藉由結構特性促進蛋白質之交互作用位置預測與分析

蘇郅挺 蘇義斌 游景盛
逢甲大學資訊工程學系
yucs@fcu.edu.tw

摘要

蛋白質間交互作用位置發生在兩蛋白質間的接觸表面的胺基酸位點，並深受交互作用區的胺基酸組成與結構影響，而確認交互作用位置亦有助於了解蛋白質功能。在本論文中我們利用支援向量機(support vector machine)演算法以及蛋白質序列上特徵的方法預測蛋白質間交互作用，在 PDB 資料庫中選取的 333 個蛋白質複合體資料集中，其 1134 條蛋白質鏈用以評估訓練和測試。不同於以往預測使用的序列演化的資訊，我們以各種結構特性分數矩陣作為特徵，如常見的二級結構(secondary structure)和暴露表面積(solvent-accessible surface area)等統計的機率分布形成之分數矩陣，與前人研究相比較，結果顯示整合結構特性分數矩陣後的預測結果有效將蛋白質交互作用位置預測準確率提升至 70%。

1 導論

1.1 研究動機

研究蛋白質交互作用(Protein-Protein Interaction; PPI)可以區分成兩大部分：其一是找出蛋白質間交互作用關係的網路(interaction network)，另一則是研究兩蛋白質間物理上的接觸所形成交互作用的機制以及其接觸胺基酸位點，而生物分子中常見的複合體即為數個蛋白質間彼此互相靠近並以特定的結合順序或位置所形成的功能性組合物。因此了解蛋白質交互作用的胺基酸位置和機制能夠幫助了解此蛋白質實際的功能；另一方面，由於透過結構生物學實驗解開蛋白質的三級結構相當耗費時間和金錢，若能藉由計算的方式獲得相關資訊，不僅可以快速的了解兩蛋白質交互作用的 3D 結構傾向，對蛋白質間交互作用網路的拼湊也有所助益。

1.2 研究目的

本研究的目的，希望能夠藉由引入更多的蛋白質結構特性資訊，輔助以往單獨使用序列資訊預測蛋白質交互作用位置，結構特性可藉由統計分析短胺基酸序列可能的典型結構區塊並分數矩陣化，因此即使缺乏蛋白質真實結構資料，仍舊能夠進行預測。

2 文獻研究與回顧

預測蛋白質交互作用的研究目前並無直接有效方式，多以機器學習演算法為主[1-12] ，如類神經網路(neural network)[2,4,7]、支援向量機(support vector machine) [3,5,6,10]、隨機森林演算法(random forest algorithm)[12]等。進行分類學習的過程所使用的特徵，大致可分為三類，第一類是單獨使用蛋白質序列的特徵[1-4]，如過去 Res 等人使用支援向量機預測 50 條蛋白質交互作用位置[3] ，結合序列相似度和演化上的資訊，並將演化資訊和多重序列比對的結果做排序(real value evolutionary trace)，最高達到 64%準確度；第二類是利用結構的資訊去提煉出序列上的資訊作為特徵[5-7]；第三類是直接使用結構的特徵，或者結合序列和結構的特性[8-11]。

由於蛋白質交互作用和其本身結構有很大的關係，使用結構特徵以改進預測方法在許多研究中可見明顯提升準確度，然而，目前蛋白質結構資料遠少於蛋白質序列本身的資訊，因此不少研究著重在於根據序列以及預測出的結構特性，進一步預測蛋白質交互作用位置。如 Ofran [4]研究中根據胺基酸序列、從而衍生的演化資訊和預測的二級結構資訊，以類神經網路演算法進行機器學習預測研究。單純使用序列可達 58%的準確率，使用演化上的資訊可達到 60%的準確率；結合演化資訊和預測的二級結構以及胺基酸暴露表面積當作特徵，達到 68%準確率。

3 材料與方法

3.1 資料集(datasets)

為了評估預測結果，我們使用與前人預測蛋白質交互作用位置的相同資料集，先由蛋白質結構(Protein Data Bank; PDB)資料庫挑選經結構確認共含 2283 條蛋白質鏈的複合體資料，為了減少影響預測評估的變因，經由以下程序逐步篩選過濾：每條篩選出來的蛋白質鏈必須和其他蛋白質鏈具有暫時性交互作用的胺基酸位點、任兩條的序列相似度必須小於 25%且長度須大於 100 個胺基酸。為了避免爭議性，排除由 NMR 方法解出的蛋白質結構、或根據理論建立的模組、以及序列長度小於 30 個胺基酸的所有序列。

為了取出具有生物功能交互作用的蛋白質，同時確認 PQS Server (Protein Quaternary Structure Server)[14]資料庫，計算交互作用時暴露表面積的減少量，並排除掉不是由於交互作用而有功能的蛋白質。最後的資料集，包括了從 PDB 資料庫中取出的 333 個蛋白質複合體，總共有 1134 條蛋白質鏈。這個資料集也是 Ofran 和 Rost 所使用的資料集[2,4]。

另外，過去研究上對於交互作用位置的定義，有些使用兩個胺基酸的 Cα 或 Cβ 距離為基準，小於門檻值(從 4 到 12Å)的區塊定義為交互作用位置，然而二十種胺基酸的大小不相同，因此我們依照 Ofran 和 Rost 的定義[2]：當胺基酸上任原子，若和另條蛋白質序列上的任意胺基酸的任意原子距離小於 6Å 的話，那麼這個胺基酸就是有交互作用的胺基酸。

3.2 預測流程

如圖 3.1，我們使用序列上獲得演化的資訊(PSSM)，以及藉由統計方式獲得的二級結構(secondary structure)和暴露表面積(solvent-accessible surface area)等結構特徵，利用支援向量機(support vector machine)預測蛋白質交互作用位置。

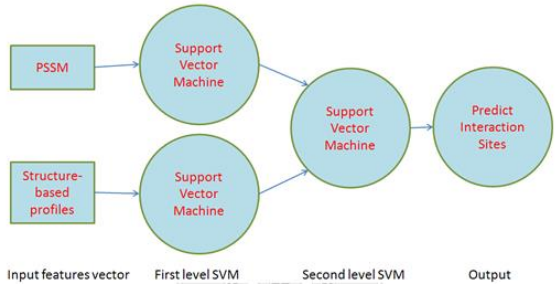


圖 3.1 預測流程圖

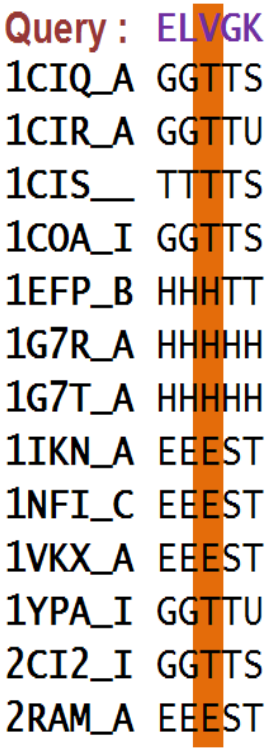
當要預測一個胺基酸是否參與交互作用時，除了這個胺基酸本身的特性，周邊的胺基酸特性也會影響。Sikic 等人藉由計算 entropy 的值發現滑動視窗大小為九的時候周圍胺基酸的特性對於欲預測的胺基酸影響最大[12]。

3.3 位置加權矩陣(Position Specific Scoring Matrix)

我們使用 PSI-BLAST 產生位置加權矩陣(Position-Specific Scoring Matrix)當作特徵預測蛋白質交互作用的比較基準。這個分數矩陣是將經由和不重複性蛋白質資料庫(non-redundant)進行遞迴序列比對而來，根據比對的結果可以統計出序列上每個胺基酸變異的機率。許多預測蛋白質交互作用相關研究皆以位置加權矩陣為預測蛋白質交互作用的基礎，Res[3]和 Ofran 等研究顯示以位置加權矩陣預測的整體準確度皆達到 60%。

3.4 結構特性分數矩陣

本論文使用 Chan[15]所提出的數種蛋白質結構特性分數矩陣，這些特性是根據序列 $\alpha = (\alpha 1, \alpha 2, \cdots, \alpha l)$和 PDB 資料庫比對序列長度為 l 的胺基酸片段，利用機率向量表達計算所得某種結構特性定義的分數矩陣，以圖 3.2 為例，如欲預測的胺基酸是”ELVGK”的”V”胺基酸，必須去資料庫查詢具有相同的短序列的蛋白質資料，並統計這些序列資料當中相同短序列的”V”胺基酸的結構，並計算出結構特性機率分布，因此在學習預測的過程中，並未直接使用到欲預測序列的真實結構資訊。本研究共使用八種結構特性所產生的分數矩陣當作預測的特徵值，這八種結構特性資訊分述如下。



$$P_3^{ELVGK} = (P_B, P_E, P_G, P_H, P_I, P_S, P_T, P_U)$$
$$= \left(\frac{0}{13}, \frac{4}{13}, \frac{0}{13}, \frac{3}{13}, \frac{0}{13}, \frac{0}{13}, \frac{6}{13}, \frac{0}{13}\right)$$
$$= (0, 0.3077, 0, 0.2308, 0, 0, 0.4615, 0)$$

圖 3.2 產生結構特性分數矩陣

3.4.1 二級結構 (Secondary Structure; SS)

蛋白質序列由於氫鍵和各種引力，會形成蛋白質二級結構，常見的三種蛋白質二級結構為：螺旋狀的 α-helix、摺板狀的 β-sheet、以及其他捲曲狀的 coil。根據 Dictionary of Protein Secondary Structure(DSSP)[16]，可進一步細分成為螺旋狀的 3_{10}-helix、α-helix、π-helix，摺板狀的 isolated β bridges、extended β sheet，以及其他捲曲狀的 bend、turn、和 others 等八種型態。

3.4.2 ALPHA 鍵角

如圖 3.5，雙平面夾角 ALPHA 是第 i-1、i、i+1 個 Ca 形成的平面和第 i、i+1、i+2 個 Ca 形成的平面，這兩個平面之間的雙平面夾角，角度範圍從-180°到 180°，區分為 12 個區間，每個區間範圍為 30°。

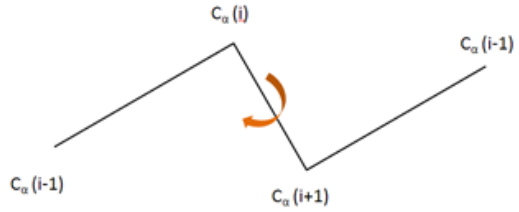


圖 3.3 雙平面夾角 ALPHA

3.4.3 TCO 鍵角

如圖 3.6，cosine 函數 TCO 是將第 i 個胺基酸羰基內的 C=O 向量和第 i-1 個 胺基酸羰基內的 C=O 向量所形成的角度取 cosine，其值為-1~+1。將 FSSP 當中 X-ray 繞射的資訊將 TCO 分成分為 4 個區間，分別為 -1~-0.625、-0.625~0，0~0+0.61，0.61~+1。通常 α-helices 當中連續兩個胺基酸羰基內的 C=O 都指向同方向，因此 TCO 接近+1；而 β-sheets 當中相連的兩個胺基酸羰基內的 C=O 指向相反方向， 因此 TCO 接近 -1。

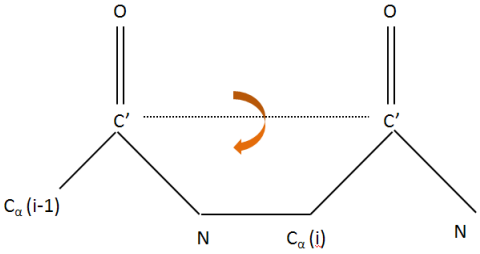


圖 3.4 Cosine 函數 TCO

3.4.4 KAPPA 鍵角

如圖 3.7，KAPPA 角度是第 i-2、i、i+2 個 Ca 形成的角度，這個角度範圍是從 0°到 180°，共分成 23 個區間。

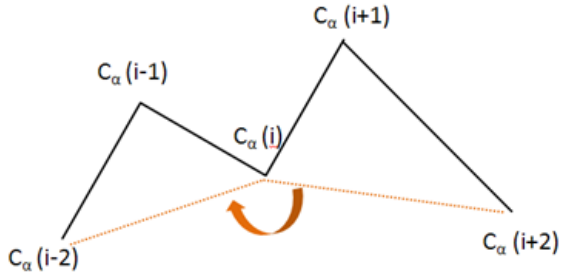


圖 3.5 KAPPA 角度

3.4.5 胺基酸暴露表面積(Solvent Accessible Surface Area; ASA)

此結構特性表示一胺基酸與水分子的可接觸的表面積，通常以與胺基酸表面積的相對比值為主[13]。表 2.1 為 20 種胺基酸的全表面積，單位為 $Å^2$。與前人研究相同，將胺基酸表面積的相對比例分為三類：0%到 9%為埋沒的(buried)，9%到 36%為居中的(intermediate)，36%到 100%為暴露的(exposed)。

表 2.1 胺基酸表面積

胺基酸	A	C	D	E	F	G	H	I	K	L
MaxAcc	106	135	163	194	197	84	184	169	205	164
胺基酸	M	N	P	Q	R	S	T	V	W	Y
MaxAcc	188	157	136	198	248	130	142	142	227	222

3.4.6 蛋白質區塊 Protein Blocks

此結構特性為 Rooman 等人[17]根據連續 5 個胺基酸的 Φ、ϕ 角度計算出的 RMSDA(Root Mean Square Deviations on Angular values)距離矩陣，分成 16 類表示不同三級結構。

3.4.7 STR

如圖 3.8，STR 是將 DSSP 分類的 8 種二級結構當中的 β-strand，根據周圍序列的序列方向再分為 6 類，如圖 3.8，周圍有兩條序列和胺基酸所屬序列平行時， 這兩條序列和胺基酸所屬序列方向相同為 P，相反為 A， 正反為 M；胺基酸所屬序列周圍只有 條平行的序列時，方向相同為 Q，相反為 Z；胺基酸所屬序列周圍沒有平行的序列則為 E。這 6 類加上 DSSP 分類的其他 7 類二級結構，共有 13 類。

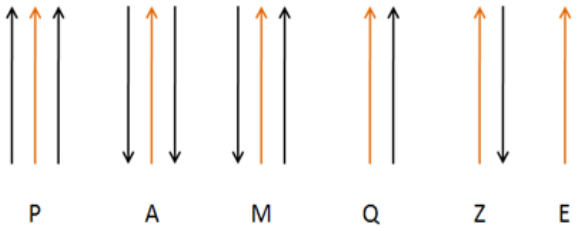


圖 3.6 STR 將β-strand 分成 6 種方向

3.4.8 HMMSTR

HMMSTR 是 Bystroff 等人將 Φ、ψ 角度所畫出的 Ramachandran plot 分成 11 類 (H,G,B,E,d,b,e,L,l,x,c)[18]

3.5 支持向量機

支援向量機是一種監督式學習的方法，將特徵向量映射到一個高維的空間中，並在這個空間當中建立一個區分這些向量並且互相平行的最大超平面。利用支持向量機可以將現有的資料進分類，在建立一個模組後，根據此模組去預測。本研究使用一整合支持向量分類、回歸、分類預測的工具 LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/)。

3.6 評估公式

本研究以三倍交叉驗證(3 fold cross-validation)進行學習與後續的測試評估，在資料集隨機等分三份後，其中兩份用以訓練支持向量機，第三份用於測試預測結果，這樣訓練和測試的資料並不會重複，接箸依序輪流，使得每個資料都能夠測試到一次。
由於參與交互作用的胺基酸數量遠少於沒有參與交互作用的，因此在評估的時候若使用整體準確度來當作評估標準，並無法完整描述準確情形，因此我們使用了以下的評估值來評估整體結果：true positive (TP，正確預測到有交互作用的胺基酸)數量，true negative(TN，正確預測到沒有交互作用的胺基酸)數量，false positive(FP，沒有交互作用的胺基酸被預測成有交互作用)數量，false negative(FN， 有交互作用的胺基酸被預測成沒有交互作用)數量。根據這四個評估值可以計算出以下四種評估公式：

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure} = \frac{2*(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}}$$

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN}$$

專一性(precision)與精確度(specificity)分別評估預測有交互作用和無交互作用的胺基酸當中，實際上也有交互作用的比例；召回度(recall)，又稱為靈敏度(sensitivity)是實際上有交互作用的胺基酸當中，被預測正確的比例；F-measure 是前兩者整合後的公式；準確度(accuracy)是所有胺基酸當中被準確預測是否有交互作用的比例。

4 結果與討論

在單獨使用位置加權矩陣(PSSM)和單一結構特性分數矩陣預測結果如表 4.1，可以發現整體的準確度(Accuracy)都在 64~66%間。但是比較 precision 後發現 PSSM 以 43%優於其他結構特性的分類器結果，而 recall 的部分，比起單一結構特性分數矩陣的 4~7%，PSSM 的 29%表現好很多，因此用來綜合評估 precision 和 recall 的 F-measure 也顯現出 PSSM 是最好的。

表 4.1 各種單 特徵預測結果

Coding scheme	Precision	Recall	F-measure	Accuracy
PSSM	0.440	0.292	0.351	0.643
ALPHA	0.362	0.051	0.089	0.657
ASA	0.360	0.058	0.099	0.655
HMMSTR	0.397	0.046	0.083	0.662
KAPPA	0.427	0.066	0.115	0.662
PB	0.401	0.055	0.096	0.661
SS	0.351	0.055	0.094	0.654
STR	0.370	0.071	0.120	0.653
TCO	0.390	0.047	0.083	0.661

4.1 預測準確度和 20 種胺基酸的傾向的相關性

表 4.2 是觀察 20 種胺基酸在 non-PPI 和 PPI 的傾向，計算公式如下：

$$P^{i}_{PPI} = \frac{\%\ \text{of residue i in PPI}}{\%\ \text{of all residues in PPI}}$$

$$P^{i}_{non-PPI} = \frac{\%\ \text{of residue i in non} - \text{PPI}}{\%\ \text{of all residues in non} - \text{PPI}}$$

我們可以發現比較傾向 non-PPI 的胺基酸，會比較不傾向 PPI，因此這樣的現象可能會影響到預測的準確度。

表 4.2 20 種胺基酸在 non-PPI 和 PPI 的傾向

PPI		Non-PPI	
V (Val)	0.785	W (Trp)	0.789
L (Leu)	0.820	H (His)	0.816
F (Phe)	0.847	Q (Gln)	0.890
G (Gly)	0.864	R (Arg)	0.909
C (Cys)	0.868	N (Asn)	0.927
A (Ala)	0.868	P (Pro)	0.933
K (Lys)	0.872	E (Glu)	0.934
T (Thr)	0.888	M (Met)	1.004
Y (Tyr)	0.937	D (Asp)	1.008
S (Ser)	0.958	I (Ile)	1.010
I (Ile)	0.971	S (Ser)	1.014
D (Asp)	0.978	Y (Tyr)	1.021
M (Met)	0.988	T (Thr)	1.037
E (Glu)	1.199	K (Lys)	1.043
P (Pro)	1.201	C (Cys)	1.044
N (Asn)	1.218	A (Ala)	1.044
R (Arg)	1.273	G (Gly)	1.045
Q (Gln)	1.331	F (Phe)	1.051
H (His)	1.550	L (Leu)	1.060
W (Trp)	1.630	V (Val)	1.072

我們使用九種特徵預測 20 種胺基酸當中最傾向於 non-PPI 的 Valine 和最傾向於 PPI 的 Tryptophan 在 PPI 當中都錯誤的情形。如圖 4.1，顯示比較傾向於 non-PPI 的 Valine 在 PPI 的時候，較容易預測錯誤；而比較傾向於 PPI 的 Tryptophan， 九種特徵都預測錯誤的比例比 Valine 少了 12%。因此可以推論出，胺基酸在 non-PPI 和 PPI 的傾向會影響到預測準確度。

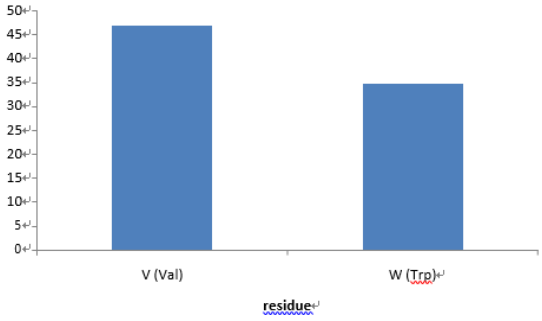


圖 4.1 PPI 當中九種特徵都預測錯誤的百分比 (%)

4.2 結合 PSSM 和八種結構特性分數矩陣預測結果

我們使用支援向量機將 PSSM 和八種結構特性分數矩陣預測蛋白質交互作用胺基酸的機率值結合起來，進行預測結果整合，所獲得的 precision-recall 對應圖如圖 4.1，紅色線條是 Ofran 和 Rost 使用序列資訊預測的結果[2]，橘色線條是我們使用 PSSM 和八種結構特性分數矩陣的預測結果。可以發現結構的特性能夠幫助預測蛋白質交互作用位置。

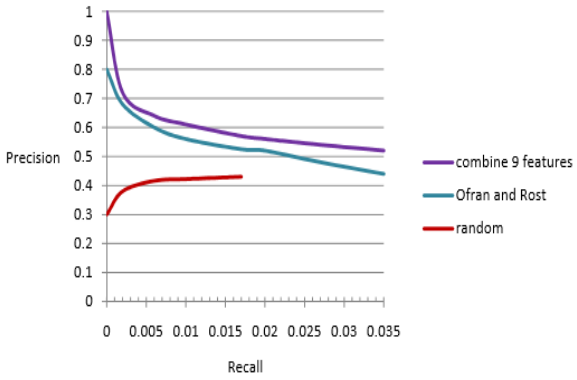


圖 4.2 結合 PSSM 和八種結構特性分數矩陣預測結果和 Rost[2]等人的比較圖

4.3 分析胺基酸和二級結構組成分對於有無交互作用胺基酸之影響

我們分析胺基酸的組成分和真實的三種二級結構(helix、sheet、coil)對於有和沒有交互作用胺基酸的組成分，分別展列如圖 4.3、4.4、4.5，以圖 4.3 當中的 helix-A (Ala)為例，表示參與交互作用並且是二級結構 helix 的 Alaine 佔全部參與交互作用胺基酸的 0.092，沒有參與交互作用並且是 helix 的 Alaine 佔全部沒有參與交互作用胺基酸的 0.133。從中我們可以發現： Alaine 在 non-PPI、helix 的結構所佔的比例是在 PPI、helix 的 1.45 倍，而 Alaine 在 non-PPI、sheet 則是在 PPI、sheet 的 1.80 倍，另外 Alaine 在 non-PPI、coil 結構是在 PPI、coil 的 1.24 倍，因此 Alaine 在 non-PPI、相對應三種二級結構當中都是比 PPI 高。另一方面，以一般偏好 sheet 結構存在的胺基酸 Valine 為例，在 non-PPI 的 sheet 是相對於 PPI 的 1.46 倍，然而將 sheet 換成 helix 和 coil 卻是差不多，因此可以推論出二級結構特性能夠幫助胺基酸組成分預測蛋白質交互作用位置。

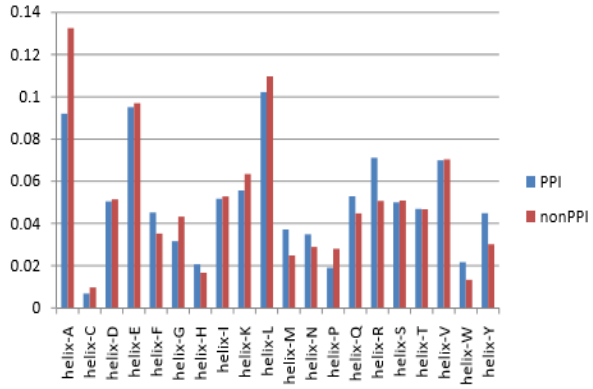


圖 4.3 helix 胺基酸在 PPI 和 non-PPI 的維成分比較

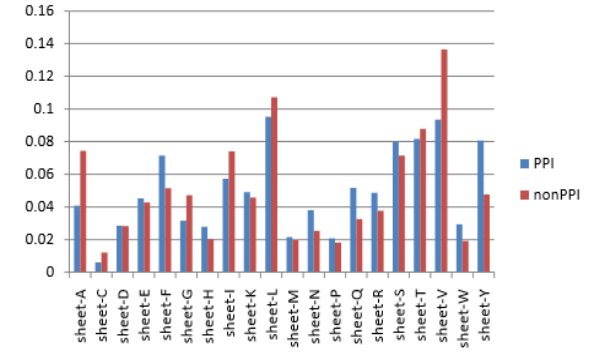


圖 4.4 sheet 胺基酸在 PPI 和 non-PPI 的組成分比較

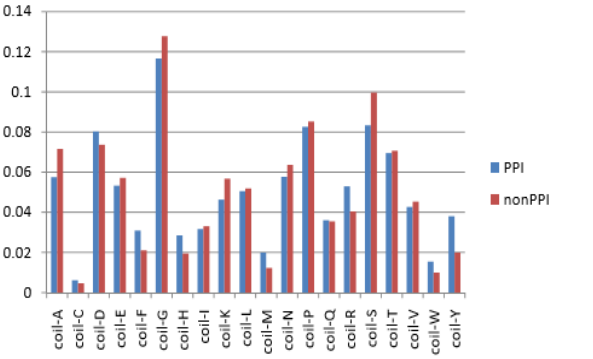


圖 4.5 coil 胺基酸在 PPI 和 non-PPI 的組成份比較

4.4 分析胺基酸和暴露表面積組成份對於有無交互作用胺基酸之影響

我們分析胺基酸的組成分和真實的三種暴露表面積(buried、intermediate、exposed)對於有和沒有交互作用胺基酸的部份，我們比較 Leucine 在三種暴露表面積當中對於蛋白質交互作用位置的組成分： Leucine 在 non-PPI、buried 是在 PPI、buried 的 1.36 倍，intermediate 是 1.22 倍，exposed 是 0.74 倍。可以發現暴露表面積越小的 Leucine 在 non-PPI 的組成分較高，暴露表面積越大的 Leucine 在 PPI 的組成分較高。因此，Leucine 的暴露表面積多寡會影響到是否參與蛋白質交互作用，可推測暴露表面積也能夠幫助胺基酸組成分預測蛋白質交互作用位置。

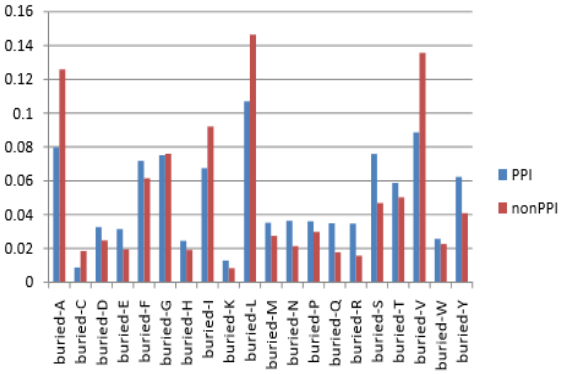


圖 4.6 buried 胺基酸在 PPI 和 non-PPI 的組成分比較

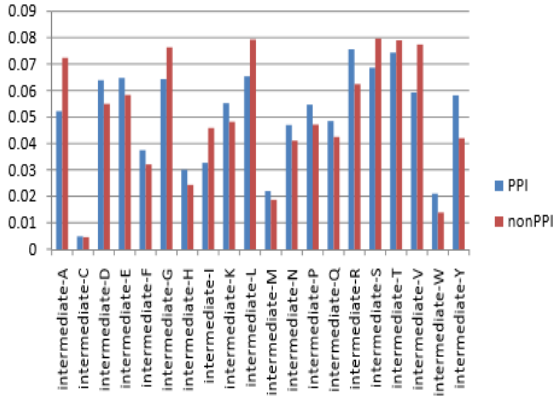


圖 4.7 intermediate 胺基酸在 PPI 和 non-PPI 的組成份比較

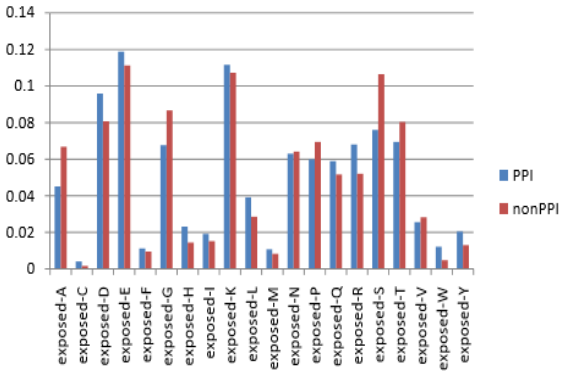


圖 4.8 exposed 胺基酸在 PPI 和 non-PPI 的組成份比較

5 結論與未來研究方向

研究結果顯示，使用結構特性分數矩陣輔助位置加權矩陣預測蛋白質交互作用位置可獲得預測結果的改進。而分析結果顯示真實結構特性和蛋白質交互作用位置有所關連，可以解釋為何結構特性分數矩陣能夠幫助預測。
未來研究方向會結合各種結構特性分數矩陣和位置加權矩陣的方式，例如：將不同特徵預測結果藉由機器學習演算法進行預測結果的整合以提升預測準確度。

6 參考文獻

[1] Xavier Gallet, B. C. A fast method to predict protein interaction sites from sequences. J Mol Biol , pp. 917–926, 2000.

[2] Yanay Ofran, B. R. Predicted protein-protein interaction sites from local sequence information. FEBS Lett 544 , pp. 236-239, 2003.

[3] Res I, M. I. An evolution based classifier for prediction of protein interfaces without using protein structures. BIOINFORMATICS , pp. 2496–2501, 2005.

[4] Yanay Ofran, B. R. ISIS: interaction sites identified from sequence. Bioinforamtics , pp. e13-e16, 2007.

[5] Changhui Yan, D. D. A two-stage classifier for identification of protein-protein interface residues. Bioinformatics , pp. 371-378, 2004.

[6] Bing Wang, P. C.-S.-j.-M. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Letters 580 , pp. 380-384, 2006.

[7] Shan, H.-X. Z. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins , pp. 336-343, 2001.

[8] Aytuna A. Selim, A. G. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics , pp. 2850-2855, 2005.

[9] James R. Bradford, C. J. Insights into protein-protein interfaces using a Bayesian network prediction method. ScienceDirect , pp. 365-386, 2006.

[10] James R. Bradford, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics , pp. 1487-1494, 2005.

[11] Nicholas J. Burgoyne, R. M. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics , pp. 1335-1342, 2006.

[12] Mile Sikic, S. T. Prediction of Protein–Protein Interaction Sites in Sequences and 3D Structures by Random Forests. PLoS Computational Biology , p. e1000278, 2009.

[13] Burkhard Rost, C. S. Conservation and Prediction of Solvent Accessibility in Protein Families. Proteins , pp. 216-226, 1994.

[14] Henrick, K., Thornton, J. M. PQS: a protein quaternary structure file server.
Trends Biochem. Sci. 23, pp. 358-361, 1998.

[15] Chen-Hsiung Chan, H.-K. L.-W.-T.-C.-K. Relationship Between Local Structural Entropy and Protein Thermostability. Proteins , pp. 684-691, 2004.

[16] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers,22, 2577-2637.

[17] Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol. 1990;213:327–336.

[18] Bystroff, C., Thorsson, V. and Baker, D. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins, Journal of molecular biology, 301,173-190.