALTQTWA
Partition: \{A\}, \{CDEHIKQY\}, \{G\}, \{LV\}, \{M\}, \{N\}, \{P\}, \{R\}, \{S\}, \{T\}, \{W\}

The probability distribution analyses: the probability \(P\) is converted to \(\log (P)\).
Figure 9

Probability distribution of SPs in verifying cipher training with mammalian HLA class I compatible antigen

Figure 10

Probability distribution of SPs in verifying cipher training with mammalian proteins
We next compared our method to the conventional alignment method, *Dotter* (http://envgen.nox.ac.uk/bioinformatics/docs/dotter.html). The difference between our method and *Dotter* is that the sequences are replaceable in our method according the tailored partition, but those in the *Dotter* are not.

Result 3: Comparing the probability distributions between sequences with and without displacement using SPs in viral hemagglutinin (Fig. 11) and mammalian HLA class I compatible antigen respectively (Fig. 12).

**Comparision between with and without displacement**

![Diagram](image1)

Figure 11. Probability comparison with and without displacement in hemagglutinin.

**Comparision between with and without displacement**

![Diagram](image2)

Figure 12. Probability comparison with and without displacement in HLA class I
compatible antigen.

The compared probability data also showed that our methods are practicable and superior to the Dotter methods. Our model may offer an alternative to investigate those diverse sequences that show no significant sequence similarity using conventional alignment methods.
References


計畫成果自評部份：

本計畫預計完成下列目標：

1. 建立一個方法用於同功但具有很大歧異度的序列比對。
   我們已經完成了符合需求的數學方法。這個方法長久以來已成功地被利用在訊息的加密與解密中。然而這個方法需大量計算，所以我們也因應此難題設計了所需的演化演算法。

2. 利用所設計的方法去操作實際的例子。
   訊號序列(signal peptide)具有短且歧異度大的特性，所以非常適合用來驗證我們的想法。我們用這個方法分別測試了病毒蛋白質紅血球凝集素(hemagglutinin)及哺乳類蛋白質 HLA class I compatible antigens 的訊號序列，發現這個方法確實是顯著而可行的。

3. 我們的短期目標。
   我們準備整理現有的成果以統計，非試驗的形式發表論文。

4. 我們的中期目標。
   我們希望利用這個方法在現存資料庫中挖掘出未知的蛋白質，並以實驗加以證明。

本研究成果可提供一種新方法去尋找歧異度大的序列的結構或功能，而這是傳統的方法較難以做到的。