行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　　　　　　　　　　　　　　　　　　　　　※

※　　利用遺傳演算法推測蛋白質立體結構　　※

※　　　　　　　　　　　　　　　　　　　　　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫

計畫編號：NSC　89－2213－E－002－020－

執行期間：　88　年　8　月　1　日至　89　年　7　月　31　日

計畫主持人：陳文進

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

執行單位：台灣大學資訊工程學系

中　華　民　國　90　年　2　月　27　日

行政院國家科學委員會專題研究計畫成果報告

國科會專題研究計畫成果報告撰寫格式說明

Preparation of NSC Project Reports

計畫編號：NSC 89-2213-E-002-020

執行期限：88 年 08 月 01 日至 89 年 07 月 31 日

主持人：陳文進　台灣大學資訊工程學系

一、中文摘要

本計畫旨在討論計算生物學裡面一個極為重要的課題：蛋白質立體結構推測問題。蛋白質立體結構推測對於人類研究生物有非常重要的意義，如果能夠藉由電腦計算而推得較精準的蛋白質結構的話，將可以加速人類對於很多疾病及生物上面的研究。蛋白質的很多特性與功能是和它實際的立體結構非常相關的，然而通常直接去決定某種蛋白質的結構，通常不是不可能就是代價太高，藉由一些方法的設計與協助，生物學家可以以較低的代價求得蛋白質可能的結構，然後再用實驗加以驗證。這些推測很多是基於最低潛能函式或蛋白質序列來比較進行的。

蛋白質立體結構推測問題的主要目的如下：

1. 給定蛋白質第一級結構（氨基酸序列），求得蛋白質第三級結構（立體結構）的形狀以及各個元素的相對位置。
2. 對於所有的蛋白質存在一個可以完全敘述其能量以及形狀的潛能函式。

在本計畫中，我們將蛋白質立體結構推測的問題，轉換成為分子機制的最佳化問題，並結合遺傳演算法和生物界所提供的知識和數據，正確且快速的提供接近最佳化的解，以提高蛋白質立體結構推測的正確性，縮短時間，降低人類研究蛋白質時所花費的成本。並從已經存在的目標函式出發，尋找更適合的目標函式，再用相關發展經驗和數據，嘗試重新定義模組，以提高蛋白質立體結構推測模擬的準確性。

關鍵詞：計算生物學、蛋白質立體結構推測、遺傳演算法、最低潛能函式

Abstract

This project aims at investigating the theory of computational biology. We are particularly interested in a very important topic in computational biology, namely, protein folding problem. Protein folding problem is very important when people do researches on biological field. If we can get accurate protein structure by means of computer, the process of studying in diseases and biology will speed up. Many characteristics and function of proteins are deeply relational to their physical structure. But it would take a high computational expense to directly decide structure of a protein. By using some useful approaches, biologist can get some possible structure by paying lower cost and then design some experiments to refine the structure. Often, many approaches predict the structure base on the minimizing the potential energy function and protein sequences.

The main purpose of protein folding problem is to study:

1. Given the protein primary structure (linear, amino acids sequences), one can get the tertiary (3-dimensional) protein structure, the shape and the relative position of all atoms.
2. Find a potential energy function, which can appropriate represent the protein's energy and shapes.

In this project, we will transform protein-folding problem into optimization problems based on molecular mechanism. We also incorporate genetic algorithms and biological knowledge and experiments in order to increase the correctness of the result.

shorten the time, and provide a nearly optimal solution efficiently, lowering the cost of people doing researches or protein. Further, we will try to develop new potential functions from old functions, and construct new model for this problem. Finally, we use the relative development experiments and data to prevent from repeat testing protein structure and promote the quality.

Keywords: Computational biology、protein folding problem、Genetic Algorithm、minimum potential energy function

二、緣由與目的

演化式計算是以達爾文的「物競天擇」為基礎，主要可分為遺傳演算法(GAs: Genetic algorithm)、演化式策略(ESs: Evolution strategies)和演化式規劃(EP: evolutionary programming)等三種方法論，這些方法皆是健全且與問題獨立的機率演化式之最佳化方法論，因此他們可應用在多方面，包含組合性最佳化、函數係數的最佳化、工程控制和自動調適、機器學習等方面，而且大部份皆可求得令人滿意之近似最佳解的近似解。遺傳演算法著重染色體的運算如交配運算(crossover operator)，演化式策略和演化式規畫則強調父子代演化間的行為關係，因此著重突變運算(mutation operator)。

其中遺傳演算法是通用而且非常有用的機率性最佳化理論，其應用範圍相當廣泛，尤其是在組合性最佳化的問題上，如排程問題、VLSI 設計問題、集合涵蓋問題、旅行問題等重要的應用領域。但傳統的方法對於許多問題有些限制，且效果不佳，我們先前的研究，以改進這些缺點並明顯增加執行的效果，此方法以成功的應用於類神經網路、函數最佳化和非線性規畫等問題上。具備解決大型計算問題的能力，而計算生物學正是一種需要大型計算的領域。

計算生物學（computational biology）是一門新興的領域，主要是研究生物學應用上據計算複雜度的問題，它吸引了許多計算機科學家、分子生物學家、數學家、物理家、…等極投入的研究。由於大部份的研究主題都是關於分子生物學，所以有些學者又稱這一新興的領域為計算分子生物學（computational molecular biology）。

就計算生物學而言，大概可分為幾個範疇：建構相關基因圖譜、建構實質探針基因圖譜（DNA 片段重組）、解讀 DNA 上所有的核酸序列、推演生物演化樹、蛋白質立體結構推測、藥物設計等重要的方向。在過去的五十年中，蛋白質立體結構推測問題引發相當多的研究人員的注意。所謂的蛋白質立體結構推測問題問題即為：給定一個蛋白質的第一級結構（蛋白質序列），找出他的第三級結構（立體結構）。那為何我們需要蛋白質的立體結構呢？因為蛋白質的很多特性與功能是和它實際的立體結構非常相關的，所以要研究蛋白質就得先找出他的立體結構。就最近的生物技術來說，大都是以核磁共震技術（NMR）或是 X 光繞射來推得；然而隨著利用 DNA 推得的蛋白質序列數量暴增，這些方法不再能滿足（速度太慢），更重要的是，這些方法花費大量的人力及設備資源。此外，這些生物儀器進步的效率約為 18 個月增加一倍，電腦的硬體速度、演算法、軟體進步的速度遠比這個倍率要來的高。所以我們可以預期的是，最終以電腦為主的蛋白質立體結構預測將會成為真正的主流。

蛋白質立體結構推測問題的主要目的如下：

1. 給定蛋白質第一級結構（氨基酸序列），求得蛋白質第三級結構（立體結構）的形狀以及各個元素的相對位置。
2. 對於所有的蛋白質存在一個可以完全敘述其能量以及形狀的潛能函式。

然而以目前而言，直接去決定某種蛋白質的結構，通常不是不可能就是代價太高。我們希望藉由遺傳演算法的設計與協助，利用電腦的計算讓生物學家可以以較低的代價求得蛋白質可能的結構，然後再用實驗加以驗證，以加速人類對蛋白質的研究。

國內在分子生物方面的研究，常常因為重複許多繁複的過程，反覆的試驗，才能得

到結果。而在蛋白質結構的推測部份往往需要使用大量的資源與人力計算，故在研究蛋白質的過程中，研究他的結構便花去了大部份的時間。由於目前的計算生物學發展未臻完善，且這些問題通常需要大量的計算，所以我們必須提出新的方法解決這些大型資料比對的系統。因此持續而有效的發展計算生物學將是推動國內分子生物學茁壯的必要條件。

在過去的五年中，國外有許多的研究在各個計算型的研究團體中展開，其中包括 cubic lattice model 的簡化法、小型的元素(atom)模擬、中間產物(intermediate)的推演以及直接利用傳統的微分方程式求解，這些研究都有一些初步的成果，然而卻未能十分成功的應用在實際的例子上。而在國內的部份，對於此問題則尚未出現有關的研究，故本計畫就此問題提出一些可能的解決方法，希望能成功的定義出新的潛能函式，正確而快速的求出蛋白質的立體結構。

三、結果與討論

我們利用遺傳演算法來對蛋白質的最低潛能函式作最佳化。在 GA 的部分，我們利用數個知名的遺傳運算子，而最低潛能函式則是使用原子原子間的凡得瓦耳力公式，e 、 r0 為常數，V = e[(r0/r)e12 - 2(r0/r)e6]，簡單的說就是原子和原子之間是否發生了碰撞，並考慮氫鍵及雙硫鍵之間特殊能量。蛋白質折疊時，一開始將各個氨基酸的支鍊視為一固定大小球狀物，等到 GA 收斂之後，在將整個蛋白質作整體的最佳化。運算原子和原子之間的能量時，距離在 10 A 以外的原子就不列入計算。但我們發現使用現在的 fitness function 都無法完全使結果的 R.M.S.D 和 potential energy 成正相關。由目前的結果來看，GA 所演算出的結構和由 protein data bank 中由實驗已知結構的 R.M.S.D 約為 2 左右。

四、參考文獻

[1] E. Althaus, O. Kolbacker, H.-P. Lenhof, and P. Muller, "A combinatorial approach to protein docking with flexible side-chains," in the Fourth Annual International Conference on Computational Molecular Biology, pp. 15-24, 2000.

[2] A. R. Leach, "Ligand docking to proteins with discrete side-chain flexibility," J. Mol. Bilo., 235, pp. 345-356, 1994.

[3] R. M. Jackson, H. A. Gabb, and M. J. E. Sternberg, "Rapid refinement of protein interface incorporating solvation: Application to the docking problem," J. Mol. Biol, 276, pp. 265-285, 1998.

[4] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai, "Protein folding: The endgame," Annu. Rev. Biochem. 66, pp. 549-579. 1997.

[5] J. W. Ponder and F. M. Richards. "Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes." J. Mol. Biol., 193, pp. 775-791, 1987.

[6] A. Liwo, J. Pillardy, C. Czaplewski, J. Lee, D. R. Ripoll, M. Groth, S. Rodziewicz-Motowidlo, R. Kazmierkiewicz, R. J. Wawak, S. Oldziej, H. R. Scheraga, "UNRES -- a united-residue force field for energy-based prediction of protein structure - origin and significance of multibody terms," in the Fourth Annual International Conference on Computational Molecular Biology, pp. 193-200, 2000.

[7] J. R. Bienkowska, L. Yu, S. Zarakhovich, R. G. Rogers, T. F. Smith, "Comprehensive statistical method for protein fold recognition," in the Fourth Annual International Conference on Computational Molecular Biology, pp. 76-85, 2000.