行政院國家科學委員會專題研究計畫 成果報告

以蛋白質序列物化特性為特徵的蛋白質激kinase-specific磷酸化位置預測方法與分析

研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 99-2221-E-009-137-
執行期間：99年08月01日至100年07月31日
執行單位：國立交通大學生物科技學系（所）

計畫主持人：黃慧玲
共同主持人：范洪春
計畫參與人員：碩士班研究生-兼任助理人員：許馨云
碩士班研究生-兼任助理人員：李銘哲
博士班研究生-兼任助理人員：劉一帆

報告附件：出席國際會議研究心得報告及發表論文

公開資訊：本計畫涉及專利或其他智慧財產權，2年後可公開查詢

中 華 民 國 100年10月31日

中文摘要： 本計畫提出「以蛋白質序列物化特性為特徵的蛋白質激酶磷酸化位置預測方法與分析」。研究進行的方式為改善蛋白質激酶磷酸化位置預測為基礎，並以下列四個方向來進行研究，以達成目標: (1)目前蛋白質激酶磷酸化位置預測工具中，對穿膜蛋白磷酸化位置預測非常不準確，首先設計針對穿膜蛋白磷酸化位置預測、(2)研究穿膜蛋白的特性並正確蒐集穿膜蛋白序列，建立穿膜蛋白序列資料集、(3) 預測蛋白質相對溶劑可接觸性(RSA,代表蛋白質上某一氨基酸和溶劑接觸程度)可以改善磷酸化位置預測、(4)探討蛋白質序列的物化特性來了解蛋白質可接觸性，有助於發展蛋白磷酸化位置預測。本計畫研究進度順利，已達預期目標，並有相關期刊論文及研討會論文發表。

英文摘要： The project proposes prediction and analysis of identifying protein kinase-specific phosphorylation sites based on the features of physicochemical properties of sequences. This study bases on improving the existing methods and achieves the project goal by way of the following fours aspects. 1) The existing prediction tools to identify protein kinase-specific phosphorylation sites have very low accuracy for the transmembrane proteins. Design an identifying transmembrane protein kinase-specific phosphorylation sites system. 2) Study the special properties of transmembrane proteins and create an up to date transmembrane protein dataset. 3) The RSA value plays an important role in developing explicit models for aiding prediction of the phosphorylation sites. 4) Investigate informative physicochemical and biochemical properties of protein sequence to understand the RAS of proteins that is helpful in developing protein kinase-specific phosphorylation sites predicting method. The goal of this project is achievement conference and journal papers.

以蛋白質序列物化特性為特徵的蛋白質激酶磷酸化位置預測方法與分析

Prediction and Analysis of identifying protein kinase-specific phosphorylation sites based on the features of physicochemical properties of sequences

計畫編號：99-2221-E-009-137-

執行期限：99 年08 月01 日至100 年07 月31 日

主持人：黃慧玲 國立交通大學
hlhuang@mail.nctu.edu.tw

摘要

本計畫提出「以蛋白質序列物化特性為特徵的蛋白質激酶磷酸化位置預測方法與分析」。研究進行的方式為改善蛋白質激酶磷酸化位置預測為基礎，並以下列四個方向來進行研究，以達成目標: (1)目前蛋白質激酶磷酸化位置預測工具中，對穿膜蛋白磷酸化位置預測非常不準確，首先設計針對穿膜蛋白磷酸化位置預測、(2)研究穿膜蛋白的特性並正確蒐集穿膜蛋白序列，建立穿膜蛋白序列資料集、(3) 預測蛋白質相對溶劑可接觸性(RSA,代表蛋白質上某一氨基酸和溶劑接觸程度)可以改善磷酸化位置預測、(4)探討蛋白質序列的物化特性來了解蛋白質可接觸性，有助於發展蛋白磷酸化位置預測。本計畫研究進度順利，已達預期目標，並有相關期刊論文及研討會論文發表。

關鍵字：物化特性, 蛋白質激酶磷酸化, 基因演算法, 溶劑可接觸性, 蛋白質序列, 預測方法

Abstract

The project proposes prediction and analysis of identifying protein kinase-specific phosphorylation sites based on the features of physicochemical properties of sequences. This study bases on improving the existing methods and achieves the project goal by way of the following fours aspects. 1) The existing prediction tools to identify protein kinase-specific phosphorylation sites have very low accuracy for the transmembrane proteins. Design an identifying transmembrane protein kinase-specific phosphorylation sites system. 2) Study the special properties of transmembrane proteins and create an up to date transmembrane protein dataset. 3) The RSA value plays an important role in developing explicit models for aiding prediction of the phosphorylation sites. 4) Investigate informative physicochemical and biochemical properties of protein sequence to understand the RAS of proteins that is helpful in developing protein kinase-specific phosphorylation sites predicting method. The goal of this project is achievement conference and journal papers.

Keywords: Physicochemical properties, kinase-specific phosphorylation, genetic algorithms, relative surface area of solvent accessibility, protein sequence, prediction method.

一. 前言

蛋白質的磷酸化是在蛋白質轉譯修飾中很重要機制，在調控基本進行過程例如新陳代謝、訊號傳遞、細胞分化和細胞膜穿透性等扮演重要角色。因此預測蛋白質的磷酸化作用位置是非常重要議題。能知道蛋白質磷酸化位置，就可以測出蛋白質功能。雖然最近有許多的磷酸化位點的預測工具已經發展出來了，但卻還是沒有一個預測工具是針對膜蛋白(membrane protein)而設計，膜蛋白對於一些生理功能是很重要的，所以針對膜蛋白的磷酸化位點預測是一項重要且迫切的課題。

二. 研究目的

膜蛋白同於具有兩性(同時有書水及親水性區段)，因此要調整出適合膜蛋白結晶的環境非常困難。近年來由於蛋白質結晶技術的進步，讓研究者可以順利取得一些膜蛋白的結晶資訊。然而針對膜蛋白的

後轉譯修飾的工具卻還是不足，因此讓研究膜蛋白的實驗需要耗費大量的人力和金錢來進行實驗。

本計畫目標系以「以蛋白質序列物化特性為特徵的蛋白質激酶磷酸化位置預測方法與分析」為主題進行研究。研究進行的方式為改善蛋白質激酶磷酸化位置預測為基礎，並以下列四個方向來進行研究，以達成目標: (1)目前蛋白質激酶磷酸化位置預測工具中，對穿膜蛋白磷酸化位置預測非常不準確，首先設計針對穿膜蛋白磷酸化位置預測、(2)研究穿膜蛋白的特性並正確蒐集穿膜蛋白序列，建立穿膜蛋白序列資料集、(3) 預測蛋白質相對溶劑可接觸性(RSA,代表蛋白質上某一氨基酸和溶劑接觸程度)可以改善磷酸化位置預測、(4)探討蛋白質序列的物化特性來了解蛋白質可接觸性，有助於發展蛋白磷酸化位置預測。

三. 文獻探討

3.1 膜蛋白磷酸化位點分析

在很多的細胞上都會有許多的穿膜蛋白，例如說離子通道或接收器等等。這些蛋白質中，磷酸化扮演了很重要的角色，這些磷酸化的過程不外乎透過自身磷酸化或是一些磷酸化激酶達到目的。在球蛋白中，包埋在內部的通常都是疏水性胺基酸，而暴露在外的都是親水性胺基酸。然而在穿膜蛋白就不是這樣的情況，有時疏水的區位會暴露在外，而親水部分反而包埋在內部。在這兩種迥異的情況下，我們假設在膜蛋白和球蛋白的磷酸化位點會十分的不同。目前有存在一些效能不錯的預測工具和資料庫列表於表1。在當中，每一個資料庫都含有我們感興趣的膜蛋白，所以在研究中需要把這些膜蛋白從中擷取出來。

而現今的預測方法包括兩大項(圖1)，有直接以磷酸化位點為預測對象，或是先以特定的磷酸化激酶為標的，再預測有哪些位點會被磷酸化。

表 1. 之前所提出預測非穿膜蛋白預測器

Tools name	Category
PHOSIDA	For specific a.a
PPSP	For kinase
NetPhosK	For kinase
KinasePhos	For kinase
Musite	For specific a.a
GPS	For kinase
DISPHOS	For specific a.a
NetPhos	For specific a.a

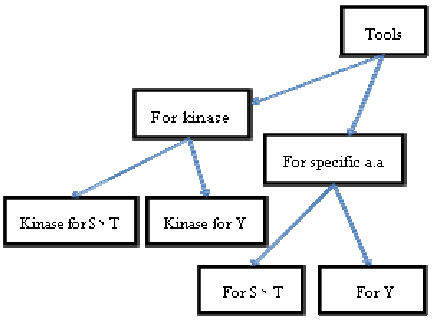


圖1. 目前磷酸化位點的預測方法

3.2 溶劑暴露面積預測

預測蛋白質的折疊無論在生物資訊或者實驗領域中都是很重要的問題。ASA 是了解蛋白質摺疊的一項重要工具。透過了解蛋白質上每一個殘基的 ASA，可以對蛋白質的構型有約略的了解，並且猜測蛋白質中殘基跟殘基間的交互作用或者是生理功能，催化位點與催化機制甚至是蛋白質-蛋白質間的交互作用。例如說 Hikijata et.al. 在預測蛋白質 3D 結構時，除了使用序列比對的結果之外，同時也使用到溶劑暴露面積輔助預測結果。在預測蛋白質網絡領域中，Martin et. al. 希望預測會進行交互作用的殘基(residues)，會同時參考蛋白質的二級結構、實驗結果跟二級結構資訊的參數，由此可知，在研究蛋白質的功能、結構等等溶劑暴露面積都扮演很重要的角色。

目前預測 ASA 的工具可以大多為 two states 的包埋暴露預測，這一類的預測器是

以預測器的效能而決定出 threshold，在某些情況下無法滿足想了解特定殘基的特定功能，因此發展出預測 RSA 實際值的預測方法，透過對 RSA 直接預測就能解決上述問題。

除此之外，雖然預測工具很多，但卻沒有一個研究有將不同胺基酸個別探討其物化特性在蛋白質摺疊上所扮演的角色。本篇研究也會針對所挑選來的 featrure 做探討，了解各個殘基的包埋或暴露的物化特性，進而發現一些原本在預測殘跡暴露與否的問題中沒有考慮到的因素。這次我們的方法不但提供了一個可以預測相對暴露面積的方法，同時我們的方法也提供了一些資訊，來了解蛋白質折疊的問題。

目前為止,預測實際相對溶劑暴露面積的研究很多如表一所示，其中的方法包括使用 NN，SVR，multi layer regression，two-stage SVR。

蛋白質的磷酸化過程中，除了將磷酸運輸至目標蛋白進行化學變化外，同時也包含有蛋白質間作用力。為了要了解蛋白質和磷酸根之間的作用力，研究需要取得一些蛋白質和磷酸根作用的資料幫助了解。然而在蛋白質中沒有一個胺基酸是帶有磷原子甚至是磷酸根為官能基的胺基酸。DNA 在骨幹(backbone)的部分是磷酸以 phospho-diesterbone 連結磷酸所構成,含有大量的磷酸根，同時在目前的資料庫裡也有足夠的資料。Kumar et al.就曾經蒐集 DNA 結合蛋白作為研究對象，使用支援向量機做出模型。研究裡我們會使用它們所使用的 dataset 作為研究對象。

四.研究方法

4.1膜蛋白磷酸化位點整理及預測

4.1.1dataset 的取得

資料從 uniprot 中擷取，為了要得到更精確的目標，因此我們只採用有真實實驗數

3.3 磷酸結合相關蛋白分析

表 2. 溶劑暴露面積相關研究和其效能比較表

Work	Regression tool	Description of features	MAE (%)	CC
Ahmad et al.,2003	NN	Amino acid composition	18.8	0.48
Yuan and Huang,2004	SVR	Amino acid composition	18.5	0.52
Adamczak et al.,2004	NN	PSSM	15.3	--
Wang et al.,2005	MLR	Amino acid composition, PSSM and sequence length	16.2	0.64
Garg et al.,2005	NN	PSSM and secondary structure information	15.9	0.65
Nguyen and Rajapakse,2006	Two-stage SVR	PSSM	15.7	0.66
Chang et al.,2008	Two-stage SVR	enhance PSSM and sequence length	14.8	0.68
ours	SVR	PSSM, Aaindex and sequence length	14.11	0.69

據驗證的資料作為磷酸化位點的參考點，如果該資料中紀錄是”potential”、”similarity”或是”probable”的資料將不會被採用。接著，我們再從PhosphoSite plus資料庫中取得各種穿膜蛋白。其中包括anion exchange、calcium transport、chloride channel、copper transport、hydrogen ion transport、ionic channel、iron transport、ligand-gated ion channel、porin potassium transport、sodium transport、viral ionic channel、voltage-gated channel、zinc transport、cobalt transport。最後得到222條蛋白質和其上總共約三萬個磷酸化位點。如表3所示。從中挑出100條作為test，剩下的122條作為training。

表 3.222 個從 phosphosite plus 和 uniprot 中所挑選出來的磷酸化位點

	S	T	Y	Total
100 proteins_positive	169	56	101	326
100 proteins_negative	6934	5199	2862	14995
122 proteins_positive	198	68	115	381
122 proteins_negative	8466	6244	3495	18205

4.1.2 磷酸化位點預測方法和效能評估

模型的建構分成兩的大部分做預測，其流程如圖1所示。首先會先分析統計該資料中顯而易見的特性，同時找尋文獻中可以使用的知識作為feature之一。而另外一部份則是使用IBCGA從531個物化特性中挑選出最具代表性的結果。最後做出恰當的模型用來進行預測。模型分為兩個預測，其一為先以磷酸激酶做分類後進行預測，首先探討以PKA作為磷酸化激酶的目標，其結果如表4所示。

結果顯示我們如果先用磷酸激酶分類後進行預測，最後最高的準確率大約是八成，而在專一性更可以達到八成五。

最後如果以磷酸化位點作為預測目標做預測，當預測serine和threonine時，結果如表5所示。結果顯示直接使用位點的結果會比以磷酸激酶分類後做出的預測好，準確率約九成，專一性更達到九成六。

表 4. 以 PKA 為磷酸化激酶所預測的磷酸化位點

	Spec	Sen	OAcc	MAcc	MCC
GPS	0.859	0.423	0.848	0.607	0.121
PPSP	0.667	0.299	0.645	0.232	-0.017
KinasePhos	0.006	0.959	0.024	0.503	-0.063
NetPhosK	0.155	0.280	0.157	0.226	-0.183

Spec = specificity、 Sen = sensitivity、OAcc = Overall accuracy 、MAcc = Mean accuracy 、MCC = Matthews correlation coefficient.

表 5. 以磷酸化位點直接進行預測之結果

	Spec	Sen	OAcc	MAcc	MCC
DISPHOS	0.958	0.141	0.941	0.455	0.069
PHOSIDIA	0.318	0.735	0.326	0.548	0.016
MuSite	0.837	0.165	0.787	0.170	0.001
NetPhos	0.016	0.938	0.034	0.548	-0.049

4.2 蛋白質與磷酸根交互作用分析

為了想了解蛋白質和磷酸根之間的作用，所以最直接的辦法就是從帶有磷酸根的分子和蛋白質間如何作用做分析。在生物體內帶有最多林酸根的分子為DNA，同時在蛋白質和DNA的共同結晶和文獻的量都足夠做為分析之用。故挑選DNA結合蛋白作為分析的目標。我們dataset使用之前Kumar et al所使用的dataset。首先我們先從531個物化特性中挑選出重要的23個，然後對這23個物化特性做fuzzy rule的分析。其流程如圖2所示。

最後fuzzy rule的判讀結果列在表6，最重要的三項因子分別為電性、凡德瓦力和溶劑暴露面積。

這項結果說明了，如果可以在與磷酸作用的位點，特別加入溶劑暴露面積，對於判斷該位點是否作用有很大的助益。

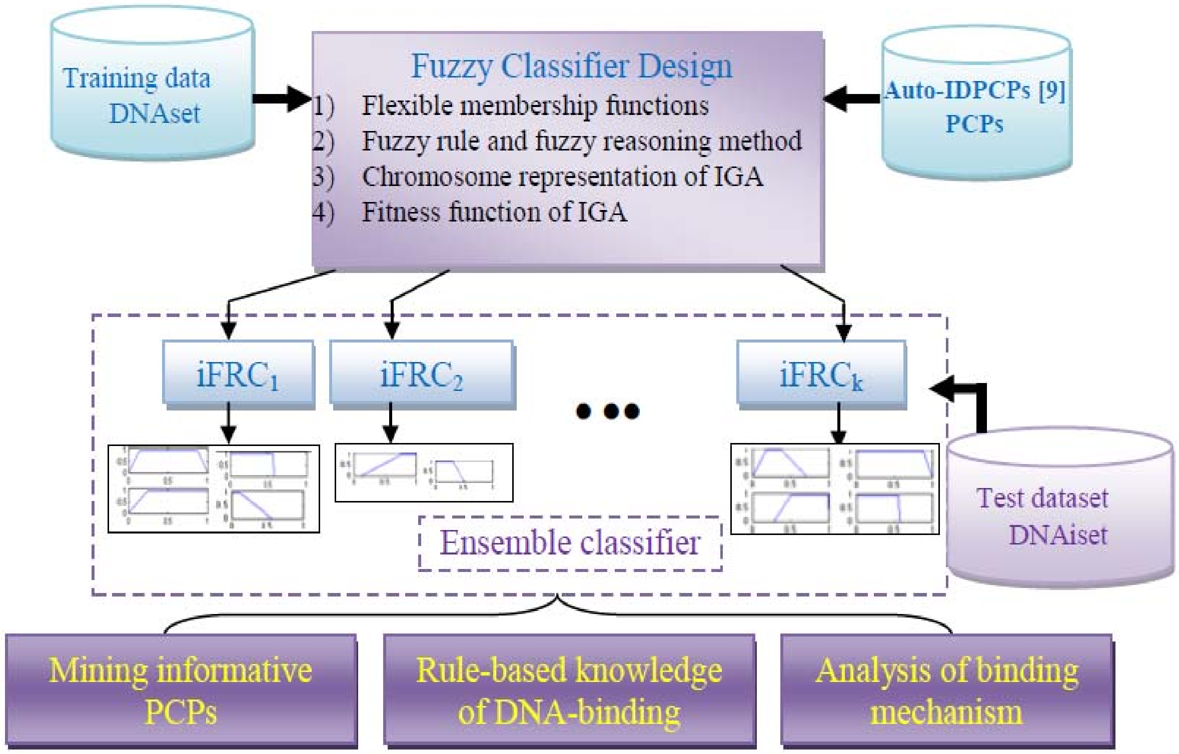
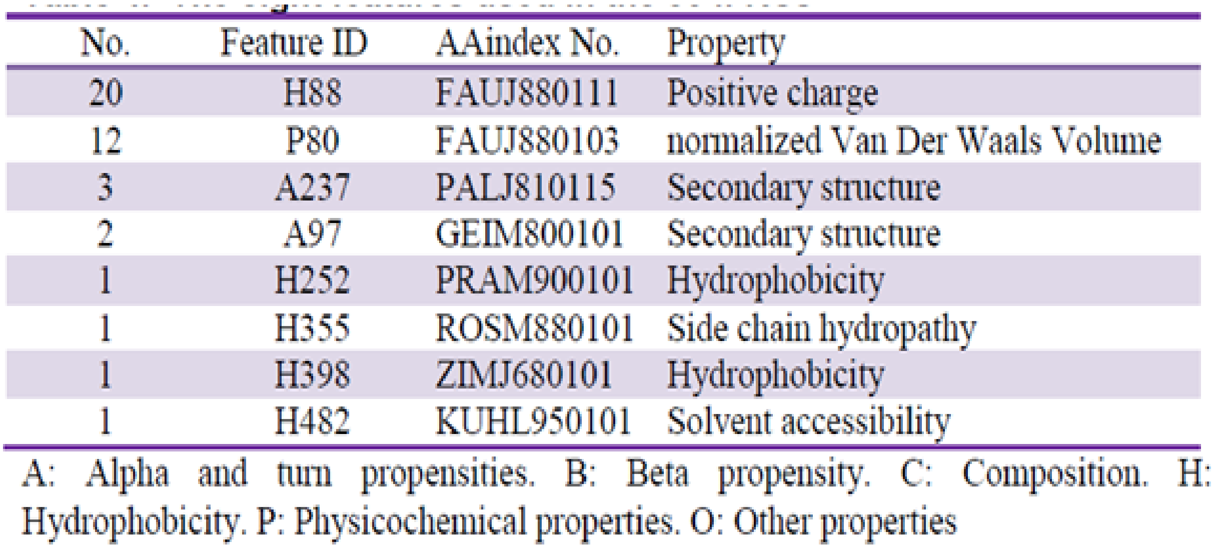

Training data
DNAset
Fuzzy Classifier Design
1) Flexible membership functions
2) Fuzzy rule and fuzzy reasoning method
3) Chromosome representation of IGA
4) Fitness function of IGA
Auto-IDPCPs [9]
PCPs
iFRC$_1$
iFRC$_2$
iFRC$_k$
Test dataset
DNAiset
Ensemble classifier
Mining informative PCPs
Rule-based knowledge of DNA-binding
Analysis of binding mechanism

圖2. FRKAS流程圖。輸入的物化特性可以經過FRKAS得到一個了解該特性值對於分類影響的規則。

表6. FRKAS挑選出來的物化特性。結果說明在磷酸化位點的結合部份，其帶有電荷的物化特性十分鐘要，其中也需要溶劑鋪暴露面積做為判斷的物化特性之一，以提高分類器的效能。



No.	Feature ID	AAindex No.	Property
20	H88	FAUJ880111	Positive charge
12	P80	FAUJ880103	normalized Van Der Waals Volume
3	A237	PALJ810115	Secondary structure
2	A97	GEIM800101	Secondary structure
1	H252	PRAM900101	Hydrophobicity
1	H355	ROSM880101	Side chain hydropathy
1	H398	ZIMJ680101	Hydrophobicity
1	H482	KUHL950101	Solvent accessibility

A: Alpha and turn propensities. B: Beta propensity. C: Composition. H: Hydrophobicity. P: Physicochemical properties. O: Other properties

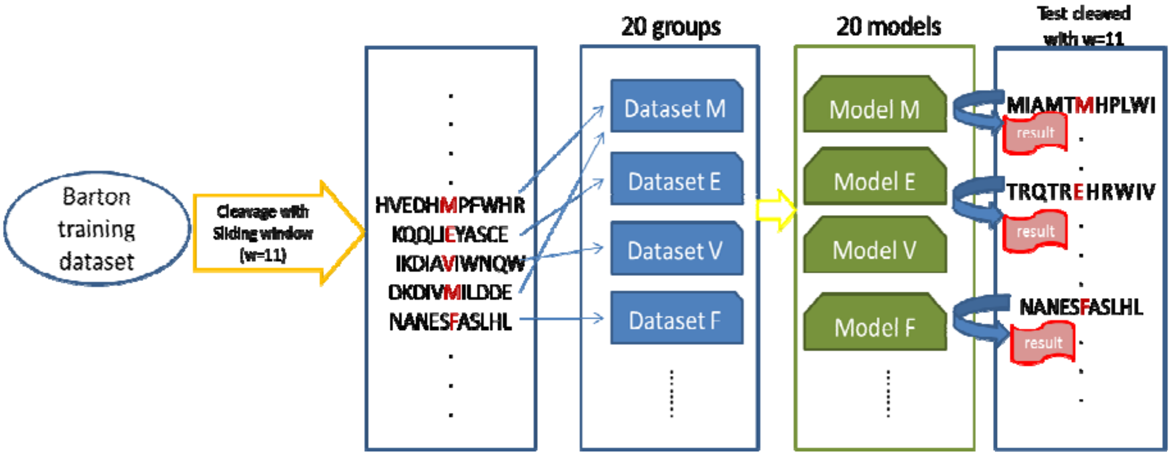


圖3. 模型建構方法

4.3 溶劑暴露面積預測器

4.3.1製作流程

先前的研究很多都使用Barton502作為預測和建構模型的資料。該資料使用的人多，因此適合用來比較預測器的效能。

同時因為資料量也夠大，因此使用該資料所得到的結果可信度也高。所有的蛋白質會按照圖3所列出來的方式，照上面的胺基酸分別建構模型和預測模型。實驗流程如圖4所示。

其中的三分之二作為traininig，剩下的三分之一作為test。各種不同的預測器效能的比較都是以test所做出來的結果。經過IBCGA挑選features後，所選出來的如表7所列之features。

4.3.2 溶劑暴露面積預測器效能評比

以MAE的平均值來看，如果將序列長度加上PSSM或是序列長度加上AAindex，表現都沒有比PSSM加上aaindex和序列長度的資訊(表8)，顯示如果PSSM搭配AAindex和序列長度在預測蛋白質溶劑暴露面積上有非常好的效果。相較於Chang et. al.和Nguyen et.al的方法，我們和他們不一樣的地方在於他們都只是使用PSSM，即使是Chang et.al.的enhance PSSM，其實也是從PSSM中擷取資訊，嚴格講起來也都是單一feature進行預測；而Wang et.al.是使用多組feature一起進行預測，但使用的方法是Multi-Learning regression的方法，相對於後來發展的方法都是使用SVR的情況來看，也許預測溶劑暴露面積的問題，使用SVR會比MLR更合適。最後比較相關係數和總MAE(這邊的MAE是把所有的胺基酸實際值跟預測值一起算)，如表8所列，使用PSSM搭配aaindex和長度資訊也確實得到不錯的結果。

4.3.3 溶劑暴露面積預測器物化特性分析

按照先前的研究將531個aaindex分成20個clusters，觀察其分布狀況(圖5)。

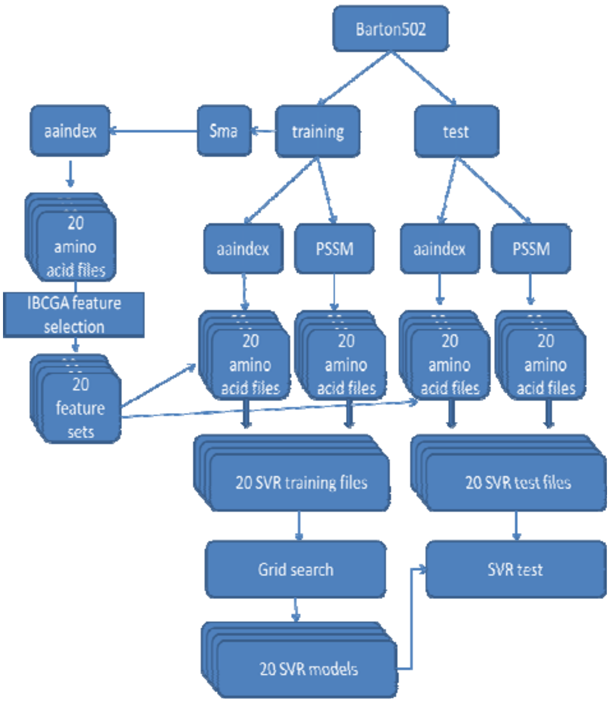


圖4. 溶劑暴露面積模型製作流程

表 7. 用來建立 model 的 feature set

amino acid	feature number	Best feature set
A	31	12,18,39,89,102,114,141,164,169,199,206,209,210,251,266,267,294,305,309,333,338,352,356,364,369,373,403,415,447,481,523,
R	21	10,41,93,105,113,152,194,259,287,317,327,334,352,358,381,387,398,404,408,409,451,
N	34	11,31,41,60,99,100,102,124,126,172,175,176,231,253,260,262,264,265,274,310,321,322,323,336,351,385,410,412,414,419,483,512,521,523,
D	25	23,94,95,113,124,146,178,179,196,210,221,257,274,275,285,295,300,318,340,352,404,442,481,490,531,
C	30	12,18,29,41,54,61,95,102,106,151,169,221,242,245,249,278,296,306,379,391,405,420,447,473,477,491,492,497,503,522,
Q	33	78,86,101,102,106,161,166,178,198,201,219,238,294,295,300,315,328,342,343,352,362,376,381,389,398,403,419,423,434,450,452,461,481,
E	28	13,44,66,71,79,108,199,200,209,211,221,247,289,302,318,332,355,371,381,387,399,404,410,414,446,501,521,522,
G	32	21,31,42,66,82,110,114,165,196,200,233,262,267,268,295,304,306,320,333,364,371,372,416,420,423,432,442,454,455,461,469,530,
H	28	3,6,18,30,69,75,83,108,113,126,200,230,252,286,300,373,374,376,402,419,420,423,447,462,467,497,515,530,
I	34	14,26,28,29,30,45,74,79,88,92,107,165,187,218,235,237,252,255,266,267,268,289,290,322,323,334,341,386,415,418,446,452,453,526,
L	26	3,10,61,69,113,168,176,194,200,236,239,305,315,318,322,323,339,349,356,370,390,419,480,508,512,518,
K	23	27,32,64,114,140,163,179,186,219,223,235,258,267,273,284,322,330,414,443,480,488,495,515,
M	29	71,90,144,147,149,151,185,221,222,272,279,284,285,294,319,348,349,352,357,386,410,422,477,481,483,484,488,516,529,
F	30	10,39,46,54,93,149,169,174,193,215,223,238,275,277,327,331,336,369,371,384,419,432,443,446,450,488,503,508,527,528,
P	21	22,40,53,61,63,80,83,110,209,228,259,260,299,300,351,361,404,418,424,443,521,
S	26	12,25,71,96,103,114,168,175,189,200,221,345,351,376,389,405,409,440,450,452,474,477,514,516,522,531,
T	27	10,11,57,58,60,98,105,106,115,159,194,224,229,231,296,309,318,339,352,362,380,383,392,400,409,422,433,
W	26	3,4,37,80,114,124,178,186,190,206,228,229,274,276,277,285,304,308,333,336,352,381,387,483,510,525,
Y	31	18,28,40,53,140,173,206,209,211,218,231,238,324,328,333,352,364,387,391,414,416,422,436,477,481,497,506,510,516,522,531,
V	12	26,117,153,235,303,316,405,447,448,450,452,509,

表8. 溶劑暴露面積預測器和相關研究之效能比較

			MAE						
no.	amino acid	pssm	+	+	+	-	Chang(2008)	Nguyen(2006	Wang(2005)
		aaindex	+	-	-	+			
		sequence length	+	-	+	+			
1	A		12.2199	13.4740	13.4723	18.8132	13.3	14.4	15.6
2	R		16.8119	16.8845	17.318	24.3333	17	17	17.1
3	N		18.4971	20.3228	20.446	25.0654	19.6	20.2	21
4	D		18.0767	20.2810	20.2245	24.5231	19.2	19.5	20.8
5	C		8.87168	10.0537	10.0537	9.61839	8.9	9.9	14.2
6	Q		16.2366	17.7884	22.1026	21.6074	17.2	17.6	18
7	E		15.9341	18.1813	18.2301	22.2619	17.8	18.3	19.3
8	G		18.0293	19.7634	19.7632	24.0462	19.5	19.6	21.1
9	H		15.8715	15.8185	20.5737	19.8794	15.1	15.4	15.7
10	I		8.09382	8.9793	8.9742	10.2656	8.7	9.7	10.6
11	L		9.79063	12.5323	10.2074	11.8409	9.8	10.8	11.6
12	K		15.7747	15.7908	15.8488	18.4192	15.8	16.4	16.3
13	M		11.3245	11.3839	11.3408	13.7533	11.3	12.1	12.9
14	F		10.0539	9.9997	9.99919	11.0815	10.2	11.2	11.9
15	P		16.6900	17.9133	17.9515	21.2484	17.4	17.7	18.2
16	S		16.0847	18.2228	18.8323	23.5514	18.3	18.8	19.8
17	T		15.8687	16.5704	16.5863	21.049	16	16.7	17.1
18	W		12.1712	12.4307	12.3633	13.4755	11.8	12.4	13.2
19	Y		11.5132	12.1179	12.1725	13.8647	13	12.9	13.3
20	V		9.89251	10.187	10.1642	12.1627	9.6	10.7	11.2
		average	13.742981	15.7183133	15.3312295	18.0430245	14.475	15.065	15.945

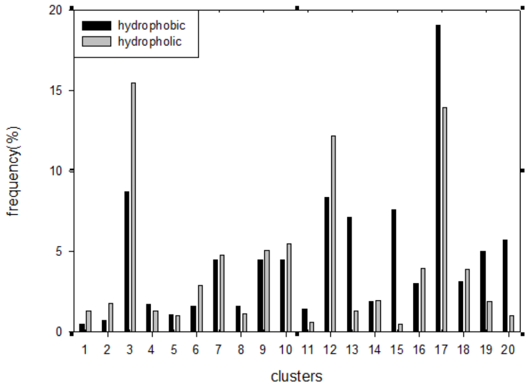
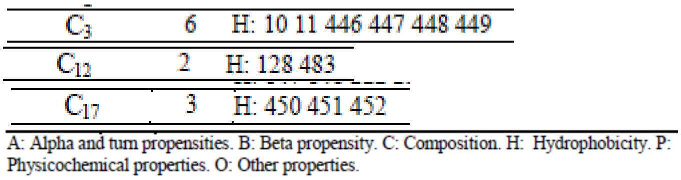


圖 5.cluster 分析

分析過後，可以發現不管是 hydrophilic 或 hydrophobic 的胺基酸，分數最高的前三名都是 cluster 3,12,17。我們分別查出餓三個 cluster 裡的成員列在圖 6，這三個 cluster 並不是 aaindex 分類中元素最多。有趣的是，這三個 cluster 當中的元素全部都是屬於 hydrophobic index。



C_3	6	H: 10 11 446 447 448 449
C_{12}	2	H: 128 483
C_{17}	3	H: 450 451 452

A: Alpha and turn propensities. B: Beta propensity. C: Composition. H: Hydrophobicity. P: Physicochemical properties. O: Other properties.

圖 6. 在第 3、12、17 群中，成員都是疏水作用力相關的物化特性

這樣證據顯示不管是在親水性和疏水性的胺基酸，hydrophobic properties 都很重要。也應證了第一張圖所提出的不管親水性胺基酸跟疏水性胺基酸，對於判斷其親疏水性是非常重要的。

四. 論文發表

這一年計劃皆有完成預定的目標。也如期投出期刊論文並且接受刊出論文。感謝國科會給予資源才能順利完成並且豐碩收穫。執行論文這一年期間，發表相關的論文如下

1 **H.-L. Huang**, F.-L. Chang, S.-J. Ho, L.-S. Shu, W.-L. Huang, and S.-Y. Ho*, "FRKAS: Knowledge Acquisition Using a Fuzzy Rule Base Approach to Insight of DNA-Binding Domains/Proteins," accepted by Protein and Peptide Letters, 2011. (SCI)

2 **H.-L. Huang**, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K-T. Hsu, W.-L. Huang, S.-J. Ho, and S.-Y. Ho*, "Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties", *BMC Bioinformatics*, 12(Suppl 1):S47, 2011.(SCI)

雖然我們致力於國科會的研究題目，但另外在其他的非直接相關的論文和研討會成果也同樣豐碩，期刊論文共一篇：

1. S.-Y. Ho,C.-Y. Chao, **H.-L. Huang**, T.-W. Chiu, P.Charoenkwan, and E. Hwang*, "NeurphologyJ: an automatic neuronalmorphology quantification method and its application in pharmacological discovery,"*BMC Bioinformatics*, 12:230, 2011. (SCI)

另外有研討會論文共八篇

1. **H.-L.Huang**, T.-F.Kao',P.Charoenkwan', W.-L. Huang, S.-J. Ho and S.-Y. Ho*, 2012, "Estimating solubility scores of dipeptides and residues for predicting proteins solubility,"The Tenth Asia Pacific Bioinformatics Conference, Melbourne, Australia, 17-19 January 2012.

2. C.-T. Tsai, W.-L. Huang, C.Liaw, C.-W. Tung, **H.-L. Huang** and S.-Y. Ho*, 2012, "Virulence-iGO: Predicting virulence factors in pathogenic bacteria using informative Gene Ontology

terms,”The Tenth Asia Pacific Bioinformatics Conference, Melbourne, Australia, 17-19 January 2012.

3. H.-C. Lee, S.-J. Ho, L.-S.Shu, F.-L. Chang, S.-Y. Ho and **H.-L. Huang***,2012, “Optimization method of predicting enzyme mutant activity from sequences by identifying a set of informative physicochemical properties,”The Tenth Asia Pacific Bioinformatics Conference, Melbourne, Australia, 17-19 January 2012.

4. **H.-L. Huang**,Y.-H. Lin, W.-L. Huang and S.-Y. Ho*, 2011, “Intelligent triple-objective genetic algorithm for selecting informative Tag SNPs,” The 22nd International Conference on Genome Informatics, Korea, Dec. 5-7, 2011.

5. **H.-L. Huang**, S.-B. C., Y.-H.Chen, and S.-Y. Ho*, 2011, “Optimization approach to estimation of kinetic parameters for modelling metabolic pathways of muscle glycogenolysis,”The 22nd International Conference on Genome Informatics, Korea, Dec. 5-7, 2011.

6. L.-S. Shu, **H.-L. Huang**, S.-J. Ho, and S.-Y. Ho*, 2011, “Establishing large-scalegene regulatorynetworks using a gene-knowledge-embedded evolutionary computation method,” IEEE International Conference on Computer Science and Automation Engineering, June 10-12, 2011, Shanghai, China.

7. **H.-L. Huang**, F.-L. Chang, S.-J. Ho, L.-S. Shu, and S.-Y. Ho*, 2011, “Interpretable knowledge acquisition for predicting DNA-binding domains using an evolutionary fuzzy classifier method,” IEEE International Conference on Computer Science and Automation Engineering, June 10-12, 2011, Shanghai, China.

8. **H.-L. Huang**, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang, S.-J. Ho, and S.-Y. Ho*, 2011, “Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties, APBC 2011, Korea, Jan. 11-14.

參考資料

[1] Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK, "KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns," Nucleic Acids Res,vol 35,(Web Server issue), pp.W588-594, 2007.

[2] Blom N, Gammeltoft S, Brunak S, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," J Mol Biol,vol 294,(5), pp.1351-1362, 1999.

[3] Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK, "The importance of intrinsic disorder for protein phosphorylation," Nucleic Acids Res,vol 32,(3), pp.1037-1049, 2004.

[4] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K et al, "Systematic discovery of in vivo phosphorylation networks," Cell,vol 129,(7),

[5] Dor O, Zhou Y, "Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties," Proteins,vol 68,(1), pp.76-81, 2007.

[6] Chen K, Kurgan M, Kurgan L, "Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values," J Biomedical Science and Engineering,vol 1,pp.9, 2008.

[7] H.-L. Huang, F.-L. Chang, S.-J. Ho, L.-S. Shu, W.-L. Huang, and S.-Y. Ho*, "FRKAS: Knowledge Acquisition Using a Fuzzy Rule Base Approach to Insight of DNA-Binding Domains/Proteins," accepted by Protein and Peptide Letters, 2011. (SCI)

[8] H.-L. Huang, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K-T. Hsu, W.-L. Huang, S.-J. Ho, and S.-Y. Ho*, "Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties", *BMC Bioinformatics*, 12(Suppl 1):S47, 2011.(SCI)

國科會補助教師出席國際會議結案心得報告

報告人姓名	黃慧玲	所屬學校 學系(所)	交通大學 生物科技學系
會議期間 及地點	2011/01/14 至 2011/01/14 南韓	補助項目 及金額	■ 機票費 □ 註冊費 ■ 生活費
會議名稱	（中文）2011 第九屆亞太生物資訊會議 （英文）2011 The Ninth Asia-Pacific Bioinformatics Conference (APBC2011)		
發表論 文題目	用一套系統化方法找一組相關的物化生化學特性集來預測與分析DNA-binding domains Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties		

報告內容：(1、參加會議經過；2、與會心得3、建議4、攜回資料)

吾人發表的論文是在01/12 日下午2 點50分的SONGDO CONVENSIA會議廳1F

1. 參加會議經過

本次會議APBC2011 為第九屆亞太生物資訊會議此次主辦單位是KSBSB (Korean Society for Bioinformatics and Systems Biology)協辦單位是KRIBB (Korean Research Institute of Bioscience and Biotechonology)、KOBIC (Korea Bioinformation Center)、 和 Chungbuk BIT Research-Oriented University Consortium。贊助機構有IBC Journal(Interdisciplinary Bio Central)以及BioMed Central。創辦人Professor Phoebe Chen (Professor and Chair & Head of Department, Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia)。亞太生物資訊會議每年都在不同國家舉辦，2011年在韓國舉辦。今年大會共有5個Room，包含口頭報告區、演講區、與海報展示走廊。本次共收55篇口頭報告論文，很榮幸我們的論文被接受口頭報告。至於海報共接受127篇。其中52個program committee來自各個國家、4個chair分為6 group負責大會工程。6個Steering committee統籌大會運作。在本次會議共有8個tutorials及六位keynote speaker及17個Session，及8場演講。會議最後共有6場IBC Journal的教授演講。

我們由首爾機場搭飯店公車抵達飯店，首爾溫度大約-5˚C~-17˚C，看著飄雪的景象，優美的雪景令疲勞的身體舒解些許。首爾的消費額比台北多些，這點個人頗有點感慨。會場的地點是在仁川所以搭地鐵大約一小時多便可以到達。仁川堆積的雪比首爾高，所以沿路的雪景比首爾更有一番風味。到會場報到後因為我們的報告是第三天，因此我們有許多時間可以聽演講。比較令人深刻映像是Keynote1 speaker Steven Jones，與chair Phoebe Chen談論生物資訊與演癌症基因體，癌症基因病變與突變，如何由運用生物資訊能力，以及資訊發現探討。覺得對於自己投入生物資訊研究，有獲得支持與可持續發展的力量。演講後有一場知性的音樂演奏做為中場結束。演講分為Biological Data Evaluation與Translational Bioinformatics二大主軸，其中交通大學生科系系主任黃憲達教授任第一場首位演講。我們在會場亦與系主任討論問題，同時也分享他們演講前必須先開會討論，再進行演講流程相當嚴謹。

第三天我們報告時間是下午2:50，這次大會將每場演講與報告都錄影起來，由網站亦可觀賞精彩演講。在我們這場報告以辨認圖樣識別為主，我們主要是提出由蛋白質序列轉物化特存在問題，對此提出解決方法並應用DNA-binding domains序列資料去分析及發現來探討問題。我對這session的一個主題發現Phosphorylation Motifs感到有興趣，因為我目前正在研究的是Phosphorylation sites研究。會議結束後今天大會的晚餐相當特別的是，搭遊覽車到仁川海港邊搭遊輪出海，並且在郵輪上享用

美味海鮮大餐，欣賞聲樂家演唱，夜色在美食美聲下越來越美，帶著滿滿的陶醉結束這一天會議行程。以下照片由右到左分別為交大生科副院長何信瑩教授、系主任黃憲達教授、APBC創辦人Professor Phoebe Chen、我、台大研究員、以及台大博士班學生，在一樓會場留影。

2、與會心得

感謝國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。此次交大生科副院長何信瑩教授、系主任黃憲達教授深為APBC創辦人Professor Phoebe Chen的愛戴，除暢談她的每年經驗以及她個人行程的滿檔，亦對此次舉辦的不充足之處作分享。並希望2015年由交大舉辦亞太生物資訊會議。

個人覺得生物資訊這領域，由此次舉辦國韓國，這國家對生物資訊投入組織相當龐大，也可見他們對這領域的企圖心與團結。反觀台灣生物資訊投入與組織結構發展還需更努力。而我們與系主任、副院長的討論彷彿將系院擴大到國際空間進行情境探討與未來計畫，相當難得的收穫。

3、建議

近年來國科會、教育部和學校積極鼓勵年輕研究人員，除鼓勵教師參與會議外，特別是博士班學生，參與大型國際會議，及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加生物資訊國際會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此交換研究心得，更可找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。目前研究生已有多管道獲(部份)補助出席國際會議，建議繼續擴大進行。而國際化的學術交流是往後的趨勢，也能有所激勵國人學界能力與國際觀。

4、攜回資料

1. 期刊一本
2. Tutorial 一本
3. 記事本一本。

日期:	Sat, 16 Oct 2010 12:15:55 +0100 [2010/10/16 19時15分55秒 CST]
寄件人:	APBC2011 <apbc2011@easychair.org>
收	Hui-Ling Huang <hlhuang@mail.nctu.edu.tw>

件人:

主旨: APBC2011 notification for paper 120 - Accept

Dear Hui-Ling Huang

Paper: 120
Title: Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties

We are very pleased to inform you that your submission has been accepted for an oral presentation at APBC 2011. The reviews are appended below. Further, the revised version of your submission, taking the comments of the reviewers into account, is invited to be published in BMC Bioinformatics. Congratulations!

Please follow these instructions carefully to avoid any possible delays or problems:

1. As per the tradition of APBC, at least one author must register by 25 Oct 2010 for the paper to be accepted. Registiation information can be found at
http://sysbio.kribb.re.kr/APBC2011/regist.php
Please note that it is APBC's policy that at least one of the authors must be able to present the paper for it appear in the proceedings.

2. Please take all the comments of the reviewers into account while preparing the revised version. If unsatisfactory, the submission may be returned to you, delaying the publication of your paper. (If substantial, include a bulleted list of changes in the body of the email.)

3. The final version in the BMC Bioinformatics format is due from you by 25 Oct 2010. This is a STRICT DEADLINE to enable timely publication. The final version should be directly emailed to s.mann@latrobe.edu.au with Subject: "APBC2011 Final Version - Paper ID". Format details are in the submission guidelines of the conference webpage (http://sysbio.kribb.re.kr/APBC2011/CFP.php).

4. To enable prompt handling of your paper publication in BMC Bioinformatics, please also include the following in the body your email:
a) Paper id (the submission id)
b) Name, address and phone number of CONTACT author
c) VAT number, if the address is in EU

5. Finally, charges - BMC Bioinformatics
The online publication fee payable to BioMed Central shall be £540 (Five Hundred and Forty Pounds Sterling) per Article. Such fees are payable regardless of any BioMed Central membership arrangements and it shall be the responsibility of the author to pay this to BioMed Central.
BMC Bioinformatics will contact you for this charge very soon.

We look forward to seeing you in Incheon, Korea, 11-14 January 2011!

Best Regards,
Phoebe Chen, La Trobe University, Melbourne, Australia
Kwang-Hyun Cho, KAIST, Korea
Program Co-Chairs of APBC2011
The 9th Asia Pacific Bioinformatics Conference
http://sysbio.kribb.re.kr/APBC2011/index.php

PREDICTING AND ANALYZING DNA-BINDING DOMAINS USING A SYSTEMATIC APPROACH TO IDENTIFYING A SET OF INFORMATIVE PHYSICOCHEMICAL AND BIOCHEMICAL PROPERTIES

Hui-Lin Huang[1,2], I-Che Lin[1], Yi-Fan Liou[2], Chia-Ta Tsai[2], Kai-Ti Hsu[2], Wen-Lin Huang[3], Shinn-Jang Ho[4], and Shinn-Ying Ho[1,2§]

[1]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan
[2]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan
[3]Department of Management Information System, Chin Min Institute of Technology, Miaoli, Taiwan
[4]Department of Automation Engineering, National Formosa University, Yunlin 632, Taiwan

[§]Corresponding author

Email addresses:
HLH: hlhuang@mail.nctu.edu.tw

ICL: augs1984@gmail.com

YFL: yifanliou@gmail.com

CTT: cttsai.bi96g@nctu.edu.tw

KTH: cadycat@gmail.com

WLH: wenlinhuang2001@gmail.com

SJH: sjho@nfu.edu.tw

SYH: syho@mail.nctu.edu.tw

I. Abstract

A. Background

Existing methods of predicting DNA-binding proteins used valuable features of physicochemical properties to design support vector machine (SVM) based classifiers. Selection of physicochemical properties and determination of their corresponding feature vectors rely mainly on known properties and experience of designers. However, there exists a troublesome problem for designers that some different physicochemical properties have similar vectors of representing 20 amino acids and closely related physicochemical properties in the same group have dissimilar vectors.

B. Methods

This study proposes a systematic approach (named Auto-IDPCPs) to automatically identify a set of physicochemical and biochemical properties in the AAindex database to design SVM-based classifiers for predicting and analyzing DNA-binding domains/proteins. Auto-IDPCPs consists of 1) clustering 531 vectors in AAindex into 20 classes using a fuzzy c-means algorithm, 2) utilizing an efficient genetic algorithm based optimization method IBCGA to select an informative set of feature vectors of representing sequences from the viewpoint of machine learning, and 3) analyzing the selected feature vectors to identify the related physicochemical properties which may affect the binding mechanism of DNA-binding domains/proteins.

C. Results

The proposed Auto-IDPCPs identified *m*=22 features of properties belonging to five classes for predicting DNA-binding domains with a five-fold cross-validation accuracy of 87.12%. If *m*=5 that one representative property is selected from each class, the accuracy of 83.59% is also promising compared with the accuracy of 82.07% of the existing method PSSM-400 using 400 features. For predicting DNA-binding sequences, three additional classes (totally eight classes) were needed, and the accuracies of 75.50% and 73.24% were obtained using *m*=28 and 8 features, respectively, where PSSM-400 has the accuracy of 74.22%. When applied on an independent test data set of DNA-binding domains, Auto-IDPCPs and PSSM-400 have accuracies of 80.73% and 82.81%, respectively. Some typical physicochemical properties discovered are hydrophobicity, secondary structure, charge, solvent accessibility, polarity, flexibility, normalized Van Der Waals volume, pK (pK-C, pK-N, pK-COOH and pK-a(RCOOH)), etc.

D. Conclusions

The proposed approach Auto-IDPCPs would help designers to investigate informative physicochemical and biochemical properties by considering both prediction accuracy and analysis of binding mechanism simultaneously. The approach Auto-IDPCPs can be also applicable to predict and analyze other protein functions from sequences.

II. Background

DNA-binding domains/proteins are functional proteins in a cell, which plays a vital role in various essential biological activities, such as DNA transcription, replication, packaging, repair and rearrangement [1]. The computational methods using support vector machine (SVM) in conjunction with evolutionary information of amino acid sequence in terms of their position-specific scoring matrices (PSSMs) for predicting DNA-binding sites were successfully developed [2]. Several methods of using machine learning approaches were developed to predict DNA-binding domains/proteins from given sequences of variable lengths [3-7], shown Table 1. Due to different design aims and data sets used, it is difficult to assess which feature type is the most informative cooperated with SVM by comparing with prediction accuracies only. The PSSM is an effective feature type of representing DNA-binding sequences, but its ability of interpretability is not satisfactory enough in analyzing the binding mechanism [5]. Besides PSSMs, the physicochemical properties with the characteristics of high interpretability were commonly used [3-4, 6-7]. Some issues are concerned in designing prediction methods, described below.

1) Selection of physicochemical properties: Generally, effective physicochemical properties of amino acids are selected as prediction features by using known properties of DNA-binding mechanism and knowledge of related binding mechanism [3-4, 6-7]. However, it is desirable to explore undiscovered properties by machine learning approaches to further advance the prediction accuracy and understand the binding mechanism.

2) Representation of sequences: How to effectively represent sequences of variable lengths as a feature vector using physicochemical properties play an important role in advancing prediction accuracy. The pseudo-amino acid

composition (PAAC) is an efficient representation method of coupling physicochemical properties, which was used to represent a sequence as a 40-dimensional feature vector for discriminating DNA-binding proteins from non-binding proteins [3]. The combined descriptor was proposed using amino acid composition and a series of associated physicochemical properties to form a 132-dimensional feature vectors [7]. The conjoint triad descriptor of 343-dimensional feature vector was proposed that 20 amino acids were clustered into seven classes according to their dipoles and volumes of side chains [6].

3) Values of amino acids for specific physicochemical properties: The AAindex database [8-9] collected 531 physicochemical properties (ignoring 13 properties without available values) with corresponding values of amino acids. Recently, some computational methods of predicting protein functions were successfully developed by mining informative physicochemical properties from AAindex [10-11].

Besides pursuit of high prediction accuracy, discovering potential properties to further understand the binding mechanism are also taken into account in this study. We present a troublesome problem in using the AAindex database and propose an effective method to solve. We found that some different physicochemical properties have similar vectors of representing the 20 amino acids and the closely related physicochemical properties in the same group have dissimilar vectors. For example, the determination of values of the 20 amino acids to represent a sequence by coupling the hydrophobicity property is highly related to prediction performance. Similarly, if a different property with a similar vector replaces the known one without significantly degrading prediction performance, it means that the replaced property may be also important to the binding mechanism from the viewpoint of machine learning. The detailed explanation by using a real quantization example is described below.

Figure 1 shows an illustration example. The 402 properties in AAindex were classified into six groups according to their biological meanings classified by Tomii et al. [9], as shown in Fig. S1 [see additional file]. According to the vectors of amino acids for 531 properties, we clustered them into 20 clusters by a fuzzy c-means algorithm [12] based on normalized Euclidean distances. The properties H88 and A392 are two different properties but their distance 0.0178 is small belonging to the same cluster 7. On the other hand, H88 and H178 belonging to the same group Hydrophobicity in AAindex have a large distance 0.0877 belonging to clusters 7 and 18, respectively. Although H88 and H151 (used in [3]) are in the same group Hydrophobicity, their distance 0.0299 is larger than that between H88 and A392.

For the aim of designing accurate prediction methods, the major concern is to identify feature vectors with high discrimination abilities for classifying positive and negative samples. This task can be done well for computational methods by an optimization approach to feature selection. If the feature vectors were identified by predetermined properties based on prior knowledge, the selected vectors of representing amino acids may be not the best. Considering the other aim of discovering potential properties to further look insight the binding mechanism, we proposed a systematic, optimization approach (named Auto-IDPCPs) to automatically identifying a set of feature vectors and analyzed the feature vectors to find properties of affecting the DNA-binding mechanism.

The proposed approach Auto-IDPCPs can identify a small number *m* of feature vectors and discover the related hydrophobicity properties with comparable performance, compared with the PSSM feature. Auto-IDPCPs would help designers to investigate informative physicochemical and biochemical properties by considering both prediction accuracy and analysis of binding mechanism simultaneously. Auto-IDPCPs can be also applicable to predict and analyze other protein functions from sequences.

III. METHODS

The system flowchart of the proposed approach Auto-IDPCPs is shown in Fig. 2. The input of the method comprises the AAindex database and three data sets, including DNA-binding domains and sequences, and one independent test data set. The output has two parts: 1) a predictor of DNA-binding domains/proteins with a set of *m* informative feature vectors and the parameter setting of SVM by an efficient feature selection algorithm IBCGA, and 2) a set of physicochemical and biochemical properties in the AAindex database for analyzing the DNA-binding mechanism.

A. *Data sets*

To evaluate effectiveness of the identified physicochemical properties by comparing with the famous PSSM features, we used the benchmark data sets used in the PSSM-400 method [5], as shown in Table 2. The data set DNAset has 146 DNA-binding domains (or protein chains) and 250 non-DNA-binding domains. No two domains have the similarity more than 25%. The data set DNAaset consists of 1153 DNA-binding proteins and 1153non-binding proteins. 3) An independent data set DNAiset is additionally used, having 92 DNA-binding domains and 100 non-DNA-binding domains [5].

B. *Feature vector representation*

All the domains/sequences have a variable length *l*. A sequence forms an *l*-dimensional profile where the value of each amino acid is obtained from the specified property in the AAindex database. The *l*-dimensional profiles are transformed into vectors with the same constant length L for utilizing SVM. The transformation can be any known effective representation [3-4, 6-7] provided that the L features can effectively classify the *l*-dimensional profiles of positive and negative sequences. The simplest feature is the mean of the profile that L=1 [10-11]. Therefore, the sequences with *m* properties are represented as an *m*-dimensional feature vectors. Finally, all values of the feature vectors are normalized into [-1, 1] for applying SVM.

C. *Feature selection by IBCGA*

Selecting a minimal number of informative features while maximizing prediction accuracy is a bi-objective 0/1 combinatorial optimization problem. An efficient inheritable bi-objective combinatorial genetic algorithm IBCGA [13] is utilized to solve this optimization problem. IBCGA bases on an intelligent genetic algorithm IGA [14] with an inheritable mechanism. The IGA algorithm uses a divide-and-conquer strategy and an orthogonal array crossover to efficiently solve large parameter optimization problems. In this study, the IGA algorithm can efficiently explore and exploit the search space of C(*n*, *r*), where *n*=531 in this study. IBCGA can efficiently search the space of C(*n*, *r*± 1) by inheriting a good solution in the space of C(*n*, *r*) [13]. Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run where *r* is specified in an interesting range such as [10, 30]. The chromosome encoding scheme of IGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters γ and C [10]. The performance of selected properties associated with the parameter values of SVM is measured by five-fold cross-validation (5-CV) for comparing with the method PSSM-400 [5].

IBCGA with the fitness function $f(\mathrm{X})$ can simultaneously obtain a set of solutions, X_r, where $r=r_{\mathrm{start}}$, $r_{\mathrm{start}}+1, \ldots, r_{\mathrm{end}}$ in a single run. In this study, the parameter settings $r_{\mathrm{start}}=10$, $r_{\mathrm{end}}=30$, $N_{\mathrm{pop}}=50$, $p_c=0.8$ and $p_m=0.05$. The output contains a set of m selected properties from AAindex and an SVM classifier with associated parameter settings. The IBCGA algorithm is given in Fig. 3. The best one of R solutions can be determined by considering the accurate one S_{a} with the highest accuracy or the robust one with the highest score S_{r} for identifying informative properties.

D. *Designing SVM classifiers*

For evaluating the effectiveness by comparing with the commonly used feature sets, we implemented the predictor using the same single-classifier SVM with the feature types, amino acid composition (AAC) and PSSM [5]. Additionally, the selected physicochemical properties (PCPs) combined with AAC and PSSM were also evaluated.

E. *Clustering properties by the FCM method*

The application of cluster approaches is to partition 531 vectors of physicochemical properties into clusters, where similar vectors are assigned to the same cluster. An index vector of amino acids is a set of 20 numerical values representing some physicochemical property of amino acids. All data were normalized in such a way that every physicochemical property had an average profile value of zero and a standard deviation equal to 1.

The fuzzy derivative of k-means, known as fuzzy c-means (FCM) [12], has an objective functional of the form, $\mathrm{J(X;U,v)}=\sum_{i=1}^{K}\sum_{j=1}^{n}u_{ij}^{s}d^{2}(v_i,x_j)$, where $n=531$ is the number of data vectors, K is the number of clusters to be found, $u_{ij}\in[0,1]$ is the membership degree of j^{th} data vector x_j in the i^{th} cluster, the i^{th}

cluster represented by the cluster prototype v_i, $s \in [1,\infty)$ is a weighting exponent called the fuzzifier and $d(v_i, x_j)$ is the distance of x_j from the cluster prototype v_i. Dembélé and Kastner [15] suggested the parameters setting s=1.12 and K=20 clusters, adopted in this study.

F. Identifying physicochemical properties

It is not easy to discover related physicochemical properties for analyzing DNA-binding mechanism by computational methods with a relatively small size of data sets. Therefore, we present a hybrid method by combining evidences from the viewpoints of both machine learning and biological meanings. Auto-IDPCPs identifies m properties belong to c of 20 clusters. We examine all properties P1 by considering the identified m properties P2 if they satisfying the criteria, P1 is a promising property to be further investigated: 1) P1 and P2 have a small distance and 2) if P2 is replaced with P1 one at a time, the prediction accuracy is not significantly decreased.

Only 402 of 531 properties were classified into six groups, (A): Alpha and turn propensities, (B): Beta propensity, (C): Composition, (H): Hydrophobicity, (P): Physicochemical properties, and (O): Other properties. We classified the other 129 properties into the six groups according their distance of vectors using a nearest-neighbor rule. The mapping of 531 feature numbers and AAindex identity and their classified result into six groups are given in Tables S1 and S2, respectively, and their statistic result is given in Table S3 [see additional file]. The statistical results of property distribution in the six groups for 531 and 402 amino acid indices are given in Fig. S2 [see additional file].

IV. RESULTS

A. Identified properties by IBCGA

The statistical result of S_r in selecting property sets from R =30 independent runs on DNAset and DNAaset are given in Fig. 4. The 18th and 6th runs for DNAset and DNAaset, respectively, are selected, and their prediction accuracies for various numbers of selected properties are given in Fig. 5. The m=22 and 28 properties selected for DNAset and DNAaset, respectively, are given in Tables S4 and S5 [see additional file] in which the AAindex identity numbers and their property description are provided. An efficient way to study the effects of several factors simultaneously is to use the main effect difference (MED) that the most effective property has the largest value of MED. The m properties are ranked by using MED is shown in Fig. 6. The properties of rank 1 are feature numbers 86 (FAUJ880109, Localized electrical effect) and 39 (CHOP780202, Normalized frequency of beta-sheet) for DNAset and DNAaset, belonging to groups Hydrophobicity and Beta propensity in the six groups, respectively.

B. *Prediction performance evaluation*

To evaluate the effectiveness of the identified *m* informative feature vectors (PCPs), three feature types were additionally evaluated, as shown in Table 3. AAC is a 20-dimensional vector of amino acid composition, PSSM is the feature representation [5] of 400 features. PCPs +AAC and PCPs +PSSM are two hybrid feature types by adding individual feature vectors. Considering the DNA-binding domain data set DNAset, the set of *m*=22 informative properties (PCPs) identified by Auto-IDPCPs performs best where the robust solution S_r with accuracy of 87.12%, compared with ACC, PSSM, PCPs+AAC and PCPs+PSSM are 80.30%, 82.07%, 81.82% and 86.62%, respectively. For the DNA-binding protein data set DNAaest, the method with PCPs and *m*=28 informative properties (75.50%) is slightly worse than that with PSSM (76.58%). However, PCPs+PSSM can improve the accuracy to 80.27%. When the predictor trained by DNAset (S_r with *m*=22 informative properties) were evaluated by the independent test data set DNAiset, the accuracy is 80.73% (=155/192), slightly worse than 82.81% (=159/192) of PSSM-400. A small, high-performance features set of size c from c clusters is given in Table 4. The properties and their descriptions are given in Tables S6 and S7 [see additional file] where c=5 and 8 for DNAset and DNAaset, respectively.

The experimental results reveal that the identified small set of *m* physicochemical properties with a simple representation performs equally well, compared with the PSSM feature type. However, the identified physicochemical properties are interpretable for further understanding the DNA-binding mechanism.

C. *Analyzing binding mechanism by physicochemical properties*

The 30 sets of *m* properties belonging to the 20 clusters from the results of 30 runs are shown in Fig. S3 [see additional file]. From the statistic result, the clusters 7, 9, 10, 16 and 18 with very high selection frequencies are more important for predicting DNA-binding domains and proteins. The *m*=22 properties (Table S4) belong to five clusters which are the same as the five clusters 7, 9, 10, 16 and 18. For predicting DNA-binding proteins, the *m*=28 properties (Table S5) belong to eight clusters with additional three clusters 3, 14 and 17.

An illustration example is given in Fig. 7. The both feature sets S1 (H88, H86, H67, C209, H178) and S2 (A392, A303, A307, C440, H178) are selected for predicting DNA-binding domains in DNAset that one properties selected from one of five clusters 7, 9, 10, 16 and 18. The identified properties H88 and A392 belong to hydrophobicity, and alpha and turn propensities groups but they belong to the same cluster 7 with a relatively small distance 0.0178. The prediction accuracy of S3 by replacing H88 with H151 is 81.05 %. On the other hand, H151 belonging to the cluster 7 and Hydrophobicity group used in [3] can be inferred from feature sets S1 and S2. After carefully analyzing all properties, we identify some properties in the five identified clusters for analyzing DNA-binding domains, shown in Table 5. Some typical physicochemical properties discovered are hydrophobicity, secondary structure, charge, solvent accessibility, polarity, flexibility, normalized Van Der Waals volume, pK (pK-C, pK-N, pK-COOH and pK-a(RCOOH)), etc. Most of identified properties were used in previous works [3-4, 6-7] but a few properties such as the flexibility property H8 BHAR880101 in cluster 7 "Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)" are not utilized yet in existing method of predicting DNA-binding domains. The correlation between protein flexibility and protein function suggests a link between DNA-binding activity and the conformational freedom of the DNA-binding domain [16].

V. Discussion

To avoid from overfitting the small-scale data sets in identifying physicochemical properties using an optimization approach, this study proposes a hybrid method of combining evidences from computational methods of considering robust factors and biological experiments from literature. The future work is to further verify these discovered properties in predicting and analyzing the DNA-binding mechanism.

VI. Conclusions

This study has proposed a systematic approach Auto-IDPCPs to automatically identify an informative set of physicochemical and biochemical properties in the AAindex database to design SVM-based classifiers for predicting and analyzing DNA-binding domains/proteins.

VII. Competing interests

The authors declare that they have no competing interests.

VIII. Authors' contributions

HLH designed the system, implemented programs, carried out the analysis, and participated in manuscript preparation. ICL provided biological knowledge and carried out the analysis. YFL developed the web server. CTT, KTH, WLH and SJH implemented programs and participated in the experimental design. SYH supervised the whole project and participated in manuscript preparation. All authors have read and approved the final manuscript.

IX. Acknowledgements

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 98-2627-B-009-004- and NSC 96-2628-E-009-141-MY3.

References

1. **Gao M, Skolnick J:** A threading-based method for the prediction of DNA-binding proteins with application to the human genome. ***PLoS Comput Biol* 2009,** 5**(11):e1000567.**

2. **Ho SY, Yu FC, Chang CY, Huang HL:** Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. ***Biosystems* 2007,** 90**(1):234-241.**

3. **Cai YD, Lin SL:** Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. ***Biochim Biophys Acta* 2003,** 1648**(1-2):127-133.**

4. **Fang Y, Guo Y, Feng Y, Li M:** Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. ***Amino Acids* 2008,** 34**(1):103-109.**

5. **Kumar M, Gromiha MM, Raghava GP:** Identification of DNA-binding proteins using support vector machines and evolutionary profiles. ***BMC Bioinformatics* 2007,** 8**:463.**

6. **Shao X, Tian Y, Wu L, Wang Y, Jing L, Deng N:** Predicting DNA- and RNA-binding proteins from sequences with kernel methods. ***J Theor Biol* 2009,** 258**(2):289-293.**

7. **Yu X, Cao J, Cai Y, Shi T, Li Y:** Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. ***J Theor Biol* 2006,** 240**(2):175-184.**

8. **Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M:** AAindex: amino acid index database, progress report 2008. ***Nucleic Acids Res* 2008,** 36**(Database issue):D202-205.**

9. **Tomii K, Kanehisa M:** Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. ***Protein Eng* 1996,** 9**(1):27-36.**

10. **Tung CW, Ho SY:** POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. ***Bioinformatics* 2007,** 23**(8):942-949.**

11. **Tung CW, Ho SY:** Computational identification of ubiquitylation sites from protein sequences. ***BMC Bioinformatics* 2008,** 9**:310.**

12. **Bezdek JC:** Pattern Recognition with Fuzzy Objective Function Algorithms. **New York: Plenum Press 1981.**

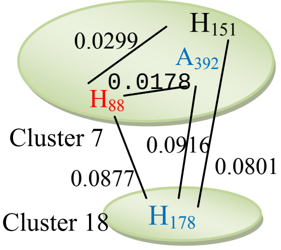
13. **Ho SY, Chen JH, Huang MH:** Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. ***IEEE Trans Syst Man Cybern B Cybern* 2004,** 34**(1):609-620.**

14. **Ho SY, Shu LS, Chen JH:** Intelligent evolutionary algorithms for large parameter optimization problems. ***Ieee Transactions on Evolutionary Computation* 2004,** 8**(6):522-541.**

15. **Dembele D, Kastner P:** Fuzzy C-means method for clustering microarray data. ***Bioinformatics* 2003,** 19**(8):973-980.**

16. **Gryk MR, Jardetzky O, Klig LS, Yanofsky C:** Flexibility of DNA binding domain of trp repressor required for recognition of different operator sequences. ***Protein Sci* 1996,** 5**(6):1195-1197.**

Figures



Six region feature ID	AAindex ID	Description
H88	FAUJ880111	Positive charge (Fauchere et al.1988)
H178	MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
H151	KYTJ820101	Hydropathy index (Kyte-Doolittle, 1982)
A392	WOLS870103	Principal property value z3 (Wold et al., 1987)

Figure 1 - Illustration example. The properties H88 and A392 are two different properties but their distance 0.0178 is small. On the other hand, H88 and H178 belonging to the same group Hydrophilicity in AAindex have a large distance 0.0877. H88 and H151 in the same group have a larger distance 0.0299 than that between H88 and A392.

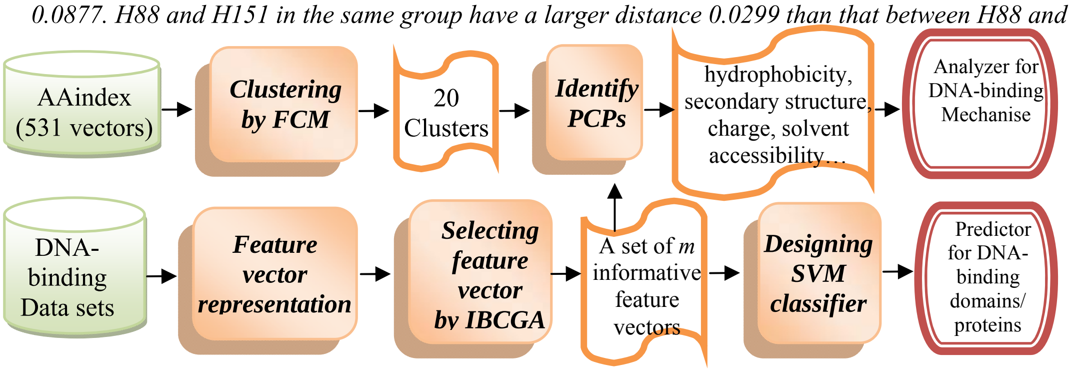

AAindex (531 vectors)
Clustering by FCM
20 Clusters
Identify PCPs
hydrophobicity, secondary structure, charge, solvent accessibility…
Analyzer for DNA-binding Mechanise
DNA-binding Data sets
Feature vector representation
Selecting feature vector by IBCGA
A set of m informative feature vectors
Designing SVM classifier
Predictor for DNA-binding domains/ proteins

Figure 2 -The system flowchart of the proposed approach Auto-IDPCPs.

Step 1) (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary genes have r 1's and n-r 0's where $r = r_{start}$.

Step 2) (Evaluation) Evaluate the fitness values of all individuals using $f(X)$.

Step 3) (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4) (Crossover) Select $p_c \cdot N_{pop}$ parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents where p_c is the crossover probability.

Step 5) (Mutation) Apply the swap mutation operator to the randomly selected $p_m \cdot N_{pop}$ individuals in the new population where p_m is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6) (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r. Otherwise, go to Step 2).

Step 7) (Inheritance) If r < r_{end}, randomly change one bit in the binary genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

Step 8) (System uncertainty) Perform Steps 6 and 7 for R=30 independent runs to obtain the best of R solutions, X_m, and the associated parameter setting of the SVM classifier. The best solution considers both high prediction accuracy and high mean of appearance frequency ratio, described as the following procedure APPF.

The procedure APPF is given as the following steps:

Step 1) Calculate the appearance frequency f(p_i) of each selected properties p_i from the R=30 sets of m_i-dimensional feature vectors, where i=1, …, 30.

Step 2) Calculate score S_r for each of R solutions, where properties p_i are in the r^{th} set and r = 1, …, R:

$$S_r = (\sum_{i=1}^{m_r} f(p_i))/m_r$$

where f(p_i) denotes the frequencies of the property p_i, m_r is the number of the selected feature in dependent run r^{th}.

Figure 3 - The algorithm IBCGA used in the proposed approach Auto-IDPCPs.

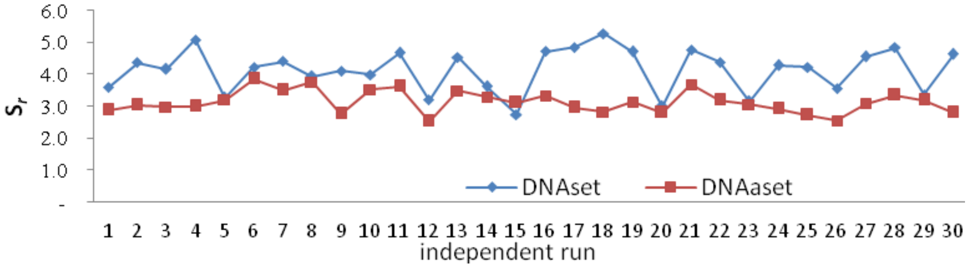
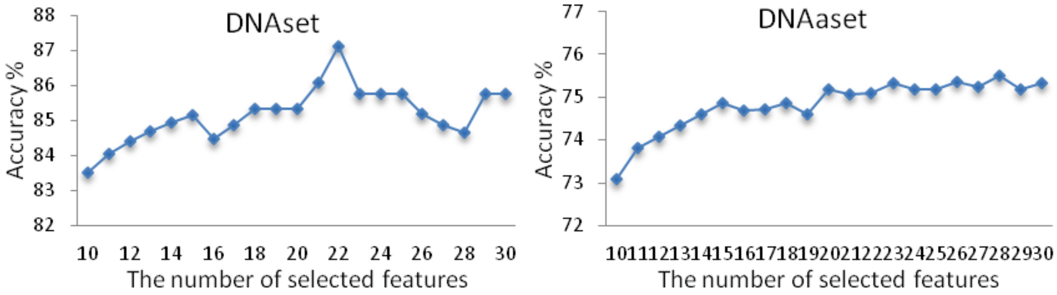
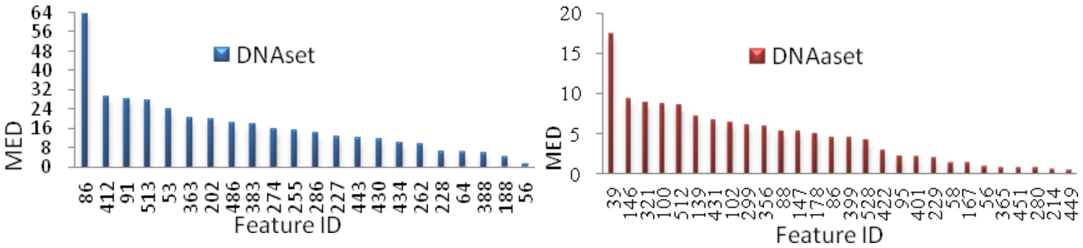


Figure 4 - The statistical result of S_r in selecting property sets from R *=30 independent runs on DNAset and DNAaset.*



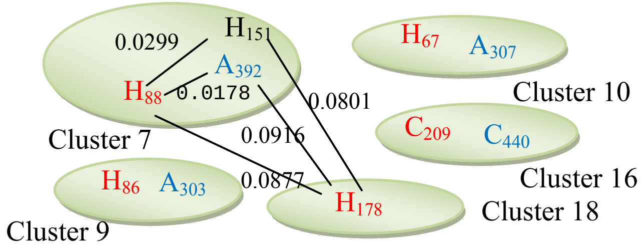
(a) (b)

Figure 5 - Prediction accuracies for various numbers of selected properties (a) DNAset and (b) DNAaset.



(a) (b)

Figure 6 - The effectiveness of properties are ranked by using the main effect difference (MED) (a) m=22 and (b) m=28.



The selected feature sets	ACC(%)
S1: H_{88},H_{86},H_{67},C_{209},H_{178}	83.59
S2: A_{392},A_{303},A_{307},C_{440},H_{178}	82.32
S3: H_{151},H_{86},H_{67},C_{209},H_{178}	81.05

Figure 7 - An illustration example for exploring properties. H151 can be inferred from feature sets S1 and S2.

Tables

Table 1 - Related works of predicting DNA-binding domains/proteins from sequences

Reference	Protein type	Identity	Feature number	Representation	Feature type	Classifier
Shao et al. 2009[6]	sequence	25%	343	Seven class Conjoint triad	PCP	SVM
Fang et al. 2008[4]	sequence	35%	40	Pseudo-AA composition	PCP	SVM
Yu et al. 2006 [7]	sequence	25%	132	Combined descriptors	PCP	SVM
Cai et al 2003 [3]	sequence	40%	40	Pseudo-AA composition	PCP	SVM
Kumar et al. 2007 [5]	Domain and sequence	25%	400	PSSM	PSSM	SVM
Ours	Domain and sequence	25%	m*	Mean value of sequence #	PCP and BCP	SVM

PCP: physicochemical property, BCP: biochemical property

*: a small number of feature vectors selected from 531 vectors

# : The averaged value of amino acids in a sequence for one property

Table 2 - The statistic of the three data sets

Datasets	Protein	No. of DNA-binding	No. of non-DNA-binding
DNAset	domain	146	250
DNAaset	sequence	1153	1153
DNAiset	domain	92	100

Table 3 - The overall accuracies (%) of 5-CV using three types of feature representations and their combination types with SVM.

Dataset		Sen.	Spe.	MCC	PCPs	AAC	PSSM*	PCPs +AAC	PCPs +PSSM
DNAset	S_a	88.89	91.20	0.76	88.89	80.30	82.07	81.57	83.59
	S_r	82.19	90.00	0.53	87.12			81.82	86.62
DNAaest	S_a	82.74	70.08	0.72	76.41	72.46	76.58	74.20	79.88
	S_r	81.96	69.04	0.51	75.50			73.59	80.27

S_a: accurate solution, S_r: robust solution, Sen.: sensitivity, Spe.: specificity, MCC: Matthew's correlation coefficient, PCPs: the *m* informative properties, PSSM*: obtained from [5] without additional fine tune of SVM

Table 4 - A small, high-performance features set of size c from c clusters. The feature number c=5 and 8 for DNAset and DNAaset, respectively

DNAset	ACC	Cluster ID	C_7	C_9	C_{10}	C_{16}	C_{18}			
	83.59%	Feature ID	H88	H86	H67	H209	H178			
DNAaset	ACC	Cluster ID	C_7	C_9	C_{10}	C_{16}	C_{18}	C_3	C_{14}	C_{17}
	73.24%	Feature ID	P159	H87	A99	C197	P63	H11	H396	H451

Table 5 - Some typical properties in the five identified clusters for analyzing DNA-binding domains

C_{id}	AAindex ID	PCP & BCP	C_{id}	AAindex ID	PCP & BCP
7	BHAR880101	Flexibility	10	FASG760105	pK-C
7	BURA740101	Secondary structure	10	JOND750102	pk- (-COOH)
7	CHOC760103	Solvent accessibility	10	RADA880108	Polarity
7	HOPT810101	Hydrophilicity	16	PRAM900101	Hydrophilicity
7	FAUJ880111	Charge	16	FUKS010104	Solvent accessibility
9	KARP850101	Flexibility	16	KUMS000103	Secondary structure
9	PALJ810115	Secondary structure	18	PONP800107	Solvent accessibility
9	ROSM880101	Hydrophilicity	18	GRAR740102	Polarity
9	KUHL950101	Solvent accessibility	18	FASG760104	pK-N
10	ZIMJ680101	Hydrophilicity	18	FAUJ880113	pK-a(RCOOH)
10	EISD860101	Solvent accessibility	18	FAUJ880103	Normalized van der Waals volume
10	GEIM800101	Secondary structure			

C_{id}: FCM cluster ID PCP & BCP: physicochemical and biochemical property

國科會補助教師出席國際會議結案心得報告

報告人姓名	黃慧玲	所屬學校 學系(所)	交通大學 生物科技學系
會議期間 及地點	2011/06/10 至 2011/06/12 上海	補助項目 及金額	■ 機票費 ■ 註冊費 ■ 生活費
會議名稱	（中文）2011 全球電機計算科學與自動工程國際研討會 （英文）2011 IEEE International Conference on Computer Science and Automation Engineering (CSAE 2011)		
發表論 文題目	用演化模糊分類器來擷取可解譯知識作DNA-binding domains預測 Interpretable knowledge acquisition for predicting DNA-binding domains using an evolutionary fuzzy classifier method		

報告內容：(1、參加會議經過；2、與會心得3、建議4、攜回資料)

吾人發表的論文是在06/12 日下午3 點50分的Session B6會議廳

1.參加會議經過

本次會議CSAE 2011由五所大學假上海合辦，五所大學分別為Beijing (China)、Tongji (China)、Xiamen (China)、Bradford (UK)和Iwate Prefectural (Japan)等。第九屆亞太生物資訊會議此次主辦單位是KSBSB (Korean Society for Bioinformatics and Systems Biology) 協辦單位是KRIBB (Korean Research Institute of Bioscience and Biotechonology)、KOBIC (Korea Bioinformation Center)、 和 Chungbuk BIT Research-Oriented University Consortium。今年大會共有4個Room，18個Section，包含口頭報告區、演講區、與海報展示。本次首先有3位Keynote Speaker，接著共161篇口頭報告論文，以及共接受490篇海報舉行。很榮幸我們的論文被接受口頭報告。此科學論壇，為助長科學、工程、生物資訊與科技的發展目前已經與眾多學術和科學界的領導組織共同合作，合作學術單位遍及日本、美國、俄羅斯、印度、馬來西亞、澳洲、埃及與伊朗等等國家，為相當國際化之會議舉辦組織。本次投稿被CSAE 2011接受之國際會議論文亦有機會被轉投至科學、工程、生物資訊國際期刊。會議參加的人員來自許多國家，包含大陸、美國、日本、澳洲、印度、新加坡、香港、韓國、台灣、馬來西亞、泰國……等等，其間大會於第二天安排午宴讓來自各個國家的學者互相交流、聯誼，期望能促進與會學者日後學術交流的機會。於此國際會議中發表之論文題目Interpretable knowledge acquisition for predicting DNA-binding domains using an evolutionary fuzzy classifier method。在電機領域報告生物資訊相關的主題跨領域的衝擊更大，但因本人亦由資訊工程跨轉生物資訊，因此能感受聽者的感受。

6月10日上午由桃園機場抵達上海的蒲東機場，然後搭地鐵到會場，行程相當順利。6月的上海也是下雨的季節，再正逢星期五的時間，由會場回國際飯店，搭計程車所花費的時間比平時多好多倍。行程中我們亦安排與上海大學生物科技研究所副所長等交流生物科技與生物資訊未來發展與發展的策略。其中蔡昱東教授對台灣生物資訊學者交流非常熱烈並且也都有期刊作品發表，此次蔡教授亦非常可望與交大生資所合作。我們一行人在交大生科院何副院長帶領之下，與蔡教

授討論一些出初步可行合作方案。蔡教授在上海長大，對上海景點介紹給我們非常好的欣賞上海的行程，此次比2009初次到上海參加會議，在學術交流與人文欣賞更有收穫。

第三天我們報告時間是下午2:50，我們主要是提出用演化模糊分類器來擷取可解譯知識作DNA-binding domains預測，這是延續我們之前發表期刊論文主要提出解決方法並應用DNA-binding domains序列資料去分析及發現來探討問題。我對這session的一個主題在動態情形追蹤行徑感到有興趣，因為我目前正在研究的是Bio-image研究，關於神經影像，拍攝神經發展也是一種動態追蹤。會議結束後上8樓享受上海美食宴會，然後帶著滿滿的陶醉結束這會議行程。以下照片由右到左分別為我、主持人、何教授信瑋、黃教授文玲、交大生科副院長何信瑩教授，另一張為我的報告影像，皆於報告會場留影。

2、與會心得

感謝國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。此次交大生科副院長何信瑩教授與上海大學作學術交流，並實質與教授群達成學術合作方案，更增加參加會議的價值。

個人覺得跨領域的交流，由此次舉辦大陸上海，他們對跨領域的企圖心由投入組織相當龐大，也可見他們對這領域的企圖心與團結。反觀台灣對跨領域投入與組織結構發展還需更努力。而我們與副院長在上海大學及會場上的討論彷彿將系院擴大到國際空間進行情境探討與未來計畫，相當難得的收穫。

3、建議

近年來國科會、教育部和學校積極鼓勵年輕研究人員，除鼓勵教師參與會議外，特別是博士班學生，參與大型國際會議，及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加生物資訊國際會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此交換研究心得，更可找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。目前研究生已有多管道獲(部份)補助出席國際會議，建議繼續擴大進行。而國際化的學術交流是往後的趨勢，也能有所激勵國人學界能力與國際觀。

4、攜回資料

1. 論文集 Proceedings of 2011 IEEE International Conference on Computer Science and Automation Engineering (CSAE 2011) VOLUME 4, July 2011
2. (ISBN:978-1-4244-8726-4)全文電子檔案
3. 紙筆記事一套。

接受函與論文

From: IEEE CSAE 2011 [mailto:csae@ieee-csae.org]
Sent: Monday, February 28, 2011 4:01 PM
To: syho.nctu@gmail.com
Cc: hlhuang@mail.nctu.edu.tw
Subject: [CSAE 2011] Acceptance Notification

Dear Authors,
Thank you for your submission to CSAE 2011, which will be held in shanghai during June 10-12, 2011. We are pleased to inform you that your paper:
ID: **11285**
TITLE:**Interpretable knowledge acquistion for predicting DNA-binding domains using an evolutionary fuzzy classifier method**
has been accepted for publication in the proceedings of 2011 IEEE International Conference on Computer Science and Automation Engineering (CSAE 2011). Congratulations! This year, we received more than 1000 submissions; only very outstanding paper can be accepted by the conference. All papers accepted will be included in IEEE Xplore and indexed by Ei Compendex and ISTP.
Here are some important issues on registration and final paper submission:
(1) At least one author of each accepted paper should register before March 30, 2011. Please visit: http://www.ieee-csae.org/index6.asp
(2) Please revise your paper in detail according to the review results, the review information can be obtained below the e-mail. The paper format can be found at: http://www.ieee-csae.org/index1.asp
(3) Please submit your final papers (pdf file) and the signed completed copyright form to csae@ieee-csae.org before March 30, 2011. The blank copyright form can be downloaded from: http://www.ieee-csae.org/IEEECopyrightForm.doc
Thank you very much for your contribution to this conference. We are looking forward to seeing you in shanghai, China.

Best Regards,
Shaozi Li, Technical Program Committee Chair
CSAE 2011 Organizing Committee.

======= Review =======
This paper proposes an interpretable physicochemical property classifier (named iPPC) with an accurate and compact fuzzy rule base using a scatter partition of feature space for DNA binding data analysis. The topic is interesting and has a practical value. But I would recommend extending it a little bit; especially comparing the experimental result with related works.

Acquisition of rule-based knowledge for predicting and analyzing DNA-binding domains

Hui-Ling Huang[1,2], Shinn-Jang Ho[3], Li-Sun Shu[4], and Shinn-Ying Ho[1,2]

[1]Department of Biological Science and Technology,
ational Chiao Tung University, Hsinchu, Taiwan
[2]Institute of Bioinformatics and Systems Biology, National
Chiao Tung University, Hsinchu, Taiwan
hlhuang@mail.nctu.edu.tw

[3]Department of Automation Engineering, National
Formosa University, Yunlin 632, Taiwan
[4]Department of Information Management, Overseas
Chinese University, Taichung 40721, Taiwan
syho@mail.nctu.edu.tw

Abstract—DNA-binding domains are functional proteins in a cell, which plays a vital role in various essential biological activities. It is desirable to predict and analyze novel proteins from protein sequences only using machine learning approaches. Numerous prediction methods were proposed by identifying informative features and designing effective classifiers. The support vector machine (SVM) is well recognized as an accurate and robust classifier. However, the block-box mechanism of SVM suffers from low interpretability for biologists. It is better to design a prediction method using interpretable features and prediction results. In this study, we propose an interpretable physicochemical property classifier (named iPPC) with an accurate and compact fuzzy rule base using a scatter partition of feature space for DNAbinding data analysis. In designing iPPC, the flexible membership function, fuzzy rule, and physicochemical properties selection are simultaneously optimized. An intelligent genetic algorithm IGA is used to efficiently solve the design problem with a large number of tuning parameters to maximize prediction accuracy, minimize the number of features selected, and minimize the number of fuzzy rules. Using benchmark datasets of DNA-binding domains, Ippc obtains the training accuracy of 81% and test accuracy of 79% with three fuzzy rules and two physicochemical properties. Compared with the decision tree method with a training accuracy of 77%, iPPC has a more compact and interpretable knowledge base. The two physicochemical properties are Number of hydrogen bond donors and Helix-coil equilibrium constant in the AAindex database.

Keywords- ***knowledge acquistion; fuzzy classifier; genetic algorithm;DNA-binding; physicochemical properties; prediction***

X. INTRODUCTION

DNA-binding domains are functional proteins in a cell, which plays a vital role in various essential biological activities, such as DNA transcription, replication, packaging, repair and rearrangement [1]. These transcription factors are mainly DNA-binding proteins (DNA-BPs) coded by 2~3% of the genome in prokaryotes and 6~7% in eukaryotes [2]. DNA-BPs play a pivotal role in various intra- and extra-cellular activities ranging from DNA replication to gene expression control. These researches reveal that the DNA-protein recognition mechanism is complicated and there is no simple rule for this recognition problem [3].

Stawiski *et al.* found that nucleic acid-binding proteins could be separated using a neural network trained that included secondary structure and charged patches, among others [4]. Ahmad and Sarai using a simple linear predictor to model a trivial system with few descriptors and they identified cutoff values for charge and dipole moment at which binding and non-binding proteins could be separated[5]. Kumar *et al.* proposed a method for predicting DNA-binding proteins using SVM and PSSM profiles [6]. The methods can fairly analyze and predict DNA-binding proteins, but suffer from obtaining human-interpretable knowledge.

Ho *et al.* [7] study aims to analyze DNA-binding proteins via acquisition of interpretable knowledge which can accurately predict binding sites in proteins to understand DNA-protein recognition mechanism. Their study investigates a novel feature set consisting of 11 features, including solvent accessibility, secondary structure, charge information near the residue, amino acid group and neighbor property. The derived binding and non-binding rules reveal that besides the well-known solvent accessibility, the electric charge distribution near the residue and the amino acid groups also play important roles in prediction of binding sites.

We have proposed Auto-IDPCPs [8] which is investigated the optimal design of predictors for DNA-DBs from amino acid sequence using both informative features and an appropriate classifier. Furthermore, we obtain a set of relevant physicochemical properties can advance prediction performance. The proposed Auto-IDPCPs identified *m*=22 features of properties belonging to five clusters for predicting DNA-binding domains with a fivefold cross-validation accuracy of 87.12%. Since the set of 22 physicochemical properties performs well, we would apply it to acquit the rule-based knowledge for predicting and analyzing DNA-binding domains.

In this paper, we propose an interpretable physicochemical properties classifier (named iPPC) with an accurate and compact fuzzy rule base using a scatter partition of feature space for DNA-binding data analysis. Because physicochemical properties from AAindex database [9] have the property of natural clustering, fuzzy classifiers using a scatter partition of feature spaces often have a smaller number of rules than those using grid partitions. The design of iPPC has three objectives to be simultaneously optimized: maximal classification accuracy, minimal number of rules, and minimal number of used physicochemical properties. In designing iPPC, the flexible membership function, fuzzy rule, and physicochemical properties selection are simultaneously optimized. An intelligent genetic algorithm IGA is used to efficiently solve the design problem with a large number of tuning parameters [10].

XI. Materials and Methods

A. *Dataset*

DNAset

This dataset also called main dataset from Kumar *et al.*, 2007 [6]. They got 146 non-redundant DNA-BPs in which no two proteins have the sequence identity of more than 25%. A non-redundant set of 250 non-binding proteins was obtained from Stawiski *et al.* [11]. They used following criteria: i) no two protein chains have similarity more than 25% and ii) the approximate size and electrostatics are similar to DNA-BPs. Final dataset called DNAset or main dataset or domain dataset, consists of 146 DNA-binding and 250 non-binding protein chains or domains.

DNAiset

We used this dataset to evaluate performance of our models and the also called DNAiset. This dataset from Kumar et al., 2007 [6] 92DNA-binding protein chains obtained from PDB and 100 nonbinding proteins picked from Swiss-Prot.

B. *Feature set*

Considering the DNA-binding domain data set DNAset, the set of *m*=22 informative properties (PCPs) identified by Auto-IDPCPs performs best where the robust solution with accuracy of 87.12% is used. The Auto-IDPCPs is a systematic approach to automatically identify a set of physicochemical and biochemical properties in the AAindex database to design SVM-based classifiers for predicting and analyzing DNA-binding domains/proteins. Auto-IDPCPs consists of 1) clustering 531 vectors in AAindex into 20 classes using a fuzzy c-means algorithm, 2) utilizing an efficient genetic algorithm based optimization method IBCGA to select an informative feature set of size *m* to represent sequences, and 3) analyzing the selected feature vectors to identify the related physicochemical properties which may affect the binding mechanism of DNA-binding domains/proteins.

The set of *m*=22 PCPs is identified by Auto-IDPCPs, we would apply it to acquit the rule-based knowledge for predicting and analyzing DNA-binding domains. The set of 22 PCPs is described in table 1.

Table 1 - The Auto-IDPCPs indented a set of *m*=22 physicolchemical properties on DNAset.

Feature ID	AAindex ID	Description
53	CHOP780216	Normalized frequency of the 2nd and 3rd residues in turn (Chou-Fasman, 1978b)
56	CIDH920103	Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)
64	DAYM780101	Amino acid composition (Dayhoff et al., 1978a)
86	FAUJ880109	Number of hydrogen bond donors (Fauchere et al., 1988)
91	FINA770101	Helix-coil equilibrium constant (Finkelstein-Ptitsyn, 1977)
188	NAGK730103	Normalized frequency of coil (Nagano, 1973)
202	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)
227	PALJ810105	Normalized frequency of turn from LG (Palau et al., 1981)
228	PALJ810106	Normalized frequency of turn from CF (Palau et al., 1981)
255	PRAM900104	Relative frequency in reverse-turn (Prabhakaran, 1990)
262	QIAN880105	Weights for alpha-helix at the window position of -2 (Qian-Sejnowski, 1988)
274	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)
286	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
363	SUEM840101	Zimm-Bragg parameter s at 20 C (Sueki et al., 1984)
383	WEBA780101	RF value in high salt chromatography (Weber-Lacey, 1978)
388	WOEC730101	Polar requirement (Woese, 1973)
412	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose, 1998)
430	MUNV940102	Free energy in alpha-helical region (Munoz-Serrano, 1994)
434	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)
443	KUMS000104	Distribution of amino acid residues in the alpha-helices in mesophilic proteins (Kumar et al., 2000)
486	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)
513	JACR890101	Weights from the IFH scale (Jacobs-White, 1989)

C. *Acquition the rule-based knowledge method*

High performance of iPPC mainly arises from two aspects. One is to simultaneously optimize all parameters in the design of iPPC where all the elements of the fuzzy classifier design have been moved in parameters of a large parameter optimization problem. The other is to use an efficient optimization algorithm IGA which is a specific variant of the intelligent evolutionary algorithm [10]. The intelligent evolutionary algorithm uses a divide-and-conquer strategy to effectively solve large parameter optimization problems. IGA is shown to be effective in the design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space [10].

1) Flexible membership function

The classifier design of iPPC uses flexible generic parameterized fuzzy regions which can be determined by flexible generic parameterized membership functions (FGPMFs) and a hyperbox-type fuzzy partition of feature space. Each fuzzy region corresponds to a parameterized fuzzy rule. In this study, each value of gene expression is normalized into a real number in the unit interval [0,1]. An FGPMF with a single fuzzy set is defined as

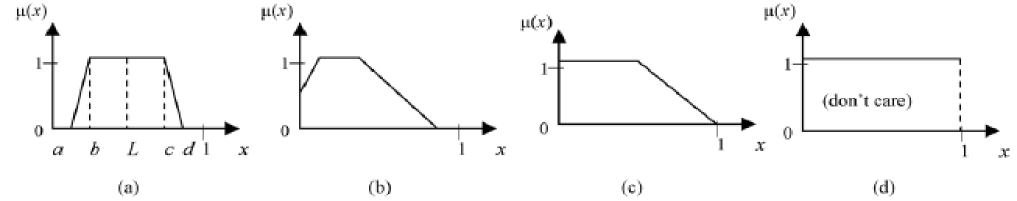


Figure 1. Illuminations of FGPMF: (a) a>0 and d< 1; (b) a<0<b, (c) b≦0; (d) b≦0 and c≧1.

$$\mu(x) = \begin{cases} 0 & \text{if } x \le a \text{ or } x \ge d \\ \dfrac{x-a}{b-a} & \text{if } a < x < d \\ \dfrac{d-x}{d-c} & \text{if } c < x < d \\ 1 & \text{if } b \le x \le c \end{cases} \tag{1}$$

where $x \in [0, 1]$ and $a \le b \le c \le d$. The variables a, b, c, and d determining the shape of a trapezoidal fuzzy set are the parameters to be optimized. It is well recognized that confining evolutionary searches within feasible regions is often much more reliable than penalty approaches for handling constrained problems [12]. Therefore, five parameters V^1, $V^2,\ldots,$ $V^5 \in [0,1]$ without constraints instead of a, b, c, and d are encoded into a GA-chromosome for facilitating IGA. Let an additional variable $L=V^1$ which determines the location of the fuzzy set characterizing the occurrence of training patterns. When V^i are obtained, variables a, b, c, and d can be derived as follows: $a=L-(V^2+V^3)$, $b=L-V^3$, $c= L+V4$, and $d=L+(V^4+V^5)$. This transformation can always make the derived values of a, b, c, and d feasible and reduce interactions among encoded parameters of GA chromosomes. Some illuminations of FGPMF are shown in Fig. 1.

2) *Fuzzy rule and fuzzy reasoning method*

The following fuzzy if–then rules for n-dimensional pattern classification problems are used in the design of iGEC:

R_j : If x_1 is A_{j1} and . . . and x_n is A_{jn} then class CL_j with CF_j, $j = 1, \ldots, N$.

where R_j is a rule label, x_i denotes a PCP variable, A_{ji} is an antecedent fuzzy set, C is a number of classes, $CL_j \in \{1, \ldots, C\}$ denotes a consequent class label, CF_j is a certainty grade of this rule in the unit interval [0, 1], and N is a number of initial fuzzy rules in the training phase.

To enhance interpretability of fuzzy rules, linguistic variables in fuzzy rules can be used. Each variable x_i has a linguistic set U= {L, ML, M, MH, H}. Each linguistic value of x_i equally represents 1/5 of the domain [0, 1]. Following the quantization criterion, we can consider PCPs to be regulated according to a qualitative level. For example, x_i is Low for down-regulated PCPs; x_i is Medium for neutral PCPs; and x_i is High for up-regulated PCPs. An antecedent fuzzy set $A_{ji} \in A_u$ where A^u denotes a set of subsets of U. Examples of linguistic antecedent fuzzy sets are shown in Fig. 2.

In the training phase, all the variables CL_j and CF_j are treated as parametric genes of GA (GA-genes) encoded in chromosomes of GA (GA-chromosomes) and their values are obtained using IGA. The following fuzzy reasoning method is adopted to determine the class of an input pattern $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ based on voting using multiple fuzzy if–then rules:

Step 1: Calculate score $S_{\text{Class } v}(v = 1, \ldots, C)$ for each class as follows:

$$S_{\text{Class } v} = \sum_{\substack{R_j \in FC \\ CL_j = Class\ v}} \mu_j(x_p) CF_j,$$

$$\mu_j(x_p) = \prod_{i=1}^{n} \mu_{ji}(x_{pi}), \tag{2}$$

where FC denotes the fuzzy classifier, the scalar value and $\mu_{ji}(\cdot)$ represents the membership function of the antecedent fuzzy set A_{ji}.

Step 2: Classify x_p as the class with a maximal value of $S_{\text{Class } v}$.

3) *Fitness function*

We define the fitness function $Fit()$ of IGA for designing iPPC as follows:

$$\max Fit(FC) = ACC - W_r N_r - W_f N_f \tag{3}$$

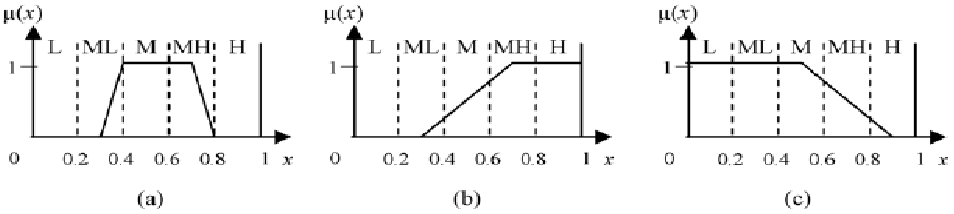


Figure 2. Examples of an antecedent fuzzy set A_{ji} with linguistic values (L: low, ML: medium low, M: medium, MH: medium high, H: high): (a) A_{ji} represents {ML, M, MH}; (b) A_{ji} represents {ML, M, MH, H}, i.e., not Low; (c) A_{ji} represents {L, ML, M, MH, H} or ALL.

where W_{r} and W_{f} are positive weights. In this study, the fitness function is used to optimize the three objectives in the following order: to maximize the accuracy rate *ACC* of correctly classified training patterns, to minimize the number N_{r} of fuzzy rules, and to minimize the number N_{f} of selected PCPs. Generally, the final number of fuzzy rules is smaller than 10. Therefore, we set $W_{\mathrm{r}} = 0.1$ to ensure that classification accuracy has the first priority to be optimized. When the two objectives *ACC* and N_{r} are simultaneously optimized for DNA-Binding data, the best number of used genes is almost determined. Hence, a very small value 0.001 is set to W_{f}. The sensitive analysis about the different settings of W_{r} and W_{f} can be referred to [10].

4) *GA-chromosome representation*

A GA-chromosome consists of control GA-genes for selecting useful genes and significant fuzzy rules, and parametric GA-genes for encoding the membership functions and fuzzy rules. The control GA-genes comprise two types of parameters. One is parameter r_j, j=1, . . ., N, represented by one bit for eliminating unnecessary fuzzy rules. If $r_j = 0$, the fuzzy rule R_j is excluded from the rule base. Otherwise, R_j is included. The other is parameter f_i, i=1, . . ., n, represented by one bit for eliminating useless genes. If $f_i = 0$, the gene *xi* is excluded from the classifier. Otherwise, x_i is included. The parametric GA-genes consist of three types: $V_{ji}^{k} \in [0, 1]$, $k = 1, \ldots, 5$, for determining the antecedent fuzzy set A_{ji}; CL_j for determining the consequent class label of rule R_j; and $CF_j \in [0, 1]$ for determining the certainty grade of rule R_j; where j=1, . . ., N and i=1, . . ., n. A rule base with N fuzzy rules is represented as an individual. The number of encoding parameters to be optimized is equal to $Np = n+3N+5Nn$. A GA-chromosome representation uses a binary string for encoding control and parametric GA-genes. There are eight bits for encoding one of parameters V_{ji}^{k} and CF_j. Since each fuzzy region defines a fuzzy rule, the initial setting of N is independent of n but dependent on the number of fuzzy regions. Generally, N is set to the maximal number of possible fuzzy regions. In this study, N=3C. The design of an efficient fuzzy classifier is formulated as a large parameter optimization problem. Once the solution of IGA is obtained, an accurate classifier with a compact fuzzy rule base can be derived.

XII. Results

The parameter settings of IGA from Ho et al. (2004a) are $N_{\mathrm{pop}} = 20$, $P_{\mathrm{c}} = 0.7$, $P_{\mathrm{s}} = 1-P_{\mathrm{c}}$, $P_{\mathrm{m}} = 0.01$, and $\alpha = 15$. Because the search space of optimal design of iPPC is proportional to the number N_p of parameters to be optimized, the stopping condition is suggested to use a fixed number $100N_p$ of fitness evaluations (Ho et al., 2004a)

A. *Performance*

The dataset all the domains/sequences have a variable length *l.* A sequence forms an *l*-dimensional profile where the value of each amino acid is obtained from the AAindex database for encoding a specific physicochemical property. The *l*-dimensional profiles are transformed into vectors with the same constant length L for utilizing classifier. The transformation can be any known effective representation provided that the L features can effectively classify the *l*-dimensional profiles of positive and negative sequences. The simplest feature is the mean of the profile that L=1. Therefore, the sequences with *m* properties are represented as an *m*-dimensional feature vectors.

The training dataset DNAset with *m*=22 properties are represented as a 22-dimensional feature vectors. This 22 physicochemical properties is pre-identified by Auto-IDPCPs. The set of 22 PCPs is described in table 1. The training accuracy is 87% and the testing accuracy is 70%. Finally, all values of the feature vectors are normalized into [0, 1] to apply iPPC.

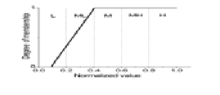
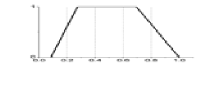
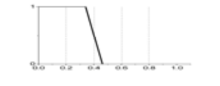
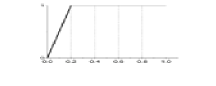
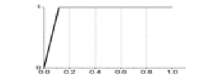
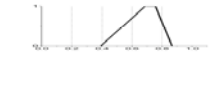
Because of the non-deterministic characteristic of GA, the experimental results are the average values of 30 independent runs. In each run, we can obtain a fuzzy classifier with the accuracy rate ACC, the number N_{r} of fuzzy rules, and the number N_{f} of selected PCPs. Using the optimal results, the test dataset DNAiset is applied to perform. The training results and testing results are shown in Table 2. The top six of high selected frequency PCPs in the 30 runs are shown in Table 3.

Table 2- The average values of 30 independent runs of the proposed iPPC.

DNAset				DNAiset	
Mean ACC%	Overall ACC%	*Avg.* N_{f}	*Avg.* N_{r}	Mean ACC%	Overall ACC%
81.39%	80.88%	2.97	2.57	67.27%	66.08%

Table 3- The top six of high selected frequency PCPs in the 30 runs.

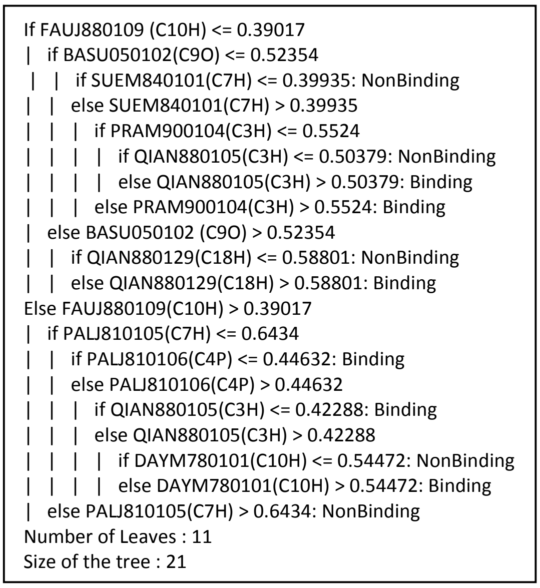
Frequency	Feature No.	AAindex No.
26	86	FAUJ880109
10	274	QIAN880117
6	255	PRAM900104
6	286	QIAN880129
6	513	JACR890101
5	91	FINA770101

	FAUJ880109 (C9H)	FINA770101(C10A)	Class	CF
R1			0	0.290
R2			1	0.325
R3			0	0.992

B. Comparison with decision tree

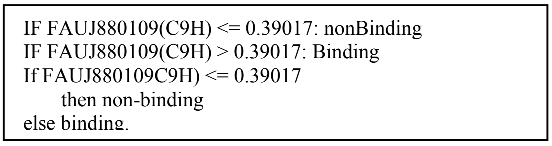
We random select one run result from iPPC independent 30 runs. The training ACC is 81%, the number N_f of selected PCPs is 2, the number N_r of fuzzy rules is 3, and testing ACC is 79%. The selected 2 PCPs are FAUJ880109 (86) and FINA770101 (91).

Using J48 in Weka3-6-4, the decision trees are built form a set 22 PCPs and the selected 2 PCPs (FAUJ880109 (86) and FINA770101 (91)) which are shown in Fig. 3(a) and Fig.3(b), respectively. The performance of decision trees, training accuracy of the 22 PCPs is 77.16% and training accuracy of the 2 PCPs is 77.67%. The decision value is fixed float value and is not easy to understand.



```
If FAUJ880109 (C10H) <= 0.39017
|  if BASU050102(C9O) <= 0.52354
 |  |  if SUEM840101(C7H) <= 0.39935: NonBinding
|  |  else SUEM840101(C7H) > 0.39935
|  |  |  if PRAM900104(C3H) <= 0.5524
|  |  |  |  if QIAN880105(C3H) <= 0.50379: NonBinding
|  |  |  |  else QIAN880105(C3H) > 0.50379: Binding
|  |  |  else PRAM900104(C3H) > 0.5524: Binding
|  else BASU050102 (C9O) > 0.52354
|  |  if QIAN880129(C18H) <= 0.58801: NonBinding
|  |  else QIAN880129(C18H) > 0.58801: Binding
Else FAUJ880109(C10H) > 0.39017
|  if PALJ810105(C7H) <= 0.6434
|  |  if PALJ810106(C4P) <= 0.44632: Binding
|  |  else PALJ810106(C4P) > 0.44632
|  |  |  if QIAN880105(C3H) <= 0.42288: Binding
|  |  |  else QIAN880105(C3H) > 0.42288
|  |  |  |  if DAYM780101(C10H) <= 0.54472: NonBinding
|  |  |  |  else DAYM780101(C10H) > 0.54472: Binding
|  else PALJ810105(C7H) > 0.6434: NonBinding
Number of Leaves : 11
Size of the tree : 21
```

(a)



```
IF FAUJ880109(C9H) <= 0.39017: nonBinding
IF FAUJ880109(C9H) > 0.39017: Binding
If FAUJ880109C9H) <= 0.39017
      then non-binding
else binding.
```

(b)

Figure 3. The decision trees are built form (a) a set 22 PCPs and (b) the iPPC selected 2 PCPs. Cid: clustering id, A: Alpha and turn propensities. B: Beta propensity. C: Composition. H: Hydrophobicity. P: Physicochemical properties. O: Other properties.

Fig. 4 shows an example of iPPC using the 2 PCPs with 3 rules. The classifier has three fuzzy rules using two PCPs FAUJ880109(C9H) and FINA770101(C10A), where The training ACC = 81% and testing ACC = 79%.

Figure 4. Fuzzy rules of the selected 2 PCPs, the training ACC is 81% and testing ACC is 79%. 0: binding, 1: non-binding.

Using the selected 2 PCPs, the proposed iPPC can obtain rule-based. The fuzzy rules are linguistically interpretable as follows:

R1 If FAUJ880109(C9H) is all and FINA770101(C10A) is all , then DNA is binding.(CF=0.290)

R2 If FAUJ880109(C9H) is {low , medium low , medium } and FINA770101(C10A) is all , then DNA is non-binding.(CF=0.325)

R3 If FAUJ880109(C9H) is all and FINA770101(C10A) is {medium low , medium , medium high , high} , then DNA is binding.(CF=0.992)

XIII. Conclusion

This paper proposes an interpretable physicochemical property classifier (named iPPC) with an accurate and compact fuzzy rule base using a scatter partition of feature space for DNA-binding data analysis. In designing iPPC, the flexible membership function, fuzzy rule, and physicochemical properties selection are simultaneously optimized. The obtained fuzzy rules are easy to interpret and analyze DNA-binding domains for biologists.

References

[1] M Gao, J Skolnick, ”A threading-based method for the prediction of DNA-binding proteins with application to the human genome.” *PLoS Comput Biol* 2009, 5(11):e1000567.

[2] D. Lejeune, N. Delsaux, B. Charloteaux, A. Thomas, R. Brasseur, “Protein-Nucleic Acid Recognition: Statistical Analysis of Atomic Interactions and Influence of DNA Structure,” *PROTEINS: Structure, Function, and Bioinformatics*, no. 61, 2005, pp. 258-271.

[3] R.A. O'Flanagan, G. Paillard, R. Lavery and A.M. Sengupta, “ Non-additivity in protein-DNA-binding.” *Bioinformatics*, 21, 2005, pp. 2254-2263.

[4] E. W.Stawiski, L. M. Gregoret, Y. Mandel-Gutfreund, "Annotating nucleic acid binding function based on protein structure" J. Mol. Biol., no. 326, 2003, pp. 1065–1079.

[5] D. C. Chan, D. Fass, J. M. Berger, “Core Structure of gp41 from the HIV Envelope Glycoprotein,” *Cell*, vol. 89, April 18, 1997, pp. 263–273.

[6] M. Kumar, MM Gromiha, GP Raghava, “Identification of DNA-binding proteins using support vector machines and evolutionary profiles” *BMC Bioinformatics* 2007, 8:463.

[7] S-J Ho, C-Y Chang, L-T Huang , S-F Hwang, and S-Y Ho, “Acquisition of Rule-based Knowledge for Analyzing DNAbinding Sites in Proteins,” Conference: Infoscale, June 6-8, 2007, Suzhou, China.

[8] H-L Huang, I-C Lin, Y-F Liou, C-T Tsi, K-T Hsu, W-L Huang, S-J Ho, S-Y Ho, ” Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties,” BMC Bioinformatics, 2010 (Accepted)

[9] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, “AAindex: amino acid index database,” progress report 2008. *Nucleic Acids Res* 2008, 36(Database issue):D202-205.

[10] S-Y Ho, L-S Shu, J-H Chen, “Intelligent evolutionary algorithms for large parameter optimization problems.” *Ieee T Evolut Comput* 2004, 8(6), pp.522-541.

[11] X. Yu, J. Cao, Y. Cai, T. Shi, Y. Li, “ Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with supportvector machines,” J Theor Biol, no. 240, 2006, pp.175-184.

[12] Z. Michalewicz, D. Dasgupta, R.G. Le Riche, M. Schoenauer, “Evolutionary algorithms for constrained engineering problems.” Comput. Ind. Eng. 30 (4), 1996, pp. 851–870.

國科會補助計畫衍生研發成果推廣資料表

日期:2011/10/31

國科會補助計畫	計畫名稱: 以蛋白質序列物化特性為特徵的蛋白質激kinase-specific磷酸化位置預測方法與分析
	計畫主持人: 黃慧玲
	計畫編號: 99-2221-E-009-137- 學門領域: 生物資訊
無研發成果推廣資料	

99年度專題研究計畫研究成果彙整表

計畫主持人：黃慧玲	計畫編號：99-2221-E-009-137-
計畫名稱：以蛋白質序列物化特性為特徵的蛋白質激kinase-specific磷酸化位置預測方法與分析	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	2	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	4	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	2	8	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

蛋白質的磷酸化是在蛋白質轉譯修飾中很重要機制，在調控基本進行過程例如新陳代謝、訊號傳遞、細胞分化和細胞膜穿透性等扮演重要角色。因此預測蛋白質的磷酸化作用位置是非常重要議題。能知道蛋白質磷酸化位置，就可以測出蛋白質功能。蛋白質常常被各式各樣的蛋白激脢(ProteinKinase) 磷酸化。用蛋白質一級序列的電腦補助預測磷酸化的地方與它們特定激脢，可以提供一個關鍵的第一步選擇，來減少候選的數目，降低昂貴的實驗成本。

近年來有許多研究用蛋白質序列資料作預測蛋白質磷酸化位置系統，為提昇正確率，許多能說明氨基酸環境特性的方法，也紛紛輔助預測。本計劃首先將蛋白質序列轉換成氨基酸物理化學特性、運用繼承式智慧型基因演算法(IBCGA)與支援向量回歸(SVR)與支援向量機(SVM)選取重要氨基酸的物化特性、研究可接觸性胺基酸物理化學特性有助於蛋白激脢磷酸化位點預測。提出一個以蛋白質序列物化特性為特徵的蛋白質激酶磷酸化位點預測系統，更精確預測位點，並且提供可以解譯蛋白質激酶磷酸化位點的物化特性知識。