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

以演算式計算推測蛋白質立體結構之研究 (2/2)

計畫類別： ☑ 個別型計畫 ☐ 整合型計畫
計畫編號： NSC91－2213－E－002－122－
執行期間： 91 年 8 月 1 日至 92 年 7 月 31 日

計畫主持人： **高成炎 教授**
共同主持人： **黃明經 博士**
計畫參與人員： **蔡元芳 博士，蔡其杭**

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：**台灣大學資訊工程系暨研究所**

中 華 民 國 92 年 10 月 1 日

關鍵詞(keywords): Human genome project, Family Competition Evolution Algorithm, Rotamer Library, Lattice Model, ab initio, Tertiary protein structure, Secondary Structure Prediction, HP model

一、 中文摘要:

蛋白質的功能主要決定於它的3D立體結構，我們嘗試以家族競爭式演化方法 (Family Competition Evolutionary Algorithms) 從蛋白質序列的觀點，來預測蛋白質的立體結構，此演算法的目的是專門解決最佳化問題。我們設計以立體晶格模型 (Lattice model) 顯示蛋白質結構，並使用氨基酸親水及疏水的特性來表示晶格上的氨基酸，這種模型稱爲 HP 模型，HP 模型即是將蛋白質簡化成兩個字母，P表示爲有極性 (polar) 且爲親水性，H表示爲非極性 (hydrophobic) 且爲疏水性。

本計畫中，我們在晶格模型中可以輸入氨基酸序列或是HP序列，家族競爭式演化方法將透過重組操做 (Recombination Operator) 及突變運算 (Mutation Operator) 產生不同的子代，我們將進行關於凝聚函數、殘基間相互影響的能量函數及親水疏水能量的計算，每個所產生的結構都將被運算，根據運算的結果，我們將取得能量最低且穩定的結構作爲我們預測的結果，經過幾代的演進之後，最後得到穩定的結果，這即是最佳的立體結構。使用者可以透過我們以圖形視覺化介面的工具看到整個折疊模擬的過程，也可以視需要調整觀看的結構角度，最後將顯示出最穩定的結構。

二、 英文摘要:

Protein functions are decided by their three-dimension structure. We have explored the application of family competition evolutionary algorithms (FCEA) to the determination of protein structure from sequence. The algorithm is proposed to solve optimization problem. The protein structures are showed with a square lattice of we design. The protein component uses Hydrophobic-hydrophilic model (HP model) to calculate structure energy. The protein components are reduced to a binary class: P for polar, hydrophilic monomers, and H for non-polar, hydrophobic monomers.

We can input amino acid sequence or HP sequence to lattice model in our system. The operation of recombination and mutation of FCEA will produce many different offspring. We will calculate structure energy that are distance function, residues-residues interaction function and HP function on every three-dimension structure of lattice model. According to the calculated result, the system will select the lowest energy. After many generations, the result will be stable status and optimization structure. The user can see the process that folding simulations and adjust the structure angle in our visual tool. The last stable structure will be show in the monitor.

三、 前言:

3.1 動機 Motivation

生物體是倚靠體內蛋白質的運作來維繫生命現象,目前對於自然界如何從蛋白序列轉變成立體結構的過程並不瞭解,因此,便有一些研究學者設計各種實驗方法,試圖瞭解蛋白質形成立體結構的過程,想藉著這些方法來進行蛋白質立體結構的預測。

基於這樣的想法,我們運用由基因演算法 (Genetic Algorithm) 所發展出的家族競爭式演算法 (family competition evolutionary approach algorithm) ,再加上親水疏水性的模型 (Hydrophobic -hydrophilic model) 來進行蛋白質折疊模擬 (Protein Folding Simulation) ,希望結合目前科技,試著以先進的電腦軟硬體技術輔助,來模擬蛋白質由序列轉變成立體結構的過程,並以最少時間及成本預測出蛋白質三維立體結構。

3.2 概論 Overview

蛋白質結構預測已經研究超過 30 年,想要藉由一些實驗來瞭解蛋白質所形成的三維立體結構,成本高又耗時。解一個蛋白質結構往往需要很長的時間,先前 IBM 的超級電腦 (Super Computer) 模擬一個小的蛋白質結構也需要一年的時間,所以準確的蛋白質結構預測是十分困難的工作,需要考量很多因素。

有一些預測結構的方法是從蛋白質序列出發,運用一些技術來作預測立體結構。以前的實驗中,蛋白質序列通常已經折疊的結構中所取得[5,6],然而當時的實驗都是以小型蛋白質切成多個小片段來取得資料,再組合這些資訊做出完整的結構,蛋白質折疊的方法便成爲蛋白質形成立體結構的關鍵[21,35],原則上,這些計算都需要靠功能強大的電腦的幫助,在折疊模擬的過程中挑出幾個可能的結構作爲預測的結果,至於預測的結果好不好,則分別計算這些結構的能量,以能量最穩定的結構作爲最後選擇的答案[2,8],爲了節省計算的時間,有些研究人員簡化了計算能量的函數及 3D 空間所帶來的複雜度,希望能解決這個關於複雜計算量的問題[19,22,23,36], 後來便有人提出以晶格(Lattice) 的表示法來表示立體結構中每一個氨基酸的位置[12,14,20],及精確的描繪出蛋白質主鍊結構預測的設計方法[11,27,28]。

目前有許多研究機構投入蛋白質結構預測 (Protein structure prediction) 的工作,每個機構都採用不同的方法加以研究,有些單位是結合蛋白質二級結構 (Protein secondary structure) 的資訊;有些使用 Lattice 模型來表示蛋白質結構;有些固定蛋白質主鍊 (backbone),使用支鍊 (side-chain) 的資訊等,選擇操作模型後,最後再以演算法來解決組合性爆炸的問題,這些演算法包含蒙地卡羅 (Monte Carlo) 隨機法、遺傳演算法 (Genetic Algorithm) 及動態規劃 (Dynamic programming) 等來進行預測,其中的 Lattice 模型即是本篇所使用的方法。

Lattice 模型是蛋白質結構的簡化,以晶格的方式來表示蛋白質在空間中的位置,這樣的想法廣泛的用在蛋白質序列設計的研究上,如 Ron Unger 及 John Moult[27]使用遺傳演算法,蒙地卡羅及使用 Lattice 模型來作蛋白質折疊模擬,遺傳演算法在裡是扮演最佳化的角色,擔任搜尋並安排最有可能的組合,在每一代的演進過程中進行,蒙地卡羅演算法則擔任隨機產生參數來選擇部分需要調整的結構,同樣的在做蛋白質結構模擬,Tianzi Jiang[31]則是使用遺傳

演算法來進行結構最佳化，塔布搜尋法 (tabu search) 紀錄先前的最佳搜尋結果及可能發生的錯誤，避免結果落在局部最佳解 (local optimization)，最後可以得到最好的蛋白質立體結構。

本報告主要架構爲黃仲銘碩士的論文研究。

四、 研究目的:

自然界生物有機體的蛋白質結構 (Protein structure) 都被賦予幫助生存的特別功能，這些蛋白質會因爲不同形狀的 3D 立體結構 (three- dimensional structures) 而有特別的功能及負責的任務，因此，3D 立體結構將決定蛋白質的功能及作用。而 3D 結構是由 1D 結構，也就是蛋白質氨基酸序列所組成的鍊狀物，由線性的形狀透過本身機制折疊(fold)成 3D 結構，折疊的機制是目前許多研究者研究的方向，希望能夠瞭解爲什麼可以蛋白質在短短的幾毫秒時間能夠形成立體結構的原因。

蛋白質結構折疊的問題在計算分子生物學裡頭是一個非常重要且有名的難題。即使簡化成 HP 模型，在計算理論上亦是 NP-Hard 的問題，每一個氨基酸可能在任意角度的位置上安排下一個氨基酸的位置，假若分成三百六十個刻度，即一度一個位置，以一個序列長度爲一百的蛋白質爲例，就有 $(360*180)^{100}$ 種角度組合，若簡化角度的表示，以晶格九十度爲一個刻度，可以旋轉的狀態一共有五個方向，也有 5^{100} 種結構組合，故用平常的演算法無法在短時間內得到好的結果。

本系統使用演化式演算法來找出最佳化的結構，再配合設計的操作，能量評估等化學特性，希望能解決組合性爆炸的問題，找出最佳解供從事蛋白質結構研究的研究員一些幫助，也可以加速及減少實驗上的成本，讓研究員可以在 X 光晶體繞射 (x-ray) 或核磁共振 (nuclear magnetic resonance ,NMR) 得到立體結構的方式外，多了一項選擇。

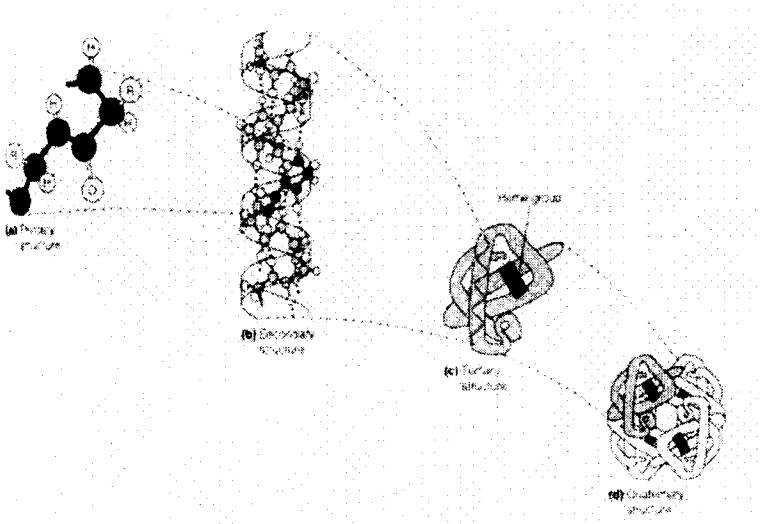
五、 文獻探討:

5.1 生物學與蛋白質 Biology and Protein

蛋白質是維持生物體的生理機制及各種催化的重要元素,大多的生命現象與新陳代謝都是仰賴有功能的蛋白質,所以蛋白質是使得細胞內各項活動得以順利進行的主要關鍵,不同的蛋白質有不同的功能,如酵素 (Enzyme),就是催化或加速生化反應的蛋白質。

蛋白質是由小分子單位,也就是氨基酸,一個一個連接而成,其結構可以分一級到四級結構,一級結構 (primary structure) 即是氨基酸序列,是根據 DNA 的核甘酸序列轉譯而來;二級結構 (secondary structure) 是由某一些固定的形狀所組成,他們主要的構成力量是氫鍵,序列中兩個 α 碳上的 R 基團 (R-group) 與前後相鄰基團的引力或斥力,使得兩相鄰胜鍵平面間的轉動,限制在一定角度範圍,因而產生兩種常見的構形:α 螺旋 (α helix)、β 帶狀 (β sheet),連接 α 螺旋或 β 帶狀時,胜肽鏈以接近 180 度的方式劇烈摺返,這些稱爲轉折點 (turns),而其他的不規則的連結片段我們稱爲任意形 (random coil);三級構造 (tertiary structure) 則爲分子內各部分的二級構造再相互組合,其構成的外型接近完整球形,有一些三級結構還會有非蛋白質的有機小分子、金屬離子等會來加以修飾,一般構成其結構的穩定作用力有靜電力 (Electrostatic interaction),如離子鍵 (ionic bond)、氫鍵 (hydrogen bond),另外有雙硫鍵 (Disulfide bond) 及分子之間的作用力 (Van der Waals forces),其中結構內部的疏水鍵就是本篇所要探討的主題;四級構造 (quaternary structure) 則是由數個相同或不同的三級構造分子,再結合成一較大的複合體,這樣組成四級結構才能進行完整的活性功能。

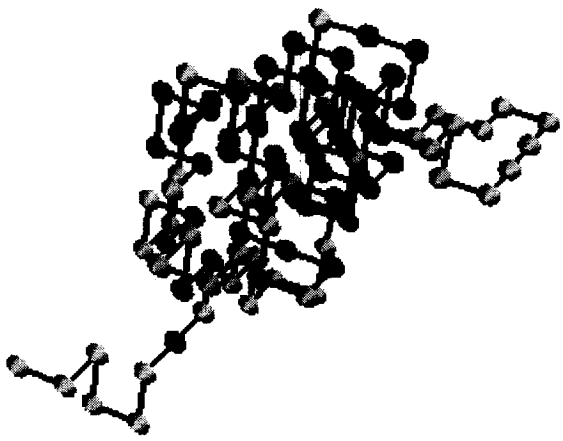
自然界中蛋白質氨基酸序列由線性形成立體結構的過程稱爲蛋白質折疊 (Protein folding),這個結構是因爲氨基酸間經由轉動或移動,也就是多肽鏈 (polypeptide chain) 折疊或聚合之後,最後形成穩定的鍵結。蛋白質折疊的過程所需的時間大約是幾毫秒(μs) 至幾微秒 (ms) 之間,如此短暫的時間便可以完成結構,而其中的折疊過程目前還在持續的研究中。



圖一:蛋白質結構

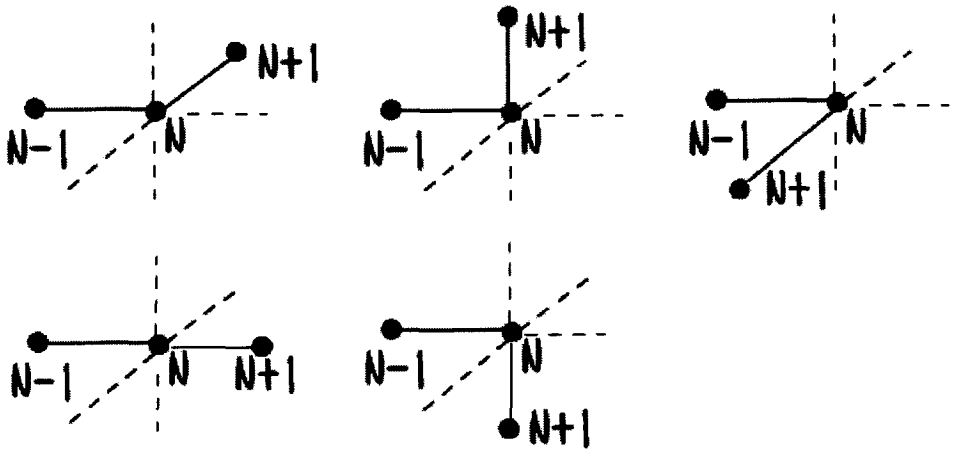
5.2 晶格模型與親水疏水模型 Lattice model and H-P model

本研究使用親水—疏水模型 (Hydrophobic-Hydrophilic Model) 簡稱 H-P 模型，這是由許多晶格（cube）所屬組成，每一個晶格爲立方體，立方體上的點代表的是蛋白質序列中該位置爲 20 種氨基酸(Amino acids)的哪一種，H-P 模型將這些氨基酸分成兩種型態，一種是疏水性(hydrophobic)，簡稱爲 H，大多存在立體結構的內部；另外一種爲親水性 (hydrophilic) 或是帶極性（polar)，簡稱爲 P，大多存在於結構的表面 (protein surface)，常與水溶液接觸，被水中的帶電離子吸引。如圖一所示，黑點代表的是 H，白點代表的是 P，蛋白質的立體結構便可由 H-P 模型所組成。



圖二：H-P 模型

H-P 的運作方式是將氨基酸放在立方體的頂點上，氨基酸變換位置的方式是以 90 度來改變，例如 N 的下一點 N＋１，可以擺放在 N 的上、下、左、右、前五個方向，如圖三所示：



圖三：Lattice 模型的操作

5.3 演化式計算 Evolutionary Computation

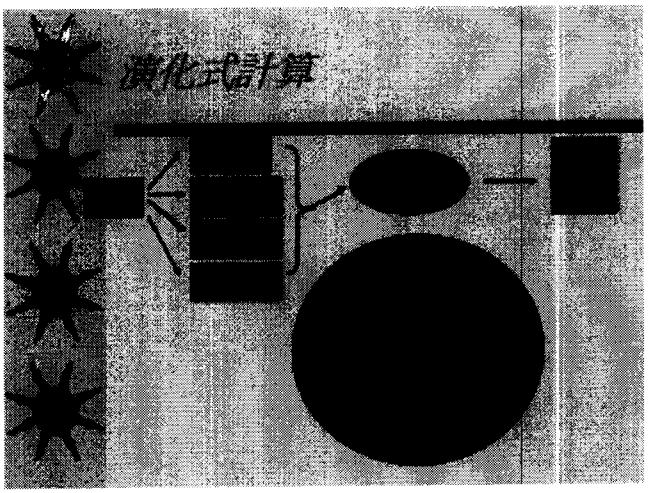
在解最佳化的問題時，往往會採用演化式演算法來計算，常見的算法包含動態環境學習的分類式系統(Classifier system)、數值分析爲主的演算法策略(Evolution Strategy)、人工智慧演算式規劃(Evolutionary Programming)、家族競爭方式的演化式方法(Family Competition Evolutionary Approach)、遺傳演算法(Genetic algorithm)、人工智慧型的遺傳規劃(Genetic programming)。

5.3.1 簡介說明及概要 Introduction

由人工智慧 (Artificial Intelligence，簡稱 AI) 研究領域所發展出來的運用生物演化方式的一項學科—演化式計算 (Evolutionary Computation，簡稱 EC) ，可以快速的找到最佳解，其原理是來自生物學家達爾文(Darwin)提出的「物競天擇，適者生存」的演化論，目前依照自然界生物演化的機制，這樣的構想發展出幾套演算法。

目前生存在地球上生物體系的生物群體 (population) 中的個體 (individual) 在有限資源的環境下，必須爲了生存 (survival) 而競爭 (competition)。在達爾文的理論架構下，只有強者或稱勝利者有生存下去的機會，也會經由繁衍後代來維繫生物型態的維繫，而失敗者將在被各種條件篩選之下被淘汰出局，勝利者將透過生物的各種遺傳(genetic) 機制繁衍後代，這樣的子代與父代有一定的相似度，也有可能因爲較大的差異而產生新的物種，因此，在這樣的機制下所產生的後代將演化成越來越能適應環境，每一個世代 (generation) 的調整也有助於適應當前的生存環境。

根據上述的演化機制，所發展的各類演化式計算方法程序大致如下 (如圖四)：
1.選擇及設定初值 (initialization) 群體。
2.評估 (evaluation) 群體中的個體，決定結束程序的條件。
3.對於群體中的個體依據適合度 (fitness function) 做選擇 (selection)，評估中適存度較高的個體較易生存，淘汰部份個體。
4.以突變(mutation)、交換(crossover)、複製(reproduction) 等遺傳運算 (genetic operation)，產生多個新的變異個體加入群體，而得到下一代的群體。
5,重複步驟 2-4。



圖四：EC 一般的方式

針對各種問題不同的特性來設計方法，以最適當的方式設計不同的行爲模式、個體結構、和群體組織，以期待能更快速地解決問題。

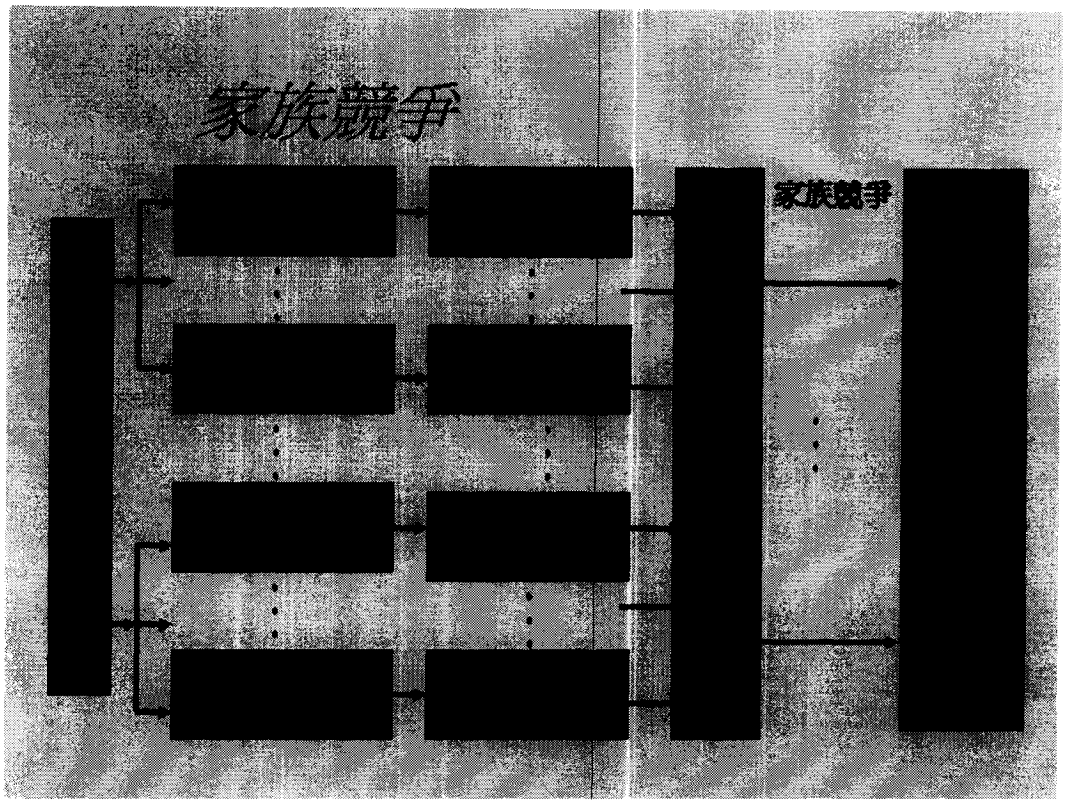
5.3.2 家族競爭式演算法 Family Competition Evolutionary Approach

何謂家族競爭演化式，在最佳化問題中最難的是解決全域性最佳化(Global optimization)的問題，家族競爭演化式(Family Competition Evolutionary Algorithm)，簡稱 FCEA，此方法以家族競爭 (family competition) 及適應規則 (adaptive rule) 爲

基礎，並整合遞減式 (decreasing-base) 及自我調整 (self-adaptive) 的突變 (mutation)，使 FCEA 能同時具有區域性最佳化及全域性最佳化兩者之長處的搜尋策略。我們將這一些策略結合之後，便具有全域性問題最佳化的解決能力，並在較短的時間內得到好的結果。

FCEA 是一個可以解決全域性最佳化問題的演算法，它整合多個突變的運算，這套演算法是由交通大學生命科學系楊進木教授於台灣大學資訊工程研究就所讀博士班時所發展出來的。FCEA 的執行，根據初值隨機產生 N 個解答，接著進入主要競爭的循環，意圖找出最佳解，每一代均會經過三次的處理，每一次的處理包含重新組合 (recombination) ，突變 (mutation)，家族競爭 (family competition) 及最後的選擇。

何謂家族競爭(family competition)？簡單的說就是在同父異母的兄弟姊妹中，只有一個能存活下來的淘汰法則。家族競爭的處理方式如圖六，一共有 L 個子代被作重組及突變的運算，而最後經過選擇後，只有最佳解會存在，其他的解將在競爭之下淘汰，如同達爾文的適者生存的理論般。



圖五：家族競爭的步驟

每一個獨立的染色體 (In)，我們稱爲這一個家族的父代，這個父代配合產生的機率 Pc 隨機產生一群解(I11...I1l)，對於這一些解在組合後新結果，這些產生的子代 (I11...I1l)及原來的家族父代再經由突變後，選出最佳解(C1_best)。

我們定義經由 FCEA 處理的資料稱做染色體(chromosome representation)，染色體的資料格式爲由使用者訂定，如果染色體序列有 n 個長度，則資料由 X1...Xn。接著重組操做(recombination operator)，FCEA 有兩種簡單的重組運算，一種是修正式離散的重組方式，以固定的比例重組。另外必須訂定不同的突變運算(mutation operator)，產生不同的子代，突變的方式有三種，分別是以遞減爲基礎的高斯突變 (decreasing-based Gaussian mutation)、高斯自我適應突變 (self-adaptive Gaussian

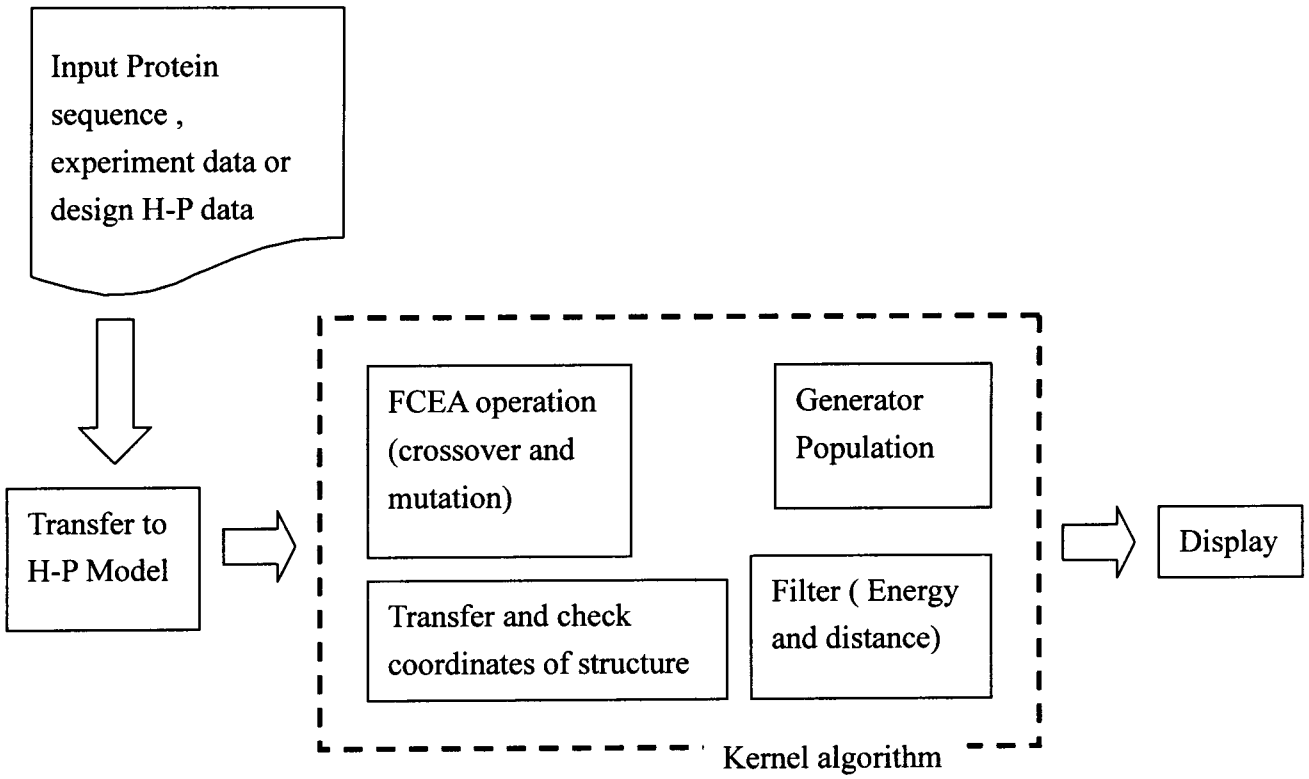
mutation)及歌西自我適應突變 (self-adaptive Cauchy mutation) ，最後還有依照不同的評估方式所規範的兩種適應規則，一是遞減規則 (A-decrease-rule)，如果父代評估結果優於子代中最好的一個，則選擇父代，並將突變率降低，使得內容不至於變的太快，另外一個是遞增規則 (D-increase-rule)，如果父代評估結果比子代的差，則在最好的子代與平均的父代中選擇較佳的，成爲下一個高斯自我適應突變參數，不斷產生子代進行演化，最後趨於穩定則此即爲最佳解的答案。

六、 研究方法與討論:

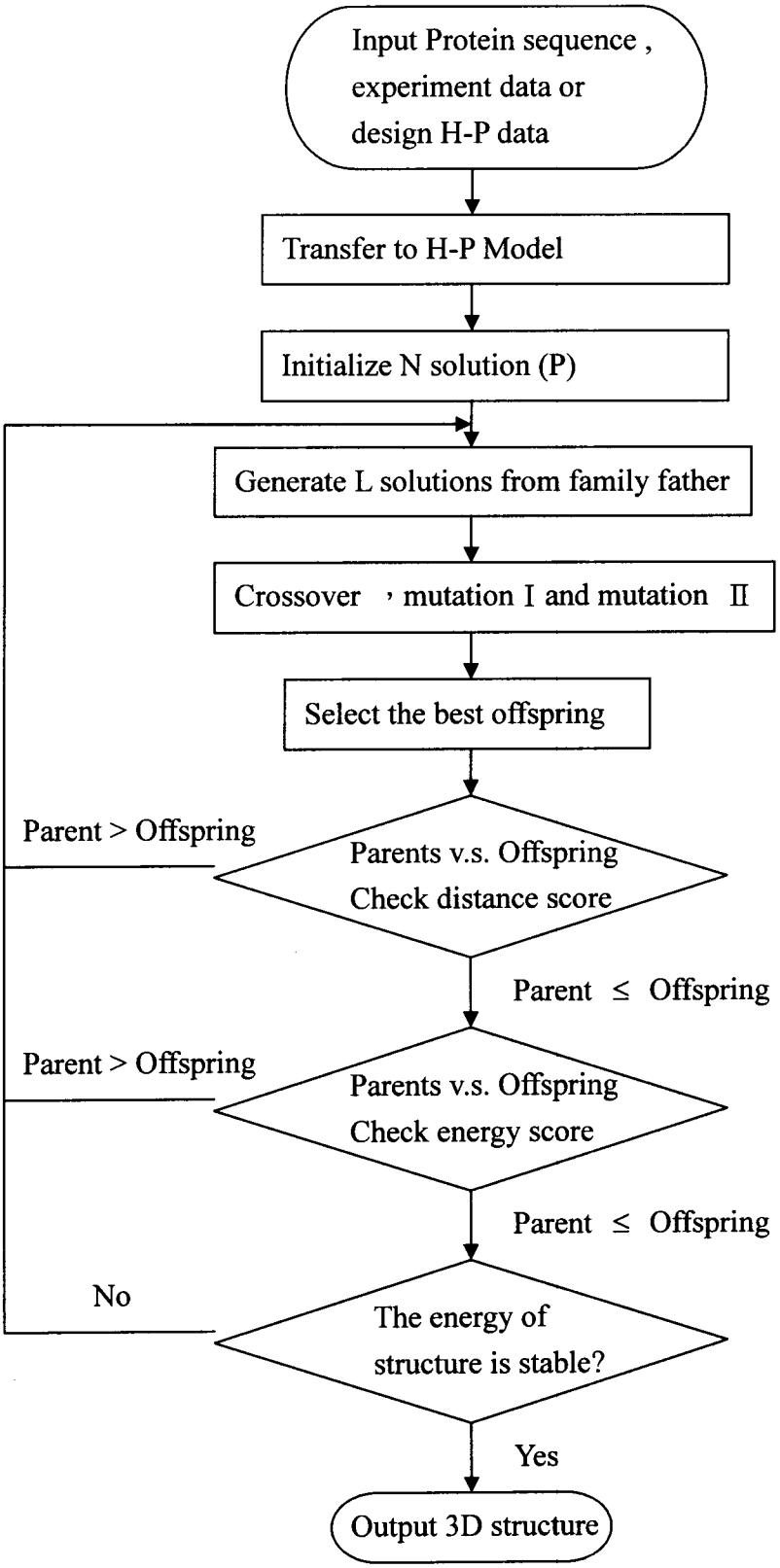
本計畫主要目的是作蛋白質折疊模擬，其中核心的演算法及架構模型分別爲 FCEA 及 Lattice 模型，其中並融入氨基酸的特性加以評估篩選。此系統可以輸入生物實驗所得的氨基酸序列，或是實驗者自行設計的 HP 序列，希望藉由快速的方法模擬並得到最後簡略的蛋白質立體結構。

6.1 演算法流程 Algorithm flowchart

本實驗方法是可以輸入一串蛋白質序列或是 HP 的資料，透過 20 種氨基酸不同的特性，轉換成 HP 模型，經由 Lattice 模型的定位及 FCEA 的核心演算法執行後，最後會得到一組蛋白質 3D 結構圖。以下是本系統的組織圖(圖六)及程式流程圖(圖七)：



圖六：系統架構圖


Input Protein sequence , experiment data or design H-P data
Transfer to H-P Model
Initialize N solution (P)
Generate L solutions from family father
Crossover ，mutation I and mutation II
Select the best offspring
Parents v.s. Offspring Check distance score
Parent > Offspring
Parent ≤ Offspring
Parents v.s. Offspring Check energy score
Parent > Offspring
Parent ≤ Offspring
The energy of structure is stable?
No
Yes
Output 3D structure

圖七：程式流程圖

6.2 晶格模型 Lattice model

利用狄卡兒座標系來定義 Lattice 模型中的位置，並按照順序將氨基酸填入空格中，定義 Non-polar(hydrophobic)爲 1 ，Polar (hydrophilic)爲 0 ，當輸入氨基酸序列時，系統依不同的氨基酸特性指定不同的值，如表一所示：

Amino Acids: Their Properties and Structures					
Polar (hydrophilic)			Non-polar (hydrophobic)		
Amino acid	three letter code	single letter code	amino acid	three letter code	single letter code
Serine	Ser	S	Glycine	Gly	G
threonine	Thr	T	Alanine	Ala	A
cysteine	Cys	C	Valine	Val	V
tyrosine	Tyr	Y	Leucine	Leu	L
asparagine	Asn	N	Isoleucine	Ile	I
glutamine	Gln	Q	Methionine	Met	M
Electrically Charged (negative, hydrophilic)			Phenylalanine	Phe	F
aspartic	Asp	D	Tryptophan	Trp	W
glutamic	Glu	E	Proline	Pro	P
Electrically Charged (positive, hydrophilic)					
Lysine	Lys	K			
arginine	Arg	R			
histidine	His	H			

表一：氨基酸與 HP 轉換表

每一個氨基酸都有 HP 轉換表轉成相對應的 0,1 值，若輸入是 HP 序列，則直接以輸入的資料作爲運算的系統設定值，最後經由計算得到的 HP 的能量估計值，這裡計算的方式是如果是疏水性氨基酸，且其他的疏水性氨基酸相鄰(兩點距離爲 1)不相接，則記爲-1，所以如果一個疏水性氨基酸均同爲疏水性氨基酸，最多可計-4 分。

Lattice 模型的運作方式是將 0,1 值放在每一個點爲結構位置的座標上，而整體形狀改變的時候，也就是座標位置更改，這個座標點只能往前一個座標點的上、下、左、右及前方等五種方向前往，即以九十度或一百八十度的運作方向變化。爲了加快運算的效率，減少所費的時間，此時需將迪卡兒座標系轉換成球面座標，公式如下：

$$\gamma = \sqrt{(x^2 + y^2 + z^2)}$$

$$\varphi = \text{Arctan}\left(z \Big/ \sqrt{x^2 + y^2} \right) \text{ ，其中 } x^2 + y^2 \neq 0$$

$\lambda = \text{Arctan}(y/x)$ ，其中 $(x \succ 0, y \succ 0)$

或是 $\lambda = \pi + \text{Arctan}(y/x)$，其中 $(x \prec 0)$

或是 $\lambda = 2\pi + \text{Arctan}(y/x)$，其中 $(x \succ 0, y \prec 0)$

公式一：迪卡兒座標轉換成球面座標

γ 是代表球面座標的向量長度，λ 是 x-y 軸的平面角度，範圍由 0 度到 360 度，φ 是 z 軸與 xy 平面的角度，範圍爲 0 到 180 度。如果是轉 90 度，只要加上 $\pi/2$，180 度加上 π，270 度加上 $3\pi/2$ 或是-$\pi/2$，即可快速轉換新的座標，要注意的是當 $x=0$ 時，也就是座標點落在 y-z 平面時，必須定義 $\lambda=0$，而 $x=0$ 及 $y=0$ 時，則 $\lambda=0$，φ 則會隨 z 值的正負而爲 $\pi/2$ 或是 $3\pi/2$。座標位置計算完畢後，則透過下面的公式轉換回迪卡兒座標即可。

$x = \gamma \cos\varphi \cos\lambda$

$y = \gamma \cos\varphi \sin\lambda$

$z = \gamma \sin\varphi$

公式二：球面座標轉換成迪卡兒座標

x，y，z 分別是迪卡兒座標系的座標值。

另外 Lattice 模型也存放氨基酸序列資料，我們可以透過表一中 20 種氨基酸相互能量影響對照表，套用能量計算公式(公式三)來計算整體結構的能量，便可以得到最佳化的立體結構。

6.3 家族競爭式演算法 FCEA algorithm

本系統套用最佳化程式演算法來進行立體結構的調整，會就 FCEA 內部參數的設定調整及基本的運算單元加以說明。

6.3.1 染色體代表 Chromosome representation

FCEA 演算法中不可或缺的就是演化中的成員，也就是染色體，本系統中我們對於染色體的設計爲氨基酸的名稱、所在空間中的位置及 HP 代表值、各項需計算及待評估的能量數值，如下表：

每一個氨基酸的資訊					計算及評估值		
氨基酸的名稱	X 座標	Y 座標	Z 座標	HP 代表值	凝聚評估	能量評估	HP 評估

表二：染色體格式

氨基酸名稱及 HP 代表值由使用者輸入，使用者可以選擇其中之一爲輸入，便利使用者使用系統，位置座標資訊的預設值則設定爲線性直線，評估值預設值則爲 0。

6.3.2 家族競爭 Family competition

家族競爭的方式是由使用者設定一共有多少家族(family)，每個家族有多少成員(solution)，並設定在合理演化條件下，一共進行幾代(generation)的演化。

在這裡，家族之間的，每一個家族都有自己的家長，由家長代表繁衍出一定數目的家族成員，再透過演算法的運算，進行最佳化的組合，在家族競爭的環境下，爲了避免

強者恆強，弱者恆弱的情況，本系統內設定隨機參數，可以跳躍家族關係，在跨家族的情形下進行交配(crossover)，讓每個家族經過競爭下，更具有代表性。

6.3.3 適應函數 Fitness function

演化的過程中需要有評估演化結果的評估函數，這裡一共有三種評估函數，包含凝聚評估函數、氨基酸能量評估函數、HP 評估函數。

凝聚評估函數：

考量疏水性的氨基酸是形成蛋白質結構形狀的主要因素，所以疏水性作爲評估的依據之一，評估的方法是考量不同遠近的疏水氨基酸對於該疏水性氨基酸的能量影響，除了氨基酸彼此之間有鍵結(bone)不予計算外，其餘依據距離遠近加以計算，求出總和即爲本系統的凝聚評估參數 ED_{ij}。

公式三：凝聚評估公式

$$ED_{ij}(\text{Distance})=\sum_{i\neq j}^{n}\sigma\frac{1}{\sqrt{AA_i^2-AA_j^2}}\quad ,j\in(0\ldots i\text{-}1,i\text{+}1\ldots n\text{-}1)$$

AAi 及 AAj 是不同位置的疏水性 AA，σ 是調整係數。

能量評估函數：

根據 David A. Hinds and Michael Levitt[12] 從 PDB 裡頭挑出了 246 個蛋白質的鍊結構(chain)，60721 個殘基(residues)，挑選的條件爲超過 60% 序列相異度(sequence identity)，解析度(resolutions)不超過 3.0A，製作了一個 20 個氨基酸的能量對照表(如表一)，透過此對照表在配合能量評估公式 EEij (公式四)，並可評估這樣氨基酸的組合對於整個立體結構是不是最佳化。

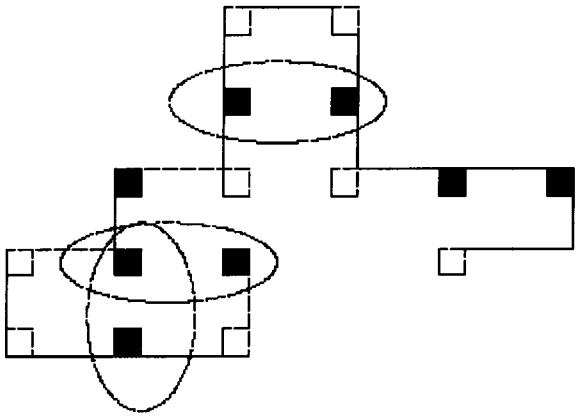
公式四：能量評估公式

$$EEij(\text{Energy})=\sum_{i\neq j}^{n}\frac{2e_{ri,rj}+e_{r(i-1),rj}+e_{r(i+1),rj}+e_{ri,r(j-1)}+e_{ri,r(j+1)}}{6}$$

$e_{ri,rj}$ 是位置 i 與 j 兩個氨基酸之間的能量影響，此公式計算以 i 爲中心的前後 i+1 及 i-1 的兩個位置，與跟 i 距離爲 1 相對 j 前後 j+1 及 j-1 兩個位置 M，能量的相互影響，所以評估能量的計算必須總和五個相互影響的作用力爲最後的答案。

HP 評估函數：

HP 評估爲最常見評估方式，只要是疏水性的位置，相鄰不相接、兩個氨基酸距離爲一且同爲疏水性氨基酸者（如圖八），則計分爲-1 分，統計系統中所有的疏水性氨基酸所形成的能量即爲 HP 評估值。

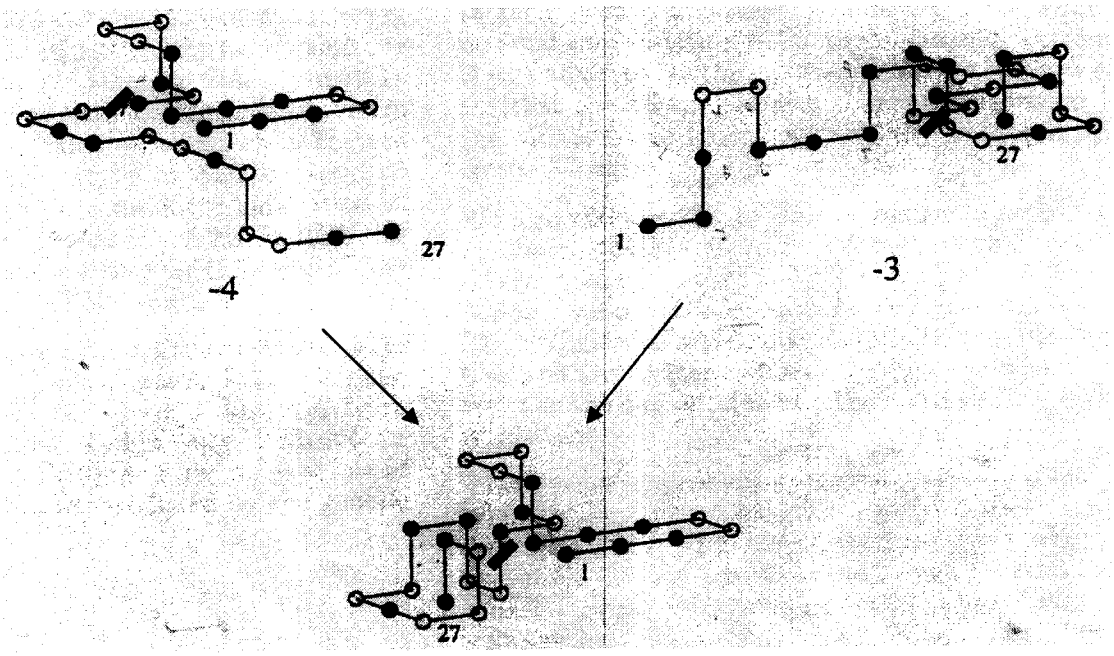


圖八：深色代表疏水性，此HP評估參數爲-3

6.3.4 交配運算 Crossover operation

由隨機函數產生一個值，這個值代表的是欲交換的關鍵位置，再從不同家族中隨機挑選的兩個結構，由這兩個結構中的關鍵點進行交換，如圖九所示。接著進行下一個突變的運算。

其中要注意的是結構碰撞的問題，因爲任意兩個結構進行交配運算時，結合結構之後，都有可能在目前的結構產生碰撞，當產生碰撞時，系統將隨機選擇轉一個方向，再確認結構是否產生碰撞，如果依然無法得到合理的結構，則放棄這個組合，繼續其他結構的交配運算。



圖九：結構交換

6.3.5 突變運算 Mutation operation

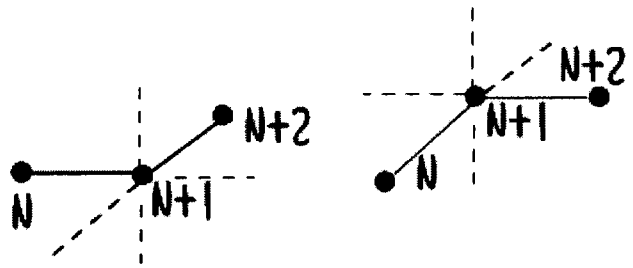
本系統的突變運算的方式有兩種，一種是對於整體作大幅度改變，另外一種是細部結構的微調。

整體突變運算：

隨機選擇一個點，由此點進行結構轉向，在 Lattice 結構中，除了原先的位置外另外有四個方向可以選擇改變(如圖三)，系統在突變運算的過程中，依然會確認結構是否產生碰撞，最後產生沒有碰撞的合理結構。

細部微調運算一：

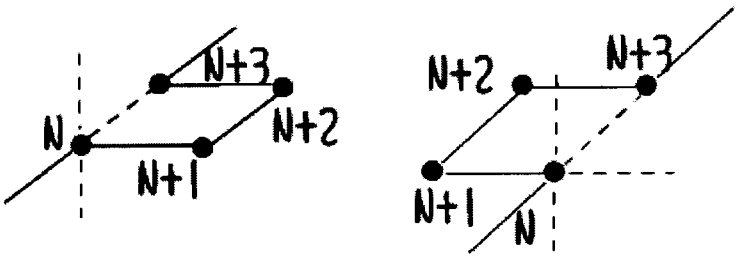
有三個座標點，比方有結構中 N，N+1，N+2 三個點，其中 N+1 要改變位置，N 及 N+2 固定，這樣的運算即可讓 N+1 點就會跑到對角線點的位置上(如圖十)。



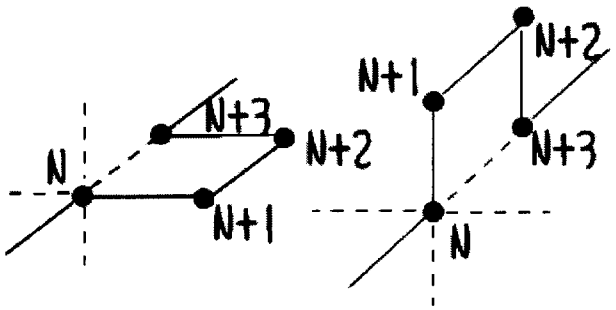
圖十：細部微調運算一(位置改變到對角線上)

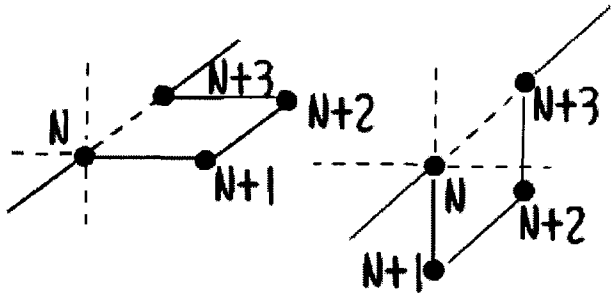
細部微調運算二：

有四個座標點，比方有結構中 N，N+1，N+2，N+3 四個點，其中 N+1 及 N+2 要改變位置，N 及 N+3 固定，這樣的運算就像是以 N 及 N+3 的連線形成圓心軸，N+1 及 N+2 跟這旋轉，就會形成不同的結構位置(如圖十一)。



圖十一之一：細部微調運算二(轉向 180 度)

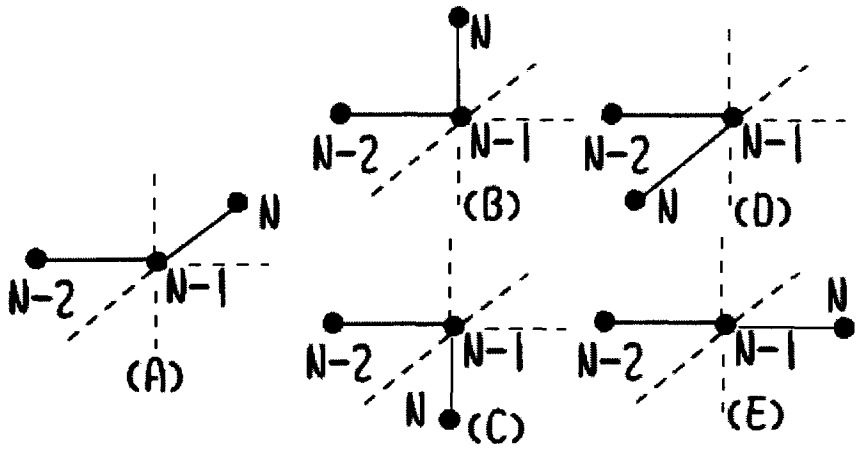




圖十一之二：細部微調運算二(轉向±90 度)

細部微調運算三：

此微調的目的在於改變第一點或是最後一個座標點的位置，比方有結構中 N 爲最後一個座標點，所以隨機在五個方向改變座標位置(如圖十二)。



圖十二：細部微調運算三(位置改變到對角線上)

6.3.6 選擇並產生下一代 Select next population

由適應函數(fitness function)對於結構產生凝聚評估、氨基酸能量評估、HP 評估三項參數，由這些參數在眾多後代中選擇一個最佳的結構成爲下一代的代表，我們評斷的方式先以凝聚評估參數爲考量的第一順位，即先以整體結構的考量爲依據，希望藉由系統的運算將疏水性氨基酸往內部移動，再以提高結構的穩定度，其次再考量氨基酸能量評估及 HP 評估，以化學的角度修正立體結構，使得整體結構符合化學意義及讓結構能量達到最低。

6.4 其他運算 Other operation

因爲蛋白質結構的化學特性，在本系統中另外有兩項設計。

6.4.1 路徑確認 Self-Avoiding Walk

由於蛋白質結構在調整構形時，會產生碰撞的現象，我們必需設計一個函數確認目前結構是否產生碰撞，所以每作一次結構調整的動作時，如交配運算，突變運算，

產生新的子代結構時，就必須確認每個座標點是否重複，也必須確認由序列的第一點座標走到最後一點座標，其中的路線都是往上、下、左、右，0^0、90^0、180^0、270^0的轉角度路徑，在任一點座標不會有經過兩次以上的可能性。

6.4.2 蛋白質折疊性質 Characteristic of protein folding

根據蛋白質結構特性，蛋白質在折疊的時候，疏水性的氨基酸會往內部集中，親水性的氨基酸會往結構表面 (surface) 移動，根據這樣的特性，在突變運算中，爲了要決定轉角的方向，我們先求出結構的中心點，確認目前隨機選取的座標點特性，決定轉角移動的方向，疏水性氨基酸往內部，親水性氨基酸往表面。

6.5 結論與討論 Conclusion and Discussion

測試資料(Test Data)有兩筆，模擬資料(Simulation Data)一共有三筆，分別使用 Kaizhi Yue[18]及 Ugo Bastolla[30]在 PROTEIN 發表的文章數據(如表格四) ，其中數字代表重複的個數，例如 P_2 就是 PP，括弧中代表的是同一組值，括弧外面的數字代表重複的個數，例如$(PH_2)_3$即爲 PH_2 PH_2 PH_2。

Sequence name	Sequence length	Sequence configuration
hp481	48	$HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3HP_8H_2$
hp482	48	$H_4PH_2PH_5P_2HP_2P_2HP_5HP_2HP_3HP_2H_2P_2H_3PH$
hp483	48	$H_3P_3H_2(PH)_2HP(H_2P)2HP_7$ $(HP)_2PHP_3HP_2H_6PH$
hp484	48	$(PH)_2P_4(HP)_3(PH)_2H_5P_2H_3PHP_2HPH_2P_2HPH_3P_4H$
hp60	60	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$
hp1001	100	$P_6HPH_2P_5H_3PH_5PH_2P_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_2HPH_3P_6HPH_2$
hp1002	100	$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2P_9H_9H_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$

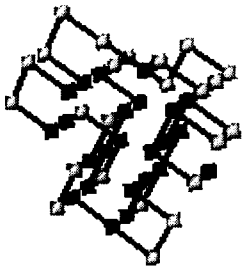
表格四：實驗數據資料

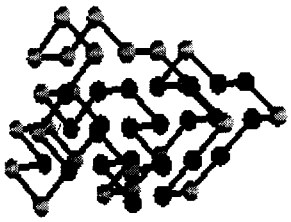
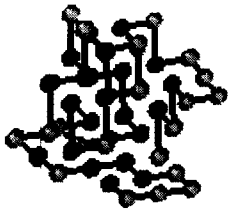
6.5.1 實驗結果 Experiment result

Epns [18]	En (native state energy) [18]	Sequence name	Sequence length	Our experiment result
-30	-32	hp481	48	-31
-30	-34	hp482	48	-30
-30	-32	hp483	48	-31
-30	-34	hp484	48	-31

表格五：實驗結果

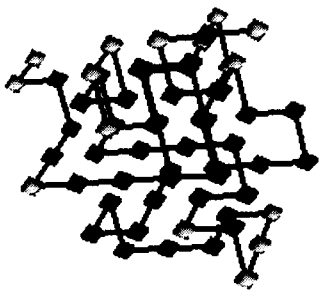
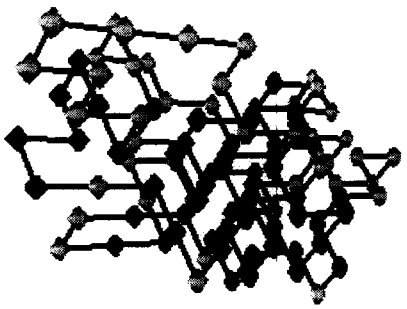
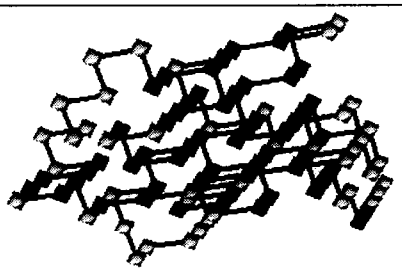
測試實驗穩定蛋白質立體結構：

hp481	hp482
	

hp483	hp484
	

表格六：測試實驗蛋白質立體結構

模擬實驗穩定的蛋白質立體結構：

Hp60	hp1001	hp1002
Generation：115	Generation：601	Generation：258
Distance score：-208.026	Distance score：-290	Distance score：-271
HP score：-44	HP score：-59	HP score：-52
		

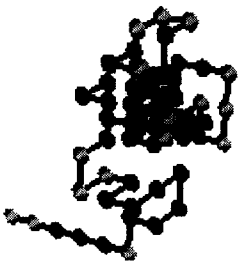
表格七：模擬實驗蛋白質立體結構

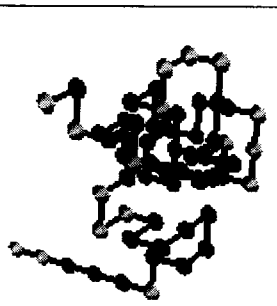
實驗模擬的過程(以 Hp60 爲例)

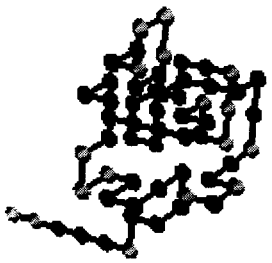
Generation：1 Distance score：-129 HP score：-17	

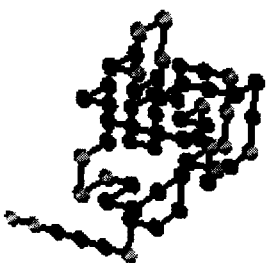
Generation：2 Distance score：-133,947 HP score：-19	

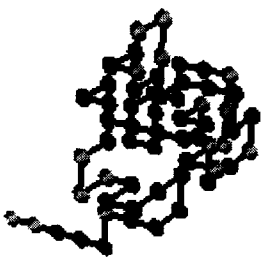
Generation：3 Distance score：-137 HP score：-19	

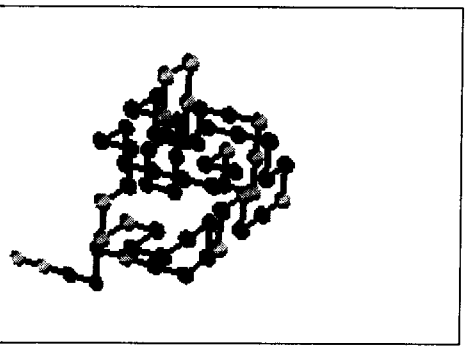
Generation：4 Distance score：-141.252 HP score：-18	

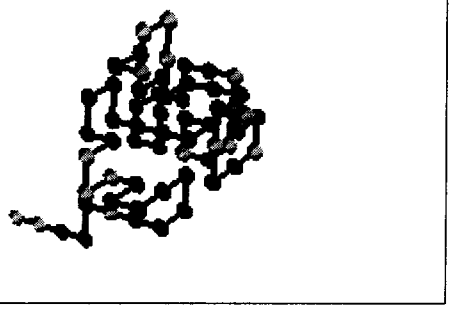
Generation：5 Distance score：-145.988 HP score：-18	

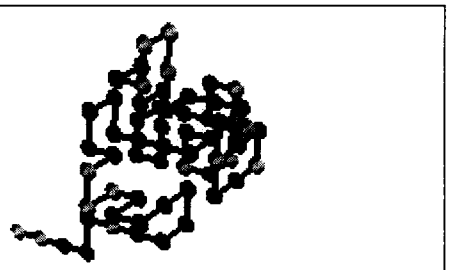
Generation：6 Distance score：-160 HP score：-23	

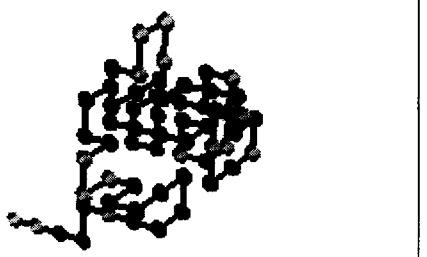
Generation：7 Distance score：-162.481 HP score：-23	

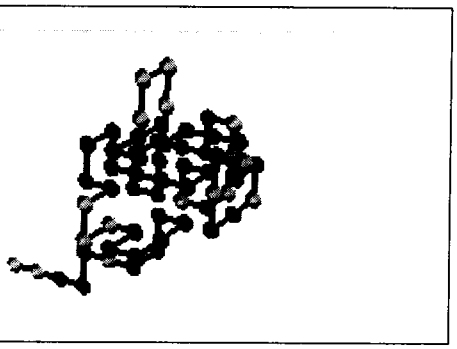
Generation：8 Distance score：-165 HP score：-25	

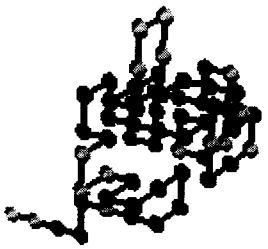
Generation : 11 Distance score : -166 HP score : -25	

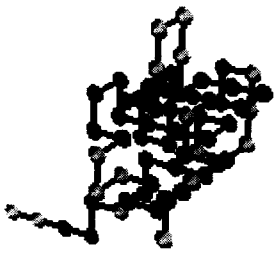
Generation : 18 Distance score : -168 HP score : -28	

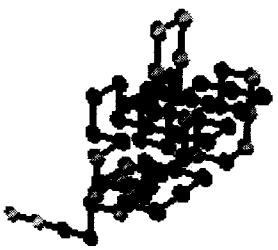
Generation : 19 Distance score : -169 HP score : -28	

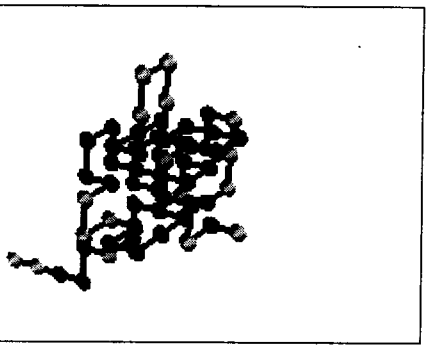
Generation : 20 Distance score : -171 HP score : -31	

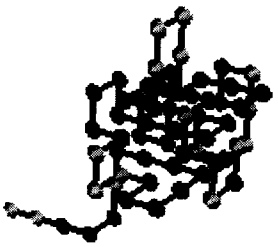
Generation : 22 Distance score : -179 HP score : -32	

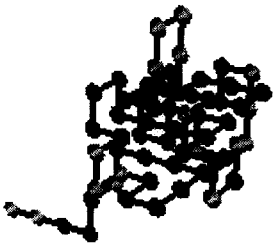
Generation : 23 Distance score : -183 HP score : -33	

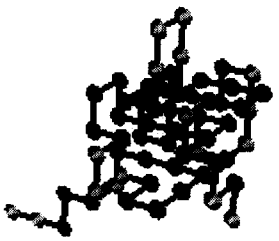
Generation : 54 Distance score : -185.131 HP score : -33	

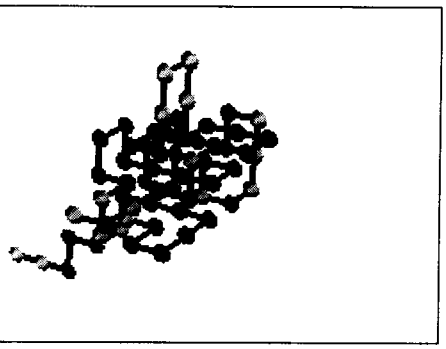
Generation : 55 Distance score : -187.386 HP score : -34	

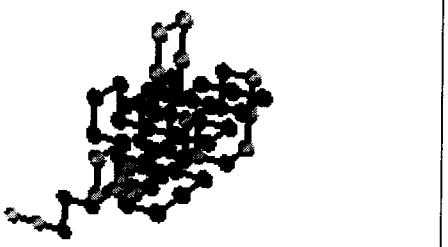
Generation : 56 Distance score : -189,742 HP score : -34	

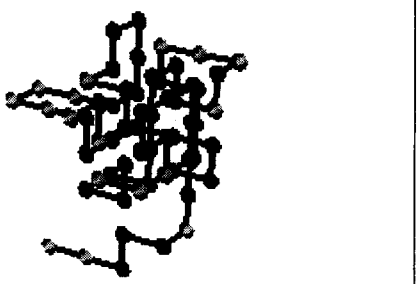
Generation : 80 Distance score : -190 HP score : -36	

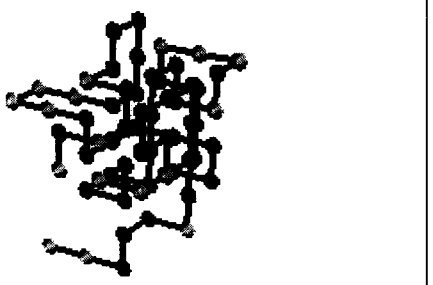
Generation : 81 Distance score : -191 HP score : -36	

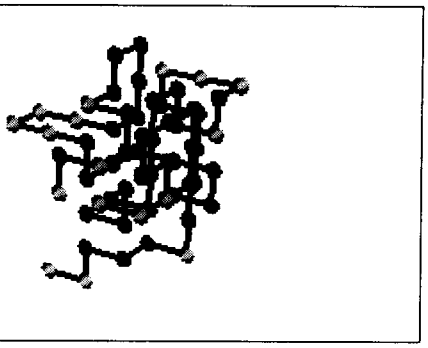
Generation : 83 Distance score : -192 HP score : -36	

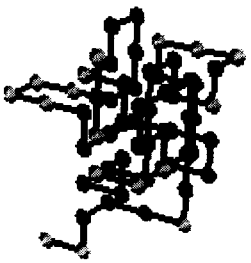
Generation : 93 Distance score : -195.79 HP score : -39	

Generation : 95 Distance score : -199 HP score : -41	

Generation : 107 Distance score : -202 HP score : -42	

Generation : 108 Distance score : -204 HP score : -42	

Generation：110	
Distance score：-207	
HP score：-43	

Generation：115	
Distance score：-208.026	
HP score：-44	

6.5.2 結論與討論 Conclusion and Discussion

模擬蛋白質折疊情況

由於加入一些化學特性，使得研究者可以看到蛋白質從序列一直演變成立體結構的過程，也就是一代代演化過程中，都是依據這些特性演變，但由於目前對於蛋白質折疊的規則還不清楚，而且也少了對環境條件的認識，像自然界蛋白質是在水溶液(solution)中折疊，而目前生物技術中使用 X 光繞射來解析蛋白質結構也不是在水溶液中解析，所以目前僅能在所知的規則中套用到模擬的情況，希望儘量接近自然界蛋白質真實的折疊情況，其次，也許可以在模擬的過程中找出一些規則或是特性，能夠幫助研究者對於蛋白質有更多的認識。

模擬過程的研究

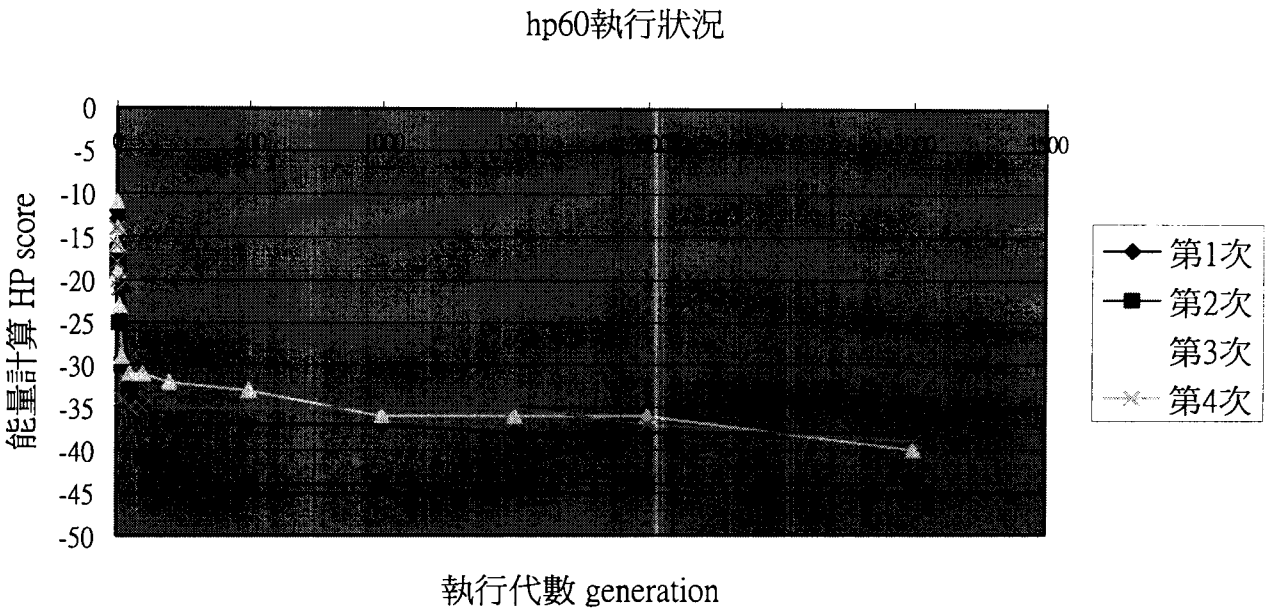
由模擬的結果可以看出，親水性的氨基酸分佈在結構的外圍，疏水性氨基酸大多都在內部，所以最後得到結構都會接近真實的情況，這是由於演算法中套用了氨基酸的化學特性，讓親水的氨基酸產生突變時向表面走，疏水的氨基酸向內走，在模擬的過程中，如果加入更多足以影響實驗結果的參數或是更多的特性，相信可以慢慢接近真實世界的折疊模式。

速度快，成本低

由氨基酸序列到形成蛋白質三級結構，序列越長組合的可能越多，例如氨基酸長度爲 100 的序列中，如果純粹以 Lattice 模型及 HP 的設定模擬，就有可能 5^{100} 的可能發生，不是目前任何一台電腦可以在短時間內考量所有可能性且計算完畢，所以我們運用演化式的

演算法協助我們快速的剔除一些比較不可能發生的結構。

由於演算法的特性是會保留好的結果，由好的結果繼續發展，所以，我們可由序列長度爲 60 的執行結果(圖十三)看出，在程式執行 500 代就已經產生穩定的結果，就已經收斂的固定值。



圖十三：HP60 的執行情況

圖十四及圖十五分別爲序列長度爲 100 的兩組值所做出的圖表，由圖十四中第一次實驗所得的數據曲線可看出，由於蛋白質結構不斷改變，在估計 HP 值時，有時候反而會比較小，這是因爲系統是凝聚參數爲主，要把所有的氨基酸放在該放的位置上，比方說左右各有兩團疏水性氨基酸集團，因爲這兩團個別非常靠近，所以 HP 的值加起來比較大，而系統必須將這兩團氨基酸和在一起，便必須以凝聚函數使這兩團氨基酸相互靠近，所以，在移動的時候，整體 HP 估計值就會比前一次高，可是最後還是會回到最佳的結構解，因而最後穩定時的 HP 估計值就會達到最低值。

修正 Lattice 模型

使用晶格模型即是把角度限定在九十度，也就是蛋白質主鍊(Main chain)僅有九十度、一百八十度、兩百七十度這幾種變化，在計算上減少了許多構形(Conformation)，簡化了形式也減少計算量，可以加快模擬的速度及減少計算的負擔，但是自然界蛋白質的主鍊結構，其角度絕對不可能僅有這幾種變化，而且真實的蛋白質結構時時刻刻都會有小幅的變動，所以可以針對真實的結構加以修正，增加主鍊旋轉的角度，可以讓整體結構更加接近真實結構。

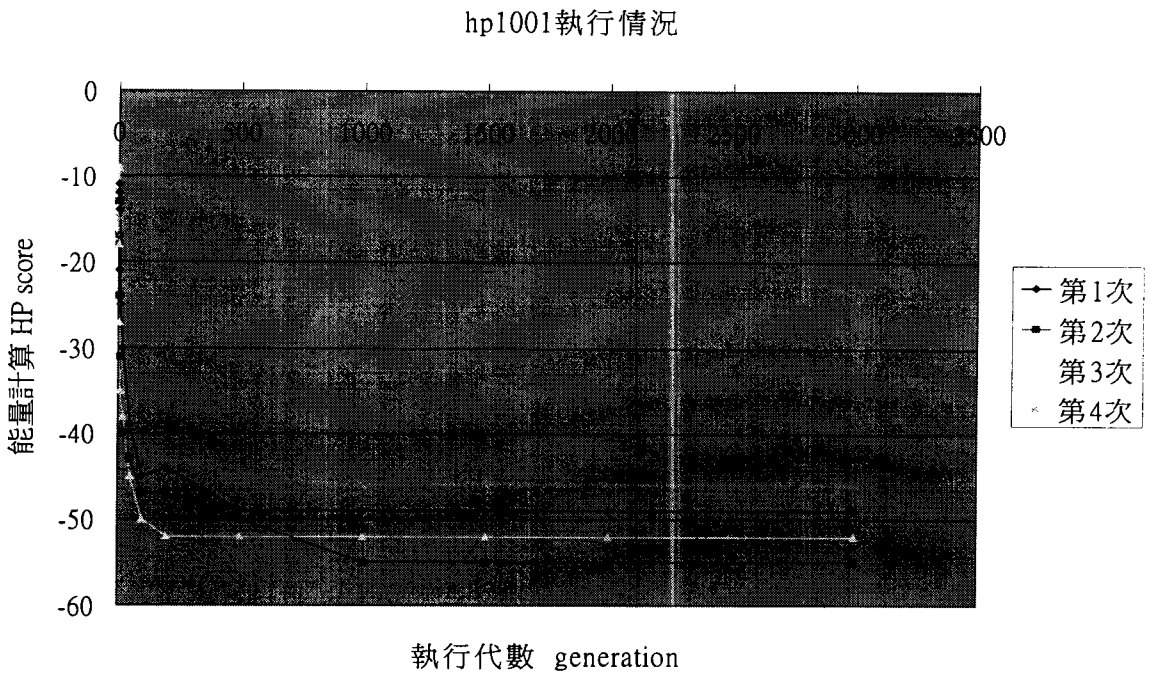
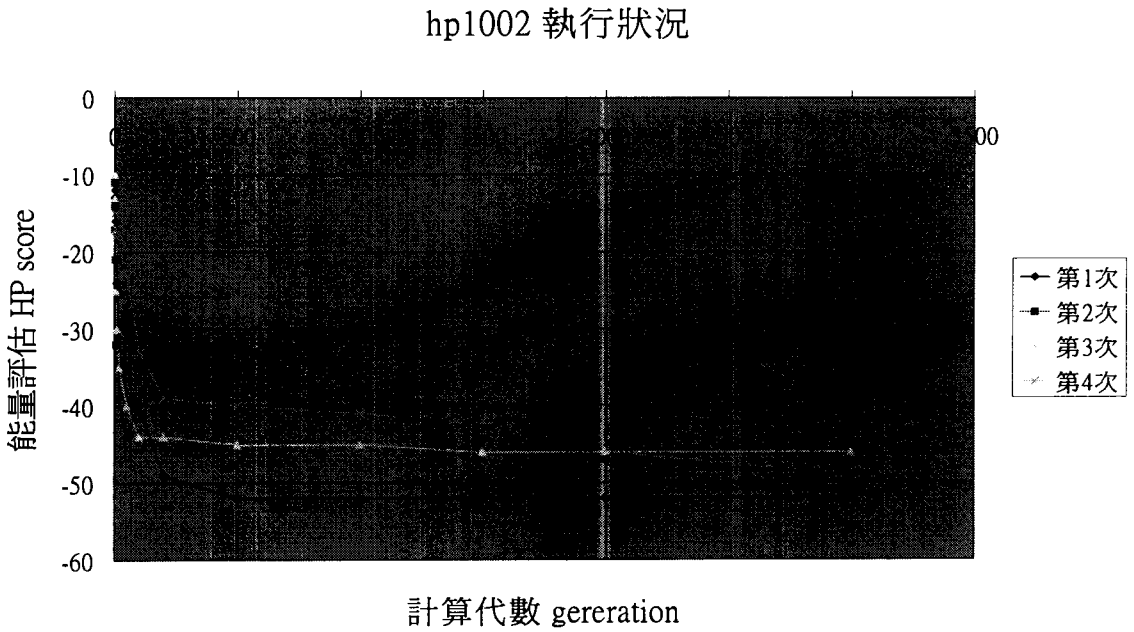


圖 十

四：：HP1001 的執行情況

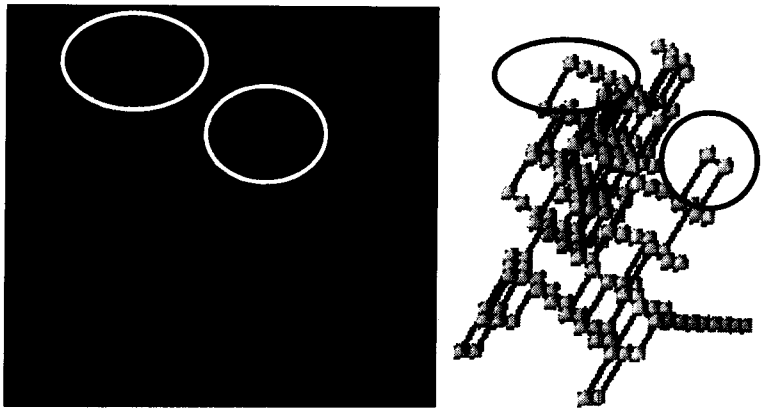


圖十

五：：HP1002 的執行情況

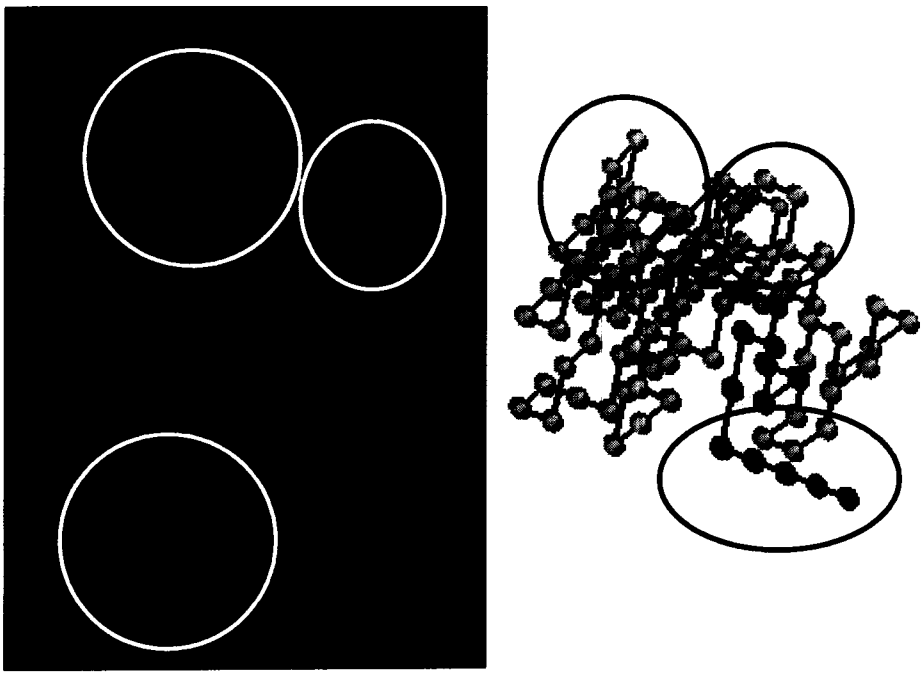
真實蛋白質結構與預測結果比較

由於系統有加入一些化學特性，所以嘗試以真實的氨基酸序列模擬出穩定的氨基酸結構，與真實的蛋白質結構相比較，圖十六爲蛋白質 1E31，左邊是由 RasMol Ver2.70 所產生的結構，右邊則是系統所產生的結構，在圖中可以發現某些區域的形狀蠻類似的，例如圈起來的區域形狀有些相似。



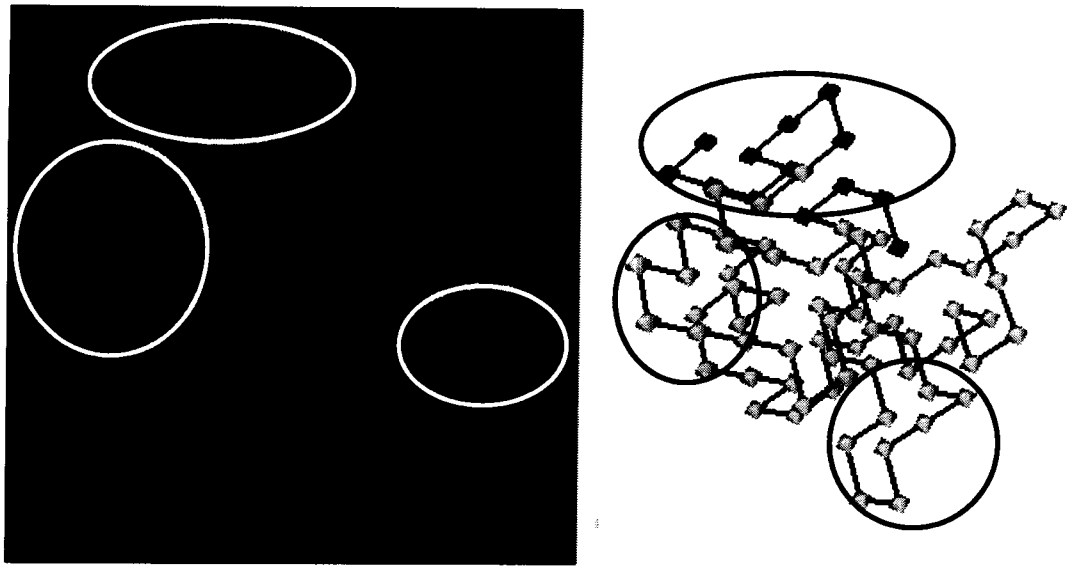
圖十六：1E31 結構

圖十七爲蛋白質 1E31，左邊是由 RasMol Ver2.70 所產生的結構，右邊則是系統所產生的結構，在圖中可以發現序列是由下往上走，由於系統的鍵長都是一樣，所以原本拉長的結構在我們系統中便縮在一起，在轉折的地方也有些相像。



圖十七：1N3G 結構

圖十八爲蛋白質 2DVH，左邊是由 RasMol Ver2.70 所產生的結構，右邊則是系統所產生的結構，在圖中序列是由上往下走，最後又回到上面，在上面的圈圈則是代表序列起點與終點，幾個轉折的構形有些相像。



圖十八：2DVH 結構

6.5.3.未來發展 Future Work

可以加入更多的規則

目前科學家對於蛋白質折疊的過程還是一知半解，爲了讓預測最後出來的結構更準確，可以加入更多的化學特性，比方說氨基酸的大小，支鍊(side chain)的結構，鍵長及主鍊(main chain)二面角的資訊，雙硫鍵(Disulfide bond)的資訊等，可以減少搜尋組合(search space)所花的時間，解決組合爆炸的問題，加速預測的速度及提高準確度，另一方面我們也可以嘗試考慮氨基酸的大小，主鍊的鍵長與鍵角等資訊，可以讓預測出來的結構更加接近自然。

改進氨基酸能量對應表(Energy Table)

透過目前 MMR 及 X-ray 的技術，蛋白質結構在許多研究機構的努力之中已經在蛋白質資料庫(Protein data bank)中存有 21838 筆資料，可以透過統計的方式來計算更新目前系統所使用的氨基酸能量表，讓整體的評估更加準確。

蛋白質家族分類

由於本系統可以快速的預測出蛋白質立體結構，可以根據目前蛋白質家族(Protein family)分類，建立蛋白質結構與 Lattice 模型之間的關係，也就是建立足以代表此家族的 Lattice 模型，當實驗得到的氨基酸序列可以透過此系統的快速結構預測及結構比對，可以找到相對應的蛋白質家族，取得更多的相關資訊，如果欲使用其他方法作結構預測，即可以此找到的蛋白質家族資料作爲評估學習資料(train data)，將此序列當作(test data)，必能得到準確率高的結構。

增加二級結構資訊

可以增加二級結構資訊及一些空間資訊，因爲純粹以晶格結構不易表示所有二級結構的

構形，必須以大致的形狀再加上空間資訊函數的輔助，才能以較完整的結構呈現。例如知道哪些序列可能形成二級結構，便可以將其片段序列直接以二級結構表示，再將其二級結構組合起來，就會形成準確度高的三級結構，如圖十九所示，由 Lattice 模型至做出的螺旋體結構(α helix)及折版結構(β sheet)，至於其他常見的結構也可以 Lattice 模型來製作，累積這些資訊所模擬出來的立體結構相信更能接近真實的蛋白質。

圖十九：由 Lattice 模型至做出的螺旋體結構(α helix)及折版結構(β sheet)

蛋白質構形比較

由實驗中得到的序列經由系統可以得到立體結構，如果能夠同時放上兩個蛋白質，便可以嘗試蛋白質結合(Docking)的工作，因爲是 Lattice 架構，所以，兩個蛋白質比較容易找到可能的結合區(Binding site)，同時亦可以作蛋白質相互之間的影響(Protein-Protein interaction)的研究。

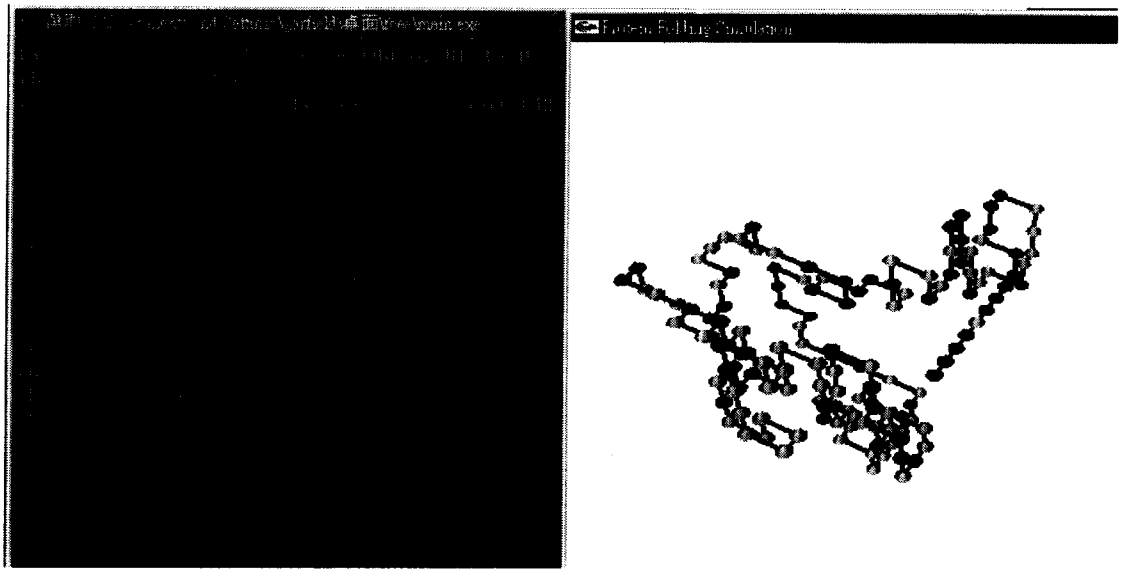
七、 參考文獻：

1. Aaron R. Dinner and Martin Karplus. (1999) Is Protein Unfolding the Reverse of Protein Folding? A Lattice Simulation Analysis. *J. Mol Biol.* ,**292**, 403-419
2. Abagyan,R.A. (1993) Towards protein folding by global energy optimization. *FEBS Letters*, **325,** 17-h.
3. Abkevich,V.I. and Shakhnovich,E.I. (2000) What can Disulfide Bonds Tell Us about Protein Energetics, Function and Folding: Simulations and Bioinformatics Analysis. *J. Mol. Biol.* , **300**, 975-985
4. Andrej Sali, Eugene Shakhnovich and Martin Karplus,(1994) Kinetics of Protein Folding: A Lattice Model Study of the Requirements for Folding to the Native State. *J. Mol. Biol.*, **235**, 1614-1636
5. Anfinsen,C.B.(1973) Principles that govern the folding of protein chains. *Science.*, **181**, 223-230.
6. Anfinsen,C.B., Haber,E., Sela,M. & White,F.H. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Nat. Acad. Sci.*, U.S.A. **47**, 1309-1314.
7. Broglia,R.A. and Tiana,G. (2001) Reading the Three-Dimensional Structure of Lattice Model-Designed Proteins from Their Amino Acid Sequence. *PROTEINS: Structure, Function, and Genetics*, **45**, 421-427
8. Brower,R.C., Vasmatzis,C., Silverman,M. & Delisi,C. (1993) Exhaustive conformational search and simulated annealing models of lattice peptides. *Biopolymers*. **33**, 329-334.
9. Carl Branden and John Tooze, Introduction to Protein Structure.(Second Edition), GARLAND.
10. Covell,D.G. (1994) Lattice Model Simulations of Polypeptide Chain Folding. *J.Mol.Biol.*,**235**,1032-1043
11. Covell,D.G. and Jernigan,R.L. (1990) Conformations of folded proteins in restricted spaces. *Biochemistry*, **29**, 3287-3294
12. Crippen. G. M. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232-4237.
13. David A. Hinds and Michael Levitt (1994) Exploring Conformational Space with a Simple Lattice Model for Protein Structure. *J. Mol. Biol.* **243**, 668-682
14. Go,N. and Taketomi,H. (1978) Respective roles of short-and long-range interactions in protein folding. *Proc. Nat. Acad. Sci.*, U.S.A., **75**, 559-563.
15. Hongyu Zhang (2002) Protein Tertiary Structures: Prediction from Amino Acid Sequence. *ENCY CLOPEDIA of LIFE SCIENCES.*
16. Jinn-Moon Yang (2001) A Family Competition Evolutionary Approach of Global Optimization in Neural Networks, Optical Thin-film Design, and Structure-based Drug Design.
17. Junni L. Zhang and Jun S. Liu (2002) A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model. *Journal of Chemical Physics*, Volume **117**, Number 7 , 3492-3498
18. Kaizhi Yue, Klaus M. Fiebig, Paul D. Thomas, Hue Sun Chan. (1995) A test of lattice protein folding algorithms. *Proc. Natl. Acad.* U.S.A., Vol.**92**, 325-329.

19. Kuntz,I., Crippen,G., Kollman,P., Kimelman,D. (1976) Calculation of protein tertiary structure. *J. Mol. Biol.* ,**106**:983-994.
20. Lau,K.F. & Dill,K.A. (1989) Lau KF, Dill KA. Lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* ,**22**:3986-3997.
21. Levinthal,C. (1968) Are there pathways for protein folding ? *J. Chem. Phys.*, **65**, 44-45.
22. Levitt,M. (1976) Simplified representations of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* ,**104**, 59-107.
23. Levitt,M. & Warshel,A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694-698.
24. MaiSua Li, D. K. Klimov, D. Thirumalai (2002) Folding in lattice models with side chains., *Computer Physics Communications*., **147** , 625-628
25. Oswin Aichholzer, David Bremner, Eric D. Demaine, Henk Meijer, Vera Sacristan, Michael Soss (2003) Long proteins with unique optimal foldings in the H-P model. *Computational Geometry*, **25** , 139-159
26. Ramakrishnan,R., Ramachandran,B., Pekney,J.F. (1997) A dynamic Monte Carlo algorithm for exploration of dense conformational spaces in heteropolymers. *J. Chem. Phys.*,**106**:2418-2425
27. Rolf Backofen, Sebastain Will and Erich Bornberg-Bauer, (1999) Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets., *BIOINFORMATICS*, Vol. **15** no.3, 234-242
28. Ron Unger and John Moult (1993) Genetic Algorithms for Protein Folding Simulations. *J.Mol. Biol,* **231**, 75-81
29. Skolnick,J. & Kolinski.A. (1990) Simulations of the folding of a globular protein. *Science*, **250**, 1121-1125.
30. Skolnick,J., Kolinski,A., Brooks,C.L. III. Godzik,A. & Rey,A. (1993) A method for predicting protein structure from sequence. *Curr. Biol.* ,**3**, 414-423.
31. Tianzi Jiang and Qinghua Cui, (2003) Protein folding simulations of the hydrophic-hydrophlic model by combining tabu search with genetic algorithm. *Journal of Chemical physics*. Volume **119**,Number 8
32. Ugo Bastolla, Helge Frauenkron, Erwin Gerstner, Peter Grassberger and Walter Nadler (1998) Testing a New Monte Carlo Algorithm for Protein Folding. *PROTEINS: structure, Function and Genetics* , **32** , 52-66
33. Unger,R. and Moult,J. (1993) A Genetic Algorithm for 3D Protein Folding Simulations. *Proceedings of the Fifth International Conference on Genetic Algorithms (ICGA'93)*, 581-588
34. Unger,R. and Moult,J. (1993) Genetic Algorithms for Protein Folding Simulations. *J. Mol. Biol.* **231**, 75-81
35. Wetlaufer,D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci.* U.S.A., **70**,697-701
36. Wilson,C., Doniach,S. (1989) A computer model to dynamically simulate protein folding: Studies with crambin. Proteins Struct Funct Genet 6:193-209.
37. Zhdannov,V.P. and Kasemo,B. (1997) Monte Carlo Simulation of Protein Folding Whit Orientation-Dependent Monomer-Monomer Interaction. *Protein : Structure, Function, and Genetics* **29**:508-516

附錄 A　Appendix A

操作環境簡介



1.左邊爲輸入環境參數，可以選擇輸入的序列爲氨基酸序列或是 HP 序列。
2.可以設定結構穩定時停止的條件，比方說經過 5000 代沒有任何能量的改變，即停止程式運作。
3.可以立即顯示程式執行的結果，可以看到演化到第幾代，凝聚函數的數值，氨基酸能量的改變及 HP 的估計值。
4.初始設定值，序列空間位置爲線性，其餘估計值爲 0。
5.可以隨時觀看空間位置，任意觀看調整角度，更新最近的演化結構，方便對於蛋白質模擬的觀測。

附錄 B　Appendix B

程式演算法 Program algorithm

```
輸入氨基酸序列或是 HP 序列
If ( 輸入是氨基酸序列 ) Then 轉換氨基酸序列爲 HP 序列
設定所有家族及成員 Lattice 的空間位置爲線性

While( 所有成員都作交配運算 )
{
    隨機選取要作運算的其他家族家族成員
    隨機選取空曠空間的位置，決定交換點
    If( 成員結構有碰撞 ) Then 放棄改變
    選擇下一個家族成員
}

While( 所有成員都作突變運算 )
{
    // ---- 註解：整體突變運算 ------
    隨機選取要作運算點位置
    計算構形中心點
    If( 選擇點性質是 H ) Then 往構形中心移動
    If( 選擇點性質是 P ) Then 往構形表面移動
    If( 成員結構有碰撞 ) Then 放棄改變

    // ---- 註解：細部突變運算一 ------
    While( 隨機選擇運算次數 )
    {
        隨機選取要作運算點位置
        進行細部突變運算一
        If( 成員結構有碰撞 ) Then 放棄改變
    }
    // ---- 註解：細部突變運算二 ------
    While( 隨機選擇運算次數 )
    {
        隨機選取要作運算點位置
        進行細部突變運算二
        If( 成員結構有碰撞 ) Then 放棄改變
    }
    // ---- 註解：細部突變運算三 ------
    While( 隨機選擇運算次數 )
    {
```

```
            隨機選取要作運算點位置
            進行細部突變運算三
            If( 成員結構有碰撞 ) Then 放棄改變
        }
    }

    計算所有家族及成員凝聚評估函數 ED 值
    計算所有家族及成員能量評估 EE 值
    計算所有家族及成員能量評估 HP 值

    // ---- 註解：選出能量最低值爲下一代結構 ------
    While( 所有成員都作比較 )
    {
        選擇 ED 值最大爲子代代表
        If( ED 值相等 )  Then  選擇 EE 值大的爲子代代表
    }

    // ---- 註解：與父代相比選出更佳
    If ( 產生的子代優於父代 )
        Then  If( 結構能量穩定 )
                    Then  結束程式
                    Else  子代變成下一代的父代，繼續演化。

    If ( 產生的父代優於或等於子代 )
        Then  If( 結構能量穩定 )
                    Then  結束程式
                    Else  有現代的父代，繼續演化。
```

附錄 C　Appendix C

圖表十三　HP 實驗數據 (序列長度 60)

	第 1 次	第 2 次	第 3 次	第 4 次
第 1 代	-12	-17	-11	-14
第 2 代	-13	-19	-14	-16
第 3 代	-16	-19	-16	-18
第 5 代	-19	-18	-19	-20
第 10 代	-19	-25	-23	-21
第 20 代	-21	-30	-29	-34
第 50 代	-33	-33	-31	-35
第 100 代	-35	-41	-31	-35
第 200 代	-37	-44	-32	-35
第 500 代	-37	-44	-33	-35
第 1000 代	-37	-44	-36	-35
第 1500 代	-38	-44	-36	-35
第 2000 代	-38	-44	-36	-37
第 3000 代	-38	-44	-40	-39

圖表十四　HP 實驗數據 (序列長度 100)

	第 1 次	第 2 次	第 3 次	第 4 次
第 1 代	-12	-13	-9	-12
第 2 代	-11	-17	-18	-15
第 3 代	-14	-18	-18	-17
第 5 代	-21	-24	-27	-25
第 10 代	-25	-31	-35	-29
第 20 代	-30	-40	-38	-35
第 50 代	-39	-43	-45	-39
第 100 代	-45	-47	-50	-42
第 200 代	-44	-47	-52	-43
第 500 代	-49	-50	-52	-45
第 1000 代	-49	-55	-52	-46
第 1500 代	-49	-55	-52	-46
第 2000 代	-49	-55	-52	-46
第 3000 代	-49	-55	-52	-46

圖表十五　HP 實驗數據 (序列長度 100)

	第 1 次	第 2 次	第 3 次	第 4 次

附錄 D　Appendix D

簡要的介紹幾個運用在蛋白質折疊上的演算法。

系統搜尋法 Systematic search

Ponder & Richards 發展第一個解決蛋白質折疊問題的方法，是假設在蛋白質內部已經被折疊，也就是殘基(residues)的位置已經固定。他們以有系統的規劃，並考慮空間位置，氫鍵及包裝密度的規則，將支鍊(side-chains) 和小的蛋白質的內部殘基作有系統的組合，並將將這樣的組合定義爲三級模版 (tertiary template)。不過，因爲組合過多，在考慮計算能力的原因，所以殘基(residues)的數目限制在 10 個以下，也由於這樣的限制，後來就沒有被繼續發展下去。

蒙地卡羅法 Monte Carlo methods

Hellinga and Richards 應用蒙地卡羅演算法來模擬來解決這個蛋白質折疊問題，這個方法是以能量公式計算維繫結構空間所需要的最低能量。實驗的設計是蛋白質 3D 結構的主鍊(backbone)是固定的，並透過隨機突變(random mutations)的方式來推測最有可能的產生這樣結構的序列，接著考量主鍊上的可轉動的角度(torsional angles)隨機旋轉(random rotations)，調整此組合至最佳結構。蒙地卡羅演算法不但可以使用在晶格模型(Lattice model)上，也可以使用在其他的模型上，所以有許多人也使用蒙地卡羅演算法，配合能量的計算，或是波次曼公式來解決關於蛋白質的研究問題。

困境排除定理 The dead end elimination theorem

Mayo 發展出一種困境排除定理(The dead end elimination theorem)簡稱 DEE 來進行蛋白質設計的研究，這個技術最初發展是在相似模型(homology modeling)安排支鍊(side chain)在空間中的位置而設計的。DEE 主要辨別哪一種全域性最小能量的構形(global minimum energy conformation (GMEC))能夠對於蛋白質及支鍊產生最低能量的可能狀態，也就是建立在支鍊構形資料庫(rotamer library)中找尋最有可能的答案，以求最佳解。

基因演算法 Genetic algorithms

基因演算法(Genetic algorithms)簡稱 GA，是解決最佳化問題非常棒的演算法，是演化式演算法(evolutionary algorithm)的一種，基因演算法有包含重組（recombination）和突變（mutation）這兩個運算，而重組是由父代產生許多子代後再以相互組合，形成新的個體，而突變是將個體內部作細微的改變，以此運算的方式選擇出最佳的個體，選擇的條件(fitness function)由設計者自訂，在條件的選擇下，選擇最佳的答案成爲下一代，再繼續進行演化的工作，直到達到設計者的需求爲止。

分支及臨界方法 Branch and bound methods

分支及臨界方法(Branch and bound method)專門針對處理將問題對應到樹狀結構，並且沿

著樹的分支搜尋出最佳的答案，蛋白質結構中，包含主鍊及支鍊的樣子，就像是樹枝的形狀，由樹的根部也就是主鍊往支鍊移動，產生幾個可能爲支鍊位置的解，這些原子位置就像樹葉般等待使用者選擇，我們不需要計算所有的分支節點，因爲這樣的問題是屬於爆炸性的組合，首先計算分支中所有節點中的最低能量，如果所得的能量高於已知支鍊構形組合中的任何一組，其分支則必須作調整，依此原則完成整個蛋白質結構的計算。