行政院國家科學委員會專題研究計畫　成果報告

人工智慧應用於計算生物學之研究(I)

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-029-008-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：東海大學工業工程與經營資訊學系

計畫主持人：張炳騰

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 8 月 17 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

人工智慧應用於計算生物學之研究(I)

計畫類別：■ 個別型計畫　□ 整合型計畫
計畫編號：NSC　94－2213－E－029－008－
執行期間：　94 年 08 月 01 日至 95 年 07 月 31 日

計畫主持人：張炳騰
共同主持人：
計畫參與人員：

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：東海大學工業工程與經營資訊學系

中　華　民　國　九十五　年　八　月　十六　日

中文摘要

本研究提出一個新的方法，用來建立蛋白質成對序列排比。目前在蛋白質序列排比上一直存在一個致命的問題，那就是當使用明確值資料去進行分析時，太多的不確定因子與敏感性資訊的遺失，導致序列排比問題出現瓶頸。由於使用不同的軟體和演算法會造成不同的結果，對於正在研究生物基因的科學家們，不同演算法亦很難廣泛地運用於基因序列上。因此基於這個最重要的前提下，本研究提出模糊的概念，將250單點突變矩陣(point accepted mutations, PAMs)與62區塊突變矩陣(blocks substitution matrix, BLOSUM)利用基因演算法(genetic algorithm, GA)於序列排比上。最主要的目的是用來減少不確定因子的影響，避免利用明確值或權重的方式，造成重要資訊的遺失，以及增加解的正確性與適用性。實驗果顯示，不論是利用PAM250還是BLOSUM62矩陣，利用GA演算法皆能找到更長且配對的蛋白質序列，並在不同矩陣的運用上，利用模糊矩陣所產生解的變動性要比明確值小，也就是說，將模糊概念運用於序列排比，確實能夠減少不確定性的影響並且增加解在區域相似上的利用性。
關鍵詞：蛋白質序列排比、基因演算法、模糊理論、仿射性間格懲罰函數

Abstract

In this paper a novel way to construct pairwise alignment of protein sequence is proposed. Currently in protein sequence alignment the vital problem is having too many uncertain factors and causes significant data loss while using crisp data. Due to using different software and algorithms that will bring about different results, for scientists researching in protein, different algorithms will be difficult to use widespread. Therefore, for this important premise, fuzzy concept is introduced and fuzziness is implemented in the matrix for 250 point accepted mutations (PAMs) and matrix for 62 blocks substitution matrix (BLOSUM) in sequence aligning, and integrated with the Genetic algorithm (GA). The purpose for this implementation is to reduce the effects of uncertain factor, avoid making use of crisp values or weights resulting in significant data loss, and increase solution accuracy and method suitability.

Results of experiment shows that application of fuzzy matrix to sequence alignment could find more continuous and identical protein sequence. Furthermore, this research used different matrix to sequence alignment. The result shows that variation of fuzzy matrix is smaller than crisp matrix. Therefore, application of fuzzy matrix certainly can reduce the effects of uncertain factor and increase solution accuracy. Hence, these results of experiment evidenced fuzzy logic useful to dealing with the uncertainties problem, and applied to protein sequence alignment successfully. The new method can provide different viewpoint for related research.

Keywords: protein sequence alignment; genetic algorithm; fuzzy theory; affine gap cost.

1. 前言

序列排比在分子生物學分析上，扮演了一個很重要的腳色。根據不同序列間的排比，可以幫助預測出未知的蛋白質功能或結構，若找出相似的部份越多，所能提供的資訊也就越多，因此兩條或多條以上的序列排比主要是盡可能找出最大的序列間的排比相似。兩條序列的排比我們通常稱為成對序列排比(pairwise)，而其他的則稱為多條序列排比(Hung et al., 2002)，圖1成對序列排比的例子取自Thompson et al.於1999年所發表的文章。

在圖1中，間格(gap, '-')是用來插入在兩條序列間，目的是為了找兩條序列最大的相似範圍，為了解決這個問題，我們需要有計量的標準來衡量排比出的序列。一般最常用的方法是利用成對序列排比總和分數/成本(sum-of-pairs score/cost) 或突變機率矩陣(mutation probability matrix)。

```
1   -NLFVALYDfvasgdntlsitkGEKLRVLgynhn - - - - - - - - -- -- - -gE
36   WCEAQt--kngqGWVPSNYITPVN- - -- - -
1   kGVIYALWDyepqnddelpmkeGDCMTIIhrede-- - -- - -- - -- - deiE
39   WWWARl- -ndkeGYVPRNLLGLYP-- - - - -
```

圖 1 成對蛋白質序列排比

在1970年，Needleman and Wunsch提出了一種演算法來解決2條蛋白質序列的排比。這個演算法是藉由計算找出排比最大的得分，而利用的衡量標準則是將所有排比對應的得分總和減去所有間格的懲罰分數總和。成對序列排比可以利用動態規劃(dynamic programming, DP)正確的解決。然而，由於排比問題屬於NP-hardness問題，因此最常使用啟發式的演算法來進行排比問題。從1970年到最近，許多的分析技術相繼被提出(Altschul et al., 1989; Altschul et al., 1990; Huang and Miller, 1991; Chao et al., 1994; Chao and Miller, 1995; Altschul et al., 1997; Chao,1998)。有一些技術能用來替代像蛋白質的成本矩陣，例如Dayhoff et al., 1978的PAM(Point Accepted Mutation)矩陣和，PAM矩陣是用來描述胺基酸在不同演化距離間的取代機率，通常結合間格成本(gap cost)或權重來產生成本函數。一般來說，不同成本函數會造成不同的最佳排比。因此，除了可利用已知的成本函數發展演算法去搜尋最佳解或近似最佳解外(Hung et al., 2002)，另一方面也可朝更接近真實生物學的替代矩陣發展(Henikoff and Henikoff, 1992)。由於矩陣的建立是利用實驗的方式找出胺基酸自然突變的比率，藉此來接近實際現象。因此本研究利用混合式演算法加上得分矩陣來找尋最佳解排比時，必定會包含一些不確定因素在內。因此本研究利用模糊集合與邏輯將得分矩陣內的每個元件模糊化。

模糊集合和邏輯最早是由Zadeh於1965年所提出用來解處理不確定環境的理論，尤其適用在複合系統(complex system)上。在模糊算數的領域裡，Zadeh的擴張原則(extension principle)被廣泛的應用近年來，有很多分子生物學上的文章都有應用模糊理論(Torres and Nieto, 2003; Carleos et al., 2003; Heger and Holm, 2003; Nieto and Torres, 2003; Huang and Li, 2004)。在序列排比方面，Blankenbecler et al. (2003)提出結構排比(structure -alignmen)的方法。這個問題所利用的成本函數包含模糊配對變數與自動配位結合。結構排比是根據Needleman-Wunsch的演算法，利用連續變數取代2元決策樹。

在本研究中，我們根據PAM與BLOSUM矩陣提出了一個新觀點。我們使用模糊數去取代矩陣中的元素。在下一個章節我們將介紹模糊數與計算，並且根據模糊後的矩陣，使用基因演算法求解。

2. 模糊算數、模糊PAM與模糊BLOSUM矩陣

2.1 模糊算數操作

模糊算數是根據Zadeh在模糊集理論中的擴張原則(Zadeh, 1965)，最初是被使用在Dubois and Prade (1980), Nahmias (1978), Mizumoto and Tanaka (1976)等的文章中,其他還有將擴張原則應用在approximate and exact manners (e.g., see Chang (2005) for a review)。假設有一個模糊集合或模糊數在實線$\Re$上，以下將介紹模糊定義與模糊算數。

模糊歸屬函數—有一個模糊集合或模糊數（FN）在實線$\Re$上，可表示為(a_1, a_2, a_3),a_2表示眾數，a_1與a_3分別表示$\tilde{A}(x)$*的左界與右界*，而歸屬函數$\tilde{A}(x)$*可被定義為*$x \in \Re$ to $\tilde{A}$

$$\tilde{A}(x) = \begin{cases} 0, & x < a_1, \\ L((x-a_1)/(a_2-a_1)), & a_1 \le x \le a_2, \\ R((a_3-x)/(a_3-a_2)), & a_2 \le x \le a_3, \\ 0, & x > a_3, \end{cases} \tag{1}$$

L和R分別表示$\tilde{A}(x)$ *中左邊與右邊的(shape functions)*。若以三角型隸屬度函數或三角模糊數（triangular membership function or called triangular fuzzy numbers(TFN)）可表示成如圖2

$$\tilde{A}(x) = \begin{cases} 0, & x < a_1, \\ (x-a_1)/(a_2-a_1), & a_1 \le x \le a_2, \\ (a_3-x)/(a_3-a_2), & a_2 \le x \le a_3, \\ 0, & x > a_3. \end{cases} \tag{2}$$

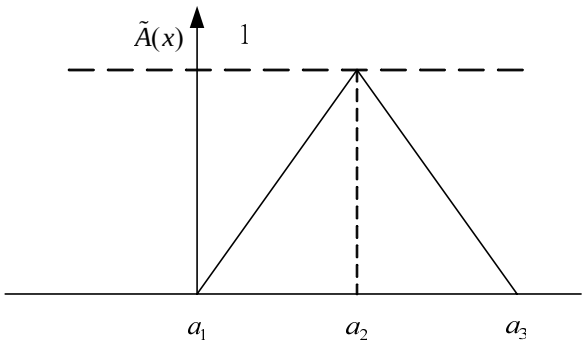

圖 2 三角隸屬度函數 (a_1, a_2, a_3)

在一信賴區間α水準，$\alpha \in (0, 1]$，我們定義A_α：

$$A_\alpha = \left\{ x \middle| \tilde{A}(x) \ge \alpha \right\}, \tag{3}$$

三角型隸屬度函數以α-截集表示為：

$$A_\alpha = [a_1^{(\alpha)}, a_2^{(\alpha)}] = [a_1 + \alpha(a_2 - a_1), a_3 - \alpha(a_3 - a_2)], \tag{4}$$

$a_1^{(\alpha)}$表示A_α的下界，而$a_2^{(\alpha)}$表示A_α的上界。

α-截集（α-cut）模糊算術**(***The α-cut fuzzy arithmetic)* 根據模糊算術方法，Zadeh's sup-min方法表示如下：

$$(\tilde{A} \circ \tilde{B})(z) = \sup_{x \circ y = z} \min(\tilde{A}(x), \tilde{B}(y)), \tag{5}$$

其中，$\circ$表示任何的算數操作(arithmetic operation)。公式(5)代表Mizumoto and Tanaka於（1976）使用相同方法在α-截集模糊數和區間算術。模糊算數的結果可被稱作α-截集算

數，並且α-截集算數的發展在Chang (2005)有做介紹

α-截集*模糊*算術中，模糊數基本運算公式（加法、減法、乘法、除法）可以計算的更快，在每一個α-level區間中，使用區間算術（Kaufmann and Gupta, 1988）以下為本研究所利用的模糊算數。

模糊數加法—令A和B分別為兩模糊數$\tilde{A}$和$\tilde{B}$，在一信賴區間α水準，$B_\alpha = [b_1^{(\alpha)}, b_2^{(\alpha)}]$, $\alpha \in (0, 1]$, $\forall \tilde{A}, \tilde{B} \subset \Re$，表示如下：

$$\begin{aligned} A_\alpha + B_\alpha &= [a_1^{(\alpha)}, a_2^{(\alpha)}] + [b_1^{(\alpha)}, b_2^{(\alpha)}] \\ &= [a_1^{(\alpha)} + b_1^{(\alpha)}, a_2^{(\alpha)} + b_2^{(\alpha)}], \forall \alpha \in (0, 1]. \end{aligned} \tag{6}$$

模糊數減法—在一信賴區間α水準，$\forall \tilde{A}, \tilde{B} \subset \Re$ $\alpha \in (0, 1]$，表示如下：

$$\begin{aligned} A_\alpha - B_\alpha &= [a_1^{(\alpha)}, a_2^{(\alpha)}] - [b_1^{(\alpha)}, b_2^{(\alpha)}] \\ &= [a_1^{(\alpha)} - b_2^{(\alpha)}, a_2^{(\alpha)} - b_1^{(\alpha)}] \end{aligned} . \tag{7}$$

模糊數乘法—在一信賴區間α水準，$\forall \tilde{A}, \tilde{B} \subset \Re$ $\alpha \in (0, 1]$，表示如下：

$$\begin{aligned} A_\alpha \times B_\alpha &= [a_1^{(\alpha)}, a_2^{(\alpha)}] \times [b_1^{(\alpha)}, b_2^{(\alpha)}] \\ &= [\min(a_1^{(\alpha)} b_1^{(\alpha)}, a_1^{(\alpha)} b_2^{(\alpha)}, a_2^{(\alpha)} b_1^{(\alpha)}, a_2^{(\alpha)} b_2^{(\alpha)}), \\ &\max(a_1^{(\alpha)} b_1^{(\alpha)}, a_1^{(\alpha)} b_2^{(\alpha)}, a_2^{(\alpha)} b_1^{(\alpha)}, a_2^{(\alpha)} b_2^{(\alpha)})]. \end{aligned} \tag{8}$$

模糊數除法—在一信賴區間α水準 $\forall \tilde{A}, \tilde{B} \subset \Re$ $\alpha \in (0, 1]$，表示如下：

$$\begin{aligned} A_\alpha \div B_\alpha &= [a_1^{(\alpha)}, a_2^{(\alpha)}] \div [b_1^{(\alpha)}, b_2^{(\alpha)}] \\ &= [\min(a_1^{(\alpha)} / b_1^{(\alpha)}, a_1^{(\alpha)} / b_2^{(\alpha)}, a_2^{(\alpha)} / b_1^{(\alpha)}, a_2^{(\alpha)} / b_2^{(\alpha)}) \\ &\max(a_1^{(\alpha)} / b_1^{(\alpha)}, a_1^{(\alpha)} / b_2^{(\alpha)}, a_2^{(\alpha)} / b_1^{(\alpha)}, a_2^{(\alpha)} / b_2^{(\alpha)})], \\ &\text{for } b_1^{(\alpha)}, b_2^{(\alpha)} > 0, \alpha \in (0,1]. \end{aligned} \tag{9}$$

模糊數解模糊化是重要的，它可被定義為將模糊數對應到一個最有可能的明確值上，這個觀念跟隨機變數取平均數是相似的。本研究所使用的解模糊化技術為重心法（the *center of area* (COA) or *COA*(x)）。其他可利用的技術可參考Leekwijck and Kerre (1999)。而COA可被定義如下：

$$COA(x) = \frac{\int_{\Re} \tilde{A}(x) x dx}{\int_{\Re} \tilde{A}(x) dx}. \tag{10}$$

2.2 模糊PAM250

在上述的討論中，我們有介紹到PAM250矩陣，1個PAM的演化距離表示100個残基中發生一個残基突變的機率，而PAM250矩陣則是將1PAM矩陣自乘250次得知。由於PAM矩陣是建立在胺基酸的改變上，並且資訊是來自許許多多的蛋白質實驗上，因此我們可以知道，其中必定包括一些不確定的因子，因此模糊歸屬函數被使用來替換PAM250矩陣中的元素，而模糊PAM250矩陣可被定義如下：

$$M_{ij} = (m_{ij} - l_{ij}, m_{ij}, m_{ij} + r_{ij}) \tag{11}$$

其中M_{ij}為 PAM250 矩陣中的元素，代表在給定演化區間的情況下，第 i 行胺基酸被第 j 列胺基酸所取代的機率，l_{ij} 表示在M_{ij} 中，演化距離的下界值，r_{ij} 表示在M_{ij} 中，演化距離的上界值，m_{ij} 則代表眾數，亦即 PAM 矩陣中第 i 行與第 j 列對應的明確值。

2.3 模糊BLOSUM62矩陣

Henikoff 和Henikoff 在1992年11月發展一種運算胺基酸序列排比時計算替換發生機率的BLOSUM矩陣。它是以比對蛋白質序列的BLOCKS資料庫為基礎，從其中擷取出來的序列推演而得。因此以BLOcks SUbstitution Matrix 來命名，目的在改善PAM矩陣的缺點。Henikoff 和 Henikoff計算出任何位置成對的胺基酸數目與全部期望出現成對的胺基酸數目的比率，結果以對數來表示。Henikoff考慮到遺傳距離相近的蛋白質序列中，彼此間胺基酸的差異不大，因此在計算每一點的各種胺基酸比例時，在遺傳距離相近的蛋白質間共有的同一種胺基酸在其點上所佔的比例會偏高，而導致計算上的偏差。所以布洛森矩陣的方法是將各個序列片段依照最低相同比例(minimum percentage identity)給予分組，每組分別計算其中每個點各種胺基酸的比例，事實上就像在計算一條序列的每個點各種不同的胺基酸出現的機率一樣，這樣就可以修正因為某些蛋白質彼此間遺傳距離較近而造成的偏差。因此本研究也將此矩陣納入模糊的考量，而模糊方式與上述的模糊PAM250矩陣相同，差別只在於所使用的 m_{ij} 在此將以BLOSUM矩陣為基礎，其他步驟同上將不再贅述。

3. 以PAM、BLOSUM矩陣和基因演算法為基礎的序列排比

在本研究中，我們利用基因演算法去找出最大的成對分數加總(sum of pairs)，但由於模糊矩陣與明確值矩陣所計算出的成對分數加總因基礎不同，不能用來比較，因此我們利用行分數(column score (CS))來作為明確值矩陣與模糊矩陣結果比較的依據，而CS公式可參考Thompson et al. (1999)。

3.1.參數編碼

本研究參考Hung (2002)中所用的編碼方式，這個方法是用來解決k緯度(k條序列)的最短路徑問題，利用向量的方式將編碼範圍分成3個緯度去進行求解搜尋。若以兩條列排比為例如圖3，圖3(a)為2條序列，利用參數編碼的方式可表示成圖3(b)，S為起點，T為終點，KPDN與KDN為兩條要進行比對的基因，從起點至終點皆有3個方向可前進，因此我們在每個x與y軸的交點上可發現皆須進行方向的抉擇，以圖3(b)中的P點為例，若向緯度1前進，則P點並不能對應到y軸的基因，因此P點會對應到一個gap，若向緯度2前進，則P點會保留，因此D點會對應到一個gap，若向緯度3前進，則P點會與y軸的基因D相對應，因此P點會對應到一個D，如此便能從起點至終點完整的對兩條基因行進方向進行編碼。

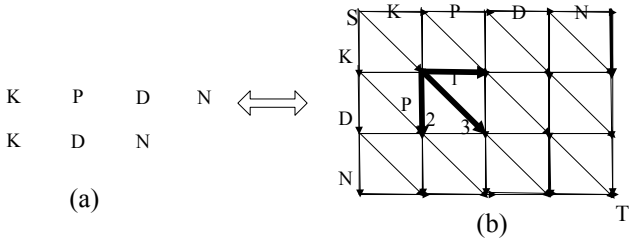

圖 3 成對排比表示方式 Hung (2002)

3.2初始母體設定

本研究的初始解(popsize)採隨機方式選擇10條進行求解，若以上圖3（a）為例，編碼位元為4，假設隨機產生的初始解中的一組解為KDN-，其中-所代表的意義為gap在進行編碼後即變成3331，若初始解為-KDN，進行編碼後即變成1333。

在基因序列排比上我們利用 PAM25 矩陣與仿射性間格懲罰函數(affine gap cost) Hung D.N (2002)來進行適應函數值的計算，而開啟成(open gap cost)設為 5，延伸成本(extend gap cost) 設為 2，以下為適應函數值計算方式：

$$f(\tilde{M}_{ij}) = COA\left(\frac{The\ best\ of\ \tilde{S}_j}{\sum \tilde{S}_j}\right) \tag{12}$$

其中，$\tilde{S}_j$ 為加總的 $\tilde{M}_{ij}$，並且利用解模糊化方法求出的明確值的來計算適應函數值。

3.3基因運算子

1) 輪盤法—假設利用每一組解所計算出的適應函數值為$f(M_i)$, i=1,…,k.，每組解的適應函數值 $f(M_i)$, i=1,…,k，則每一組解的複製個數計算如下：

$$n_i = \frac{f(M_i)}{\sum_{i=1}^{k} f(M_i)} \times k \tag{13}$$

這個方法可使較佳適應含數值被選擇的機率增加。

2) 局部保留—我們使用這步驟的原因是為了讓求解收斂快速。由於基因演算法屬於全域搜尋的方法，因此本研究希望除了依靠交配(crossover)來進行解的收斂外，另外再加上其他原則以便使收斂能夠更快速。若以上圖 3(a)為例，若初始解為 KDN-則與 KPDN 比較後發現，第一個位置的基因 K 與初始解相同，因此我們將其保留，原因在於當相同基因對應時所獲得的分數，遠高於不同基因對應時所獲得的分數，因此若將其保留，可加速求解收斂的速度。

3) 交配—由於我們所使用的參數編碼是以行進方向為基礎，因此不須利用傳統兩兩交配方式，而是父代與母代皆為相同染色體(圖 4)

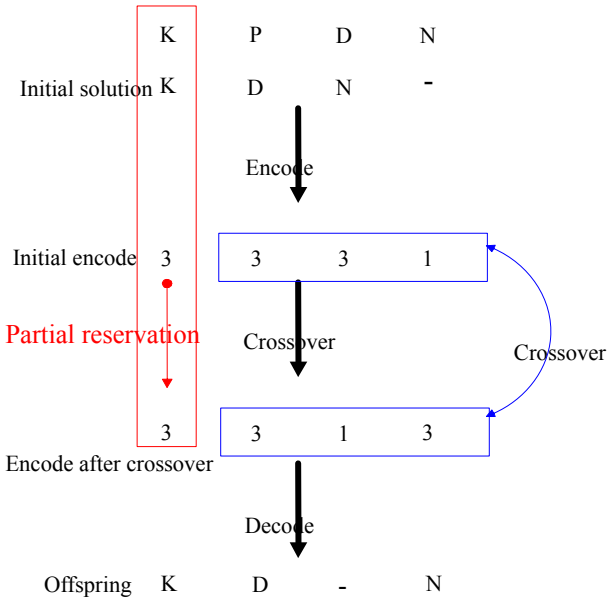

圖 4 局部保留與交配操作說明

4) 突變— 傳統的基金演算法為了增加收斂速度，通常是使用較小的突變機率。然而，在本研究中我們使用較高的突變機率(設為0.2)，因為若使用較低的突變機率，將會因收斂太過快速而導致無法搜尋到最佳解。除此之外，在局部保留時所留下的基因也有可能發生突變，這個條件是為了增加搜尋最佳解的機率。在圖5，我們可知道如何去進行突變的操作。

4. 序列排比分析說明

在眾多的資料中，我們利用BAliBASE(Thompson et al., 1999)所提供的蛋白質序列作為序列排比的樣本，其中利用reference 1 中的SH3來進行排比與分析，並且使用模糊PAM矩陣作為評分依據。除此之外，為了解決排比序列長度不同的問題，本研究利用間隔性懲罰函數(affine gap penalties)，主要原因在於此函數有較低的複雜性，因此求解快速。而在本研究所使用的間隔性懲罰函數中，開啟成本(open gap cost)設5，延伸成本(extended gap cost)則設2。.

在本研究中，我們使用了不同的三角歸屬度函數，分別為右偏、對稱、左偏的三角歸屬度函數。而右偏歸屬度函數為 $M_{ij}=(m_{ij}-1,\ m_j,\ m_{ij}+3)$，對稱歸屬度函數為 $M_{ij}=(m_{ij}-3,\ m_{ij},\ m_{ij}+3)$，左偏歸屬度函數為 $M_{ij}=(m_{ij}-3,\ m_{ij},\ m_{ij}+1)$。基因演算法求解代數為500代，重複次數為20次。基因演算法利用明確值求解(GA-crisp)與基因演算法利用模糊值求解(GA-fuzzy)比較結果參考表1、表2、圖6、圖7與圖8 。表1中我們包含平均CS、最低CS、最高CS以及連續最長配對CS。其中，平均CS與GA-crisp 沒有顯著差異，而最低CS與最高CS皆優於GA-crisp，而在表2中，平均CS與GA-crisp 沒有顯著差異，最低CS優於GA-crisp，最高CS在序列lpht-lihvA與lpht-lvie有顯著差異，圖6則是針對最高CS所畫出的圖形，其中藍色與紅色線分別代表模糊PAM與模糊BLOSUM矩陣應用於基因演算法，而綠色與粉紫色線分別代表明確值PAM與明確值BLOSUM矩陣應用於基因演算法。

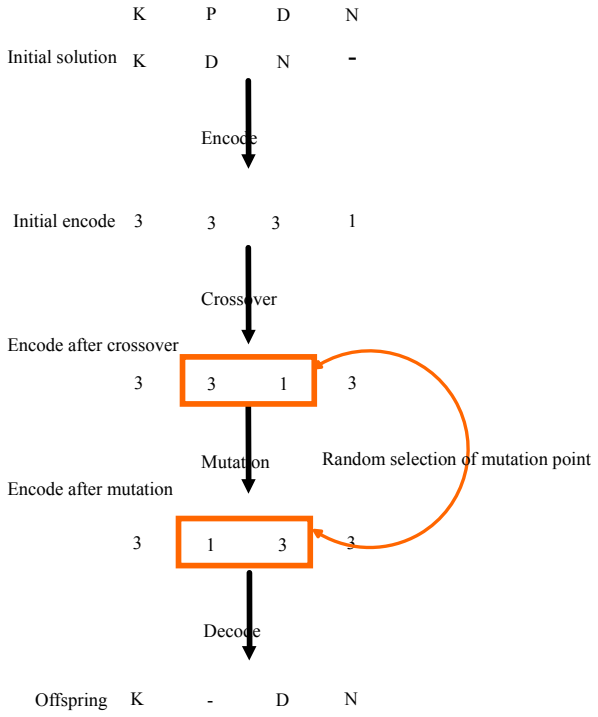

圖 5 突變操作說明

除此之外，連續最長配對CS(The longest continued sequence CS)也是一項非常重要的指標，若我們能找出越長的連續配對基因，代表兩序列間的特徵、功能較有可能出現相似的情形。不管從表1或表2中，顯示出我們的方法所獲得的連續最長配對CS，皆優於傳統的基因演算法，圖7則是針對連續最長配對畫出的圖形，且由圖8可得知，應用不同矩陣時，模糊矩陣所得的結果波動要比利用明確值矩陣來的小，這証明模糊理論確實可以降低不確定性環境的影響，其中圖6是利用表1與表2中的連續最長配對CS所畫出，以lpht-lycsB為例，兩模糊矩陣的誤差值為35%-33.75%=1.25%，而兩明確值的誤差值為28.75%-25%=3.75%，以此類推畫出lpht-lycsB、lpht-laboA、lpht-lihvA及lpht-lvie4點。

表 1 蛋白質序列利用GA-crisp與GA-fuzzy比較結果(PAM矩陣)

Sequence name	Method	Epoch	TFNs	Avg. CS(%)	The best CS(%)	The worst CS(%)	The longest continued sequence CS(%)
lpht-lycsB	GA-fuzzy	500	Slant-right	35	65	21.25	33.75
lpht-laboA	GA-fuzzy	500	Symmetric	33.75	48.75	20	20
lpht-lihvA	GA-fuzzy	500	Slant-left	22.5	35	11.25	16.25
lpht-lvie	GA-fuzzy	500	Slant-right	27.5	47.5	12.5	25
lpht-lycsB	GA-crisp	500	None	36.25	51.25	18.75	25
lpht-laboA	GA-crisp	500	None	32.5	45	18.75	17.5
lpht-lihvA	GA-crisp	500	None	22.5	33.75	11.25	10
lpht-lvie	GA-crisp	500	None	27.5	50	10	18.75

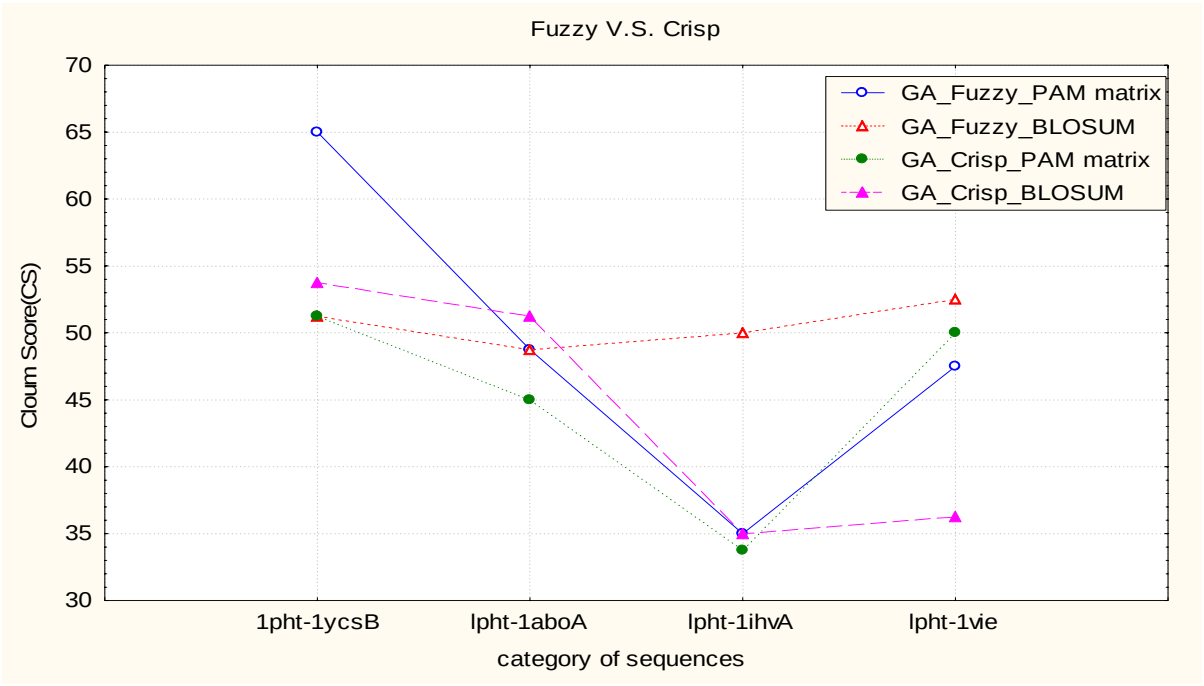

圖 6 GA-crisp與GA-fuzzy應用不同矩陣比較圖(最高CS)

表 2 蛋白質序列利用GA-crisp與GA-fuzzy比較結果(BLOSUM矩陣)

Sequence name	Method	Epoch	TFNs	Avg. CS(%)	The best CS(%)	The worst CS(%)	The longest continued sequence CS(%)
lpht-lycsB	GA-fuzzy	500	Slant-right	36.25	51.25	30	35
lpht-laboA	GA-fuzzy	500	Symmetric	35	48.75	30	26.25
lpht-lihvA	GA-fuzzy	500	Slant-left	25	50	10	22.5
lpht-lvie	GA-fuzzy	500	Slant-right	27.5	52.5	13.75	25
lpht-lycsB	GA-crisp	500	None	36.25	53.75	27.5	28.75
lpht-laboA	GA-crisp	500	None	36.25	51.25	25	25
lpht-lihvA	GA-crisp	500	None	21.25	35	10	15
lpht-lvie	GA-crisp	500	None	26.25	36.25	13.75	18.75

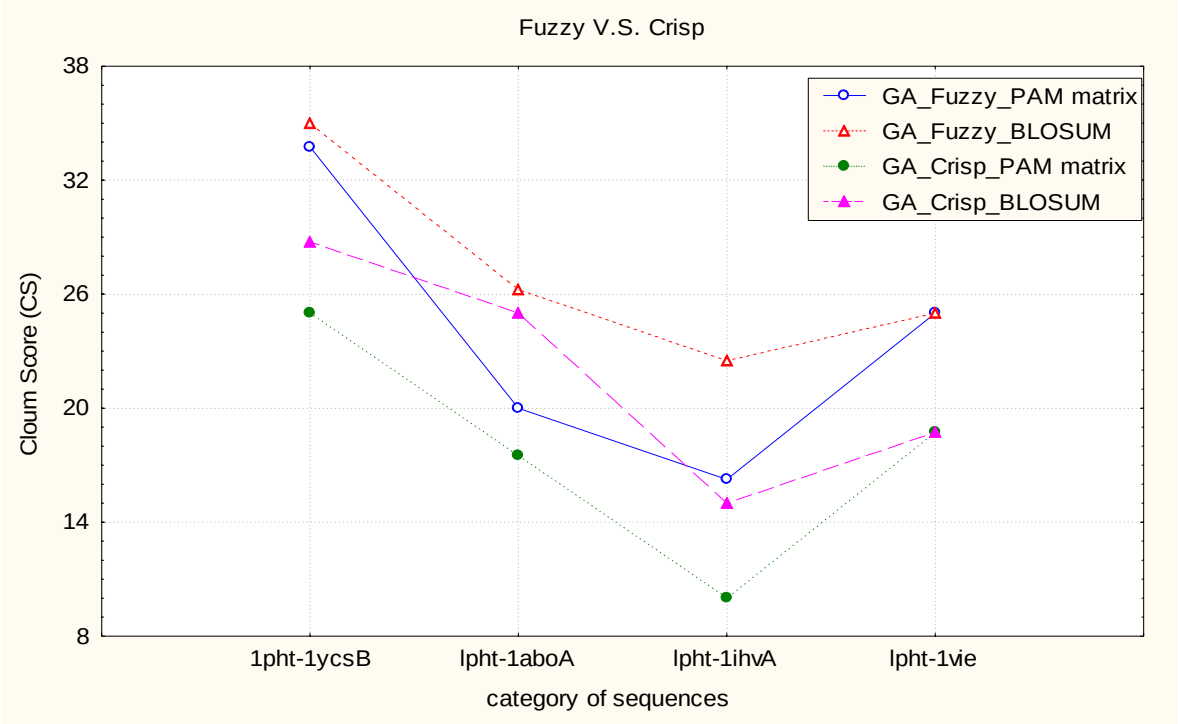

圖 7 GA-crisp與GA-fuzzy應用不同矩陣比較圖(連續最長配對CS)

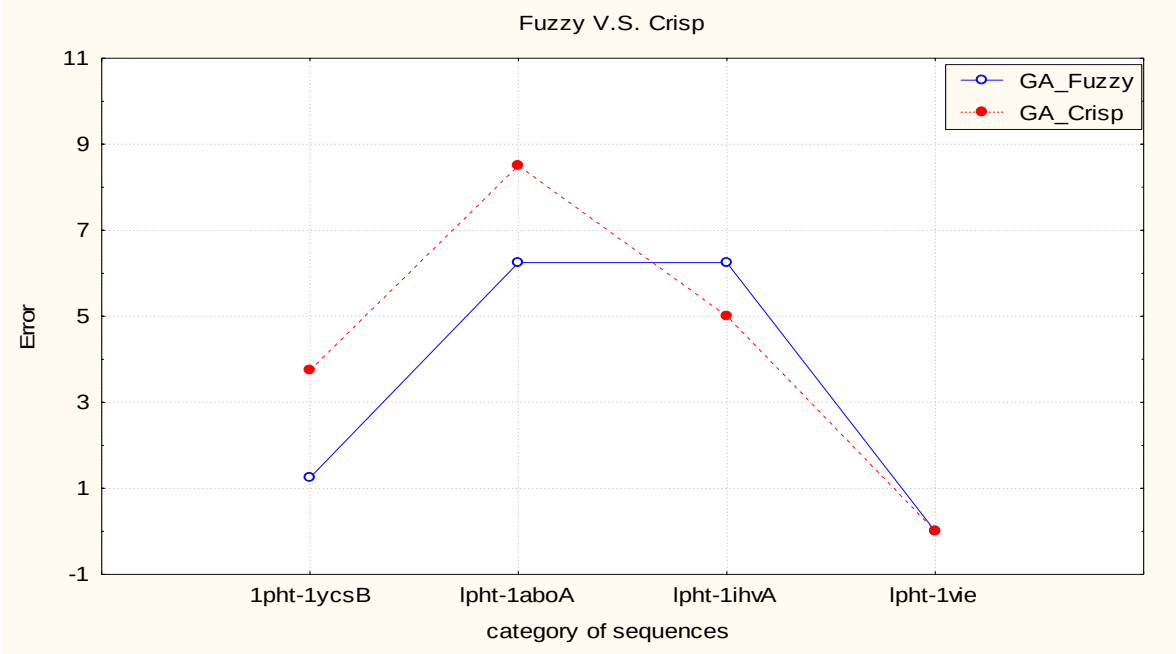

圖 8 GA-crisp與GA-fuzzy誤差比較圖(連續最長配對CS)

5. 結論

在本研究中提出一種先進的方法來進行蛋白質序列排比。首先，我們使用模糊邏輯去建立模糊PAM與BLOSUM矩陣，接著我們使用模糊算數所估計出的分數去計算基因演算法中的適應函數值。實驗結果發現，不論應用哪種矩陣，GA-fuzzy皆能夠找到較長的連續基因配對序列，且由於PAM與BLOSUM矩陣中包含一些不確定因子，因此本實驗結果證明模糊邏輯確實能有效的處理不確定性問題並成功的應用在蛋白質序列排比上。這個先進的方法可在相關研究中，提供不同的觀點。

附錄

PAM250 矩陣

		ORIGINAL AMINO ACID																			
		C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
R	C Cys	12	0	-2	-3	-2	-3	-4	-5	-5	-5	-3	-4	-5	-5	-2	-6	-2	-4	0	-8
E	S Ser	0	2	1	1	1	1	1	0	0	-1	-1	0	0	-2	-1	-3	-1	-3	-3	-2
P	T Thr	-2	1	3	0	1	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	-3	-3	-5
L	P Pro	-3	1	0	6	1	-1	-1	-1	-1	0	0	0	-1	-2	-2	-3	-1	-5	-5	-6
A	A Ala	-2	1	1	1	2	1	0	0	0	0	-1	-2	-1	-1	-1	-2	0	-4	-3	-6
C	G Gly	-3	1	0	-1	1	5	0	1	0	-1	-2	-3	-2	-3	-3	-4	-1	-5	-5	-7
E	N Asn	-4	1	0	-1	0	0	2	2	1	1	2	0	1	-2	-2	-3	-2	-4	-2	-4
M	D Asp	-5	0	0	-1	0	1	2	4	3	2	1	-1	0	-3	-2	-4	-2	-6	-4	-7
E	E Glu	-5	0	0	-1	0	0	1	3	4	2	1	-1	0	-2	-2	-3	-2	-5	-4	-7
N	Q Gln	-5	-1	-1	0	0	-1	1	2	2	4	3	1	1	-1	-2	-2	-2	-5	-4	-5
T	H His	-3	-1	-1	0	-1	-2	2	1	1	3	6	2	0	-2	-2	-2	-2	-2	0	-3
	R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6	3	0	-2	-3	-2	-4	-4	2
A	K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5	0	-2	-3	-2	-5	-4	-3
M	M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6	2	4	2	0	-2	-4
I	I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	2	4	1	-1	-5
N	L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	2	2	-1	2
O	V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	-1	-2	-6
	F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	7	0
A	Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	0
C	W Trp																				
I		-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
D																					

BLOSUM62矩陣

		ORIGINAL AMINO ACID																			
		C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
REPLACEMENT AMINO ACID	C Cys	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
	S Ser	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
	T Thr	-1	1	5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-2
	P Pro	-3	-1	-1	7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
	A Ala	0	1	0	-1	4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3
	G Gly	-3	0	-2	-2	0	6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2
	N Asn	-3	1	0	-2	-2	0	6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4
	D Asp	-3	0	-1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
	E Glu	-4	0	-1	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3
	Q Gln	-3	0	-1	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
	H His	-3	-1	-2	-2	-2	-2	1	-1	0	0	8	0	-1	-2	-3	-3	-3	-1	2	-2
	R Arg	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
	K Lys	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-2	-3	-2	-3
	M Met	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	1	0	-1	-1
	I Ile	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	0	-1	-3
	L Leu	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	1	0	-1	-2
	V Val	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
	F Phe	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
	Y Tyr	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
	W Trp	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

參考文獻

[1] S.F. Altschul, R.J. Carroll, and D.J. Lipman, "Weights for data related by a tree," *Journal of Molecular Biology*, Vol. 207, 1989, pp 647-653.

[2] S.F. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "A basic local alignment search tool," *Journal of Molecular Biology*, Vol. 215, 1990, pp 403-410.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, Z.W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, Vol. 25, 1997, pp 3389-3402.

[4] R. Blankenbecler, M. Ohlsson, C. Peterson, and M. Ringner, "Matching protein structures with fuzzy alignments," *PNAS*, Vol. 100, 2003, pp 11936-11940.

[5] C. Carleos, F. Rodriguez, H. Lamelas, and J.A. Baro, "Simulating complex traits influenced by genes with fuzzy-valued effects in pedigreed populations," *Bioinformatics*, Vol. 19, 2003 pp 144-148.

[6] P.-T. Chang, "Fuzzy strategic replacement analysis," *European Journal of Operational Research*, Vol. 160, 2005 pp 532-559.

[7] K.M. Chao, "On computing all suboptimal alignments," *Information Sciences*, Vol. 105, 1998, pp 189-207.

[8] K.M. Chao, and W. Miller, "Linear-space algorithms that build local alignments from fragments," *Algorithmica*, Vol. 13, 1995, pp 106-134.

[9] K.M. Chao, R.C. Hardison, and W. Miller, "Recent developments in linear-space alignment methods: A survey," *Journal of Computational Biology*, Vol. 1, 1994, pp 271-291.

[10] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, *National Biomedical Research Foundation*, Vol. 5, 1978, pp 345-352.

[11] D. Dubois, and H. Prade, *Fuzzy Sets and Systems, Theory and Applications*, Academic Press, 1980.

[12] A. Heger, and L. Holm, "Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins," *Bioinformatics*, Vol. 19, 2003, pp i130-i137.

[13] S. Henikoff, and J. G. Henikoff, "Amino acid substitution matrices from protein blocks,"

Proceedings of the National Academy of Sciences of the United States of America, Vol. 89, 1992, pp 10915-10919.
[14] X. Huang, and W. Miller, "A time-efficient, linear-space local similarity algorithm," *Advances in Applied Mathematics*, Vol. 12, 1991, pp 337-357.
[15] Y. Huang, and Y. Li, "Prediction of protein subcellular locations using fuzzy k-NN method," *Bioinformatics*, Vol. 20, 2004, pp 21-28.
[16] N.D. Hung, Y. Ikuo, Y. Kunihito, and Y. Moritoshi "Aligning multiple protein sequences by parallel hybrid genetic algorithm," *Genome Informatics*, Vol. 13, 2002, pp 123-132.
[17] A. Kaufmann, and M.M. Gupta, *Fuzzy Mathematical Models in Engineering and Management Science*, Elsevier, 1988.
[18] W.-V. Leekwijck, and E.-E. Kerre, "Defuzzification: criteria and classification," *Fuzzy Sets and Systems*, Vol. 108, 1999, pp 159-178.
[19] M. Mizumoto, and K. Tanaka, "The four operations of arithmetic on fuzzy numbers," *Systems Computers Controls*, Vol. 7, 1976, pp 73-81.
[20] S. Nahmias "Fuzzy variables," *Fuzzy Sets and Systems*, Vol. 1, 1978, pp 97-111.
[21] S.B. Needleman, and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, Vol. 48, 1970, pp 443-453.
[22] H.D. Nguyen, I. Yoshihara, K. Yamamori, and M. Yasunaga, "Aligning multiple protein sequences by parallel hybrid genetic algorithm," *Genome Informatics*, Vol. 13, 2002, pp 123-132.
[23] J.J. Nieto, and A. Torres, "Midpoints for fuzzy sets and their application in medicine," *Artificial Intelligence in Medicine*, Vol. 27, 2003, pp 81-101.
[24] J.D. Thompson, F. Plewniak, and O. Poch, "BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs," *Bioinformatics*, Vol. 15, 1999, pp 87-88
[25] A. Torres, and J.J. Nieto, "The fuzzy polynucleotide space: basic properties," *Bioinformatics*, Vol. 19, 2003, pp 587-592.
[26] L.A. Zadeh, "Fuzzy sets," *Information and Control*, Vol. 8, 1965, pp 338-353.