樹 德 科 技 大 學

資訊管理研究所
碩士學位論文

以支持向量機預測蛋白質三級結構

Prediction of Protein Tertiary Structure-Using Support Vector Machine

研　究　生：黃 進 南
指 導 教 授：董 信 煌

中 華 民 國 九十五 年 一 月

以支持向量機預測蛋白質三級結構

Prediction of Protein Tertiary Structure-Using Support Vector Machine

研　究　生：黃 進 南
指 導 教 授：董 信 煌

樹德科技大學
資訊管理系碩士班
碩士論文

A THESIS
SUBMITTED TO
INSTITUTE OF INFORMATION MANAGEMENT
SHU-TE UNIVERSITY OF SCIENCE & TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER
IN
INFORMATION MANAGEMENT

JANUARY 2006

中華民國 九十五 年 一 月

摘 要

扭角(Torsion Angle)是影響蛋白質結構的重要關鍵因素之一。若能精準的預測出扭角，則對於決定蛋白質的結構具有極大的幫助，一旦決定了蛋白質的結構後便能清楚的了解其功能。

傳統的 X 光繞射與核磁共振(NMR)可以正確無誤的找出蛋白質結構，但此兩種方法必需花費許多的時間與成本方能完成。於是人類利用電腦來計算與預測，便可減少許多時間與成本的花費。

本研究的目的為預測蛋白質主鏈的三級結構，首先是利用 BLAST(Basic Local Alignment Search Tool)找出測試蛋白質(測試集)的同源序列(訓練集)，並建立其 PSSM(Position Specific Scoring Matrix)，經過編碼後(PSSM+二級結構)，以 SVM(Support Vector Machine)進行訓練與預測其扭角 ϕ (PHI)、ψ (PSI)、ω (OMEGA)和三個鍵角 $\Delta CNCA$、$\Delta NCAC$、$\Delta CACN$，接著將預測的結果代入旋轉公式，計算測試蛋白質主鏈上的原子 3 維座標，評估方式為計算主鏈上 CA 原子的 RMSD(Root Mean Square Deviation)。

最後，取其它文章的實驗蛋白質來進行預測並比較，結果顯示測試集與訓練集之間的序列相似度越高效果越好，這也顯示出未來仍然有許多可以改善的地方。

Abstract

Torsion angle is important factor of influence the protein structure. If we can predict torsion angle correct. It is useful for determining the structure of the protein. We can understand the protein function after determined structure of the protein.

Traditional X-rays diffraction and nuclear magnetic resonance (NMR) can find out the protein structure correctly, but they must spend a lot of time and cost. Then people use computer to calculate and predict, reduce a lot of time and cost.

The purpose of this research is that predict the tertiary structure of main chain of protein. First, use BLAST(Basic Local Alignment Search Tool) to find out the homology sequences(train sets) which target protein(test set) and create PSSM (Position Specific Scoring Matrix). After coded(PSSM and second structure), use SVM(Support Vector Machine) to train and predict torsion angle ϕ(PHI)、ψ(PSI)、ω(OMEGA) and three bond angle $\Delta CNCA$、$\Delta NCAC$、$\Delta CACN$. Then take the result of predicted into the rotation formula and calculate the 3D coordinate of atom. Evaluate the experiment, calculate the RMSD(Root Mean Square Deviation) of CA atom.

Final take experiment proteins of other paper to predict and compare. The results show, sequence's identity between test set and train set more high the results more better. And still there are a lot of places that can be improved in the future.

誌　謝

預測蛋白質三級結構是一項困難的工作，單憑個人之力絕對無法完成，幸好在這努力的背後，有著許多人的幫助與支持，在此僅能以言語來表達深深的謝意。

首先，感謝父母親，提供了一個讓學生無憂無慮的研究環境，並給予深深的支持。接著是指導教授董老師，董老師的知識學問實在是讓學生大開眼界，並且常常替學生解決各種疑問與傳授相關知識，可謂是本研究最大的推手。不過最讓學生在意的是董老師的待人處事之道，是相當值得學習與效法，光是這幾點就足以讓學生對自己的父母親以及指導教授獻上最深的謝意。

另外，在口試期間，吳委員與賴委員的寶貴意見，更是讓學生對於論文的寫作更上一層樓，吳委員那種啟發式的教導方式以及賴委員清楚、詳細的說明和鼓勵，對學生而言都是一大助力，在此對二位委員說聲幸苦你們了，非常感謝。

額外值得一提的是吉原學長，吉原學長搜尋資料的能力一流，而且對於研究的態度相當認真、努力，是值得學習的對象。在這整個過程中，感謝學長提供了許許多多的幫忙，不甚感激。

最後，對於那些無時無刻互相勉勵、互相關懷的同學們以及”煩事”找她就對的系助育貞，感謝你們，感謝大家，本人在此獻上最深的謝意，謝謝！

目 錄

表 目 錄

圖 目 錄

第一章　前言

1.1 研究動機

近年來，生物資訊成為越來越多人投入的一個領域，其原因在於越來越多的生物資料已經被解碼出來了，但是人們對於這些資料所代表的意義，卻依然非常的陌生。舉例來說，截至不久前為止，已經解出序列的蛋白質就有 283,416 筆(PIR-PSD，2005/3/1)，但是已經解出結構的蛋白質，卻只有 34,303 筆(PDB，2005/12/20)。若將生物資訊結合電腦快速的運算能力，配合資訊科學設計有效的演算法，以及生物科技提供豐富的生物知識背景，期望可以揭開生命的奧秘。

人類基因定序計劃(human genome project)只是一個開始，在產生的大量基因體序列之後，另一波的研究風潮將是蛋白質的三度空間立體結構，因為蛋白質三度空間立體結構的決定將有助於新藥的開發，透過分子藥物構造的搜尋，將大幅降低新藥開發所需的時間以及投資【13】。

1.2 研究目的

蛋白質三度空間立體結構的決定遠比 DNA 序列定序困難，自 1957 年第一個蛋白質 myoglobin 的立體結構被決定之後，至今四十八年間總共有 34,303 個蛋白質的立體結構已經被決定，也就是說，平均一年只解出七百多個蛋白質立體結構，即使現在的技術已經進步很多，但是蛋白質立體結構決定的速度仍然無法滿足需求。

採用實驗方法來決定蛋白質立體結構的速度非常慢，如核磁共振以及 X-ray 繞射結晶，此二種方法都可以正確無誤的找出蛋白質結構，但是並非所有的蛋白質都適合使用結晶的方法。因此，許多生物資訊公司或學者開始致力於研究預測蛋白質立體空間結構的方法。

蛋白質立體空間結構也可以稱為蛋白質三級結構。其實蛋白質三級結構說穿了就是原子的 3 維座標，決定蛋白質三級結構就是決定這條蛋白質上所有原子的 3 維座標，只要事先假設第一個胺基酸的原子座標，再利用機器人學的旋轉公式配合鍵長將其它胺基酸的原子座標一一計算出來，便能得到一個蛋白質三級結構。在公式中需要代入一些角度，而蛋白質三級結構裡則有扭角、鍵角等等。如果可以知道這些扭角、鍵角，並將這些數值代入旋轉公式中並配合鍵長的計算，

便能計算出原子的 3 維座標，也就能決定蛋白質的三級結構。於是本研究便以預測出這些扭角、鍵角為目的。關於蛋白質結構會在第二章加以說明。

1.3 研究目標

蛋白質三級結構是由一條主鏈加上多個側鏈組合而成的；主鏈上的原子比較固定，例如：N、CA、C、O 等。側鏈上的原子則會依胺基酸的不同而有所不同，也是比較難預測的。因此，本研究決定以預測蛋白質的主鏈為主，期望能夠達到，準確的預測出蛋白質的三級結構這個目標。

1.4 實驗流程

預測蛋白質的三級結構有好幾種方式，較常見的有：同源模擬法(Homology Modelling)、摺疊辨識法(Fold recognition)、重新計算法(Ab initio)等【25】【27】。不管用何種方法，最後都是要決定原子的 3 維座標。如果直接預測原子的 3 維座標的話，會產生一些問題，例如：維度、座標值等。座標的 3 個維度本身就是一個問題，接著就算蛋白質之間結構相似，也不能保證它們之間相對應的原子座標會接近，底下圖 1.1、1.2 就是一個例子。

```
 85  ATOM      1  N    LEU     1      12.275 -10.917  -3.962  1.00  1.69    1  2IGH  86
 86  ATOM      2  CA   LEU     1      11.733  -9.866  -4.870  1.00  1.15    1  2IGH  87
 87  ATOM      3  C    LEU     1      10.779  -8.939  -4.094  1.00  1.04    1  2IGH  88
 88  ATOM      4  O    LEU     1       9.851  -9.401  -3.459  1.00  1.18    1  2IGH  89
 89  ATOM      5  CB   LEU     1      10.953 -10.520  -6.056  1.00  1.78    1  2IGH  90
 90  ATOM      6  CG   LEU     1      11.874 -11.441  -6.925  1.00  2.13    1  2IGH  91
 91  ATOM      7  CD1  LEU     1      11.033 -12.137  -8.022  1.00  2.90    1  2IGH  92
 92  ATOM      8  CD2  LEU     1      12.947 -10.592  -7.636  1.00  1.95    1  2IGH  93
 93  ATOM      9 1H    LEU     1      11.859 -10.805  -3.015  1.00  2.00    1  2IGH  94
 94  ATOM     10 2H    LEU     1      12.039 -11.854  -4.345  1.00  2.17    1  2IGH  95
 95  ATOM     11 3H    LEU     1      13.310 -10.826  -3.899  1.00  1.53    1  2IGH  96
 96  ATOM     12  HA   LEU     1      12.559  -9.280  -5.241  1.00  0.74    1  2IGH  97
 97  ATOM     13 1HB   LEU     1      10.134 -11.101  -5.659  1.00  2.22    1  2IGH  98
 98  ATOM     14 2HB   LEU     1      10.538  -9.746  -6.686  1.00  1.78    1  2IGH  99
 99  ATOM     15  HG   LEU     1      12.344 -12.191  -6.307  1.00  2.18    1  2IGH 100
100  ATOM     16 1HD1  LEU     1      10.259 -12.745  -7.580  1.00  3.22    1  2IGH 101
101  ATOM     17 2HD1  LEU     1      10.574 -11.404  -8.669  1.00  2.99    1  2IGH 102
```

圖 1.1　2IGH 的 PDB 檔片斷，與 2IGG 結構相似

96	ATOM	1	N	LEU	1	-0.188	-1.206	2.037	1.00	2.07	1	2IGG 97
97	ATOM	2	CA	LEU	1	0.027	0.266	1.949	1.00	1.61	1	2IGG 98
98	ATOM	3	C	LEU	1	0.583	0.577	0.546	1.00	1.51	1	2IGG 99
99	ATOM	4	O	LEU	1	1.016	-0.312	-0.162	1.00	1.76	1	2IGG 100
100	ATOM	5	CB	LEU	1	1.037	0.698	3.062	1.00	1.36	1	2IGG 101
101	ATOM	6	CG	LEU	1	1.225	2.249	3.152	1.00	1.05	1	2IGG 102
102	ATOM	7	CD1	LEU	1	-0.095	2.954	3.544	1.00	1.25	1	2IGG 103
103	ATOM	8	CD2	LEU	1	2.287	2.587	4.219	1.00	1.10	1	2IGG 104
104	ATOM	9	1H	LEU	1	0.085	-1.651	1.138	1.00	2.20	1	2IGG 105
105	ATOM	10	2H	LEU	1	0.393	-1.596	2.807	1.00	2.12	1	2IGG 106
106	ATOM	11	3H	LEU	1	-1.193	-1.399	2.224	1.00	2.29	1	2IGG 107
107	ATOM	12	HA	LEU	1	-0.924	0.759	2.078	1.00	1.67	1	2IGG 108
108	ATOM	13	1HB	LEU	1	0.692	0.318	4.015	1.00	1.56	1	2IGG 109
109	ATOM	14	2HB	LEU	1	1.993	0.243	2.850	1.00	1.45	1	2IGG 110
110	ATOM	15	HG	LEU	1	1.574	2.634	2.207	1.00	1.02	1	2IGG 111
111	ATOM	16	1HD1	LEU	1	-0.447	2.596	4.501	1.00	1.48	1	2IGG 112
112	ATOM	17	2HD1	LEU	1	0.079	4.019	3.616	1.00	1.32	1	2IGG 113

圖 1.2　2IGG 的 PDB 檔片斷，與 2IGH 結構相似

為了簡化預測蛋白質三級結構的問題，本研究提出了與其它方法比較不一樣的做法。基本上，下一個原子座標是可以由上三個原子座標計算而出，其中必須利用機器人學的旋轉公式配合鍵長才能達成；第三章將有詳細的說明。但先決條件有二：

(1)　預先假設第一個胺基酸原子 N、CA、C 的 3 維座標

(2)　旋轉的角度

這裡的重點是在條件二。因此，預測蛋白質三級結構的的問題就直接簡化到預測這些角度即可，這些角度包括了主鏈上的扭角(Torsion Angle)、鍵角(Bond Angle)。

目前的問題已經轉換到預測蛋白質主鏈上的扭角與鍵角；至於預測工具，本研究採用 SVM(Support Vector Machine)進行預測，雖然還有其它預測方法，如類神經網路等。不過本研究還是對 SVM 比較感興趣。同時，它也具有統計學習理論的優點【23】。SVM 的簡易流程是訓練→模組→預測，訓練一些歷史資料成為一個模組，再以這個模組對測試資料進行預測。一般稱歷史資料為訓練集，測試資料為測試集。SVM 對於訓練集與測試集有自己的文件格式，為了能夠讓 SVM 認識這些資料，必須在訓練之前將資料進行編碼的動作。那麼，應該用哪些資料來當訓練集與測試集？在這部份，本研究使用 PSI-BLAST 軟體來完成。

假設要預測一條結構未知的蛋白質 P(測試集)，將蛋白質 P 交給 PSI-BLAST 來處理，PSI-BLAST 會找出與蛋白質 P 具有同源性質的其它蛋白質 P1、P2、P3…等等(訓練集)。接著，將測試集和訓練集的 PSSM 以及二級結構進行編碼，然後計算出訓練集的 6 種角度，再用 SVM 將訓練集的編碼和已知角度訓練出 6 個模組，再以這 6 個模組跟測試集 P 去預測出測試集的 6 種角度。最後，把這些角度代入公式裡並配合鍵長的計算，原子的 3 維座標就逐一被計算出來，一條蛋白質 P 的 3 級結構就此產生。圖 1.3 為實驗流程圖。

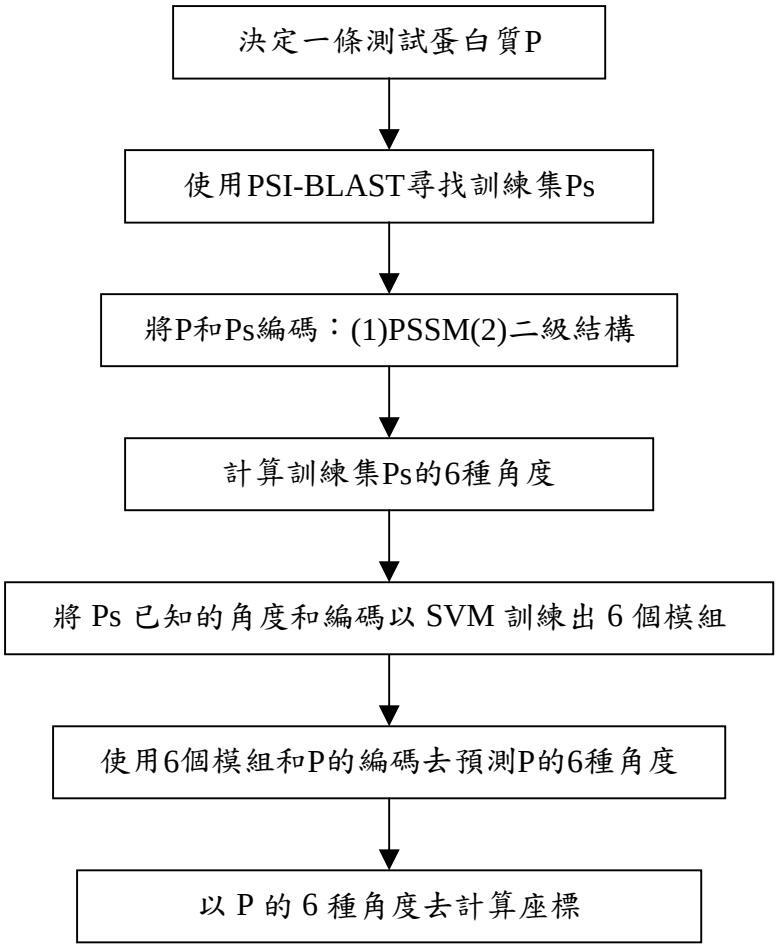

圖1.3　實驗流程圖

Ps 有可能為一條或多條蛋白質。6 種角度，其中包括 3 種扭角、3 種鍵角。以上是屬於比較簡單的介紹，詳細的情形會在第二、第三章節加以說明。

第二章　文獻探討

蛋白質的功能主要是決定於它的三級結構【13】，而蛋白質研究最終的目標是要確認蛋白質的功能，蛋白質功能的確認可以幫助藥物的開發或是醫學上的研究等，因此便有許多的學者投入於蛋白質三級結構的研究。

決定蛋白質三級結構的方法大致上可分成實驗法與預測法。實驗法則包含了X光繞射與核磁共振(NMR)，這二種方法可以準確的決定蛋白質結構，但相對的必須花費許多的時間以及成本。預測法方面，則有同源模擬法、摺疊辨識法、重新計算法等【25】【27】，這是結合了電腦的快速計算能力和有效的演算法以及生物學的專業知識。本研究則以預測法為主，不管使用何種方法，目的只為了要決定蛋白質的三級結構。

在第二章裡，將從蛋白質結構開始介紹，接著則是目前預測蛋白質三級結構的方法，例如：同源模擬法、摺疊辨識法、重新計算法等，並比較本研究的方法與它們之間的差異。最後則是針對研究裡會用到的演算法SVM做一探討。

2.1 蛋白質結構

蛋白質共分成4種結構，分別是一級結構(Primary structure)、二級結構(Secondary structure)、三級結構(Tertiary structure)、四級結構(Quaternary structure)等【21】，底下將針對這4種結構做一描述。

2.1.1　一級結構(Primary structure)

一級結構是胺基酸(Amino Acid)序列，是由二十種不同的胺基酸單字縮寫排列而成，二十種胺基酸縮寫如下表2.1所示，第一個欄位是胺基酸的中文名稱，第二個欄位是英文全名，第三個欄位是3個字母的縮寫，最後一個欄位則是單字縮寫。而每個胺基酸都有相同的基本結構，胺基酸的基本結構如圖2.1，圖中N、CA、C、O、H稱為原子，R則是側鏈(Side Chain)，側鏈中又有多個原子，胺基酸的差異就在於R的不同【31】。當二個胺基酸連結時，上一個胺基酸的O、H原子會跟下一個胺基酸的H原子形成H2O水分子而脫離這條序列，如圖2.2、圖2.3。蛋白質最後的結構請參考圖2.3，圖中胺基酸均以N、CA、C原子互相連結而成一條長鏈，稱為主鏈。

表 2.1　二十種胺基酸縮寫

表格來源：http://juang.bst.ntu.edu.tw/BCbasics/Amino1.htm

名　稱		縮寫	
甘胺酸	Glycine	Gly	G
丙胺酸	Alanine	Ala	A
纈胺酸*	Valine	Val	V
白胺酸*	Leucine	Leu	L
異白胺酸*	Isoleucine	Ile	I
苯丙胺酸*	Phenylalanine	Phe	F
酪胺酸	Tyrosine	Tyr	Y
色胺酸*	Tryptophan	Trp	W
組胺酸*	Histidine	His	H
天冬胺酸	Aspartic acid	Asp	D
天冬醯胺酸	[Asparagine]	Asn	N
麩胺酸	Glutamic acid	Glu	E
麩醯胺酸	[Glutamine]	Gln	Q
離胺酸*	Lysine	Lys	K
精胺酸*	Arginine	Arg	R
絲胺酸	Serine	Ser	S
蘇胺酸*	Threonine	Thr	T
OH-脯胺酸	Hydroxy Pro		
甲硫胺酸*	Methionine	Met	M
胱胺酸 雙胱胺酸 (3)	Cysteine Cystine	Cys	C
脯胺酸 (3)	Proline	Pro	P

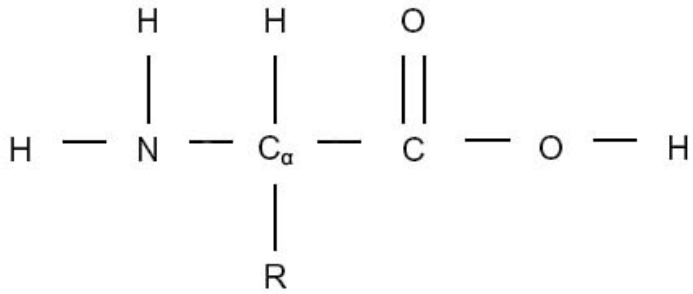

圖 2.1　胺基酸基本結構

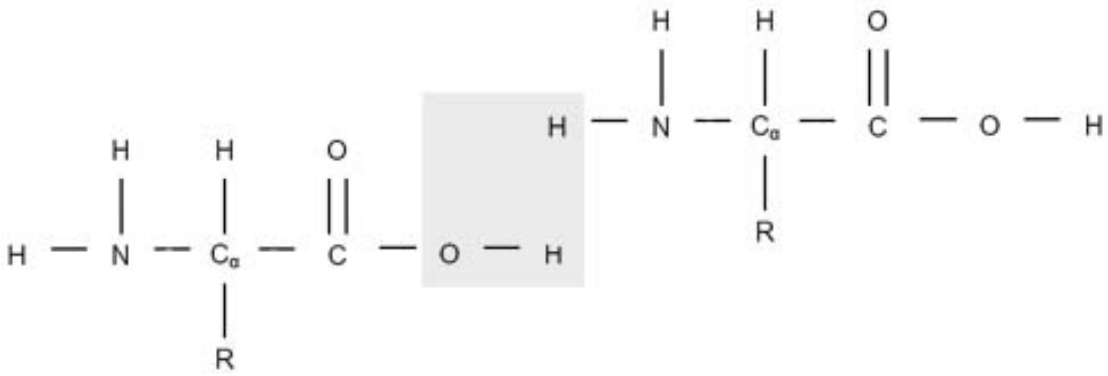

圖 2.2　胺基酸連結圖

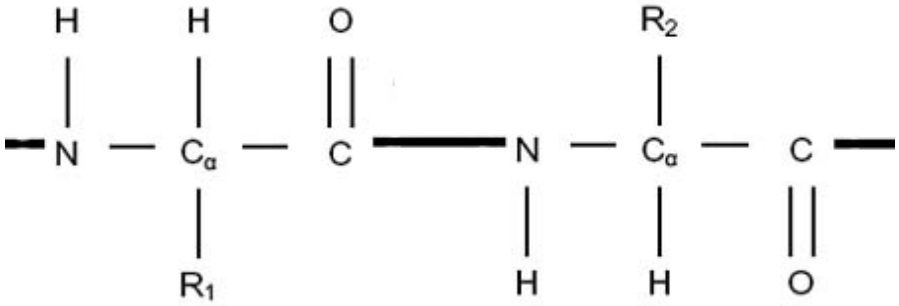

圖 2.3　蛋白質結構

舉個例子，底下是蛋白質 1P7E 的一級結構：

MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATK
TFTVTE

很清楚的可以看到 1P7E 是由 MQY…胺基酸單字縮寫所組合而成的一條長度為 56 的蛋白質，對照表 2.1 便可以了解這條蛋白質是由哪些胺基酸所組成。

2.1.2　二級結構(Secondary structure)

蛋白質二級結構是指相鄰胺基酸在空間中排列的情形，在不同的胺基酸序列片段，會形成一個固定的二級結構【13】。一般把二級結構分成3類，例如阿法螺旋(Alpha Helix)可參考圖2.4、貝塔摺板(Beta Sheet) 可參考圖2.5，剩下的部份則歸類成線圈(Coil) 【30】。通常以H代表阿法螺旋、E代表貝塔摺板、C代表線圈。

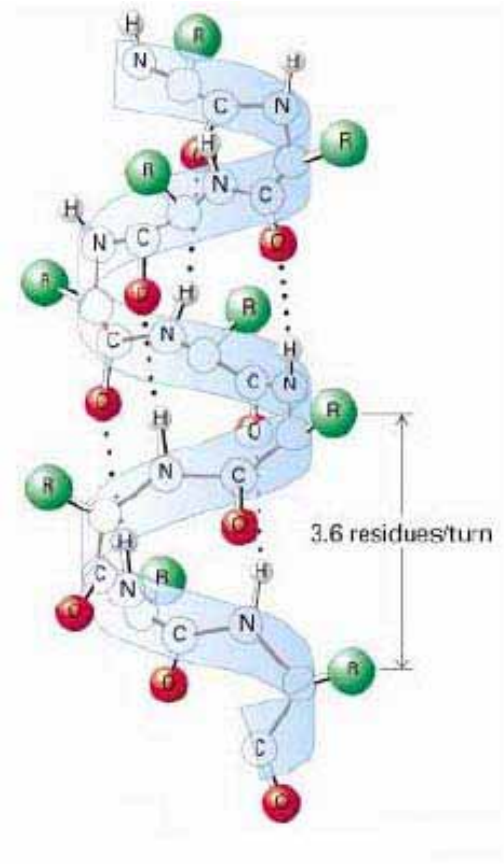

圖 2.4　阿法螺旋圖

圖片來源：【20】

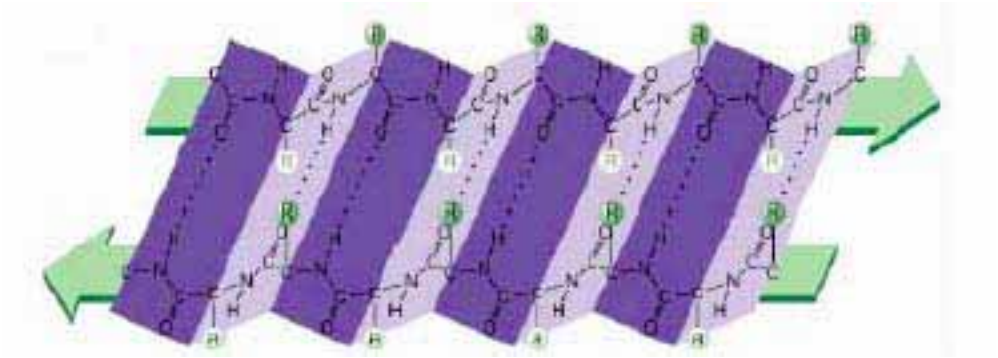

圖 2.5　貝塔摺板圖

圖片來源：【20】

其中結構最穩定的是阿法螺旋(H)，其次是貝塔摺板(E)，結構比較不規則的則是線圈(Coil)的部份。實驗中將以 H、E、C 三種分類來表示二級結構，底下例子用來表示每一個胺基酸所對應的二級結構。

一級結構：MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVW
二級結構：CEEEEEEECCCCCCCCCCEECCHHHHHHHHHHHHHHCCCCCEE

上述例子中可以清楚看到從第 2 個胺基酸 Q 開始一直到 N 胺基酸，這段的二級結構是貝塔摺板(E)；第 23 個胺基酸 A 開始一直到 D 胺基酸，這段則是阿法螺旋(H)。如此，以 H、E、C 來表示胺基酸是位於哪一種二級結構區段。

2.1.3 三級結構(Tertiary structure)

蛋白質三級結構是整個二級結構在 3 維空間中的表現，如圖 2.6。蛋白質的功能是決定於三級結構。因此，目前有許多的學者正投入於蛋白質三級結構的研究中。只要能夠知道結構裡原子的 3 維座標，再透過分子結構檢視軟體，便能觀察到蛋白質的三級結構。

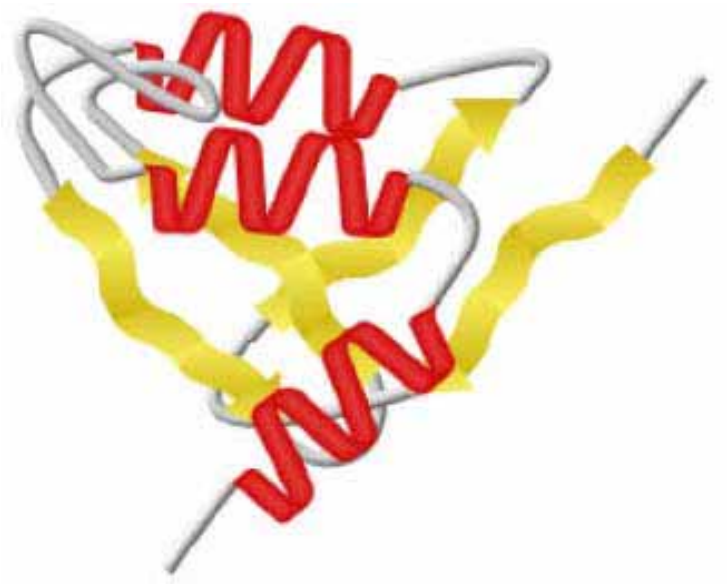

圖 2.6　蛋白質三級結構
圖片來源：【3】

2.1.4 四級結構(Quaternary structure)

蛋白質四級結構包含了多個三級結構，因此更複雜。圖 2.7 為蛋白質四級結構。

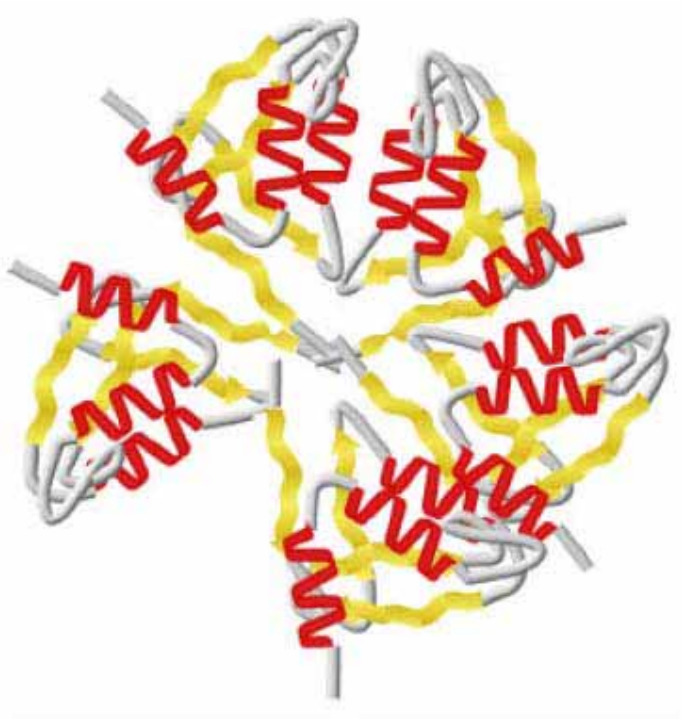

圖 2.7　蛋白質四級結構

圖片來源：【3】

2.2 蛋白質三級結構預測

目前對於蛋白質三級結構預測方面，比較常見的有同源模擬法(Homology Modelling) 【7】【25】【1】、摺疊辨識法(Fold recognition) 【9】【24】、重新計算法(Ab initio) 【28】，接著將針對這三種預測法做一介紹。

2.2.1　同源模擬法(Homology Modelling)

首先決定一條目標蛋白質(target)，接著從已知結構的蛋白質資料庫裡搜尋其相似序列，並以這些相似序列當成模板(template)。將目標蛋白質與模板做多重序列排比(MSA)。利用模板來決定目標蛋白質的結構，最後以分子動力學的能量最小化來做後續處理，步驟如下：

(1) 以一條未知結構的目標蛋白質(target)去搜尋出已知結構的模板(template)。
(2) 比對目標蛋白質和模板。
(3) 從模板決定目標蛋白質的骨幹結構。
(4) 針對缺口(gap)的部份做處理。

(5) 加入側鏈的結構。
(6) 最佳化側鏈的位置。
(7) 使用能量最小化來調整結構。

同源模擬法的第一個步驟是到結構資料庫去搜尋模板，結構資料庫如 PDB 或 Swiss-Prot 等。而資料庫的搜尋可以使用相似性搜尋工具，例如 PSI-BLAST。接著則是以多重序列排比(multiple sequence alignment)的方式對目標蛋白質和模板進行排比。多重序列排比的方式能夠獲得可信度比較高的結構資訊，最主要的目的是要得到同源蛋白質的結構排列。結構排列完之後則是決定目標蛋白質的骨幹結構，這裡可以以繼承的方式直接將模板的主鏈結構指定給目標蛋白質或是利用其它方法來決定目標蛋白質的骨幹結構。對於缺口或是環狀結構的部份可以運用 ab initio 或搜尋結構資料庫的方式來建置。一般說來，環狀結構的長度若是超過 5 個胺基酸，則困難度會增加以及準確性會降低，而用來評估環狀結構的計分函數，其品質決定環狀結構預測的準確性。若目標蛋白質的骨幹是從模板直接繼承，那側鏈結構亦可從模板直接繼承給目標蛋白質，否則就必須至資料庫搜尋可能的側鏈結構。最後則是利用能量最小化來修正不利的非共價碰觸，並滿足空間限制的條件。能量最小化是一連串的計算，包括鍵長，鍵角，二面角，靜電作用力(electrostatics)和 VdW 作用力等位能參數，它的能量計算公式：

$$E_{total} = E_{stretching} + E_{bending} + E_{dihedral} + E_{out\text{-}of\text{-}plane} + E_{cross\ terms} + E_{VdW} + E_{coulombic} \quad (2\text{-}1)$$

能量最小化常用的力場(force fields)有 CHARMM，AMBER 或 GROMOS 等，利用其中一個最適合的力場來計算最初結構的能量，再調整結構中上述的各項參數直到最低總能量被計算出來為止。

在 Comparative protein structure modeling by iterative alignment, model building and model assessment 文章裡，作者的做法是運用同源模擬法和基因演算法來預測蛋白質三級結構。當比對完目標蛋白質和模板後，便以同源模擬法建立多個目標蛋白質的結構，並計算 GA341 的分數，如果最佳的 GA341 分數沒有小於 0.6，則選擇一個最佳的結構並結束整個流程。如果最佳的 GA341 分數小於 0.6，則會使用基因演算法再重新進行比對並以同源模擬法建立目標蛋白質的結構，當此流程循環 25 次以後，則會對剩下的目標蛋白質結構計算另一種適性值，並選出最佳的結構，結束整個流程。這是屬於同源模擬法的應用之一。

2.2.2 摺疊辨識法(Fold recognition)

摺疊辨識法或稱為穿針引線法(Threading)，它是以測試蛋白質與已知結構的摺疊資料庫進行比對，並利用判別函數或是一些特性來決定測試蛋白質可能的摺

疊結構，大致上的步驟如下：

(1) 建立摺疊資料庫。
(2) 利用某些特性，例如胺基酸間的作用力或是親水性等等來當作判別函數，判別結構的相似程度。
(3) 使用探索性的方式來搜尋最佳樣板。探索性的方式，例如動態程式(Dynamic Programming)或其它。

2.2.3 重新計算法(Ab initio)

重新計算法試著不參考已知結構的任何資料，它利用其它方法，例如自然界的法則或理論為依據，建構出蛋白質的結構。常見的技術是分子動力模擬(Molecular Dynamic)，模擬胺基酸的摺疊過程。在 Genetic Algorithms and Protein Folding【28】一文中，當它產生一個新結構時，便會利用能量最小化或是適性值的計算再配合基因演算法(Genetic Algorithm)來決定蛋白質結構。

其實，只要知道所有原子的 3 維座標就可以了解蛋白質的三級結構，所以預測原子的 3 維座標也是一種預測蛋白質三級結構的方法。不過預測座標的方法過於困難，光是要預測主鏈上每個胺基酸的 N、CA、C 原子，就有 9 個變數(一個原子 3 個維度座標 X、Y、Z)，同時這也會增加預測時所花費的時間。為了簡化問題再加上扭角與蛋白質的二級結構是有相關的【26】，而二級結構對整個立體空間的形成有很重要的影響【11】【22】，因此本研究將這 9 個變數轉換到蛋白質主鏈上的 3 個扭角、3 個鍵角和 3 個鍵長，在這裡將鍵長設為固定常數，因此變數只剩下 3 個扭角和 3 個鍵角，也就是將本來預測 9 個座標變數換成預測 6 個角度變數。接著將對扭角、鍵角和鍵長做一介紹。

2.3 扭角與鍵角

所謂的扭角指的是二個相鄰平面之間的夾角【31】，可參考圖 2.8。蛋白質主鏈上有 3 個扭角，分別是 PHI(ϕ)、PSI(ψ)、OMEGA(ω)，如圖 2.9。而每個胺基酸的 CA 原子與側鏈之間尚有一種扭角 χ，不過本研究只針對蛋白質主鏈，因此暫不考慮 χ。

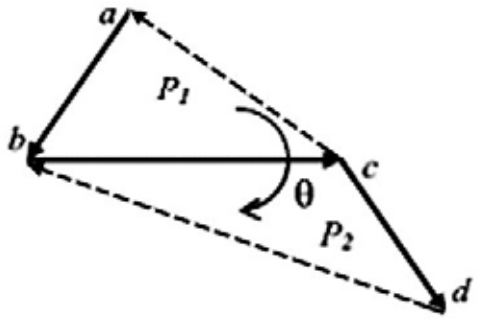

圖 2.8　扭角示意圖，P1 與 P2 二個相鄰平面之間的夾角

二點成線，三點成面。圖 2.9 中可以看到由胺基酸 i-1 的 C’原子和胺基酸 i 的 N、CA 原子三點形成平面 1；胺基酸 i 的 N、CA、C’原子三點形成平面 2。平面 1 與平面 2 之間的夾角即為扭角 PHI。同理，胺基酸 i 的 CA、C’原子和胺基酸 i+1 的 N 原子三點形成平面 3，平面 2 與平面 3 之間的夾角即為扭角 PSI。OMEGA 則是平面 3 與下一個平面 4 之間的夾角。

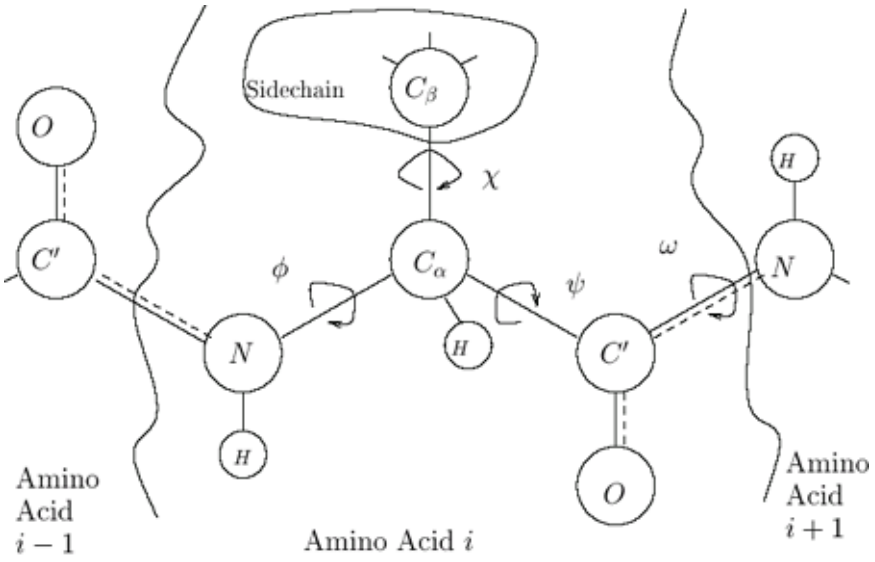

圖 2.9　胺基酸與扭角圖

圖片來源：【27】

扭角是介於-180~+180 之間，當中比較特別的是 OMEGA 扭角，在“Torsion Angle Selection and Emergent Non-Local Secondary Structure In Protein Structure Prediction”一文中提到 OMEGA 扭角有 99%是 180 度，其餘是 0 度。但事實上，

OMEGA 並非 180 度，只是很接近 180 度，甚至於會有 160 度的可能，所以 OMEGA 扭角對蛋白質結構影響也不小，因此將 OMEGA 也列為預測的對象，由表 2.2 可觀察得之。

表 2.2　1P7E 扭角統計表

表格來源：http://www.pdb.org

Dihedral Angle	Chain Id	Tot Num	Cal Ave	Cal StdDev	Std Val	Std StdDev	Minimum	Maximum
Chi1 trans	A	40	-23.18	112.030	183.6	16.8	-178.28	179.839
Omega	A	55	48.38	170.966	180	5.8	-179.96	179.955
Phi	A	41	-100.13	60.463	-65.3	11.9	-153.54	160.699
Phi helix	A	14	-64.30	3.682	-65.3	11.9	-70.75	-58.367
Psi	A	37	104.54	65.617	-39.4	11.3	-41.33	170.315
Psi helix	A	14	-40.24	5.089	-39.4	11.3	-46.46	-28.613
Psi(G)	A	4	-30.77	137.463	-39.4	11.3	-165.25	170.385

蛋白質結構裡實際上有許多鍵角存在，但本研究只針對蛋白質主鏈進行預測，也就是只考慮主鏈上的 N、CA、C 原子座標，主鏈上跟 N、CA、C 原子座標相關的鍵角是 $\Delta CNCA$ 、$\Delta NCAC$ 、$\Delta CACN$ 。可參考圖 2.10，角度分別是 122 度、111 度、116 度。圖中所顯示的鍵角、鍵長，只是一個大約值，並非所有的鍵角、鍵長都是固定的。曾有學者對鍵角跟鍵長做過統計，資料如表 2.3、表 2.4。表 2.3 顯示出各鍵長的平均數與標準差，鍵長的變動比較小，對結構所造成的影響也比較小，因此本研究將鍵長設為固定常數，常數值可參考圖 2.10；表 2.4 可以清楚了解到各鍵角的平均數與標準差，若只考慮 $\Delta CNCA$ 、$\Delta NCAC$ 、$\Delta CACN$ 三個角度，標準差從 1.5 到 5.0 不等。如果從整個結構上而言，使用固定的角度，亦會對整個蛋白質結構造成相當程度的影響，因此將鍵角也加入預測。

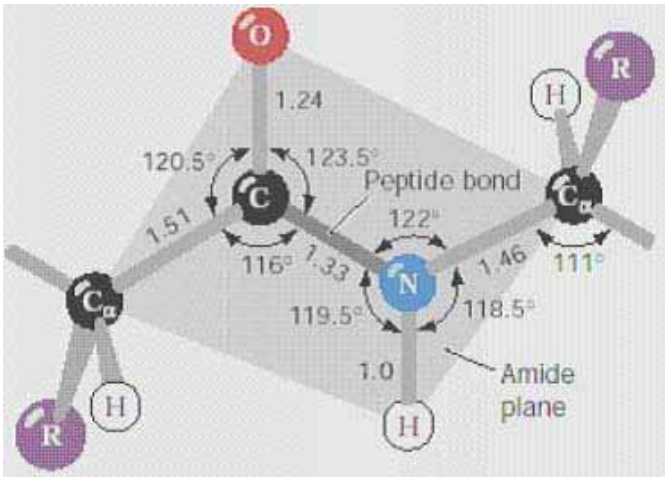

圖 2.10　鍵角鍵長示意圖

圖片來源：【33】

表 2.3　鍵長統計表

表格來源：【12】

Bond	X-PLOR labelling		Value	sigma
C-N	C-NH1	(except Pro)	1.329	0.014
	C-N	(Pro)	1.341	0.016
C-O	C-O		1.231	0.020
Calpha-C	CH1E-C	(except Gly)	1.525	0.021
	CH2G*-C	(Gly)	1.516	0.018
*Calpha-Cbeta	CH1E-CH3E	(Ala)	1.521	0.033
	CH1E-CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E-CH2E	(the rest)	1.530	0.020
N-Calpha	NH1-CH1E	(except Gly,Pro)	1.458	0.019
	NH1-CH2G*	(Gly)	1.451	0.016
	N-CH1E	(Pro)	1.466	0.015

*Calpha=CA。Cbeta 則是側鏈裡的其中一個原子，與 CA 相連接。

表 2.4　鍵角統計表

表格來源：【12】

Angle	X-PLOR labelling		Value	sigma
C-N-Calpha	C-NH1-CH1E	(except Gly,Pro)	121.7	1.8
	C-NH1-CH2G*	(Gly)	120.6	1.7
	C-N-CH1E	(Pro)	122.6	5.0
Calpha-C-N	CH1E-C-NH1	(except Gly,Pro)	116.2	2.0
	CH2G*-C-NH1	(Gly)	116.4	2.1
	CH1E-C-N	(Pro)	116.9	1.5
Calpha-C-O	CH1E-C-O	(except Gly)	120.8	1.7
	CH2G*-C-O	(Gly)	120.8	2.1
Cbeta-Calpha-C	CH3E-CH1E-C	(Ala)	110.5	1.5
	CH1E-CH1E-C	(Ile,Thr,Val)	109.1	2.2
	CH2E-CH1E-C	(the rest)	110.1	1.9
N-Calpha-C	NH1-CH1E-C	(except Gly,Pro)	111.2	2.8
	NH1-CH2G*-C	(Gly)	112.5	2.9
	N-CH1E-C	(Pro)	111.8	2.5
N-Calpha-Cbeta	NH1-CH1E-CH3E	(Ala)	110.4	1.5
	NH1-CH1E-CH1E	(Ile,Thr,Val)	111.5	1.7
	N-CH1E-CH2E	(Pro)	103.0	1.1
	NH1-CH1E-CH2E	(the rest)	110.5	1.7
O-C-N	O-C-NH1	(except Pro)	123.0	1.6
	O-C-N	(Pro)	122.0	1.4

在 2.2 節所介紹的三種預測法，以同源模擬法的效果最好，其次是摺疊辨識法，而同源模擬法是依賴已知結構的模板來決定目標蛋白質的三級結構，要是無法找到模板，便不適合用同源模擬法來預測蛋白質三級結構。這時，可以考慮摺

疊辨識法或是重新計算法。若摺疊資料庫裡的結構種類太少，也會對預測的結果產生影響。

而本研究的做法，與同源模擬法比較類似，一開始也是利用目標蛋白質的胺基酸序列到結構已知的蛋白質資料庫尋找同源蛋白質；本研究將目標蛋白質稱為測試集(Test Set)，模板稱為訓練集(Train Set)；接著同源模擬法會將目標蛋白質與模板做多重序列排比(MSA)並將模板的結構繼承給目標蛋白質，而本研究則是計算訓練集裡每個胺基酸的 6 種角度(3 個扭角、3 個鍵角)，再使用 SVM(Support Vector Machine)【8】【23】【26】去預測出測試集的 6 種角度，最後以預測出的 6 種角度代入機器人學的旋轉公式【10】再配合固定的鍵長，便可以計算出測試集的原子座標。當同源模擬法預測出目標蛋白質的結構後，會再以能量最小化來調整蛋白質三級結構。由於本研究並沒有預測蛋白質的側鏈結構，因此也沒有以能量最小化來調整蛋白質三級結構。

2.4 SVM

SVM(Support Vector Machine)是一種基於統計學習理論(statistical learning theory)基礎的學習機器。基本上 SVM 會將訓練集裡的資料都當成特徵向量(feature vectors)，並將這些特徵向量經由核心函數(kernel function)轉換到一個特徵空間(feature space)，這個空間有可能是高維度的空間，最後是在空間裡頭找出一條最佳分割線 OSH(Optimal Separation Hyperplane)或是最佳超平面，並將特徵向量分類，如圖 2.11。當有測試資料要分類時，可以直接利用以上的結果進行分類。詳細的理論可以參考 Vapnik 的文章【34】35】或是“An Introduction to Support Vector Machines and other kernel-based learning methods”一書【23】。

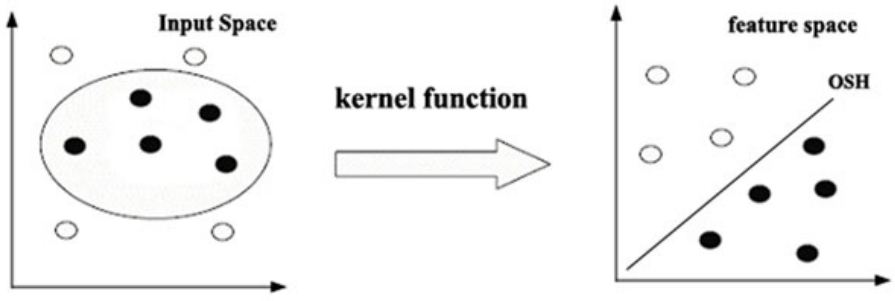

圖 2.11　SVM 示意圖，左圖無法用一直線將黑白球分類，利用 SVM 將它轉到一個維度空間，就能容易分類。

事實上，找出最佳超平面的用意就是在做最大化邊際(margin)的工作【2】

【5】。假設有一組訓練向量 S 如下，並有二種分類。

$$S = \{(x^1, y^1), ..., (x^l, y^l)\}, x \in \Re^n, y \in \{-1, 1\} \tag{2-2}$$

超平面 f(x)表示如下：

$$f(x) = \langle w, x \rangle + b = 0 \tag{2-3}$$

w 是權重向量(weight vector)，x 是輸入向量(input vector)，b 是偏差值(bias)。對於訓練集 S 而言，SVM 的應用是為了求在兩種類別資料(-1,1)之間具有最大邊際值的分割超平面，如公式(2-3)。為了計算出最佳超平面，可藉由將函數邊際值固定為 1 來計算，於是原本最大化邊際γ的公式便能被化簡，如公式(2-4)。

$$\begin{aligned} \gamma &= \frac{1}{2}\left(\left\langle \frac{w}{\|w\|_2} \bullet x^+ \right\rangle - \left\langle \frac{w}{\|w\|_2} \bullet x^- \right\rangle\right) \\ &= \frac{1}{2\|w\|_2}\left(\langle w \bullet x^+ \rangle - \langle w \bullet x^- \rangle\right) \\ &= \frac{1}{\|w\|_2} \end{aligned} \tag{2-4}$$

由圖 2.12 可以看出，最大化邊際就是最大化二條虛線間的距離，中間的實線為 OSH 也就是最佳超平面。

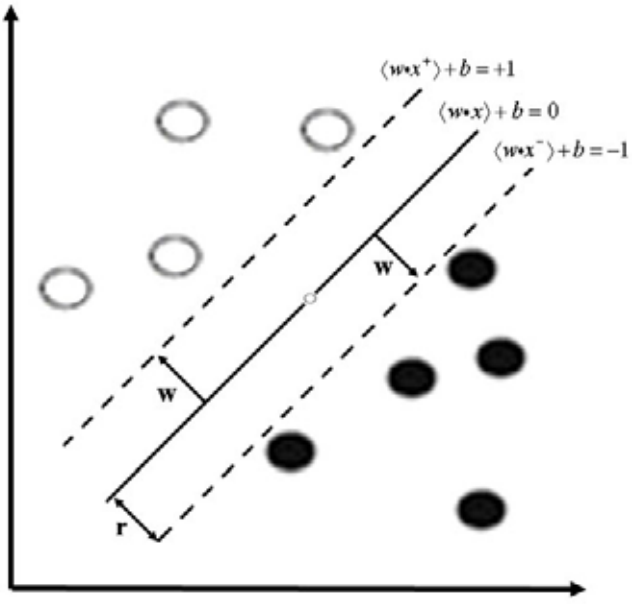

圖 2.12 邊際示意圖

尋找最大邊際是一個二次規劃的問題，這裡可以使用 Lagrangian 函數來求最佳解，Lagrangian 函數為：

$$L(w,b,a)=\frac{1}{2}\langle w\bullet w\rangle-\sum_{i=1}^{l}a_i[y_i(\langle w_i\bullet x_i\rangle+b-1)] \tag{2-5}$$

甚中，$a_i\geq 0$ 是 Lagrange multipliers。
透過不同的權重(w)和偏差值(b)，可以推得下列的式子：

$$\begin{aligned}&\frac{\partial L(w,b,a)}{\partial w}=w-\sum_{i=1}^{l}y_i a_i x_i=0\\&\Rightarrow w=\sum_{i=1}^{l}y_i a_i x_i\\&\frac{\partial L(w,b,a)}{\partial b}=\sum_{i=1}^{l}y_i a_i=0\\&\Rightarrow 0=\sum_{i=1}^{l}y_i a_i\end{aligned} \tag{2-6}$$

將公式(2-6)代入(2-5)中，則可獲得下列式子；

$$\begin{aligned}L(w,b,a)&=\frac{1}{2}\langle w\bullet w\rangle-\sum_{i=1}^{l}a_i[y_i(\langle w\bullet x_i\rangle+b)-1]\\&=\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j a_i a_j\langle x_i\bullet x_j\rangle-\sum_{i,j=1}^{l}y_i y_j a_i a_j\langle x_i\bullet x_j\rangle-\sum_{i=1}^{l}a_i\\&=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j a_i a_j\langle x_i\bullet x_j\rangle\end{aligned} \tag{2-7}$$

根據 Lagrangian 函數與公式(2-7)的結果，可將對偶模式寫成：

$$\begin{aligned}&\text{maximize}\quad W(a)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i,j=1}^{l}y_i y_j a_i a_j\langle x_i\bullet x_j\rangle\\&\text{subject to}\quad\sum_{i=1}^{l}y_i a_i=0\\&\qquad\qquad a_i\geq 0, i=1,...,l\end{aligned} \tag{2-8}$$

以 a_i 為最佳解的前提下，在上述模式中，權重向量(w)和偏差值(b)二個變數可在主要的限制中找出，其各自最佳解表示如下：

權重向量　　$w^* = \sum_{i=1}^{l} y_i a_i^* x_i$

偏差值　　$b^* = -\frac{\max_{y_i=-1}(\langle w^* \bullet x_i \rangle) + \max_{y_i=1}(\langle w^* \bullet x_i \rangle)}{2}$

邊際值　　$\gamma^* = \frac{1}{\|w^*\|_2}$ ，其中 $w = \sum_{i=1}^{l} y_i a_i x_i$　　(2-9)

SVM 除了可以做分類外，還能處理迴歸問題。迴歸函數如下：

$$f(x,w) = \sum_{i=1}^{l} w_i \phi_i(x) + b \qquad (2\text{-}10)$$

w 是權重向量，x 是輸入向量，b 是偏差值，Φ(*x*) 是為了處理線性不可分割時，對輸入向量所作的映射。在這裡需藉由不同的損失函數(loss function)來做迴歸【15】，常見的損失函數有：Quadric、Laplace、Huber 和 ε-Insensitive 等 4 種。而其中以 ε-Insensitive 最適合用於 SVM，它的表示如下：

$$|y - f(x,w)|_\varepsilon = \begin{cases} 0 & if\ |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon & otherwise \end{cases} \qquad (2\text{-}11)$$

如果損失函數等於 0，表示預測值 f(x,w)與實際值 y 的距離小於 ε ，圖 2.13 可以表示 ε-Insensitive 損失函數。接著將經驗風險最小化以及最大化邊際，於是公式(2-11)可改成：

$$R = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l} |y_i - f(x_i,w)|_\varepsilon\right) \qquad (2\text{-}12)$$

根據圖 2.13 定義鬆弛變數(slack variables) ξ 與 ξ^*

$$\begin{aligned} |y - f(x,w)| - \varepsilon &= \xi \\ |y - f(x,w)| - \varepsilon &= \xi^* \end{aligned} \qquad (2\text{-}13)$$

並將公式(2-12)改成：

$$R_{w,\xi,\xi^*} = \left[\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l} \xi + \sum_{i=1}^{l} \xi^* \right) \right] \qquad (2\text{-}14)$$

並滿足下列限制：

$$\begin{aligned} & y_i - f(x,w) \le \varepsilon + \xi, i = 1,...,l \\ & f(x,w) - y_i \le \varepsilon + \xi^*, i = 1,...,l \\ & \xi \ge 0 \\ & \xi^* \ge 0 \end{aligned} \qquad (2\text{-}15)$$

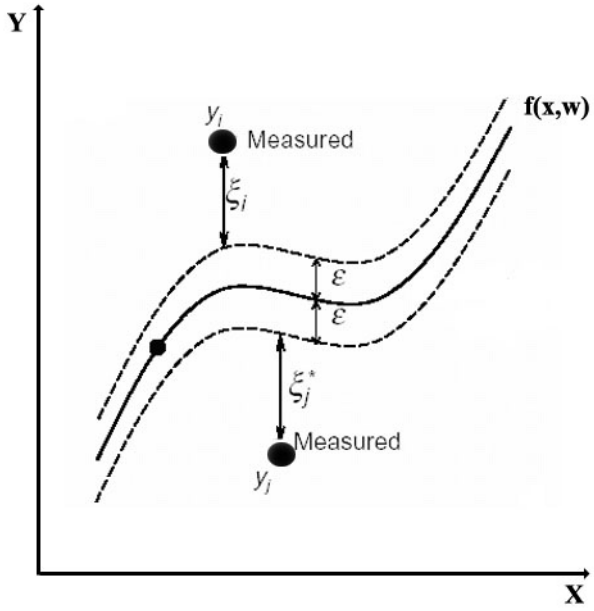

圖 2.13　一維支持向量迴歸

再將最佳化問題轉換成 Lagrangian：

$$
\begin{aligned}
L(w,b,\xi,\xi^*,\alpha_i,\alpha_i^*,\beta_i,\beta_i^*) = \frac{1}{2}w^T w + C\left(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^*\right) \\
-\sum_{i=1}^{l}\alpha_i^*\left[y_i - w^T x_i - b + \varepsilon + \xi_i^*\right] \\
-\sum_{i=1}^{l}\alpha_i\left[w^T x_i + b - y_i + \varepsilon + \xi_i\right] \\
-\sum_{i=1}^{l}(\beta_i^*\xi_i^* + \beta_i\xi_i)
\end{aligned}
\tag{2-16}
$$

為了求出α、α^*、β和β^*，可以將問題轉換成對偶問題(dual problem)：

$$
L_d(\alpha,\alpha^*) = -\varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)x_i^T x_j \tag{2-17}
$$

並滿足下列限制：

$$
\begin{aligned}
&\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i, i = 1,...,l \\
&0 \le \alpha_i^* \le C \\
&0 \le \alpha_i \le C
\end{aligned}
\tag{2-18}
$$

α可以利用對L_d做偏微分加上解聯立方程式的方式分別求出，最後權重向量(w)和偏差值(b)為：

$$
w_{optimal} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi(x_i) \tag{2-19}
$$

$$
b_{optimal} = \frac{1}{l}\left(\sum_{i=1}^{l} y_i - x_i^T w_{optimal}\right) \tag{2-20}
$$

最後，最佳迴歸超平面則為：

$$
f(x,w) = w^T\phi(x) + b \tag{2-21}
$$

SVM比較常被用在分類上，例如文件的分類【18】。但也有學者將它運用在

蛋白質相關方面，例如摺疊辨識(fold recognition) 【9】、蛋白質結構預測【30】、扭角預測【26】等等，效果都相當不錯。其中與本研究較有相關的是在扭角預測【26】這篇文章中，作者對於扭角的預測與本研究不同，他們將Ramachandran平面圖分成好幾個區域，如圖2.14；Ramachandran平面圖的橫軸代表PHI扭角，縱軸為PSI扭角，範圍是正負180度；再利用SVM跟類神經網路(Neural Network)等預測扭角是位於哪一塊區域中，結果是以SVM的效果較佳。因此本研究便以SVM預測主鏈上的扭角、鍵角。本研究與他們不同之處是他們預測扭角的類別，並不是真的預測出角度值。而本研究是要預測出扭角的值，扭角是屬於連續的數值，因此必須使用迴歸預測的方式，也就是運用SVR(Support Vector Regression)來預測。

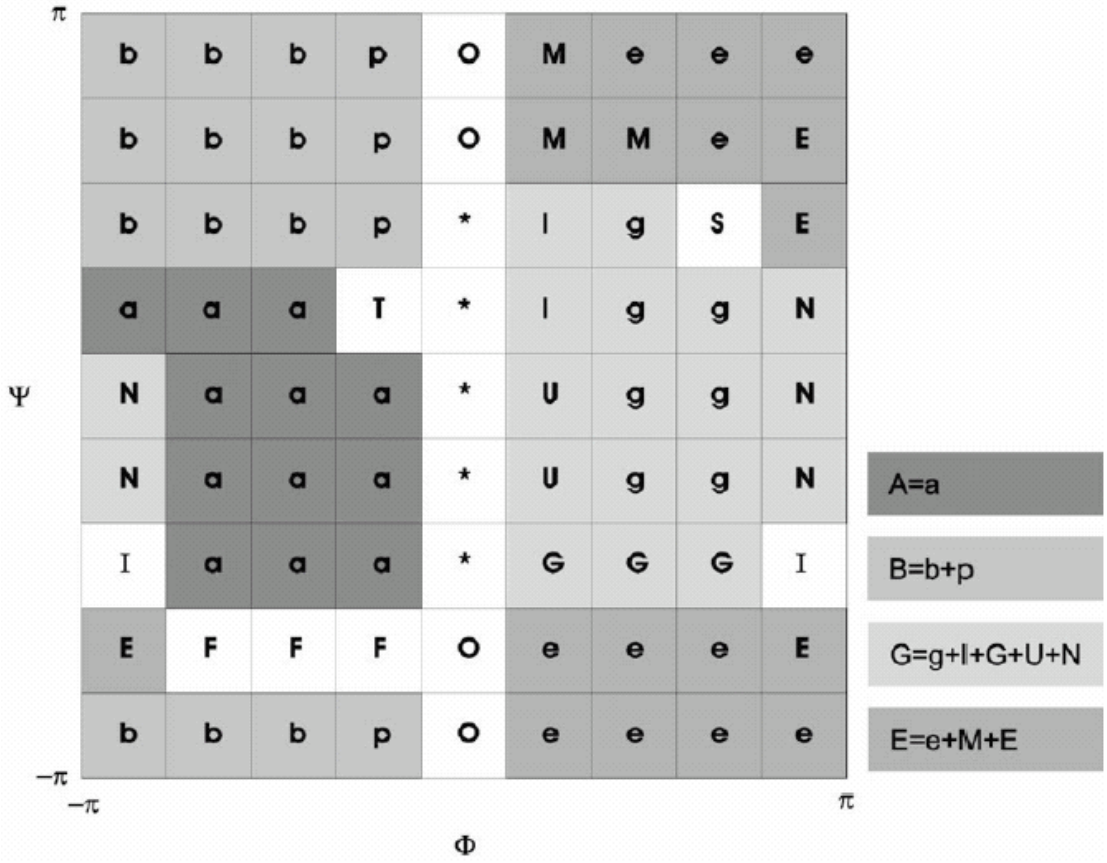

圖 2.14　Ramachandran 平面圖

圖片來源：【26】

本研究所選擇的 SVM 預測工具是 LIBSVM(A Library for Support Vector Machines)，LIBSVM 軟體提供了 SVR 的功能。它整合了 SVC(Support Vector Classification)、SVR 和 distribution estimation，同時也提供了多類別的分類【8】。這是一套於 2001 年由 Chang C-C 和 Lin C-J 等學者所發展出來的工具，LIBSVM

最大的目的是要讓使用者能夠輕易的上手。LIBSVM 的輸入格式如下：
[label] [index1]:[value1] [index2]:[value2] [index3]:[value3] ~
[label] [index1]:[value1] [index2]:[value2] [index3]:[value3] ~
[label] [index1]:[value1] [index2]:[value2] [index3]:[value3] ~

label 或說是 class，如果要用 SVM 來做分類的工作，label 則是指要分類的種類，通常是一些整數，而本研究是要預測扭角、鍵角，因此在 label 部份必須放上扭角、鍵角。index 是有順序的索引，通常是放連續的整數。value 就是用來訓練的資料，通常是實數。[index1]:[value1] [index2]:[value2] [index3]:[value3] ~ 整個可稱為屬性，本研究以 PSSM(position specific scoring matrix)跟二級結構來當屬性。

第三章　研究方法

根據 1.4 節，圖 1.3 可以了解實驗流程，文章接著會針對實驗流程的步驟、方法加以說明。在此先假設要預測的蛋白質為 P(測試集)，取測試集的一級結構給 PSI-BLAST 到結構已知的 PDB 資料庫搜尋同源蛋白質(訓練集)，接著準備測試集和訓練集的 PSSM 跟二級結構，開始進行編碼，編碼完後計算訓練集的 6 種角度(3 個扭角、3 個鍵角)，並將這些角度和訓練集的編碼一起丟入 SVM 訓練出 6 個模組，再用這 6 個模組跟測試集的編碼以 SVM 預測出測試集的 6 種角度，最後將預測出的角度代入機器人學的旋轉公式並配合固定的鍵長計算出原子 3 維座標，也就是測試集的三級結構。

3.1 PSI-BLAST 尋找同源序列

PSI-BLAST(Position specific iterative BLAST)常常在蛋白質研究方面被使用，PSI-BLAST 是以 BLAST 為基礎再加上反覆式搜尋的概念所設計而成的。此種搜尋方法比其他方法更敏感，可以更有效的找到更多序列相似度低而結構功能相似的蛋白質，而這些蛋白質同時也成為本研究中的訓練集。

研究中所用的 PSI-BLAST 軟體，版本為 2.2.9 for win32 版。資料庫為 pdbaa。
PSI-BLAST下載處 ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.9/
資料庫下載處 ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/
之所以會使用 pdbaa 資料庫是因為 PDB 資料庫裡的結構都是已知的，實驗的目的是要預測蛋白質三級結構，利用已知的結構來預測未知的結構是合理的。

在使用 PSI-BLAST 軟體進行搜尋之前，必須先將資料庫格式化。可使用 PSI-BLAST 軟體的指令 formatdb 來執行，研究中對於參數的設定-i 為 pdbaa、-o 為 T、-p 為 T。詳細的說明可參考 PSI-BLAST 文件檔。

將資料庫格式化完成後，便可以使用 pdbaa 資料庫來搜尋同源序列。在搜尋前必須先準備測試集的一級結構檔，而且要是 FASTA 格式。上面下載的資料庫 pdbaa 是 FASTA 格式。測試集的一級結構可至 PDB 網站上下載。
PDB網站 http://www.rcsb.org/pdb/
底下是一個 FASTA 格式的樣本：蛋白質 1P7E，檔名 1P7E ，副檔名 FASTA

>1P7E:A IMMUNOGLOBULIN G BINDING PROTEIN G
MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATK
TFTVTE

有了測試集的一級結構以後，接著可以使用 PSI-BLAST 軟體來尋找測試集的同源序列，這裡所用的指令是 blastpgp，研究中資料庫設定為 pdbaa，迭代 5 次。底下圖 3.1、圖 3.2 是一個結果樣本的片斷。

```
BLASTP 2.2.9 [May-01-2004]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 1P7E:A IMMUNOGLOBULIN G BINDING PROTEIN G
         (56 letters)

Database: pdbaa
           19,111 sequences; 4,264,239 total letters


Results from round 1

                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

pdb|1PN5|A Chain A, Nmr Structure Of The Nalp1 Pyrin Domain (Pyd)      94   4e-021
pdb|1IBX|B Chain B, Nmr Structure Of Dff40 And Dff45 N-Terminal ...    94   8e-021
pdb|1PGX|  Protein G Type 7 (B2 Domain)                               81   4e-017
pdb|1P7F|A Chain A, Gb3 Solution Structure Obtained By Refinemen...   79   1e-016
pdb|2IGH|  Protein G (Third Igg-Binding Domain) (Nmr, 24 Structu...   76   2e-015
pdb|2IGD|  Anisotropic Structure Of Protein G Igg-Binding Domain...   75   2e-015
pdb|1QKZ|A Chain A, Fab Fragment (Mn14c11.6) In Complex With A P...   72   2e-014
```

圖 3.1　PSI-BLAST 搜尋 1P7E 的同源序列結果檔片斷 1

```
>pdb|1PN5|A Chain A, Nmr Structure Of The Nalp1 Pyrin Domain (Pyd)
          Length = 159

 Score = 94.4 bits (233), Expect = 4e-021
 Identities = 50/56 (89%), Positives = 53/56 (94%)

Query: 1   MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATKTFTVTE 56
           MQYKL++NGKTLKGETTT+AVDA TAEK FKQYANDNGVDG WTYDDATKTFTVTE
Sbjct: 1   MQYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE 56


>pdb|1IBX|B Chain B, Nmr Structure Of Dff40 And Dff45 N-Terminal Domain
          Complex
          Length = 145

 Score = 93.6 bits (231), Expect = 8e-021
 Identities = 50/56 (89%), Positives = 53/56 (94%)

Query: 1   MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATKTFTVTE 56
           MQYKL++NGKTLKGETTT+AVDA TAEK FKQYANDNGVDG WTYDDATKTFTVTE
Sbjct: 1   MQYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE 56
```

圖 3.2　PSI-BLAST 搜尋 1P7E 的同源序列結果檔片斷 2

圖 3.1 裡是記錄同源序列、比對分數和 E-VALUE 值。圖 3.2 則是查詢序列與同源序列的詳細比對結果，Identities 表示相似度，查詢序列與第一條同源序列比對的長度是 1~56。根據這個結果，便可以決定測試集的同源序列蛋白質 Ps(訓練集)。研究中對於訓練集的記錄格式如下：

蛋白質 PDB 編號：Chain 序列比對啟始編號 序列比對結束編號
例如圖 3.1、圖 3.2 的同源序列可記錄成：
1PN5：A 1 56
1IBX：B 1 56
1PGX 1 56

1PN5：A，A 指的是蛋白質 1PN5 的 Chain A，有些蛋白質本身會由多條 Chain 組成，記錄 Chain 的用意是指編碼時只對 1PN5 的 Chain A 進行編碼，1PN5 在 PSI-BLAST 跟測試集 1P7E 的一級結構比對後，胺基酸 1 至 56 與 1P7E 的胺基酸序列比較相似，因此整個記錄成 1PN5：A 1 56。編碼時則會針對 1PN5 的 Chain A 胺基酸範圍 1 至 56 來編碼。決定測試集與訓練集之後，即可開始建立其 PSSM 和二級結構。

3.2 建立 PSSM 和二級結構

本研究以 PSSM 和二級結構當做 SVM 的輸入屬性，因此在編碼之前必須準備好測試集跟訓練集的 PSSM 和二級結構。

3.2.1 PSSM

特定位置評分矩陣 PSSM 經常被使用，例如參考文獻【17】一文中，作者使用 SVM 和 PSSM 來進行研究。PSSM 是由 PSI-BLAST【29】所產生的，底下將介紹 PSI-BLAST 產生 PSSM 的流程。

首先將目標蛋白質序列(測試集)用 PSI-BLAST 來搜尋沒有重複序列之資料庫(non-reduent database)或其它資料庫，例如 pdbaa。找出相似度較高的蛋白質序列，然後將這些序列做多重序列排比(MSA)，接著建構一個查詢序列長度的 profile，而這 profile 就是所謂的 PSSM。接著用上述步驟所產生的 PSSM，再以局部排比方式搜尋蛋白質資料庫，把低於期望值的序列保留下來，然後將這些保留下來的序列做多重序列排比，接著再建構一個新的 PSSM。如此反覆搜尋直至沒有新的結果產生為止。反覆的搜尋會增加結果的敏感性，可以找出更多同源的序列，最後的 PSSM 將包含更多有用的演化資訊【4】，整個流程如圖 3.3 所示。

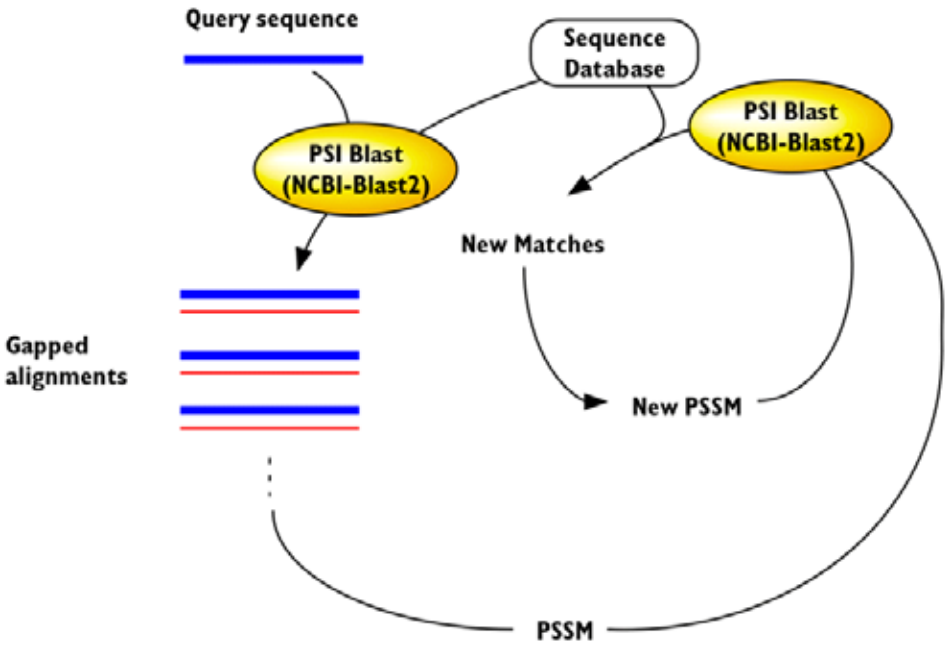

圖 3.3　PSI-BLAST 流程圖

圖片來源：http://www.ch.embnet.org

PSSM 是根據氨基酸在每個位置上的出現頻率個別加重計分，矩陣的橫軸是 20 個氨基酸出現在此位置的機率，而縱軸則是氨基酸序列的長度，矩陣上每個位置都代表著氨基酸在該位置上出現機率的分數，如圖 3.4。

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	Q	-1	1	0	0	-3	6	2	-2	0	-3	-2	1	-1	-3	-1	0	-1	-2	-2	-2
3	Y	-2	-2	-3	-4	-3	-2	-3	-4	1	-2	-1	-3	-1	4	-4	-2	-2	2	7	-2
4	K	-1	3	-1	-1	-4	1	0	-2	-1	-3	-3	5	-2	-4	-2	-1	-1	-4	-2	-3
5	L	-2	-3	-4	-4	-2	-3	-3	-4	-3	1	5	-3	2	0	-3	-3	-2	-2	-2	1
6	V	-1	-3	-4	-4	-2	-3	-3	-4	-3	4	1	-3	0	2	-3	-3	-1	-2	-1	4
7	I	-2	-3	-4	-4	-2	-3	-4	-4	-4	5	2	-3	1	0	-3	-3	-1	-3	-2	2
8	N	-1	-1	5	0	-3	-1	-1	-2	-1	-1	-2	1	-1	-3	-3	0	-1	-4	-2	2
9	G	-1	-3	-2	-2	-3	-3	-3	5	-3	1	0	-2	-1	-2	-3	-1	-2	-3	-3	-1
10	K	-1	1	2	-1	-3	0	0	2	-1	-2	-3	4	-2	-3	-2	-1	-1	-4	-2	0
11	T	-1	-2	0	2	-2	-1	-1	1	-2	-2	-2	-1	-2	-3	-2	1	5	-3	-3	-1
12	L	-2	2	-3	-3	-2	-2	-3	-4	-2	0	3	-1	1	2	-3	-2	1	-2	-1	0
13	K	-1	1	-1	-1	-2	0	0	-2	-2	-1	-2	4	-1	-3	-2	1	2	-3	-2	1
14	G	-1	-3	-1	-2	-3	-3	-3	5	-3	-2	-2	-2	-2	2	-3	-1	-2	-2	-1	0
15	E	-1	-1	-1	0	-3	1	5	-3	-1	-1	-2	0	-2	-3	-2	0	1	-3	-2	1

圖 3.4　PSSM 矩陣

PSSM 的建立，依舊要由 PSI-BLAST 來完成，可使用指令 blastpgp，研究中的資料庫設定為 pdbaa、迭代次數設為 5 次。PSSM 雖然包含了許多的演化資訊，但只使用 PSSM 來當屬性是不夠的；許多三級結構的預測經常會用到二級結構的知識【24】，因此本研究將二級結構的資訊也納入。

3.2.2 二級結構

接著是二級結構的部份，測試集和訓練集對於二級結構的取得方法不同。測試集是一條結構未知的蛋白質，因此它的二級結構也是未知的。於是本研究對測試集的二級結構便使用預測的，可至 SCRATCH servers 預測二級結構，參數的設定完全使用系統預設。SCRATCH servers 上的輸入為一級結構，輸出為預測的二級結構。

SCRATCH servers：http://www.igb.uci.edu/tools/scratch/

預測的二級結構，範例如下：

Query_name: 1P7E

Query_length: 56

Prediction:

MQYKLVINGKTLKGETTTKAVDAETAEKAFKQYANDNGVDGVWTYDDATK

CEEEEEEECCCCCCCCCCEECCHHHHHHHHHHHHHCCCCCEEEECCCCC

(底下省略)

上述範例中，第五行 HEC 是相對應於第四行胺基酸的二級結構，這是 SCRATCH servers 所預測出來的結果。

至於訓練集的二級結構, 訓練集裡的所有蛋白質是屬於結構已知的蛋白質。因此，訓練集的二級結構是可以得知的。本實驗是使用 DSSPCMBI 軟體來取得訓練集的二級結構。DSSP 軟體是由 Wolfgang Kabsch 和 Chris Sander 所設計的【36】，可至 CMBI 網站上免費下載。DSSPCMBI 軟體的功能可以將 PDB 檔轉換成 DSSP 檔，二級結構就記錄在 DSSP 檔裡，而 PDB 檔可以至 PDB 網站上免費下載。

DSSPCMBI網站：http://swift.cmbi.ru.nl/gv/dssp/

PDB網站：http://www.rcsb.org/pdb/

DSSP 檔案片斷如下圖 3.5，圖中灰色範圍代表每個胺基酸的二級結構。DSSP 對於二級結構以 8 種類別表示，分別是 H、G、I、E、B、S、T、空白等 8 種。而測試集的二級結構預測完後是以 3 種類別表示，因此本研究將統一測試集和訓練集的二級結構，均以 3 種類別來代表。編碼時，二級結構是用 H、E、C 來表示，所以必須將 8 種類別轉換成 3 種類別，這裡使用的是一般最常見的轉換方式，分別是：

H、G、I → H

E → E

B、S、T、空白 → C

最後 SVM 的輸入屬性則是 PSSM+二級結構。為了符合 LIBSVM 的輸入格式，必須將屬性做一個編碼的動作。

```
#  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O
  1    1 A M              0   0  122      0, 0.0    19,-2.2     0, 0.0
  2    2 A Q  E     -A   19   0A 134     17,-0.2     2,-0.3    19,-0.1
  3    3 A Y  E     -A   18   0A  16     15,-3.1    15,-2.2    -2,-0.6
  4    4 A K  E     -Ab  17  51A  78     46,-2.0    48,-3.0    -2,-0.3
  5    5 A L  E     -Ab  16  52A   0     11,-2.9    11,-2.1    -2,-0.4
  6    6 A V  E     -Ab  15  53A  29     46,-2.5    48,-2.3    -2,-0.5
  7    7 A I  E     +Ab  14  54A   4      7,-2.7     7,-2.0    -2,-0.5
  8    8 A N  E     +Ab  13  55A  77     46,-3.0    48,-2.0    -2,-0.6
  9    9 A G          -    0   0   5      3,-1.9    30,-0.1    -2,-0.6
 10   10 A K  S    S+   0   0  167     -2,-0.1     3,-0.1     1,-0.1
 11   11 A T  S    S+   0   0  142      1,-0.2     2,-0.5     0, 0.0
 12   12 A L         +    0   0   62     25,-0.1    -3,-1.9     2,-0.0
 13   13 A K  E     +A   8   0A 152     -2,-0.5     2,-0.3    -5,-0.2
 14   14 A G  E     -A   7   0A  36     -7,-2.0    -7,-2.7    -2,-0.3
 15   15 A E  E     +A   6   0A 124     -2,-0.3     2,-0.3    -9,-0.2
 16   16 A T  E     -A   5   0A  45    -11,-2.1   -11,-2.9    -2,-0.3
 17   17 A T  E     -A   4   0A  88     -2,-0.3     2,-0.3   -13,-0.2
 18   18 A T  E     -A   3   0A  40    -15,-2.2   -15,-3.1    -2,-0.3
 19   19 A K  E     +A   2   0A 154     -2,-0.3     2,-0.3   -17,-0.2
 20   20 A A         -    0   0   16    -19,-2.2     3,-0.1    -2,-0.4
 21   21 A V  S    S+   0   0  130     -2,-0.3     2,-0.3     1,-0.1
 22   22 A D  S  > S-   0   0   62    -21,-0.1     4,-1.6     1,-0.1
```

圖 3.5　DSSP 檔案片斷

3.3 編碼

本研究是以 PSSM 跟二級結構做為 SVM 的輸入屬性。為了讓 PSSM 跟二級結構符合 LIBSVM 的文件格式，必須進行編碼的動作。在編碼前已前將「測試集」的「PSSM」和「預測的二級結構檔」、「訓練集」的「PSSM」和「DSSP 檔」準備齊全。

在編碼時，考慮到週圍胺基酸所造成的影響【6】【14】【16】，因此加入了 Sliding Window【32】的觀念。並且讓二級結構的 Sliding Window 為 PSSM 的 2 倍。如果 Sliding Window 設定為 11，則 PSSM 的 Sliding Window=11，二級結構的 Sliding Window=21，這樣的做法是考慮到更大範圍的影響。接著將針對 PSSM 跟二級結構的編碼分別說明。

3.3.1 PSSM 編碼

參考圖 3.4 的 PSSM 片斷，縱軸是一級結構，也就是胺基酸序列，依序是 MQYKL…等等。當加入 Sliding Window 後，圖 3.4 可想像成圖 3.6。

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	X																				
	X																				
	X																				
	X																				
	X																				
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	Q	-1	1	0	0	-3	6	2	-2	0	-3	-2	1	-1	-3	-1	0	-1	-2	-2	-2
3	Y	-2	-2	-3	-4	-3	-2	-3	-4	1	-2	-1	-3	-1	4	-4	-2	-2	2	7	-2
4	K	-1	3	-1	-1	-4	1	0	-2	-1	-3	-3	5	-2	-4	-2	-1	-1	-4	-2	-3
5	L	-2	-3	-4	-4	-2	-3	-3	-4	-3	1	5	-3	2	0	-3	-3	-2	-2	-2	1
6	V	-1	-3	-4	-4	-2	-3	-3	-4	-3	4	1	-3	0	2	-3	-3	-1	-2	-1	4
7	I	-2	-3	-4	-4	-2	-3	-4	-4	-4	5	2	-3	1	0	-3	-3	-1	-3	-2	2
8	N	-1	-1	5	0	-3	-1	-1	-2	-1	-1	-2	1	-1	-3	-3	0	-1	-4	-2	2
9	G	-1	-3	-2	-2	-3	-3	-3	5	-3	1	0	-2	-1	-2	-3	-1	-2	-3	-3	-1
10	K	-1	1	2	-1	-3	0	0	2	-1	-2	-3	4	-2	-3	-2	-1	-1	-4	-2	0
11	T	-1	-2	0	2	-2	-1	-1	1	-2	-2	-2	-1	-2	-3	-2	1	5	-3	-3	-1
12	L	-2	2	-3	-3	-2	-2	-3	-4	-2	0	3	-1	1	2	-3	-2	1	-2	-1	0
13	K	-1	1	-1	-1	-2	0	0	-2	-2	-1	-2	4	-1	-3	-2	1	2	-3	-2	1
14	G	-1	-3	-1	-2	-3	-3	-3	5	-3	-2	-2	-2	-2	2	-3	-1	-2	-2	-1	0
15	E	-1	-1	-1	0	-3	1	5	-3	-1	-1	-2	0	-2	-3	-2	0	1	-3	-2	1

圖 3.6　加入 Sliding Window 後的 PSSM 想像圖

圖 3.6 中，灰色選取範圍是 Sliding Window 的範圍，長度 11。中間位置，胺基酸 M，表示正在對 M 進行編碼。由於 M 是起始胺基酸，也就是說 Sliding Window 位置 1~5 是沒有胺基酸的。因此，X 後方相對應於橫軸的 20 個欄位 ARN…WYV 全以 0 代替。編碼時，則是從 Sliding Window 位置 1 開始，將相對應於橫軸的 ARN…WYV 的 20 個值逐一取出，接著是位置 2，將相對應於橫軸的 ARN…WYV 的 20 個值逐一取出，位置 3 的 20 個值、位置 4 的 20 個值、…、位置 11(V 胺基酸)的 20 個值，全部取出後並加上索引以符合 LIBSVM 的屬性格式。目前屬性長度則為 Sliding Window*20(欄位 ARN…WYV) = 11*20 = 220，編碼後的內容如下：

1:0.0 2:0.0 3:0.0 ~ 218:-2.0 219:-1.0 220:4.0
1:0.0 2:0.0 3:0.0 ~ 218:-3.0 219:-2.0 220:2.0
1:0.0 2:0.0 3:0.0 ~ 218:-4.0 219:-2.0 220:2.0
.
.
.

第一行代表胺基酸 M 的 PSSM 編碼，第二行代表胺基酸 Q 的 PSSM 編碼，依此類推。

3.3.2 二級結構編碼

參考圖 3.5 的 DSSP 片斷，欄位 AA 是一級結構，也就是胺基酸序列，依序是 MQYKL…等等。當加入 Sliding Window 後，圖 3.5 可想像成圖 3.7。

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O
		X	2^0						
		X	2^1						
		X	2^2						
		X	2^3						
		X	2^4						
		X	2^5						
		X	2^6						
		X	2^7						
		X	2^8						
		X	2^9						
1	1 A	M	2^10	0	0	122	0, 0.0	19,-2.2	0, 0.0
2	2 A	Q	E -A	19	0A	134	17,-0.2	2,-0.3	19,-0.1
3	3 A	Y	E -A	18	0A	16	15,-3.1	15,-2.2	-2,-0.6
4	4 A	K	E -Ab	17	51A	78	46,-2.0	48,-3.0	-2,-0.3
5	5 A	L	E -Ab	16	52A	0	11,-2.9	11,-2.1	-2,-0.4
6	6 A	V	E -Ab	15	53A	29	46,-2.5	48,-2.3	-2,-0.5
7	7 A	I	E +Ab	14	54A	4	7,-2.7	7,-2.0	-2,-0.5
8	8 A	N	E +Ab	13	55A	77	46,-3.0	48,-2.0	-2,-0.6
9	9 A	G	-	0	0	5	3,-1.9	30,-0.1	-2,-0.6
10	10 A	K	S S+	0	0	167	-2,-0.1	3,-0.1	1,-0.1
11	11 A	T	S S+	0	0	142	1,-0.2	2,-0.5	0, 0.0
12	12 A	L	+	0	0	62	25,-0.1	-3,-1.9	2,-0.0
13	13 A	K	E +A	8	0A	152	-2,-0.5	2,-0.3	-5,-0.2
14	14 A	G	E -A	7	0A	36	-7,-2.0	-7,-2.7	-2,-0.3
15	15 A	E	E +A	6	0A	124	-2,-0.3	2,-0.3	-9,-0.2

圖 3.7　加入 Sliding Window 後的 DSSP 想像圖

圖 3.7 中，灰色選取範圍是 Sliding Window 的範圍，長度 21。二級結構的 Sliding Window 是 PSSM 的 2 倍，所以是 21。中間位置，胺基酸 M，表示正在對 M 進行編碼。由於 M 是起始胺基酸，也就是說 Sliding Window 位置 1~10 是沒有胺基酸也沒有二級結構的。實驗中對於二級結構編碼的做法是分別計算 Sliding Window 範圍裡 H、E、C 的值。規則如下：

H = 啟始值 * H 出現次數
E = 啟始值 * E 出現次數
C = 啟始值 * C 出現次數

要計算 H、E、C 的值之前要先決定啟始值。在 Sliding Window 裡每一個位置都會有一個數值，如位置 1 就是 2 的 0 次方，也就是 1。位置 2 是 2 的 1 次方，

也就是 2。位置 3 是 4，位置 4 是 8，依此類推。首先，先決定在 Sliding Window 範圍裡第一個胺基酸的出現位置，依圖 3.7 而言，第一個胺基酸的出現位置在 11，於是便將 2 的 10 次方 1024 設為啟始值，接著計算 Sliding Window 裡 H、E、C 出現的次數，沒有胺基酸的部份不納入計算。H 出現 0 次，E 出現 7 次，C 出現 4 次，整理如下表 3.1。

表 3.1　HEC 整理

	啟始值	出現次數	值
H	1024	0	0
E	1024	7	7168
C	1024	4	4096

最後，胺基酸 M 的二級結構編碼完，H 是 0，E 是 7168，C 是 4096。將這三個值接在 PSSM 編碼後方，內容如下所示：

1:0.0 2:0.0 3:0.0 ~ 218:-2.0 219:-1.0 220:4.0 221:0 222:7168 223:4096
1:0.0 2:0.0 3:0.0 ~ 218:-3.0 219:-2.0 220:2.0 221:0 222:3584 223:2560
1:0.0 2:0.0 3:0.0 ~ 218:-4.0 219:-2.0 220:2.0 221:0 222:2048 223:1280

第一行代表胺基酸 M 的 PSSM+二級結構編碼，第二行代表胺基酸 Q 的 PSSM+二級結構編碼，依此類推。每行長度為 Sliding Window*20(PSSM)+3(HEC 二級結構) = 11*20+3 = 223。不管是預測的二級結構檔或是 DSSP 檔，編碼方法均是如此。

以上只是 LIBSVM 裡的屬性，LIBSVM 的文件格式是 Label+屬性。測試集的 Label 與訓練集的 Label 是不同的；測試集的 Label 是實驗中要預測的角度。因此，是未知的，實驗裡將測試集的 Label 全設為 1。測試集的文件範例如底下圖 3.9 所示。訓練集的 Label 則有 6 種，分別是 PHI 角度、PSI 角度、OMEGA 角度、CNCA 角度、NCAC 角度、CACN 角度。這 6 種角度可由公式計算而得，分別為公式(3-1)、(3-2)。

PHI、PSI、OMEGA 是扭角，PHI 扭角的計算公式如下：

$$\pm\cos\phi = \frac{\vec{P}n \bullet \vec{P}c'}{\left|\vec{P}n\right|\left|\vec{P}c'\right|} \qquad (3\text{-}1)$$

$$\vec{P}n = \vec{n} \bullet \vec{a}$$

其中

$$\vec{P}c' = \vec{a} \bullet \vec{c}'$$

公式(3-1)會解出二個值，所以必須加入行列式 D 來判斷

$$D = \left(\vec{P}n \bullet \vec{P}c'\right) \bullet \vec{a}$$

假如 D>0，則公式(3-1)取正值，反之取負值。n、a 與 c 表示向量，可參考圖 3.8，Pn 向量為 C、N、CA 平面的法向量，Pc’向量為 N、CA、C 平面的法向量，這二個法向量之間的夾角為 PHI 扭角。

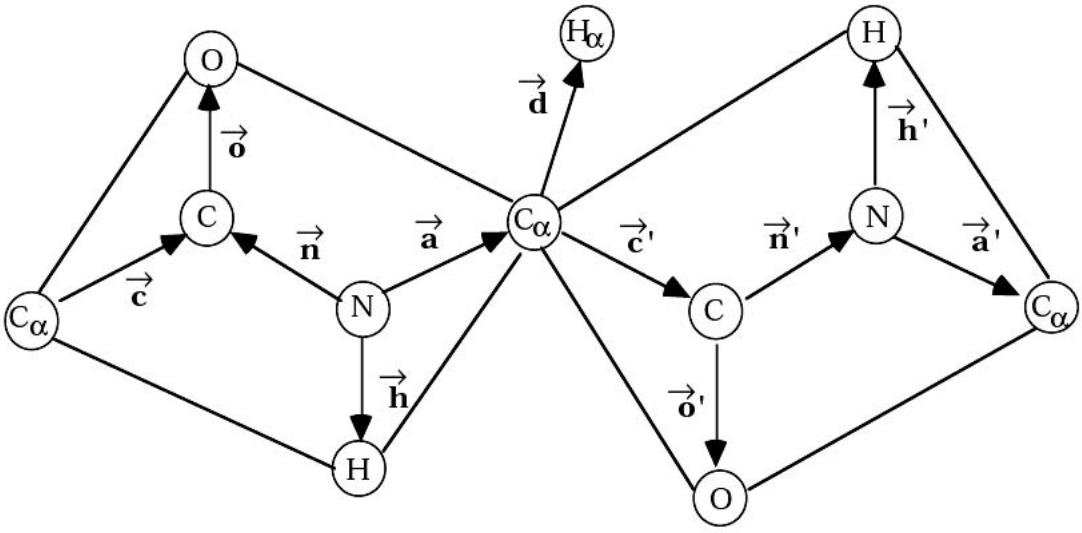

圖 3.8　單位向量表示圖

圖片來源：【19】

CNCA、NCAC、CACN 是鍵角，CNCA 鍵角的計算公式如下：

$$\cos\phi = \frac{\vec{n} \bullet \vec{a}}{\left|\vec{n}\right|\left|\vec{a}\right|} \qquad (3\text{-}2)$$

由公式(3-1)、(3-2)可計算出所有扭角、鍵角。計算出所有角度後，接著將每種角度與編碼檔合拼，每一種角度都會與編碼後的資料儲存成一個文件檔，因此訓練集總共有 6 個文件檔，範例如圖 3.10。

```
1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
1 1:-2.0 2:6.0 3:-1.0 4:-2.0 5:-4.0 6:1.0 7:0.0 8:-3.0 9:-1.0 1
1 1:-2.0 2:3.0 3:-2.0 4:-2.0 5:-4.0 6:-1.0 7:1.0 8:-3.0 9:-2.0
1 1:-2.0 2:-1.0 3:0.0 4:5.0 5:-4.0 6:-1.0 7:1.0 8:-2.0 9:-2.0 1
1 1:-2.0 2:-4.0 3:-4.0 4:-4.0 5:-3.0 6:-3.0 7:-4.0 8:-4.0 9:-2.
1 1:0.0 2:-5.0 3:-4.0 4:-5.0 5:10.0 6:-4.0 7:-5.0 8:-4.0 9:-4.0
1 1:-2.0 2:-1.0 3:1.0 4:1.0 5:-3.0 6:1.0 7:-1.0 8:-1.0 9:-2.0 1
1 1:-2.0 2:-2.0 3:-2.0 4:-1.0 5:-3.0 6:0.0 7:4.0 8:-4.0 9:-2.0
1 1:-2.0 2:-3.0 3:-2.0 4:-2.0 5:-4.0 6:1.0 7:0.0 8:0.0 9:3.0 10
1 1:1.0 2:1.0 3:-3.0 4:-3.0 5:-3.0 6:-1.0 7:0.0 8:-3.0 9:-2.0 1
1 1:-2.0 2:2.0 3:2.0 4:4.0 5:-4.0 6:-1.0 7:2.0 8:-3.0 9:-1.0 10
```

圖 3.9　測試集的範例文件片斷

```
-82.8 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
-62.6 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
-75.1 1:0.0 2:0.0 3:0.0 4:0.0 5:0.0 6:0.0 7:0.0 8:0.0 9:0.0 10:0.0
-91.8 1:-2.0 2:6.0 3:-1.0 4:-2.0 5:-4.0 6:1.0 7:0.0 8:-3.0 9:-1.0 1
-80.5 1:-2.0 2:3.0 3:-2.0 4:-2.0 5:-4.0 6:-1.0 7:1.0 8:-3.0 9:-2.0
-64.1 1:-2.0 2:-1.0 3:0.0 4:5.0 5:-4.0 6:-1.0 7:1.0 8:-2.0 9:-2.0 1
-60.4 1:-2.0 2:-4.0 3:-4.0 4:-4.0 5:-3.0 6:-3.0 7:-4.0 8:-4.0 9:-2.
-113.7 1:0.0 2:-5.0 3:-4.0 4:-5.0 5:10.0 6:-4.0 7:-5.0 8:-4.0 9:-4.
-71.1 1:-2.0 2:-1.0 3:1.0 4:1.0 5:-3.0 6:1.0 7:-1.0 8:-1.0 9:-2.0 1
72.5 1:-2.0 2:-2.0 3:-2.0 4:-1.0 5:-3.0 6:0.0 7:4.0 8:-4.0 9:-2.0 1
-77.5 1:-2.0 2:-3.0 3:-2.0 4:-2.0 5:-4.0 6:1.0 7:0.0 8:0.0 9:3.0 10
-71.1 1:1.0 2:1.0 3:-3.0 4:-3.0 5:-3.0 6:-1.0 7:0.0 8:-3.0 9:-2.0 1
-110.0 1:-2.0 2:2.0 3:2.0 4:4.0 5:-4.0 6:-1.0 7:2.0 8:-3.0 9:-1.0 1
-89.4 1:1.0 2:0.0 3:-2.0 4:0.0 5:-3.0 6:-2.0 7:0.0 8:-3.0 9:-3.0 10
-119.3 1:-1.0 2:-3.0 3:-2.0 4:-2.0 5:-4.0 6:-3.0 7:-3.0 8:7.0 9:-3.
```

圖 3.10　訓練集的範例文件片斷

3.4 使用 LIBSVM 預測

當一切的準備工作都完成後(1 個測試集的編碼檔+6 個訓練集的編碼檔)，便可開始進行訓練模組。這裡所使用的 LIBSVM 工具，版本是 2.81 版。可於下列網址中下載。

LIBSVM：http://www.csie.ntu.edu.tw/~cjlin/libsvm/

由於編碼後的二級結構數值過大，因此要先做一個 Scale 的動作，這個動作有點類似將屬性的值正規化，可預防數值過大或過小，能夠有效減少 SVM 所花費的時間，這裡要注意的是測試集做 Scale，訓練集也要做 Scale，訓練集有 6 個編碼檔，因此要做 6 次 Scale。可使用 LIBSVM 的 svmscale 指令來完成，本研究將 scale 範圍設定在正負 5 之間。

Scale 結束後，接著是訓練模組(model)，由 LIBSVM 的 svmtrain 指令完成，本研究將 SVM 種類設為 3，也就是 epsilon-SVR，這是因為角度屬於連續數值，必須使用迴歸預測的方式，參數 c 則設為 300，參數 g 為 0 或 0.9。最後由訓練集的編碼與 LIBSVM 訓練出 6 個模組。

6 個模組都訓練完後，接著分別用這 6 個模組和 Scale 過的測試集編碼檔，進行預測，這裡可使用 LIBSVM 的 svmpredict 指令來預測測試集的角度。

3.5 代入旋轉公式計算原子 3 維座標

當預測完角度後，必須將它換算成原子座標，在“Introduction to Robotics”【10】一書中提到旋轉公式，恰巧可以運用在角度換算成座標這方面。圖 3.11 所顯示的是 $\vec{k}$ 垂直於底下四角型平面，而且 $\vec{\beta}$ 是由 $\vec{\alpha}$ 繞著 $\vec{k}$ 旋轉 θ 度。

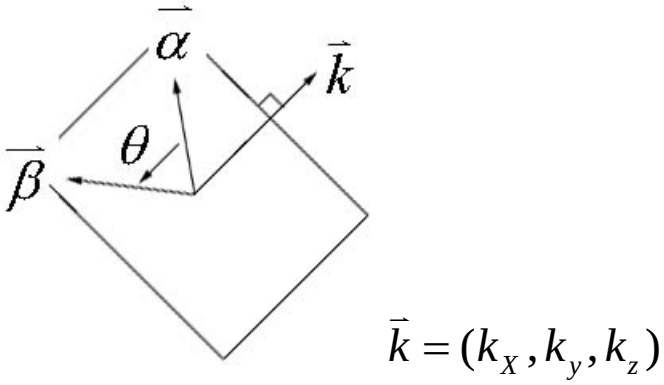

圖 3.11　旋轉示意圖，單位向量

因此 $\vec{\beta} = R_{\vec{k}}(\theta) \bullet \vec{\alpha}$ (3-3)

$$R_{\vec{k}}(\theta) = \begin{pmatrix} k_X^2 \upsilon\theta + c\theta & k_X k_y \upsilon\theta - k_z s\theta & k_X k_z \upsilon\theta + k_y s\theta \\ k_X k_y \upsilon\theta + k_z s\theta & k_y^2 \upsilon\theta + c\theta & k_y k_z \upsilon\theta - k_x s\theta \\ k_X k_z \upsilon\theta - k_y s\theta & k_y k_z \upsilon\theta + k_x s\theta & k_z^2 \upsilon\theta + c\theta \end{pmatrix}$$

其中 $\upsilon\theta = 1 - \cos\theta, c\theta = \cos\theta, s\theta = \sin\theta$

從公式中可以了解到除了θ角度，還必須知道$\vec{\alpha}$、$\vec{k}$才能求出$\vec{\beta}$。這時，參考圖 3.11 和圖 3.8，先設定好第一個胺基酸的 N、CA、C 座標，並計算出$\vec{P}c'$，$\vec{P}c'$是垂直於 N、CA、C 平面的法向量，所以也與$\vec{c}'$互相垂直。在這裡讓$\vec{\alpha} = \vec{P}c'$，$\vec{k} = \vec{c}'$，θ=預測的 PSI 扭角，$\vec{\beta}$=CA、C、N 平面的法向量。將$\vec{\alpha}$、$\vec{k}$、θ的值代入公式後，便求出$\vec{\beta}$，可參考圖 3.12。接著讓$\vec{\alpha} = \vec{c}'$，$\vec{k} = \vec{\beta}$，θ=預測的 CACN 鍵角，$\vec{\beta} = \vec{n}'$，將$\vec{\alpha}$、$\vec{k}$、θ的值代入公式後，便求出$\vec{\beta}$，也就是圖 3.8 的$\vec{n}'$被計算出來，參考圖 3.13。接著將 C 原子座標+$\vec{n}'$*CN 鍵長 = 下一個胺基酸的 N 原子座標。(鍵長可參考圖 2.10)

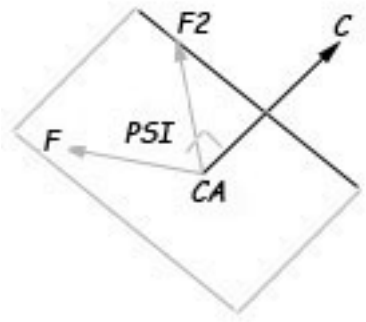

圖 3.12　旋轉示意圖 2，F、F2 表示法向量

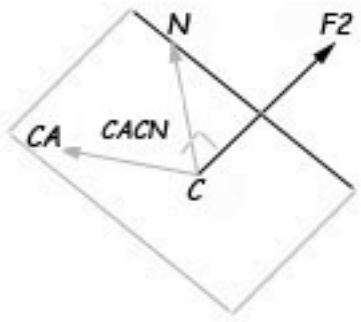

圖 3.13　旋轉示意圖 3，F2 表示法向量

當 N 原子座標計算出來後，再用 CA、C、N 平面的法向量繞著 CN 向量轉 OMEGA 度，求出 C、N、CA 平面的法向量。下一步用 CN 向量繞著 C、N、CA 平面的法向量轉 CNCA 鍵角，求出 NCA 向量。N 原子座標+NCA 向量*NCA 鍵長 = CA 原子座標。

當 CA 原子座標計算出來後，再用 C、N、CA 平面的法向量繞著 NCA 向量轉 PHI 度，求出 N、CA、C 平面的法向量。下一步用 NCA 向量繞著 N、CA、C 平面的法向量轉 NCAC 鍵角，求出 CAC 向量。CA 原子座標+CAC 向量*CAC 鍵長 = C 原子座標。

重覆以上的步驟將之後的所有原子座標逐一算出，最後儲存成 PDB 檔，即為預測的蛋白質三級結構。

3.6　以 RMSD 評估

為了觀察本研究的實驗方法是否適用於蛋白質的三級結構預測，因此測試集蛋白質必須選擇結構已知的蛋白質並與實驗最後預測出來的蛋白質結構做一個比較。關於蛋白質三級結構的評估方式，最常被使用的應該屬於RMSD(Root Mean Squared Distance)，RMSD值是越低越好，公式如下。

RMSD 公式：

$$\sqrt{\frac{\sum_{i=1}^{N} d_i^{\,2}}{N}} \qquad (3\text{-}4)$$

其中 d_i=2 個原子間的距離，N=原子數量

許多文章只針對每個胺基酸的 CA 原子計算 RMSD 值，因此實驗最後也只計算每個胺基酸的 CA 原子 RMSD 值。而且在計算 RMSD 值之前要先進行結構重疊的動作，這是為了避免二條結構相似的蛋白質因為距離太遠所以計算出來的 RMSD 效果太差，這跟 RMSD 的公式有關。結構重疊則是將其中一條蛋白質經過旋轉、位移後與另一條蛋白質重疊，當 RMSD 值達到最低時，結構重疊就會停止。因此，將二條蛋白質結構重疊，縮小他們之間的距離。CCP4MG 軟體可以做結構重疊的動作，實驗中使用的版本是 0.12。
CCP4MG網站：http://www.ysbl.york.ac.uk/~ccp4mg/

當 CCP4MG 將預測的 PDB 檔經過旋轉、位移後與正確的 PDB 檔結構重疊，並產生新的 PDB 檔。最後，用新的 PDB 檔跟正確的 PDB 檔比較，計算其 RMSD，最小的 RMSD 就這樣被計算出來。

第四章　實驗結果

預測蛋白質三級結構有多種方法，例如：同源模擬法、摺疊辨識法、重新計算法等，而本研究的做法與同源模擬法較類似，一開始也是利用目標蛋白質的胺基酸序列到結構已知的蛋白質資料庫尋找同源蛋白質；本研究將目標蛋白質稱為測試集(Test Set)，模板稱為訓練集(Train Set)，使用的相似性搜尋工具為PSI-BLAST。當決定測試集與訓練集後，還必須準備 PSSM 跟二級結構，PSSM 一樣由 PSI-BLAST 產生，測試集的二級結構使用預測的二級結構，訓練集的二級結構則由 DSSP 工具轉換而得。接著依照 3.3 節所介紹的方法進行編碼，編碼結束後，計算訓練集裡每個胺基酸的 6 種角度(3 個扭角、3 個鍵角)，再將訓練集的 6 種角度與它的編碼合拼，產生 6 個 LIBSVM 輸入文件，然後使用 LIBSVM 訓練出 6 個模組，再以這 6 個模組和測試集的編碼預測出測試集的 6 種角度，最後以預測出的 6 種角度代入機器人學的旋轉公式再配合固定的鍵長，便可以計算出測試集的原子座標，將座標存成 PDB 檔，並與原本測試集的 PDB 檔互相比較、結構重疊、計算 RMSD。

實驗一開始要先決定要預測的蛋白質，也就是決定測試集，實驗裡會針對 1P7E、1ABA、1LTS 等蛋白質進行預測。1P7E 是自行至 PDB 資料庫裡找到的，是一條最近幾年才被發現的蛋白質。1ABA 和 1LTS 是 Comparative protein structure modeling by iterative alignment, model building and model assessment 文章裡的實驗蛋白質。

接著，調整實驗中的輸入、輸出變數和鍵長，將同一條蛋白質的預測分成多組，並觀察之中的差異。輸入變數的調整指編碼時二級結構的部份，測試集是使用預測的二級結構來進行編碼，這裡可以將它調整成正確的二級結構，或是編碼時只使用 PSSM 來編碼，不考慮測試集和訓練集的二級結構，藉此觀察二級結構對蛋白質三結構預測的影響。輸出變數的調整指的是針對預測後的鍵角進行調整，鍵角可使用預測的鍵角、正確的鍵角、固定的鍵角等三種。目的是要了解鍵角對蛋白質三級結構的影響以及鍵角應該適用何種方式來決定。最後是鍵長的調整，鍵長可以調整成固定的鍵長或是正確的鍵長，目的與鍵角的調整一樣。測試集裡的蛋白質均是已知結構的蛋白質，因此二級結構、鍵角、鍵長也是已知的，調整時正確的二級結構、正確的鍵角、正確的鍵長指的就是這些已知的資料。

所有組別如下：

第一組：輸入變數-測試集使用預測的二級結構進行編碼

輸出變數-使用預測的鍵角計算座標

鍵長-使用正確的鍵長計算座標

第二組：輸入變數-測試集使用預測的二級結構進行編碼
輸出變數-使用預測的鍵角計算座標
鍵長-使用固定的鍵長計算座標

第三組：輸入變數-測試集使用預測的二級結構進行編碼
輸出變數-使用固定的鍵角計算座標
鍵長-使用固定的鍵長計算座標

第四組：輸入變數-測試集使用預測的二級結構進行編碼
輸出變數-使用正確的鍵角計算座標
鍵長-使用固定的鍵長計算座標

第五組：輸入變數-測試集使用正確的二級結構進行編碼
輸出變數-使用預測的鍵角計算座標
鍵長-使用固定的鍵長計算座標

第六組：輸入變數-測試集和訓練集只使用 PSSM 進行編碼
輸出變數-使用預測的鍵角計算座標
鍵長-使用固定的鍵長計算座標

第一組和第二組可以觀察鍵長的影響，第二組、第三組和第四組可以了解鍵角的影響，第二、五、六組能觀察二級結構的影響。接著針對 1P7E、1ABA、1LTS 等蛋白質進行預測。值得一提的是，最後除了比較原文章的預測結果外還加入了另一種預測方法的結果，所用的方法也是同源模擬法，流程如圖 4.1。它的做法是先輸入一組排比資料到實驗中，然後使用演化式計算結合同源模擬法的方式來塑模，最後以適性值來判斷並決定結構。利用演化式計算的特性，可以搜尋到相當多的序列排比空間，而塑模的工具則是使用 MODELLER。

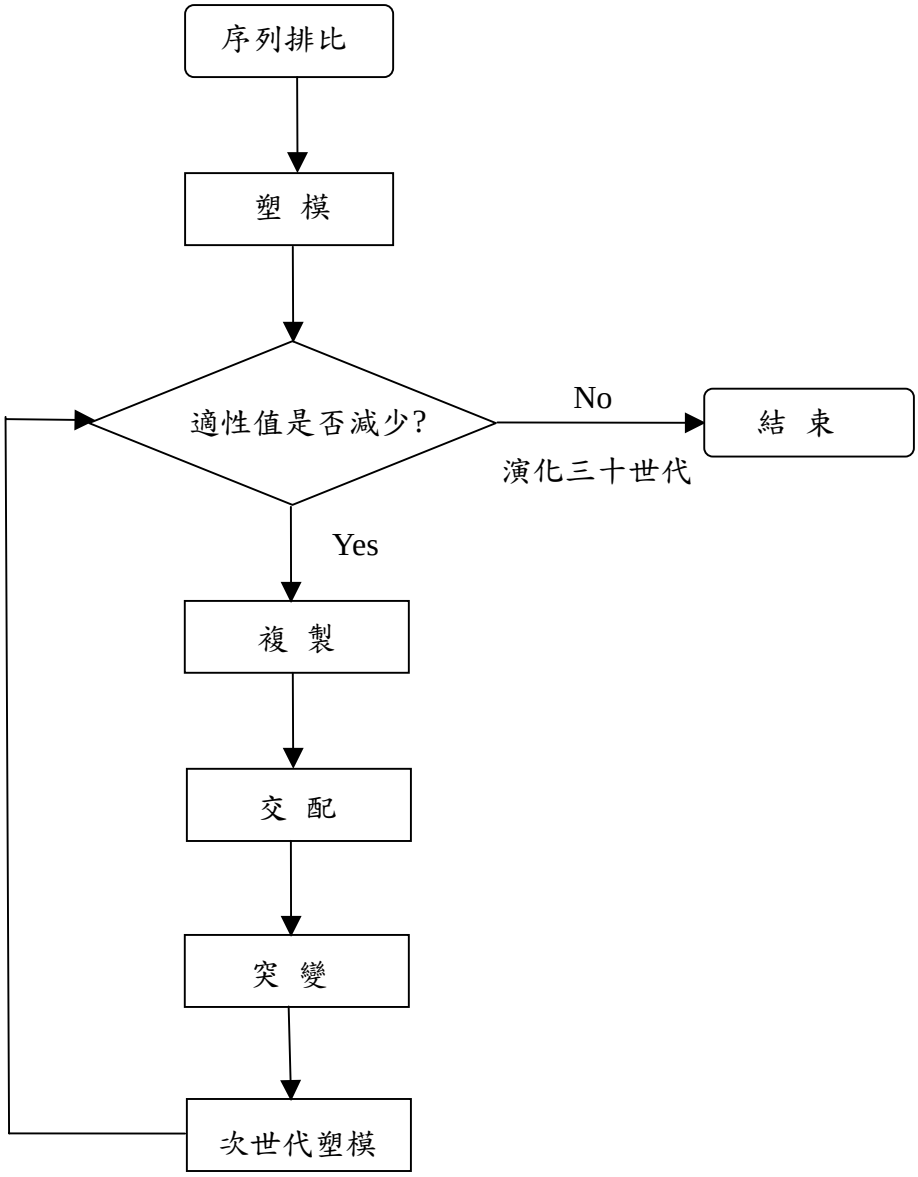

圖4.1　其它方法流程圖

4.1 1P7E

測試集：1P7E　　長度：56　　Chain：A
訓練集 1：　相似度(Identity)：72%~100%
1PGX 16 69
1P7F:A 1 56
2IGH 8 61
2IGD 8 61
1QKZ:A 8 61
2IGG 8 61
1FCC:C 3 56
1PN5:A 1 56
1IBX:B 1 56
2GB1 1 56
1Q10:B 1 56
1MVK:L 1 56
1FD6:A 4 57
1GB4 4 57
1FCL:A 3 56
1EM7:A 3 56
1MHX:A 26 65
1WND:D 22 42
1MI0:B 12 65

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0.9
Scale 範圍：-5 ~ 5

訓練集 2：　相似度(Identity)：72%~89%
1PN5:A 1 56
1IBX:B 1 56
2GB1 1 56
1Q10:B 1 56
1MVK:L 1 56
1FD6:A 4 57
1GB4 4 57
1FCL:A 3 56
1EM7:A 3 56

1MHX:A 26 65
1WND:D 22 42
1MI0:B 12 65

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0
Scale 範圍：-5 ~ 5

1P7E 是一條近期被發現的蛋白質，序列長度 56。在經過 PSI-BLAST 決定訓練集後，本研究將訓練集分成 2 組，分別是訓練集 1 與訓練集 2，訓練集 2 是使用相似度比訓練集 1 差的蛋白質，當決定訓練集後，便開始編碼→訓練→預測。實驗中將對 1P7E 分成 6 組進行實驗，每次都會針對輸入、輸出變數或是鍵長加以變動，用以了解這些項目對實驗結果的影響，結果列於表 4.1。

表 4.1　1P7E 實驗結果表

	訓練集 1 CA RMSD	**訓練集 2 CA RMSD**
第一組	0.6	
第二組	0.6	14.6
第三組	1.7	
第四組	0.7	
第五組	0.4	
第六組	0.4	14.5

觀察表 4.1，首先是訓練集 1 的實驗結果。第一組和第二組在於鍵長的不同，結果顯示 1P7E 鍵長對於 1P7E 實驗結果的影響並不大。第二組、第三組跟第四組在於輸出變數-鍵角的不同，結果顯示 1P7E 使用固定鍵角來計算座標的效果最差，這也說明了鍵角對結構的影響。接著第二組與第五組差別在於輸入變數，測試集二級結構的使用，第二組採用預測的二級結構進行編碼，第五組使用正確的二級結構進行編碼，結果顯示測試集二級結構預測的正確性與否會產生影響，1P7E 預測的二級結構有 9 個位置被預測錯誤，因此第二、五組 RMSD 有些微差異。最後，第六組則是在輸入變數上不加入二級結構的編碼，只使用 PSSM 編碼來進行實驗，這再次證明了二級結構對實驗結果的影響。

訓練集 1 與訓練集 2 的結果以訓練集 1 的結果為佳，也就是說，使用相似度較高的同源蛋白質當訓練集會得到較佳的結果。

計算 RMSD 之前必須進行結構重疊的動作，底下圖 4.2 至圖 4.5 為 1P7E 蛋白質三級結構與預測後的 1P7E 蛋白質三級結構使用 CCP4MG 軟體進行結構重疊的情形。

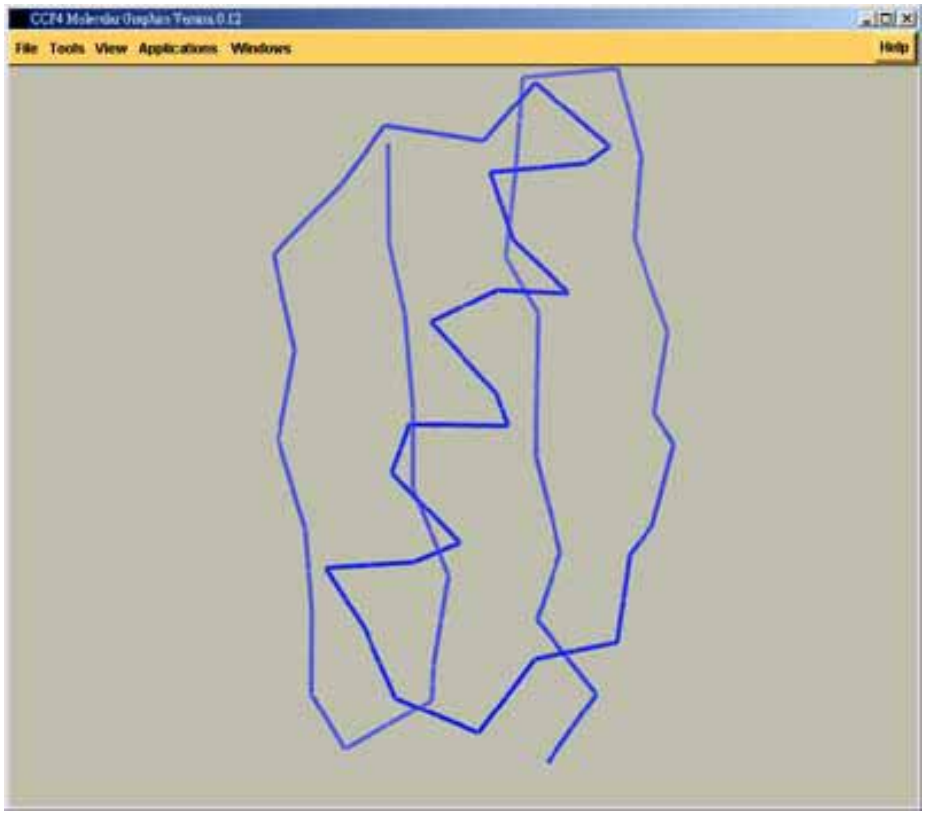

圖 4.2　1P7E 結構—正確的 PDB 檔

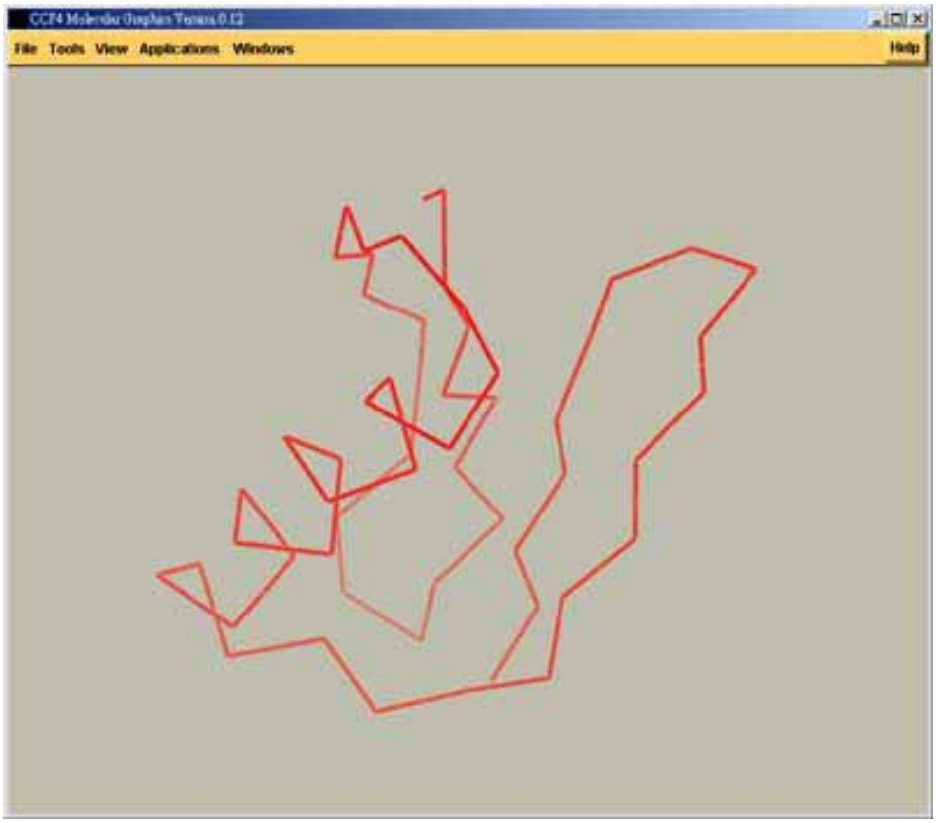

圖 4.3　預測的 17PE 結構—預測的 PDB 檔

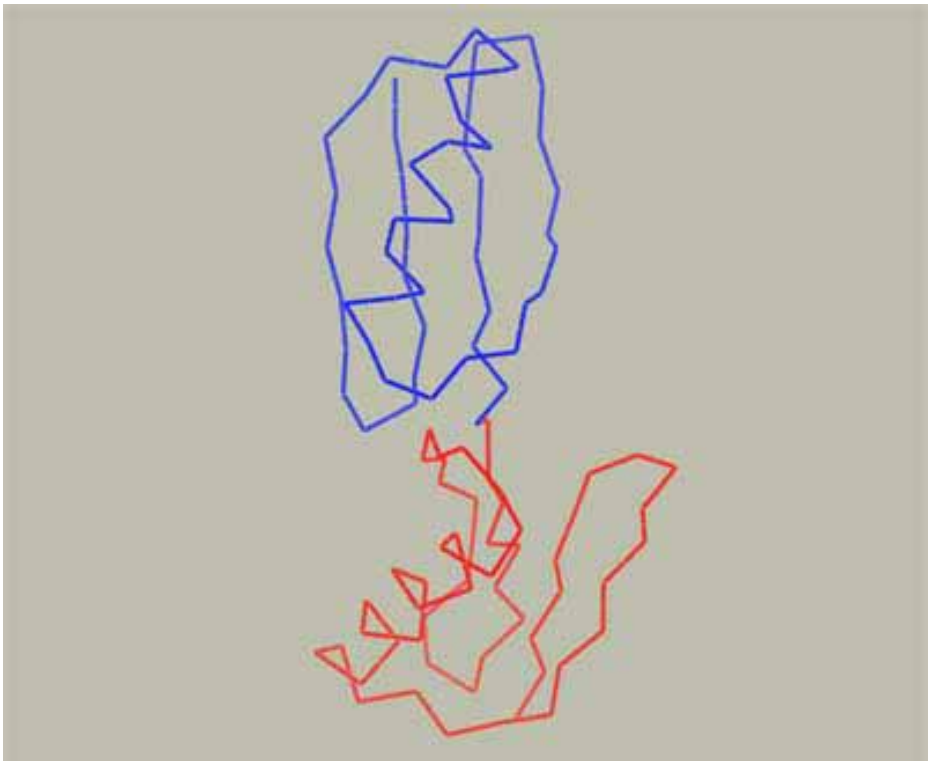

圖 4.4　正確跟預測的結構尚未進行結構重疊前

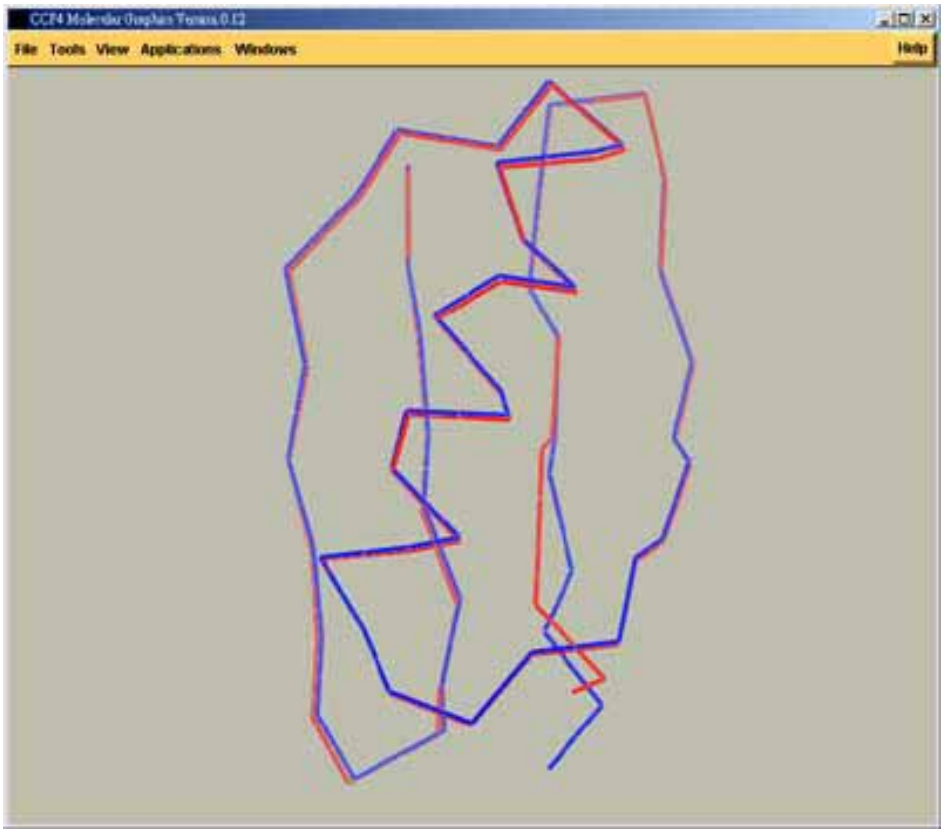

圖 4.5　正確跟預測的結構進行結構重疊後

4.2 1ABA

測試集：1ABA　　長度：87

訓練集 1：　　相似度(Identity)：22%~100%

1DE2:A 1 87

1FOV:A 11 68

1NM3:B 180 240

3GRX 11 68

1EGR 11 80

1QFN:A 11 80

1KTE 22 90

1JHB 23 86

1G7O:A 2 35

1H75:A 11 33

1VD5:A 227 262

1S3X:A 267 313

1HJO:A 265 311

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0
Scale 範圍：-5 ~ 5

訓練集 2：　相似度(Identity)：22%~38%
1FOV:A 11 68
1NM3:B 180 240
3GRX 11 68
1EGR 11 80
1QFN:A 11 80
1KTE 22 90
1JHB 23 86
1G7O:A 2 35
1H75:A 11 33
1VD5:A 227 262
1S3X:A 267 313
1HJO:A 265 311

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0
Scale 範圍：-5 ~ 5

1ABA是“Comparative protein structure modeling by iterative alignment, model building and model assessment”【7】的實驗蛋白質，文章中表2記錄1ABA最好的實驗結果RMSD為4.1，序列長度87。

實驗中一樣對1ABA分成6組進行實驗並將訓練集分成2組，每次都會針對輸入、輸出變數或是鍵長加以變動，用以了解這些項目對實驗結果的影響，結果如下表4.2。

表 4.2　1ABA 實驗結果表

	訓練集 1 CA RMSD	訓練集 2 CA RMSD
第一組	1.5	
第二組	1.5	10.7
第三組	2.0	
第四組	3.2	
第五組	1.5	
第六組	1.5	10.7

觀察表 4.2，第一組和第二組在於鍵長的不同，結果顯示 1ABA 鍵長對於 1ABA 實驗結果的影響並不大。第二組、第三組跟第四組在於輸出變數-鍵角的不同，結果顯示 1ABA 使用正確鍵角來計算座標的效果最差，這應該是受到扭角的影響。但表中依舊可以了解到鍵角的差異對於結構的影響。接著第二組與第五組差別在於輸入部份測試集二級結構的使用，第二組採用預測的二級結構進行編碼，第五組使用正確的二級結構進行編碼，表中的結果是出乎意料之外，1ABA 預測的二級結構有 11 個位置被預測錯誤，但實驗結果卻一樣，唯一的解釋是研究方法中對於二級結構的編碼是適用於 1ABA 蛋白質。最後，第六組則是在輸入變數中不加入二級結構的編碼，只使用 PSSM 編碼來進行實驗。當訓練集的相似度降至 22%~38%時，效果就比較差。

4.3 1LTS

測試集：1LTS　　長度：103　　Chain：D
訓練集 1：　相似度(Identity)：80%~100%
1PZI:H 1 103
1LTR:H 1 102
1B44:H 1 102
2CHB:H 3 104
1S5F:H 2 103
3CHB:H 3 104
1CHQ:H 2 103
1CHP:H 2 103
1G8Z:H 2 103
1CT1:H 2 103
1XTC:H 2 103

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0.9
Scale 範圍：-5 ~ 5

訓練集 2：　相似度(Identity)：80%~82%
2CHB:H 3 104
1S5F:H 2 103
3CHB:H 3 104
1CHQ:H 2 103
1CHP:H 2 103
1G8Z:H 2 103
1CT1:H 2 103
1XTC:H 2 103

Sliding Window：11
SVM 參數設定：-s 3 -c 300 -g 0
Scale 範圍：-5 ~ 5

1LTS 是“Comparative protein structure modeling by iterative alignment, model building and model assessment”【7】的實驗蛋白質，文章中表 1 記錄 1LTS 最好的實驗結果 RMSD 為 3.1，序列長度 103。

實驗中一樣將對 1LTS 分成 6 組進行實驗並將訓練集分成 2 組，每次都會針對輸入、輸出變數或是鍵長加以變動，用以了解這些項目對實驗結果的影響，結果列於下表 4.3。

表 4.3　1LTS 實驗結果表

	訓練集 1 CA RMSD	**訓練集 2 CA RMSD**
第一組	13.5	
第二組	13.5	17.4
第三組	12	
第四組	13.8	
第五組	9.5	
第六組	0.9	17.1

由表 4.3 可以得知輸入變數加入二級結構編碼的效果較差，原因有可能是實驗中對二級結構的編碼並不適合 1LTS 的預測，而且 1LTS 預測的二級結構有 25 個位置被預測錯誤，這也會對結果造成影響。表 4.3 中比較特別的是當輸入變數使用正確的二級結構進行編碼時，RMSD 達到 9.5，這是因為測試集本身的二級結構與訓練集的二級結構有些許不同所造成的結果。如果降低訓練集的相似度，如訓練集 2，則結果依舊是表現較差。

整合以上實驗結果並與其它文章的實驗結果做一比較，列表於 4.4。由表 4.4 的實驗結果可以看出，鍵長對於結構的影響並不大，反而是鍵角影響較大，而使用預測的鍵角，可得到不錯的結果，因此輸出的部份可以考慮使用預測的鍵角。之所以會有固定鍵角是因為許多文章都將鍵角、鍵長以及 OMEGA 扭角設為固定值，但事實上同一條蛋白質的鍵角、鍵長及 OMEGA 並未相同，是有差異的。不過鍵長的差異較小，可以使用固定的鍵長來表示，鍵角及 OMEGA 的差異較大，因此全都加入預測。

至於輸入變數部份，使用預測的二級結構來編碼，或許是不太適當的，但有些蛋白質雖然胺基酸序列可能不同，但結構本身卻很相似，這時有二級結構多少都會有助於這方面蛋白質的預測。測試集的蛋白質本該是一條只有一級結構的蛋白質，它所提供的資訊有限，因此二級結構是有它存在的必要。

當訓練集的相似度降低後，預測的效果也跟著降低，這方面或許可以從編碼來改善，可以考慮增加其它屬性或是以其它方式進行編碼。而實驗後的結果都是略差於對照組 2，本研究只針對蛋白質主鏈預測，並未預測側鏈的部份，因此也沒有進行後續的處理，如能以能量最小化來處理，相信結果應該是會改善的。

表 4.4　實驗結果對照表

	第一組	**第二組**	**第三組**	**第四組**	**第五組**	**第六組**	**對照組 1**	**對照組 2**
1P7E	0.6	0.6	1.7	0.7	0.4	0.4		0.2
1P7E*		14.6				14.5		
1ABA	1.5	1.5	2.0	3.2	1.5	1.5	4.1	1.3
1ABA*		10.7				10.7		
1LTS	13.5	13.5	12	13.8	9.5	0.9	3.1	0.3
1LTS*		17.4				17.1		

*表示使用訓練集 2

對照組 1：原文章的實驗結果。

對照組 2：同源模擬法加演化計算所得的結果，1P7E 的模板是 1P7F(1~56)，1ABA 的模板是 1DE2:D(1~87)，1LTS 的模板是 1PZI:H(1~103)。

4.4 實驗限制

依據第三章的實驗方法，對於某些蛋白質依然無法有效的預測。由於訓練集的取得是決定於 PSI-BLAST 的搜尋結果，這點和同源模擬法找模板的方式一樣，如果無法搜尋到同源序列或模板則無法預測其三級結構。

第五章　結論與建議

5.1 結論

從決定一條測試集開始，中間經過尋找訓練集、準備 PSSM 和二級結構、編碼、計算訓練集角度、SVM 訓練、預測，最後以旋轉公式配合鍵長計算出原子 3 維座標，得到蛋白質三級結構。整個流程之中，最需要改進的應該是編碼部份。從之前的實驗結果可以看出，本研究的預測方法對於相似度較低的訓練集而言，效果並不是很好，這點或許可以從編碼方式來改善，或則是增加編碼時的屬性。至於扭角、鍵角與鍵長，扭角的變化較大，並不適合用固定的常數來代替，只能以預測的方式產生。鍵角使用 SVM 預測可以得到不錯的結果。鍵長則可設為固定常數。從實驗結果可以得知，未來還是有很大的改善空間，尤其在編碼、尋找同源序列及後續處理方面。

5.2 未來研究方向

未來的研究，應該針對 4 點著手進行。

(1) 相似度不高的訓練集處理方法
(2) 編碼時的屬性選擇
(3) 側鏈的預測
(4) 後置處理

5.2.1 相似度不高的訓練集處理方法

實驗方法對於相似度不高的訓練集無法有效的處理，這方面或許可以利用其它方法，例如使用預測的二級結構來尋找訓練集，或是以 CATH 的分類方式訓練出幾個大樣本的模組，用來初步分析測試集的結構等等。

5.2.2 編碼時的屬性選擇

屬性方面，目前只有 PSSM 和二級結構，PSSM 跟二級結構所提供的資訊也是有限的，應該還可以再加入其它資料來編碼當做屬性。至於編碼方法或許可以考慮其它編碼方式。

5.2.3 側鏈的預測

側鏈的預測是比較困難的，但沒有側鏈的蛋白質是不完整的，它不像主鏈一樣那麼有規律，20 種胺基酸就有 20 種不同的側鏈，因此對於側鏈的預測要考慮其它的方法才行。

5.2.4 後置處理

實驗最後計算出原子座標後，並未加以後置處理，例如將結果放置分子動力軟體，解決空間上重疊的問題，這個步驟必須有側鏈才能進行。或是比對訓練集蛋白質的原子距離，再將預測的結果加以調整等等。

參考文獻

【1】 李錦和，2003，一個以分子動力模擬為基礎之蛋白質摺疊平行演算法，東吳大學商學院資訊科學所碩士論文。

【2】 林長青，2003，支撐向量機應用於科學探索，雲林科技大學電子與資訊工程研究所碩士論文。

【3】 陳奎昊，2003，蛋白質問題之研究，國立暨南國際大學資訊工程學系，碩士論文。

【4】 葉吉原，2004，使用混和式方法預測蛋白質二級結構，樹德科技大學資訊管理學所碩士論文。

【5】 韓歆儀，2004，應用兩階段分類法提昇SVM法之分類準確率，成功大學工業與資訊管理研究所碩士論文。

【6】 Asai, K., Haymizu, S., Handa, K., 1993, "Prediction of protein secondary structure by the hidden Markov model", CABIOS 2, 141-146.

【7】 Bino J., Andrej S., 2003, "Comparative protein structure modeling by iterative alignment, model building and model assessment".

【8】 Chang C-C, Lin C-J, 2001, "LIBSVM：a library for support vector machines".

【9】 C.H.Q. Ding, I. Dubchak, 2001, "Multi-class protein fold recognition using support vector machines and neural networks", Bioinformatics.

【10】Craig, "Introduction to Robotics", 2nd ed. Page 52.

【11】Daggett, V., Levitt, M., 1993, "Protein unfolding pathways explored through molecular dynamics simulations", J. Mol. Biol. 232, 600- 619.

【12】Engh R A., Huber R., 1991, "Accurate bond and angle parameters for X-ray protein structure refinement", Acta Cryst., A47, 392-400.

【13】F-K Lin, 2002, "Protein Secondary Structure Prediction Using Genetic Algorithm", Institute of Computer and Information Science National Chiao-Tung University.

【14】Gibrat J. F., Robson, B., Garnier, J., 1987, "Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs", J. Mol. Biol. 198, 425-443.

【15】Gunn, Steve R., 1998, "Support Vector Machine for Classification and Regression", Technical Report, Faculty of Engineering and Applied Science, Department of Electronics and Computer Science, University of Southampton.

【16】Holley, L. H., Karplus, M., 1989, "Protein secondary structure prediction with a neural network", Proc Natl. Acad. Sci. USA 86, 152-156.

【17】Hyunsoo K., Haesun P., 2003, "Prediction of Protein Relative Solvent Accessibility with Support Vector Machines and Long-range Interaction 3D Local Descriptor", Department of Computer Science and Engineering,

University of Minnesota.
【18】Joachims, T., 1998, "In Proceedings of the European Conference on Machine Learning", Springer-Verlag, Berlin/New York.
【19】J. R. Quine Matt Brenneman, T. A. Cross, 1997, "Protein Structural Analysis From Solid State Nmr Derived Orientational Constraints", Department of Mathematics Florida State University.
【20】Marc S. Wold, 2001, "Principles of Protein Structure", Biochem 34, 10703. August 27.
【21】Mihai P., 1996, "Protein Folding: Computational Challenges", May.
【22】Myers, J. K., Oas, T. G., 2001, "Preorganized secondary structure as an important determinant of fast protein folding", Nature Structure Biology 8, 552-558.
【23】Nello C., John S. T., 2000, "An Introduction to Support Vector Machines and other kernel-based learning methods".
【24】R. B. Russell, R. R. Copley, G. J. Barton, 1996, "Protein fold recognition by mapping predicted secondary structures", Journal of Molecular Biology, 259:349-365.
【25】R-S Cheng, 2003, "Protein Structure Prediction Based on Secondary Structure Alignment", Department of Computer Science and Engineering National Sun Yat-sen University.
【26】Rui K., Christina S. L., A-S Yang, 2004, "Protein backbone angle prediction with machine learning approaches", BIOINFORMATICS, Vol.20 no. 102004, pages 1612-1621.
【27】Sean L. F., 2001, "Torsion Angle Selection and Emergent Non-Local Secondary Structure In Protein Structure Prediction", University of Iowa.
【28】Steffen S. K., 1996, "Genetic Algorithms and Protein Folding".
【29】Stephen F., Altschul et al., 1997, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, Vol. 25, No. 17, 3389-3402.
【30】Sujun H., Zhirong S., 2001, "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach", J.Mol. Biol,308:397-407.
【31】Thomas C. K. T., 2004, "Discovering Protein Sequence-Structure Motifs and Two Applications to Structural Prediction".
【32】Vitali P., Edward C., 1998, "SWAN: sliding window analysis of nucleotide sequence variability", The Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Road Oxford OX1 3PS, UK, BIOINFORMATICS APPLICATIONS NOTE,

Pages 467-468, Vol. 14 no. 5.
【33】Voet , 1995, "Biochemistry", John Wiley & Sons, 2nd edition.
【34】V. Vapnik, 1998, "Statistical Learning Theory", Wiley-Interscience, New York.
【35】V. Vapnik, 1995, "The Nature of Statistical Learning Theory", Springer, New York.
【36】Wolfgang K., Chris S., 1983, "Description of the DSSP program", MPI MF, Heidelberg.
【37】Y-d Cai, S-L Lin, 2003, "Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence".