

# 行政院國家科學委員會專題研究計畫 成果報告

## 最佳化演算法在蛋白質結構分析上的應用(2/2)

計畫類別：個別型計畫

計畫編號：NSC93-2112-M-034-002-

執行期間：93年08月01日至94年07月31日

執行單位：中國文化大學物理學系

計畫主持人：鄒忠毅

計畫參與人員：管怡和，曾建誌

報告類型：完整報告

處理方式：本計畫可公開查詢

中 華 民 國 94 年 10 月 30 日

# 最佳化演算法在蛋白質結構分析上的應用

## 中英文摘要

我們在 2000 年時，發展出了一種名為導引模擬退火法的新的最佳化演算法，並且已經將這個方法應用在數個不同的問題上。在本研究計畫中，我們計畫繼續改進這個演算法，並將它運用在蛋白質結構分析問題上。

在計畫中，我們大幅改進了演算法，並稱它為“知識進化演算法”。新的演算法嘗試模擬，人類透過知識的累積而促進文明演進的過程。整個演算程序由數個不同世代的許多搜尋過程所組成。在演算程序中，前一代搜尋過程裡，以數種不同的局部搜尋方法所得到的資訊，透過整理與分析後，成為有用的知識。我們將這些知識儲存到一個我們稱為導引函數的資料庫中。這個導引函數常常顯現出系統的某種整體性質。有了前一代的知識（導引函數）後，我們就可以利用它來協助下一代的更有效率的搜尋。

We developed a new algorithm, the guided simulated annealing method, for several optimization problems in 2000. The goal of our proposal is to improve the new algorithm, and to apply this algorithm to the protein structure analysis problem.

In our project, we have improved our algorithm substantially and called it a new name “The Knowledge-based Evolution Algorithm (KEA)” .

This KEA method tries to simulate the process of the knowledge development. The whole searching procedure is composed of several generations of local search. At each generation of sampling, we retrieve certain useful knowledge and construct a knowledge database (called guiding functions). These functions are then used to guide the search in the next generation in addition to the consideration of minimum energy. The evolutionary guiding function will help to direct searching processes and to reach the global minimum faster.

## 關鍵詞(keywords)：

最佳化演算法，蛋白質結構分析，導引模擬退火法，知識進化演算法。

Optimization algorithm, protein structure analysis, guided simulated annealing method, knowledge-based evolution algorithm.

## 目錄：

- 一、 前言
- 二、 研究目的
  - 甲、 計畫的背景
  - 乙、 計畫的核准與修正
- 三、 文獻探討
  - 甲、 最佳化演算法介紹
  - 乙、 蛋白質結構分析相關問題介紹
- 四、 研究方法
  - 甲、 導引模擬退火法
  - 乙、 知識進化演算法
- 五、 結果
  - 甲、 演算法上的改進
  - 乙、 蛋白質結構分析上的應用
- 六、 討論
  - 甲、 自評
  - 乙、 建議

## 一， 前言

本文為專題計畫“最佳化演算法在蛋白質結構分析上的應用”的成果報告。我們將介紹此研究計畫的目的、方法及一些結果。下文總結了這一年半來我們研究組的一些心得，也彙集了近年來我們在這個專題項目上，所發表的一些研究結果<sup>[12, 27, 28, 29, 32, 41, 42]</sup>。

## 二， 研究目的。

### 甲， 計畫的背景

我們在 2000 年發展出了一種名為導引模擬退火法 (guided simulated annealing)<sup>[12][27][28][29][32]</sup> 的較有效的最佳化演算方法 (optimization algorithm)，並且已經將這個方法應用在數個不同的問題上。這個方法與傳統的模擬退火法最大的不同處，在於引進了一個我們稱為導引函數的新性質。使得在對於整體極小值的搜尋過程中，有了一個整體的方向的指導。這使得新的方法能更有效率地找到極小值。我們已經將這個方法運用在數個不同的問題上，並且已發表的一些結果。但是這個方法仍然需要改進，也需要進一步地測試它的應用範圍。在這個計劃的第一部分，就是要對這個演算法作更深入的研究與改進，希望能加快它的運算速度。

在這個計劃的第二部分，我們計畫拓展它的應用範圍，希望將這個方法更有效率地應用在幾個目前廣泛受大家注目的方向上。其中一個大方向，就是與蛋白質結構分析有關的幾個問題。

計劃中所指的蛋白質結構分析有關的問題，包含三個層次。第一個是所謂的蛋白質摺疊問題 (protein folding problem)<sup>[18][19][20][21]</sup>，也就是蛋白質如何由無規律的初始狀態，逐漸摺疊成為某一具有最低能量狀態的穩定結構。第二個是蛋白質結構測定的問題，也就是如何能從實驗 (X-ray 繞射或 NMR 實驗) 所取得的資料中，決定蛋白質分子結構<sup>[8][9][10][11][12]</sup>。第三個是蛋白質結構預測的問題<sup>[33]</sup>，也就是是否能由目前已知的知識或資料庫，預測出未知的蛋白質結構。這些與蛋白質結構分析有關的問題，它們一直受到科學研究人員的注意。相關的研究結果，也將能夠被利用在醫學及生物科技等領域上。

我們的目標就是利用現有的，由其他科學家或由我們自行發展出的模型作為標的，將這一些模型轉換成為最佳化問題。再利用我們所發展的較有效率的最佳化演算法，嘗試加速解決這些問題，並將這個最佳化演算法推廣到更多的研究領域中。

以上是報告人在提出計畫申請時的原始目標。

### 乙， 計畫的核准與修正

經過審核後，最後核准的計畫做了一些修正。首先是計畫時間由三年減縮為

一年半，另外經費也有一些縮減。這代表實際計畫的目標必須更為集中，所以我將計畫目標專注在最佳化演算法發展及應用在前述的蛋白質結構分析三個層次中的第一層次，也就是蛋白質摺疊問題上。

### 三，文獻探討

#### 甲，現有最佳化方法的介紹

在於許多不同領域與學科中，研究者所關心的問題，常常具有共同的數學特性。例如原子團簇問題 (atomic cluster problem)<sup>[16]</sup> 主要研究的是，眾多原子聚集時的最低能量結構為何？蛋白質摺疊問題 (Protein Folding Problem) 研究的是蛋白質分子鏈如何由無序的形狀轉變為有序的最穩定結構。旅行商問題 (Traveling Salesman Problem)<sup>[17]</sup> 指的是，旅行商如何能用最短的路程，拜訪他的全部客人所在的城市？基本上，在研究者處理這些問題時，大都希望將關注的目標 (或希望得到的結果) 先轉化成為某個數學上的函數。接下來這些問題也就轉化為函數求極大或極小值的問題。或者說，這些問題也就變成了所謂的“最佳化問題”了。

但是想要求解這些問題，卻不是一件容易的事。許多最佳化問題，隨著問題中的變數逐漸增多，問題的複雜度與計算難度將以驚人的速度在增加。這表示我們很難用人工的方式求解。我們必須求助於電腦及使用有效的演算法。

現在已經有了許多不同的最佳化演算法<sup>[1-6]</sup>。但是，當我們仔細回想這些演算法的精神時，我們可以發現，有許多的演算法其實是在模擬或學習一些自然界中已經存在的現象。其中有學習物質世界演化的方法，例如模擬石頭從山上順著山勢滾下的梯度法及最陡下降法，模擬水從山上流下並改變地貌的能量鋪砌法，模擬物質由液體凝結為固體或是結晶過程的模擬退火法，及模擬一群質點彼此施力並且因此運動的分子動力學方法等等。還有學習生物世界演化的方法。例如模擬物種演化的遺傳演算法，模擬生物覓食與運動的試錯法及貪心算法，模擬生物認知與學習過程的人工神經網路法，及模擬蟻群覓食的螞蟻群聚算法等等。最後也有學習人類文明發展的演算法，例如模擬文化形成的文化算法，及報告所介紹的嘗試模擬知識發展的知識進化演算法 (原先稱為導引模擬退火法<sup>[27]</sup>)。

#### (1) 現有蛋白質結構分析相關問題研究情況介紹

##### 蛋白質摺疊問題

這個問題用最簡單的說法，就是在我們知道一個蛋白質內胺基酸 (amino acid) 的排列順序後，可否決定其立體結構？假設蛋白質內原子之間的作用力可以用一

個有效位能來描述，則蛋白質在自然環境中的結構應具有最低的自由能 (Free energy)。

要直接計算真實的蛋白質結構並不是件簡單的事，所以科學家便想出了一些

名稱及來源	測試序列長度及所得最低能量									
	N=20	N=24	N=25	N=36	N=48	N=50	N=60	N=64	N=100,a	N=100,b
<b>Genetic Algorithm</b> R. Unger and J. Moult, J. Mol. Biol. 231, 75 (1993)	-9	-9	-8	-14	-22	-21	-34	-37	---	---
<b>Contact interactions method</b> L. Toma and S. Toma, Protein Sci., 5, 147 (1996)	-9	-9	-8	-14	-23	-21	-35	-40	---	---
<b>Core-directed chain Growth model</b> T. C. Beutler and K. A. Dill, Protein Sci., 5 2037 (1996)	-9	-9	-8	-14	-23	-21	-35	-42	---	---
<b>dynamic Monte Carlo method</b> R. Ramakrishnan, B. Ramachandran and J. F. Pekny, J. Chem. Phys. 106, 2418 (1996)	---	---	---	---	---	---	---	---	-44	-46
<b>Pruned-enriched Rosenbluth method</b> H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger and W. Nadler, Phys. Rev. Lett., 80, 3149 (1998)	---	---	---	---	---	---	-36	---	-47	-49
<b>Multi-Self-Overlap Ensemble method</b> G. Chikenji, M. Kikuchi and Y. Iba, Phys. Rev. Lett, 83, 1886 (1999)	---	---	---	---	---	---	---	-42	-47	-50
<b>Evolutionary Monte Carlo method</b> F. Liang and W. H. Wong, J. Chem. Phys. 115, 3374 (2001)	-9	-9	-8	-14	-23	-21	-35	-39	---	---
<b>導引模擬退火法 (Guided Simulated Annealing Method)</b>	-9	-9	-8	-14	-23	-21	-36	-42	-48	-50

表 (一)，GSA 的結果與現有結果的比較表

簡化模型來模擬真實的蛋白質分子。其中最簡單的一種就是二維 HP 晶格模型<sup>[18]</sup> (2D HP Lattice model)。

我們用新的方法，得到了一些初步的結果。除了能重複得到與目前已知最低能量結構相同的結果外，在某組 100 個胺基酸序列上，我們得到了更新的最低能量結構。表 (一)

晶格模型只是蛋白質結構的最簡單近似，接下來我們也發展了一個更接近真實的模型 (Off-lattice model)，在此模型中，粒子鏈中的胺基酸不再被限定在二維格子點上。我們使用的能量形式，由氫鍵能量與具有方向性的疏水親水胺基酸交互作用能量所組成。配合利用導引模擬退火法，我們得到傳統簡化模型<sup>[34-40]</sup>所不易得到的具有  $\alpha/\beta$  二級結構形式的蛋白質結構。但這個模型還有許多問題等待解決。

## 四，研究方法。

### 甲，導引模擬退火法介紹

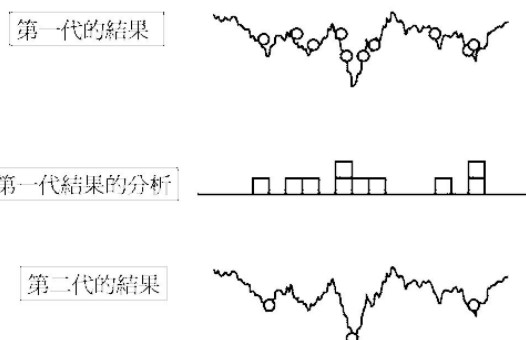
接下來介紹本計畫所將使用及繼續改進的最佳化演算法 —— 導引模擬退火法。

我們在 2000 年中以前，一直在嘗試如何改良傳統的模擬退火法，使它能更有效率地應用在多變數系統的求解基態問題上。但是我們發現，要使模擬退火法一

次就求出整體的最佳解是非常困難的。在一次次模擬計算中，絕大多數的試驗結果總是得到能量較高的亞穩態解。可是我們觀察這些能量較基態為高的亞穩態解時，卻發現到這一些亞穩態解之間有一些關聯性。所以我們便有了一些想法，並藉此發展出了一個最佳化演算法。

我們的基本想法有三點：第一，我們不再期望能由傳統的模擬退火法一次就找到基態解，改為在較短時間內用它來找到許多亞穩態解。第二，藉由分析這些亞穩態解的性質，找到有用的資訊。第三，想辦法利用這些資訊，來幫助下一代的搜尋過程。這些想法使我們的演算法不僅包含了傳統模擬退火法單純地模擬物質世界中材料的冷卻與結晶過程的表現，更具有了生物世界世代交替逐漸演化的精神。

接下來是要如何用數值方法來實踐我們的基本想法？首先我們要使用統計方法來分析亞穩態解的性質。我們將這些亞穩態解做出統計分布圖也就是所謂的直方圖 (histogram)。再來觀察亞穩態解的 histogram 可以知道這些解的可能出現區域，並由這些解的共有性質來去除不相關的自由度。接下來我們在下一代的



搜尋過程中增加 histogram 中，高可能性區域的搜尋機會，如此便可能找到更低能量的解。我們用以下的示意圖來說明，圖 (一) 中圓點代表在能量曲面中亞穩態解的位置，然後將亞穩態解做出統計分布圖，藉由增加搜尋分布圖高可能性區域，在第二代中找到基態。當然這只是最簡單的示意圖，實際的問題通常不能簡單地將高維度的能量曲面及分布圖繪出來。

圖 (一) 導引模擬退火法的基本想法

由以上簡單的示意圖，我們也可以看出，這個新的演算法的關鍵，就在於如何用上一代有用的資訊來導引下一代的搜尋過程。對於簡單的問題，我們當然可以用個別變數位置的 histogram 來引導下一代的搜尋。但是對於各種不同的較複雜的問題，上一代所共有的有用的資訊，可能不僅僅是個別變數位置那麼直觀。為了使我們的新演算法，能更廣泛的適應各種不同的問題，我們把這種用來引導下一代的搜尋過程的，上一代所共有的有用資訊，以一種函數形式來表現。我們稱它為導引函數(guiding function)，而這個新演算法，我們就稱它為導引模擬退火法( Guided Simulated Annealing )。

### 導引模擬退火法的操作程序

接著就來介紹導引模擬退火法的簡單的操作程序：

首先要針對問題選定目標函數，並且將這個目標函數視為一個廣義的能量函數  $E(x)$ 。

第一步：利用傳統模擬退火法得到第一組親代。由於我們用了很快速的退火策略，所以第一組親代群體，沒有找到整體極小值。

第二步：分析第一組群體的變數分布圖作為導引函數。對於簡單的問題，我們可以直接用個別變數位置的 histogram 來作為導引函數(以  $\rho(x)$  表示)。

第三步：利用加入導引函數的退火過程得到子代。有了導引函數後，我們就可以用它來幫助下一代群體的產生。導引函數的使用方法如下，

- (1) 與傳統模擬退火法不同，我們不再隨機的選擇變數並隨機的改變變數。在導引模擬退火法的操作中，變數的選擇機會與變數所具有的導引函數成反比，也就是在許多變數中，越與上一代整體特徵不符合的變數越容易被挑選到。這個步驟也可以用下式表示

$$P_{sel}(i) \propto 1/\rho(x_i) ,$$

其中  $P_{sel}(i)$  表第  $i$  個變數被挑選的機率。

- (2) 挑選了變數之後，變數的改變也與該變數的導引函數成正比，也就是該變數較會被改變到導引函數值高的區域中。用下式這個步驟表示

$$P_{mov}(x_i) \propto \rho(x_i)$$

其中  $P_{mov}(x_i)$  表變數  $i$  被移動到位置  $x_i$  的機率。

- (3) 接下來與傳統模擬退火法相同，用波茲曼概率分布函數來決定移動成功與否。

第四步：分析子代群體的變數分布作為新的導引函數。然後並回到前一步，直到世代次數到達限定值或目標函數值收斂到某一個極值。

### 導引模擬退火法的特色

接下來我們來看看這個方法的一些特色。首先，我們引入了一個新的性質，也就是導引函數，它使搜尋過程中，系統組態較有可能在能量區面上，數個低能量區域間變換，使得系統組態也較不易陷入局部極小。其次，導引函數可以有效地去掉不相關的自由度，使搜尋範圍逐漸減少，進而加速搜尋效率。又因為我們所得到的每一個子代都是由各自獨立的退火過程產生，使得這個演算法可以容易地在分散或平行化電腦設施上執行，可大幅加強計算效率。

對於某些問題，我們可能可以用其他的方法預測解的大致範圍。也就是說，已經有了一個粗略的導引函數。在這種情形下，導引模擬退火法很容易與其他方法配合使用，例如用其他數值或實驗方法先取得粗略的導引函數，再用本法做進一步的計算，找出更精確的解。

另外這個方法在漏斗形的能量區面問題上特別有用。這是因為這種能量區面，亞穩態解之間的關聯性較強。我們很容易就統計出親代共有的特性，到到很



好的導引函數來幫助下一代的搜尋。

### **導引模擬退火法與傳統方法的異同處**

在討論了導引模擬退火法的一些特色後，我們再來看看和這個方法有相關性的兩種方法，也就是模擬退火法與遺傳算法與導引模擬退火法的異同處。

首先是模擬退火法，這兩個方法表現出來的，都是使系統組態在能量曲面上做無規行走(random walk)，並且透過波茲曼概率分布函數對能量的選擇，使系統組態逐漸變化到能量最小值處。但是導引模擬退火法多了一個方向性，使得本方法除了保留了原先模擬退火法良好的局部搜尋能力，更藉著導引函數的幫助，加強了這個方法的全局搜尋能力。

接下來來討論導引模擬退火法與遺傳算法的異同處。明顯可見，這兩個方法都是世代演化形式的演算，並不指望在第一代搜尋中便找到基態解。而是透過世代演化的計算過程，逐漸地找到最好的解。而且，在這兩個方法的演算過程中，子代與親代也都有有一定的相似性。可是對於導引模擬退火法的計算過程，子代不是僅由一或二個的親代產生，它包含有上個世代整體特徵。或是說，親代的整體特徵透過導引函數遺傳到了子代。另外，由於在產生子代的過程中，使用了退火過程，也使得本方法比遺傳算法具有更好的局部搜尋能力。

### **乙，知識進化演算法介紹**

2004 年春天，我們對導引模擬退火法做了一些重要的測試。我們發現，模擬退火法並不是這個演算法的關鍵。例如，在蛋白質二維晶格模型的測試上，我們不再使用模擬退火法作為局部搜尋方法，改使用鏈成長方法配合導引函數來做局部搜尋。最後我們所得到的結果是，計算速度反而大幅提升。在經過我們做一些參數調整之後，對於 64 個胺基酸長度的粒子鏈，平均計算時間由二十幾分鐘減少為數秒鐘。

對我們來說，這是一個令人驚奇的本質上的大改變。新的演算法不再拘泥於使用模擬退火法作為局部搜尋方法，而可以配合許許多多的局部搜尋方法。這除了顯示出，新演算法有了更多的可能性，也代表它將較不受單一局部搜尋方法的對解的偏好性的限制。由我們目前的結果來看，混合的局部搜尋方法反而有較好的結果。這顯示出，單一局部搜尋方法的對解的偏好性，可能被平均了。總的來說，新演算法保留了世代交替的計算步驟及導引函數，可以使用不同的局部搜尋方法，並且加入了更清楚的天擇程序。在改進後，新演算法大幅提高了計算效率。

由於新演算法與模擬退火法不一定有關，我們必須給新演算法一個名字。在經過我們思考何為新演算法的最重要性質時，我們發現，這個演算法嘗試模擬，人類透過知識的累積而促進文明演進的過程。所以我們將它命名為“知識進化演算法”。

### **基本想法**

如前所述，此演算法嘗試模擬，人類透過知識的累積而促進文明演進的過程。所以它的精神主要如下：在演算程序中，前一代搜尋過程中以幾種不同的局

部搜尋方法所得到的資訊，透過整理與分析後，成為有用的知識。我們把這個步驟稱為建立導引函數。有了前一代的知識（導引函數）後，我們就可以利用它來協助下一代的更有效率的搜尋。這也使得本法較易快速地找到整體的最小值。

這個精神也如同，每一代的人類透過學習前人所流傳的知識，以各種個人所體認的方式來解決他們所面對的問題。同時他們解決問題的各種方法與心得也將彙整成新的知識，來協助更新的一代來更有效地解決問題。

### 算法的關鍵

這個演算法的關鍵，就是如何用前一代的有用知識，來幫助下一代搜尋。

在演算法中，我們將上一代所共有的有用資訊或知識，以一種統計的函數形式來表現，就稱它為導引函數(guiding function)。而導引函數的建立，通常必須對上一代的成果透過淘汰、歸納及分類等程序來完成。對於簡單的問題，當然可以直接用個別變數數值的分佈情況來作為簡單的導引函數。對於較複雜的問題，從我們的經驗的得知，如果使用比變數本身更為複雜的有用資訊來建立較高階的導引函數，常常可以加速計算。

那什麼是更為複雜的有用資訊或較高階的導引函數呢？對於一般物理系統，這常常就是系統的某種對稱性或是某種的整體性質。或者換句話說，就是系統的某種序參量（order parameter），這種序參量指出了某種演化的方向。隨著計算的進行，不但一個世代接著一個世代的搜尋結果越來越好，有用的知識(序參量)也隨著世代變化逐漸清晰。

進一步說，序參量不僅只表現了系統的某種的整體性質。它也可能顯示了複雜參數空間（parameter space or configuration space）的某種分類的依據標準。我們透過序參量的分析，也許可將複雜的能量區面劃分為數個主要的子區域。而在每個子區域裡存在一個較簡單的漏斗型的能量區面。如此，由於各個漏斗型的能量區面都較易求解。我們可以在各主要區域的極小解找出後，再決定系統的整體極小解。

### 運算程序

接下來解釋知識進化演算法的計算步驟。在計算開始前，使用者必須先決定局部搜尋法的種類與導引函數的形式，還有導引函數的使用方式。

首先，以使用最基本的導引函數的形式（變數分布）為例，它的基本計算步驟如下：

1. 利用最快的方法(或是幾種方法混用)，尋找一大堆亞穩態解。
2. 利用天擇程序，留下較優的一群解。
3. 統計這一群解的共通性，建立導引函數。
4. 利用最快的方法配合導引函數，尋找下一代的一大堆亞穩態解。
5. 回到步驟 2，直到滿足結束條件

可是，如果我們對問題的了解較深。我們可以使用較高階的導引函數，來作為參數空間的分割以加速計算。它的計算步驟如下：

1. 利用最快的方法(或是幾種方法混用)，尋找一大堆亞穩態解。
2. 利用天擇程序，留下較優的一群解。
3. 統計這一群解的共通性，並由此分割參數空間。
4. 決定目前處理的子空間。
5. 建立此一子空間的導引函數。
6. 利用最快的方法配合導引函數，尋找下一代的一大堆亞穩態解。
7. 利用天擇程序，留下較優的一群解。
8. 回到步驟 5，直到找到此子空間的最佳解。
9. 回到步驟 4，直到主要的子空間都已搜尋過，或滿足結束條件

### 與其他方法的比較

接下來，我們想說明知識進化演算法的特色，及此法與其他方法的比較。

綜合前述，知識進化演算法有以下幾點特色，

- (一) 不必局限於某種特殊的局部搜尋法，也因此可減少各種局部搜尋法的對解的偏好性的限制。這表示它是組合式的演算程序，不只是單一的搜尋法。
- (二) 如果在每一代的搜尋中，同時使用多種局部搜尋法。將會發現，不同世代中，最有效率的局部搜尋法可能不同。這表示，在本法的計算程序了，不僅每世代的解在互相競爭，各局部搜尋法也在競爭。
- (三) 遺傳的訊息不必為變數本身，也就是可直接傳遞高階訊息至下一代搜尋程序。這表示它改進了傳統遺傳算法的基因碼與變數的關係。
- (四) 遺傳的訊息不必為單一種類，可為數種型態的訊息。這點可能比其他演化型的算法，更拓展了遺傳的形式及效能。
- (五) 高階導引函數可有效畫分參數空間。這使本法更能應用在多漏斗型能量區面的問題上。

這些特色，使我們覺得，本法與其它的演算法不同，是一種更接近人類知識發展的演算法。

## 五，結果

### 甲，演算法的改進

在這個研究計畫中，我們改良原先的演算法，最後得到了一個全新的演算法類別。利用它能更有效率地找到整體的最小值。此演算法嘗試模擬，人類透過知識的累積而促進文明演進的過程。這個方法能運用在許多的問題上，並已有了初

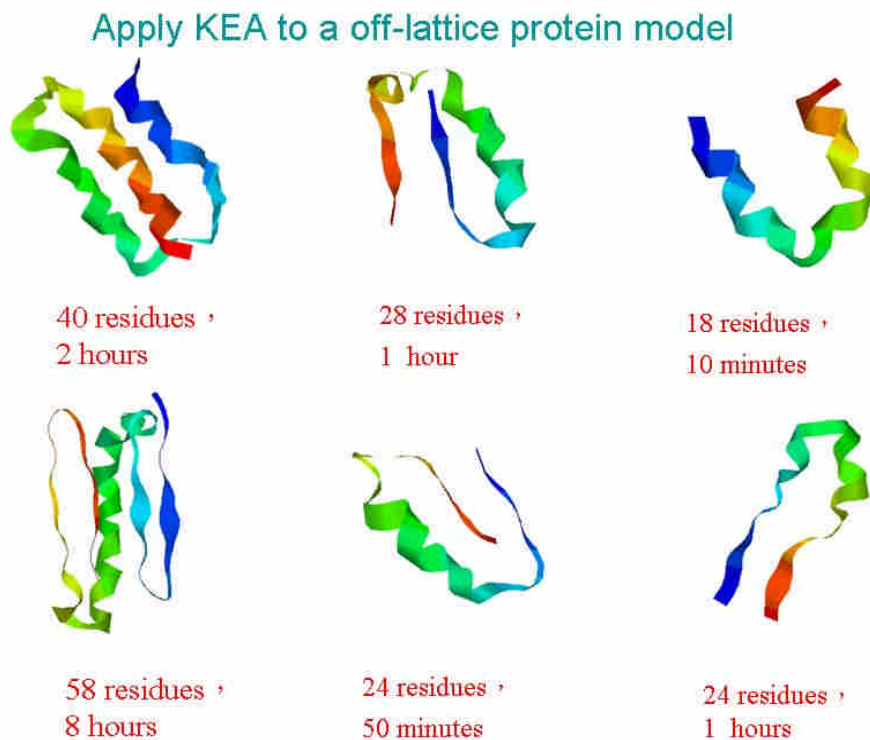
步的結果。目前對於幾種最佳化問題，知識進化演算法可以較快地找到可能的最低能量結構。新演算法的關鍵在導引函數，較高階的導引函數形式，常可加速計算。

雖然知識進化演算法嘗試模擬，知識逐漸演進的過程，但是它仍有許多限制。第一，它的局部搜尋方法及導引函數的形式與用法必須事先由研究者選定。第二，必須要對所解的最佳化問題的本質，有較深入了解，才能寫出較適當的高階的導引函數。這代表了本法仍然需要再進一步的加強。

## 乙，蛋白質結構分析上的應用

我們已經將這個演算法應用在幾個問題上並有了結果，特別在前述的蛋白質摺疊問題上，有了好結果。以蛋白質晶格模型為例，如果利用原來導引模擬退火法需要計算10分鐘的測試序列，利用知識進化演算法並使用鏈成長的局部搜尋法只需要三秒鐘。相關的論文，即將發表。

我們也將這個方法運用在蛋白質摺疊的離晶格模型，也有了初步的結果，的確也可以大大節省計算時間。以下介紹報告人與中央研究院物理所李定國教授的研究團隊，所發展出了一個更接近真實的離晶格模型 (Off-lattice model)。在此模型中，粒子鏈中的胺基酸不再被限定在二維格子點上。所使用的能量形式，由氫鍵能量與具有方向性的疏水親水胺基酸交互作用能量所組成。配合利用知識進化演算法，我們得到傳統簡化模型所不易得到的較複雜結構形式的蛋白質結構。下圖是我們的幾個測試結果。在圖中，展示了利用我們的模型所得到的幾種形式的蛋白質結構，與新演算法在一台 PC 上所需的計算時間。



## 六，討論

### 甲， 自評

- i. 研究內容與原計畫相符程度：由於在 2004 年，我們將原有演算法做了意料之外大幅度修改。於是整個研究方向便轉向為以新演算法理論的發展與測試為主，應用部分的比例變降低了。
- ii. 達成預期目標情況：我們發展了一個新概念的最佳化演算法。這個結果比我們原先計畫的修改舊演算法的目標要好的多。
- iii. 研究成果之學術或應用價值：我們以為，新演算法的也許可以在更多的需要求得最佳解的領域中使用。當然，這還要我們再加以推廣。
- iv. 是否適合在學術期刊發表或申請專利：新演算法的理論與應用已在數個研討會上發表，相關的期刊論文也進入發表的程序中。但是，我們以為在目前，演算法的開放使用及廣為介紹，遠比申請專利重要。

### 乙， 建議

這是報告人的第一個研究計畫，感謝國科會對本計畫的支持。有一個建議，希望國科會未來能再加強，對新進人員在研究計畫寫作上的協助。謝謝。

### 丙， 補充說明

本報告內容，為匯集報告人近年所公開發表的論文而成。如想要知道近一步的細節，請參考以下資料[12, 27, 28, 29, 32, 41, 42]

## 參考資料：

- [1] N. Metropolis et.al., "Equation of state calculation by fast computing machines", *Journal of Chemical Physics* **21** (1953) 1087-92.
- [2] S. Kirkpatrick et.al., "Optimization by simulated annealing", *Science* **220** (1983) 671-80.
- [3] J. H. Holland, "Adaptation in Natural and Artificial Systems", *MIT Press* (1975).
- [4] F. Glover, "Tabu search: part I", *ORSA Journal on Computing* **1** (1989) 190-206; "Tabu search: part II", *ORSA Journal on Computing* **2** (1990) 4-32.
- [5] J. J. Hopfield et.al., "Computing with neural circuits: A model", *Science* **233** (1986) 625-33.

- [6] Z. Michalewicz, “Genetic Algorithms + Data Structures = Evolution Programs”, (3rd. ed.) Springer-Verlag (1996).
- [7] B. A. Berg, “Locating global minima in optimization problems by a random-cost approach”, *Nature* 361 (1993) 708-710.
- [8] H. Hauptman, “The direct method of X-ray Crystallography”, *Science* **233** (1986) 178-183.
- [9] H. Hauptman, “Looking ahead”, *Acta Cryst. B* **51** (1995) 416-422.
- [10] W. P. Su, “Simulated annealing as a tool for *Ab initio* phasing in X-ray crystallography”, *Acta Cryst. A* **51** (1995) 845-849.
- [11] Y. Chen and W. P. Su, “A modified X-Y method”, *Acta Cryst. A* **57** (2001) 733-735.
- [12] C. I. Chou (鄒忠毅) and T. K. Lee (李定國), “Guided simulated annealing method for crystallography”, *Acta Cryst. A* **58** (2002) 42-46.
- [13] V. Z. Pletenev, V. T. Ivanov, D. A. Langs, P. Strong, & W. L. Duax, “Crystal and molecular structure of isoleucinomycin, cyclo[-(D-Ile-Lac-Ile-D-Hyi)<sub>3</sub>-](C<sub>60</sub>H<sub>102</sub>N<sub>6</sub>O<sub>18</sub>), *Biopolymers* 32 (1992) 819-827.
- [14] D. J. Wales and H. A. Scheraga, “Global optimization of clusters, crystals, and biomolecules”, *Science* **285** (1999) 1368-1372.
- [15] D. M. Deaven, N. Tit, J. R. Morris and K. M. Ho, “Structural optimization of Lennard-Jones cluster by genetic algorithm”, *Chem. Phys. Lett.* **256** (1996) 195-200.
- [16] 參照 The Cambridge Cluster Database 網站：  
<http://brian.ch.cam.ac.uk/CCD.html>
- [17] 參照 TSPLIB 網站：  
<http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/>
- [18] C. Tang. “Simple model of the protein folding problem”, *Physica A* **288** (2000) 31-48
- [19] F. Liang and W. H. Wong, “Evolutionary Monte Carlo for protein folding simulations”, *J. Chem. Phys.* **115** (2001) 3374-3380.
- [20] G. Chikenji, M. Kikuchi and Y. Iba, “Multi-self-overlap ensemble for protein folding: ground state search and thermodynamics”, *Phys. Rev. Lett.* **83** (1999) 1886-1889.
- [21] R. Unger and J. Moult, “Genetic algorithms for protein folding simulations”, *J. Mol. Biol.* **231** (1993) 75-81.
- [22] Chung-I Chou, Choon-Lin Ho and Bambi Hu, “Universality in the Frenkel-Kontorova model with a cosh-type interaction”, *Phys. Rev. E* **55**, 5092 (1997).

- [23] Chung-I Chou, Choon-Lin Ho, Bambi Hu and Hsuan Lee, “Morse-type Frenkel-Kontorova model” , *Phys. Rev. E* **57**, 2747 (1998).
- [24] C. N. Chen, C. I. Chou, C. R. Hwang, J. Kang, T. K. Lee and S. P. Li, “Monte Carlo dynamics in global optimization” , *Phys. Rev. E* **60**, 2388 (1999).
- [25] Choon-Lin Ho and Chung-I Chou, “A simple variational approach to the quantum Frenkel-Kontorova model” , *Phys. Rev. E* **63**, 016203 (2001).
- [26] Y. H. Wong, C. I. Chou and S. P. Li, “Dynamics of genetic algorithms in optimization problems” , *Chinese J. Phys.* **39**, 12 (2001).
- [27] 李世炳、鄒忠毅, “簡介導引模擬退火法及其應用” , 物理雙月刊, 24 卷 2 期 307-319 (2002 年 4 月)
- [28] C. I. Chou, R. S. Han, S. P. Li and T. K. Lee, “Guided simulated annealing method for optimization problems” , *Phys. Rev. E* **67**, 066704 (2003).
- [29] C. I. Chou, R. S. Han, T. K. Lee and S. P. Li, "A guided Monte Carlo approach to optimization problems", *LNCS* **2690**, 447 (2003).
- [30] Choon-Lin Ho and Chung-I Chou, “Simple variational approach to the quantum Frenkel-Kontorova model” , “7th International Conference on Squeezed States and Uncertainty Relations” , Jun 4-8, 2001, Boston, Massachusetts, USA.
- [31] Choon-Lin Ho and Chung-I Chou, “Quantum Frenkel-Kontorova model” , “XXIV International Colloquium on Group Theoretical Methods in Physics” , July 15-20, 2002, Paris.
- [32] 鄒忠毅、韓榕生、李定國, “導引模擬退火法在蛋白質結構分析上的應用” , “蛋白質結構分析中的一些問題” 研討會, Nov 18-22, 2002 北京。
- [33] 參照 CASP 網站：  
<http://predictioncenter.llnl.gov/>
- [34] J. Lee, A. Liwo and H. A. Scheraga, “Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K” , *Proc. Natl. Acad. Sci. USA*, **96** (1999) 2025-2030.
- [35] J. Pillardy, A. Liwo, M. Groth and H. A. Scheraga, “An efficient deformation-based global optimization method for off-lattice polymer chain: self-consistent basin-to-deformed-basin mapping (SCBDBM). Application to united-residue polypeptide chains” , *J. Phys. Chem. B*, **103** (1999) 7353-7366.
- [36] J. Pillardy, C. Czaplewski, W. J. Wedemeyer and H. A. Scheraga,

- “Conformation-Family Monte Carlo (CFMC): An efficient computational method for identifying the low-energy states of a macromolecule” , *Helv. Chim. Acta*, **83** (2000) 2214-2230.
- [37] A. Irback, F. Sjunnesson and S. Wallin, “Three-helix-bundle protein in a Ramachandran model” , *Proc. Natl. Acad. Sci. USA*, **97** (2000) 13614-13618.
- [38] G. Favrin, A. Irback and F. Sjunnesson, “Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space” , *J. Chem. Phys.*, **114** (2001) 8154-8158.
- [39] A. Irback, F. Sjunnesson and S. Wallin. “Hydrogen bonds, Hydrophobicity forces and the character of the collapse transition” , *J. Bio. Phys.*, **27** (2001) 169-179.
- [40] G. Favrin, A. Irback and S. Wallin, “Folding of a small helical protein using hydrogen bonds and hydrophobicity forces” , *Proteins: Struct. Funct. Genet.*, **47** (2002) 99-105.
- [41] Chou Chung-I, Lee Ting-Kuo, “Knowledge-based Evolution Algorithm for Protein Folding Problems” , The Autumn Conference 2005 of Operations Research Society of Japan (ACORSJ·05), Kobe, Japan, September 14-15, 2005。
- [42] 鄒忠毅、李定國, “知識進化演算法與最佳化問題” , 台灣作業研究學會 2005 年研討會, Nov 4, 2005, 台北。