行政院國家科學委員會專題研究計畫成果報告

以演算式計算推測蛋白質立體結構之研究(1/2)

Protein Structure Prediction by Evolutionary Strategy

計畫編號：NSC 90-2213-E-002-140
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：高成炎 教授　台灣大學資訊工程學系
計畫參與人員：xxxxxx　執行機構及單位名稱

一、中文摘要

若已知蛋白質立體結構，則可進一步推測出此蛋白質的作用物，及實際的功能。鑑於蛋白質序列與立體結構數量上的差異越來越大，如何快速的預測蛋白質立體結構已成當務之急，因此本計畫將藉由演化式計算方法來預測蛋白質結構。本計畫的前半部先著眼於蛋白質支鏈預測的子題，發展支鏈預測所需的蛋白質支鏈資料庫，以及建構演化式計算方法的雛形，以便之後推廣到預測整個蛋白質立體結構。

關鍵詞：蛋白質結構預測、演化式計算、蛋白質支鏈預測

Abstract

We can infer the function of a protein and the mechanism of reaction from the given protein structure. So, for figure out the functions of millions of proteins with sequences, the approach to predict protein structure efficiently is needed. This project aims at predicting protein structure by an evolutionary strategy method. We focus on the prediction of protein side-chain conformations in the former part of this project, to build the rotamer library, and construct the prototype of evolutionary strategy method, which will be extended to predict the whole protein structure in the later part of this project.

Keywords: Protein Structure Prediction、Evolutionary Strategy、Protein Side-Chain Prediction

二、Introduction

在人類基因體解碼計畫之後，人類對於生命的藍圖有了更進一步的認識。藉由對於基因更深入的瞭解，解開生命奧秘的一天便指日可待。而基因是藉由產生蛋白質來執行各種功能，因此蛋白質可說是一切生命現象的基礎，從生物體中的器官組織、運輸氧氣的血紅素、消化道的分解酵素、免疫系統的抗體抗原、甚至遺傳物質的複製轉譯等等，全都要靠特定的蛋白質、以及蛋白質與蛋白質之間的互動來完成。人類及所有生物可以說是由蛋白質為零件所建構而成的，所以如果能夠了解每一個蛋白質的功能與特性，科學家便可以根據這些知識，來瞭解人類的生理現象、疾病起因，而加以克服治療。甚至更進一步，科學家可以根據所需要達到的目的來設計蛋白質。

而蛋白質的功能，與蛋白質的結構有很大的關係。酵素與抗體、抗原等的專一性，全都是因為其結構上的互補所導致。所以如果能夠找出一個蛋白質的 3D 結構，我們就可以了解這個蛋白質的特性、以及它的功能。由於 DNA 解碼技術的突飛猛進，使得蛋白質序列與結構之間數量上的差異越來越大。傳統取得蛋白質結構的方式主要有 X-ray 晶體繞射與 NMR 等，但都耗時耗力，且有諸多限制。於是如何應用電腦強大的運算能力，與現有的資訊科技來預測蛋白質立體結構，成為生物資訊相關領域中一個很重要的課題。

因此本計畫主要的目的便是要發展出

一套演化式演算法，來預測蛋白質的立體結構。而在計畫的前半部份，我們首先著眼於支鏈結構預測這個子題，作為進入這個問題的踏腳石。

當蛋白質主鏈（main chain）的結構固定時，支鏈（side chain）的結構便決定了整個蛋白質的能量是否穩定。因此若能準確的預測支鏈的結構，對於預測蛋白質三級結構的問題有很大的幫助。

蛋白質是由二十種氨基酸所排列組合而成，而這二十種氨基酸就只有其支鏈之處不同。蛋白質在自然界中，會自動的摺疊（folding）成一定的形狀，因為這樣的形狀會使得整個蛋白質的能量最低，使其結構最穩定。於是我們所要預測的結構，便是所有的結構中能量最低的一個。

在本計劃中，我們使用凡德瓦爾位能（van der Waals）與單鍵的旋轉（torsion energy）位能，來描述一個結構的能量。本計劃中所採用的 FCEA 演算法，使用了 self-adaptive and decreasing-based Gaussian mutation 與 Rotamer mutation，結合了連續與離散型態的搜索，可以有效的找到全域最佳解（global optimal）。

由這個階段所建構的 Rotamer Library 以及整個演化式的演算法架構，都將在下一個階段加以延伸，用來預測整個蛋白質結構。

三、Method

我們所發展的演化式演算法能有效預測支鏈結構，演算法中使用家族競爭演化式方法（Family Competition Evolutionary Approach），加上以旋轉資料庫（Rotamer Library）為基礎所發展出來的旋轉異構物突變算子（Rotamer Mutation Operator），來尋找所要預測的蛋白質支鏈結構。

當蛋白質在自然界中摺疊（folding）時，會摺疊成為最穩定的狀態，也就是能量上最低的形狀。因此，我們若要預測蛋白質自然狀態（native state）的結構，只要在所有可能產生的結構中，找出能量最低的一個，便是我們要找的結構。

我們所要尋找的解空間，是由所有支鏈上每個可以旋轉的單鍵（single bond），任意旋轉角度所形成的結構。對於每一個結構，我們使用其凡德瓦爾位能（van der Waals energy）與單鍵旋轉的旋轉位能（torsion energy）的總和，來衡量這個結構的能量。為了要在龐大的解空間中，尋找我們所要的最穩定的結構。我們以結構的能量作為得分函數，使用由楊進木博士所開發之家族競爭演化式方法（Family Competition Evolutionary Approach）[Yang J. M, PhD thesis, 2001]來尋找能量的最低結構，也就是求最佳解（optimal）。在 FCEA 中我們除了 self-adaptive 與 decreasing-based Gaussian mutation 外，使用了以 Rotamer Library 為基礎的 Rotamer mutation operator。將無限大的連續性解空間，簡化成由數個狀態排列組合而成的離散空間，加快搜索的速度，使得收斂的時間縮短。

在本方法中，每個氨基酸的支鏈的構形（conformation），依照支鏈的長短，使用一到四個不等的旋轉角度（chi angle）來描述;FCEA 演算法藉由重組運算與突變運算不斷的產生不同旋轉角度的向量，然後計算其相對應結構的能量，一直到解收斂，便是我們所預測的結構。

在本報告中，我們將介紹家族競爭演化式方法及 Rotamer 突變運算子。另外將針對 Rotamer Mutation，介紹 Rotamer Library 的建構方式。最後將指出本方法中所使用的能量模型。

(一) FCEA

家族競爭演化式方法（FCEA）是一個多運算子（multi-operator）的方法，其中整

合了 decreasing-based Gaussian mutation，and self-adaptive Gaussian mutation。在以下的方法裡，我們加入了新的運算子，Rotamer mutation operator，使其能夠進行離散型態（discrete type）的搜索。FCEA 與其他的演化式演算法最大的差異之處，在於 FCEA 加入了家族競爭的觀念。藉由於家族競爭，保存了一開始個體之間的差異性（diversity），使得 FCEA 比起其他演化式演算法更不容易掉進區域最佳解（local optimal）的瓶頸之中，是一個有效的全域最佳解（global optimal）搜尋演算法。

(二) Rotamer Mutation

在 Rotamer 突變運算中，我們以一個氨基酸為基本單位進行突變，每個氨基酸有一到四個不等的支鏈旋轉角度變數（χ_1 ~ χ_4 角度）。我們由 Rotamer Library 中，將每種不同的氨基酸，在資料庫中出現機率最高的前十種構形（conformation），作為每種氨基酸突變的樣板（template）。當突變發生時，發生突變的氨基酸，便於十個樣板中隨機選擇一個，作為其突變後的構形。則突變的氨基酸上每一個旋轉角度變數，都以樣板的角度重新設定。

氨基酸支鏈上的單鍵，可以 360 度任意的旋轉，但為了有效搜尋解空間，我們將連續性的旋轉角度，簡化為離散的數個狀態，我們稱之為 Rotamer States。藉由統計資料庫中已知的蛋白質結構，我們得到每個氨基酸所在這些 Rotamer States 中的機率分布，並且計算所有位於這些離散區間中角度的平均值。建立這樣一個紀錄每個狀態的機率與平均角度的資料庫，我們稱之為氨基酸單鍵旋轉資料庫（Rotamer Library）。

Rotamer 突變運算子，便是以 Rotamer Library 為基礎，將每種氨基酸出現機率最高的十種狀態，作為其突變後的樣板。有些氨基酸的 Rotamer state 不到十種，如 Valine 只有三個 Rotamer State，我們便取這三種狀態為 Valine 在 Rotamer mutation 中的樣板。當氨基酸進行突變時，我們隨機選擇一個樣板，取代原來的支鏈構形。亦即將原來氨基酸支鏈上單鍵的旋轉角度，都置換成樣板 Rotamer state 由 Rotamer Library 所得到的平均角度。

與前面所述的 self-adaptive and decreasing-based Gaussian mutation 最大的不同處，在於 Rotamer mutation 並不是藉由調整步距，連續性（continuous）的搜索解空間；而是離散（discrete）的藉由數個狀態排列組合來搜尋解空間中的最佳解。雖然使用這樣的運算子並不一定能夠找到真正的最佳解，但卻能很有效的快速搜索解空間，可以加快收斂的速度。

(三) Force Field

在本方法中，我們假設所有的鍵長固定，鍵與鍵的夾角也固定，並且不考慮電荷的影響。所以計算一個蛋白質的能量，只單純的考慮凡德瓦爾位能（van der Waals energy）與單鍵旋轉的扭力（torsion energy），其式子如下：

$$E_{total} = E_{vander} + \alpha E_{torsion}$$

其中 E_{vander} 為凡德瓦爾位能，$E_{torsion}$ 為單鍵旋轉的能量，α 為一常數，在這裡我們使用 $\alpha = 2$。

由於蛋白質的主鏈（main chain）假設是已知的，所以我們將主鏈上的所有原子都固定住，並且在計算能量時不將主鏈上的原子列入計算，所以接下來的步驟都只考慮支鏈上的原子，同時我們也不考慮氫原子的能量。

我們依照蛋白質序列的順序，依序的計算每個氨基酸支鏈上的原子，與其餘的氨基酸支鏈上的原子之間的凡德瓦爾位能，並且將它們加總起來為 E_{vander}：

$$E^{vander} = \sum_{\substack{non-bonded \\ atom\ pairs\ i,j}} \varepsilon \left[\left(\frac{r^0}{R} \right)^{12} - 2 \left(\frac{r^0}{R} \right)^{6} \right] + H\_bond + disulfide\_bond$$

由於凡德瓦爾位能在距離很遠時趨近於零，所以我們只計算距離在 10Å 以內的氨基酸上的原子。另外我們也考慮氫鍵（hydrogen bond）與雙硫鍵（disulfide bond）的生成，有機會生成氫鍵或雙硫鍵的原子，我們採用特殊的能量計算公式。

然後我們依序計算每個氨基酸支鏈上，每個單鍵旋轉所造成的旋轉位能，乘上其加權，並且將之加總起來為 $E_{torsion}$：

$$E_{torsion}=\sum_{i=1}^{residue}\left[\sum_{j=1}^{chi}(1/j)A\{1-\cos[n\alpha-\alpha_0]\}\right]$$

最後乘以常數 α 加入 E_{vander} 得到 E_{total}，當成這個結構的總能量，若能量越低，便表示這個結構越穩定，也越接近自然界中的結構。

四、Results

我們使用了十二個蛋白質序列來測試本方法，並假設蛋白質資料庫中的結構為其自然的狀態（native state），與本方法的結果進行比較。我們採用了兩個評斷標準來衡量一個預測結果的準確性，一為預測的結構與資料庫中的結構兩者之間的方均根偏差（r.m.s.d）；第二個數值為預測 χ_1 角度的正確率（accuracy）。另外，我們也分別列出了考慮所有的氨基酸殘基（residues）與只考慮核心氨基酸殘基（core residues）兩種情況下的結果，並加以比較。在 Table1 中，我們分別列出了收斂時整個族群的平均值（average）以及最佳的個體（best）的值。

四、Conclusion and Discussion

由 Table1 中我們可以看到在考慮所有殘基時，RMSD 平均為 1.915，平均 χ_1 角度的正確率 70.7%。在只考慮核心殘基的狀況下 RMSD 平均為 1.294，正確率為 83%。由於核心殘基所處的環境較為緊密，其自由度較小；且與溶劑的接觸面較小，因此大多不帶電荷，所以預測的結果較準確。本方法所得到的結果與目前已發表的其他軟體有相似的結論。並且，在本計畫的下一階段，我們將會延伸整個演算法，使用到主鏈的預測上。藉由修正演算法中的染色體定義，與突變運算元，本方法的架構可以很容易的應用到整個蛋白質結構的預測上。但由於主鏈預測的自由度更高，演算法中所用到的能量公式將需要更深入的研究，才能夠得到較佳的結果。這些都是本計畫下一部努力的方向。

五、參考文獻

[1] Dumbrack,R. and Karplus,M. (1993) “Backbone-dependent Rotamer Library for Proteins: Application to Side-chain prediction.” *J. Mol. Biol.*, **230**, 543-574.

[2] Dunbrack, R. L. Jr. and Cohen. F. E. (1997) "Bayesian statistical analysis of protein sidechain rotamer preferences ." *Protein Science*, **6**, p1661-1681

[3] Laughton,C. (1994) “Prediction of Protein Side-chain Conformations from Local Three-dimensional Homology Relationships.” *J. Mol. Biol.*, **235**, 1088-1097.

[4] Lee,C. and Subbiah,S. (1991) “Prediction of protein side chain conformation by packing optimization.” *J. Mol. Biol.*, **226**, 507-533.

[5] Samudrala,R. and Moult,J. (1998) “An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction.” *J. Mol. Biol.*, **275**, 895-916.

[6] Yang J. M., Kao, C. Y., and Horng, J. T.. (1997) “A continuous genetic algorithm for global optimization” *Proc. Of the Seventh Int. Conf. On Genetic Algorithms*, p 230-237

[7] Yang J. M.. (2001) “A Family Competition Evolutionary Approach of Global Optimization in Neural Networks, Optical Thin-film Design, and Structure-based Drug Design” *PhD thesis*, NTU

Table 1 The results of side-chain prediction by using FCEA

		All Residues				Core Residues			
		average		best		average		best	
PDB code	Residues	r.m.s.d	chi1	r.m.s.d	chi1	r.m.s.d	chi1	r.m.s.d	chi1
1crn	46	1.536	0.738	1.411	0.865	1.128	0.889	1.128	0.889
1ctf	68	1.484	0.815	1.400	0.851	0.402	1.000	0.402	1.000
1lz1	130	1.881	0.748	1.741	0.790	1.248	0.833	1.191	0.870
2cro	65	2.390	0.653	2.363	0.709	1.346	0.767	1.112	0.833
2fox	138	1.901	0.745	1.869	0.771	1.161	0.888	1.067	0.920
2tmn	317	2.068	0.654	1.959	0.675	1.598	0.764	1.467	0.792
3app	323	1.725	0.680	1.591	0.722	1.541	0.820	1.431	0.828
3apr	325	1.894	0.664	1.811	0.690	1.576	0.758	1.569	0.788
3fxn	138	1.915	0.692	1.875	0.720	1.485	0.742	1.358	0.771
3tln	316	2.081	0.652	1.978	0.679	1.744	0.719	1.655	0.752
5pti	58	2.243	0.683	1.976	0.739	1.057	0.923	1.018	0.923
7rsa	124	1.861	0.717	1.703	0.752	1.242	0.861	1.041	0.907

六、相關著作

附錄刊登在 **MEDICAL INFORMATICS SYMPOSIUM TAIWAN**

國際醫學資訊研討會 **MIST2001** Conference Paper

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　以演算式計算推測蛋白質立體結構之研究(1/2)　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：■個別型計畫　　□整合型計畫

計畫編號：NSC　90－2213－E－002　－140　－

執行期間： 90 年 08 月 01 日至 91 年 07 月 31 日

計畫主持人：高成炎 教授

共同主持人：楊進木 教授

計畫參與人員：陳永松 博士
蔡懷寬
陳鍾城
林柏亨
莊涵宇
黃崇隆

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：國立台灣大學資訊工程學系

中　華　民　國　　91 年　　5 月　　31 日