應用關聯法則於辨識蛋白質交互作用之類別

郭煌政
國立嘉義大學資訊工程學系
hckuo@mail.ncyu.edu.tw

楊哲維
國立嘉義大學資訊工程學系
s0980402@mail.ncyu.edu.tw

摘要

最近幾年來，有越來越多關於辨識蛋白質交互作用類別的研究結果或論文被發表。藉由研究蛋白質複合體內部的蛋白質識別的作用原理，可以幫助進一步確定蛋白質上每個氨基酸所蘊含的生物資訊上的意義。因此，藉由使用辨識的蛋白質複合體以及非辨識的蛋白質複合體上的每個氨基酸的屬性資料，我們可以探勘出一些在蛋白質交互作用的類別上為有意義的規則資料。在資料探勘的領域上，關聯法則探勘是一個眾所周知用來發掘潛在的關聯關係的方法。本文提出了一種分類方法與關聯法則的蛋白質識別的基礎上，應用於各個氨基酸的蛋白質結合位面。可以成功地分類辨識蛋白質與非辨識蛋白質，在使用留一驗證的方法驗證可以達到平均準確率 95%。

關鍵詞：蛋白質、交互作用、分類、關聯法則

應用關聯法則於辨識蛋白質交互作用之類別

壹、緒論(Introduction)

在最近幾年，在分類辨識蛋白質此一研究主題上有許多類似的研究應用了不同的演算法。例如，基因演算法（GA）是近幾年被提出從而逐漸被使用的演算法，與其類似的演算法還有基因規劃法（GP）。兩者都是以演化論的進化觀點為基礎，兩者的提出都是試圖在關於預測未知類別的蛋白質的問題上能夠有突破性的發展 (Lan et. al 2009; Nanni & Lumini 2008; Yu et. al 2010; Zhang et. al 2007) 。不過，基因演算法與基因規劃法仍有一些待改善空間使得他們仍可有更進一步的可能。

首先，基因演算法以及基因規劃法所採用的數據資料為每個蛋白質複合體的各種生物資訊上的各種資料，需要使用非常高精確度的生物實驗儀器方可測量取得。雖然現今有很多的線上的網路資料庫網站。

然而，這些各式不同的資料庫可能會採用自己所定義的資料格式作為儲存格式。此外，即便是在相同的實驗環境下使用相同的儀器進行量測，資料的儲存格式也會因人而異。有的資料庫可能會採用較為大眾所使用的格式，有些資料庫可能會自行創建一個新的資料格式，在這種情形下，想要從各個線上資料庫之中截取所需要的特定資料的話，必須熟知各種格式的定義。

其次，基因演算法以及基因規劃法在實際的預測應用上是採取逐次漸進的方法。換句話說，這種以進化為基礎的方法可能會受限於資料上的演化極限。以演化的過程中，可能在趨近某個峰值時，就朝著該峰值收斂，而無法朝下一階段繼續進行演化。在此一情況下，將會顯見的影響預測方法。

此外，也有許多不同於前面提到的應用方法被提出。例如，支持向量機（SVM），類神經網絡（NN），以及其他應用機器學習的方法使用前述的那些特徵值或資料來進行分類蛋白質與蛋白質之間的交互作用的類別 (Nanni & Lumini 2008) 。

同樣的，在應用支持向量機以及類神經網路作為分類方法時，在演算法的設計結構上仍存在一些可能會影響結果的問題會發生。例如，若是使用支持向量機，則高維度的資料結構常常會影響到真正的重點資料的發掘，即便採取了相關的預防措施，也只能降低影響程度，要完全的避免仍是相當困難的。

近期也有一些研究是應用關聯法則的預測方法進行此一類別的研究。其中作為主要的分類方式，關聯法則之中，早期較廣為人所知的演算法就是 A-priori 演算法。利用 A-priori 演算法來產生關聯規則，並以該群關聯規則來進行蛋白質的功能的分類，此一方法仍有可供借鏡之處。此外，也有一些研究會利用關聯法則來預測蛋白質的辨識型態 (He & Loh 2010) 。

此外，在某些研究中，會先將資料以關聯法則進行處理，再將關聯法則所產生的結果作為一種新型態的資料輸入其他應用機器學習的方法。如，GP，GA，SVM 和 NN 中，再進行下一階段的實驗，以達到增加資料變化性的目的 (Zhang et. al 2010) 。

簡而言之，關聯法則可以從大量混亂的資料中找出有意義的資訊。在應用於蛋白質辨識的分類時，最重要的就是有限的資料數量找出有意義的關聯法則來作為辨識蛋白質分類的依據 (Bock & Gough 2001) 。

貳、相關的研究與探討

在討論研究方法之前，在此先簡單的介紹蛋白質辨識的相關資訊。一個蛋白質是由二十種氨基酸以不同的數量所組合而成，一個蛋白質複合體通常是由兩個或更多的蛋白質所組合而成。而本研究中，我們所討論的蛋白質與蛋白質之間的交互作用，是著眼於蛋白質複合體上的結合位面。所謂的蛋白質複合體上的結合位面指的是蛋白質複合體中兩個蛋白質互相結合的詳細細部部分，其後所討論的資料部分均是指此一部分的資料，包含有：親水性，疏水性，電性，極性等各式物理化學性質。當談到蛋白質交互作用時，這些物理化學性質是結合位面上我們可獲取的有意義的資訊 (Dohkan et. al 2004; Wang et. al 2006) 。這是時下關於蛋白質複合體一個重要的研究問題。此外，為了解說上的區分，根據蛋白質複合體的相對位置，我們可以將結合位面粗略分為凹面的部分以及凸面的部分。

近幾年，在蛋白質辨識的相關研究已經有許多的演算法被提出，單一的演算法的準確率已經不足以滿足要求了。因此，藉由結合多種的資料探勘的相關演算法的技術，近期的研究方法在變得越加複雜的同時，其所引用的各種基本的資料探勘技術的數量也隨之增加。而隨著應用的技術數量的增加，其研究結論的準確率也隨之緩步提升 (Bock & Gough 2001; Cho & Zhang 2010; Liu et. al 2001; Nanni & Lumini 2008; Thabtah & Cowling 2007; Zhang et. al 2010) 。

舉例來說，基因演算法和基因規劃法兩項演算法最初的提出都是用於模擬生物的進化。然而，近期有許多的論文都是把基因演算法以及基因規劃法作為數據挖掘技術來進行其他目的的研究。將一項技術應用於不同的領域，是近期許多研究的常見方法，此一概念在本文中也是如此運用，關聯法則被視為一種對於分類蛋白質複合體的技術。

關聯法則在最近十年已被廣泛應用於生物信息學上。作為一個相對年輕的科學技術，當研究所討論的是極大量的資料時，關聯法則常被提出作為相關概念的延伸基礎 (Cho & Zhang 2010; Leung et. al 2010) 。關聯法則最常舉的例子就是超市營銷概念中的啤酒與尿布此一案例，該案例就是討論大量資料時應用關連法則的實例。並且，在本文中，我們採取關聯法則作為我們提出方法的基本架構。

目前也有許多應用關聯法則作為分類方法依據的相關研究曾被發表，下列為相關之論述以及概念。

其中，在 He Cong (2010)所提出的論文認為，當所需要分類的類別的數量增加時，不同類型之間的資料的分類將更加顯著。然而，作為結果的關聯規則所著重的卻是不同的問題。例如，關聯法則所產生的法則數量將會影響到分類實驗的結果 (Park et. al 2010) 。

此外，Fadi A Thabtah (2007) 的論文認為若是能在不同類別所分別歸納出的法則上標記上類別標籤將會有利於分類的結果，可以提高整體的效率。

參、資料與方法

（一） 資料與使用資料庫

在本研究中，我們藉由修改關聯法則的基本演算法，A-priori 演算法，以期來達成辨識蛋白質交互作用類別的研究。在本研究中，使用關聯法則的目的是尋找出可能存在的法則，可能的話，找出辨識蛋白質複合體與非辨識蛋白質複合體所各自獨有的法則。簡而言之，當比較對象為另外一類的蛋白質複合體，不同種類的蛋白質複合物之間可能存在一些其各自所特有的法則。因此，關聯法則在此僅僅作為一項工具，用來挑選那些符合要求以及限制的關聯法則。本研究最初的目的即是找出辨識蛋白質複合體與非辨識蛋白質複合體所獨有的法則。

此外，本文使用有兩種類型的實驗資料。分別為先前所提及的辨識蛋白質複合體與非辨識蛋白質複合體兩種實驗資料。由於辨識蛋白質複合體在生物學領域上的研究為主要的研究主體，因此有相對較多的實驗的進行是針對於辨識蛋白質複合體所進行的，原因在於實驗上辨識蛋白質複合體的存在確認會相對易於非辨識蛋白質複合體。

在本文中，我們的方法主要是針對蛋白質複合體其上的結合位面中的凹面部分以及凸面部分的二十種氨基酸的特徵值。本方法所採用的資料是藉由線上資料庫網站 BOND（http://bond.unleashedinformatics.com/）所取得，在經由比對整理後，整理出本研究所需要的資料。

在本文所進行的實驗中，我們所針對的蛋白質複合體是只由兩種蛋白質所結合的複合體，所以可以刪去由三種以上蛋白質相結合的蛋白質複合體以及其他不符合的資料。因此，經由篩選後，我們最終從 BOND 的資料庫取得了 71 筆的辨識蛋白質複合體和 45 筆的非辨識蛋白質複合體。

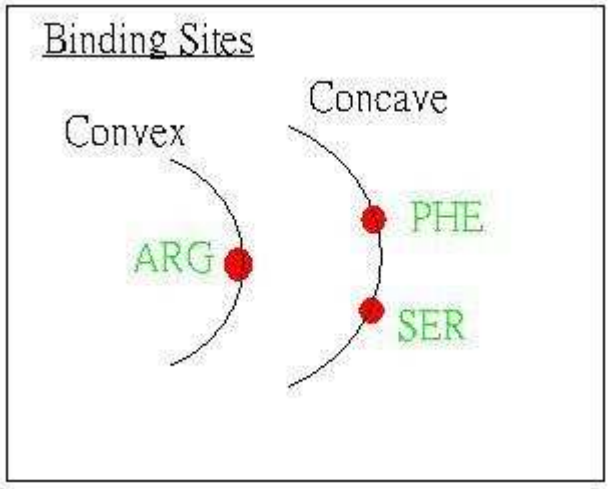


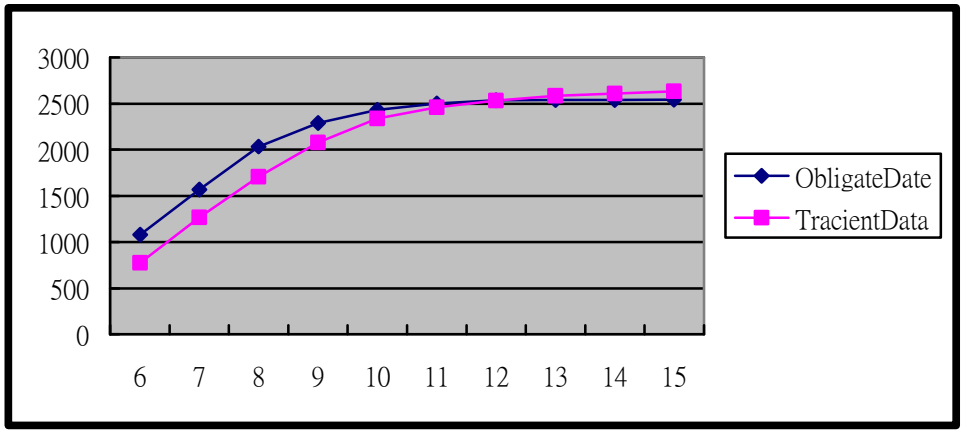
圖 1 .Binding Sites 上的氨基酸

由於胺基酸的配對要兩兩彼此在距離上相近才具有意義。在本研究中，我們引進了”Transaction”的概念，將在相鄰範圍的氨基酸劃為一組 Transaction 的簡單概念，我們可以把原本 116 筆的資料轉化為數百甚至數千筆資料。

以上圖 1 為例，以凸面部分的氨基酸 ARG 來看，假設在預設的距離內，可以在凹面部分找到氨基酸 PHE 以及氨基酸 SER，那麼，我可以將此三個氨基酸劃為一個 Transaction。

一般而言，整個結合位面的整體長度平均約為 20Å，而根據上一段中所描述的 Transaction 概念來看，若是將預設的距離設定為 20Å，則整個結合位面即視為一組資料，然而，這樣的設定對於關聯法則的應用並沒有實質的幫助。預設的距離的大小也影響著關聯法則的結果。因此，若是在由小到大逐步調整的情形下，根據預設距離的遞增，我們可以得到下表一，分別為不同距離下所產生的 Transaction 數。

表 1 不同預設距離下所產生的 Transaction 數



（二） 應用關聯法則之分類方法

我們提出的方法可分為下列三個步驟。

第一步：分別輸入的辨識蛋白質複合體和非辨識蛋白質複合體的 Transaction 資料。然後，演算法將計算對應到結合位面的兩側的二十種氨基酸分別對應另外一側的二十種氨基酸的信心值，我們稱之為 1-1 信心值。

第二步：在取得 1-1 信心值後，重複步驟一，演算法將會逐步累加的代回產生下一階段的信心值。例如，在第二輪將會產生被稱為 1-2 信心值，在第三輪將會產生被稱為 1-3 信心值，直到所有的信心已低於預先定義的門檻值。

第三步：當最後 X-Y 信心值產生後，我們將不同階段的信心值給分門別類的整理好。

在本實驗中，由於顯著差異的幅度，留一驗證法是作為用於測量準確度的最佳方案。

下列為演算法的圖示。

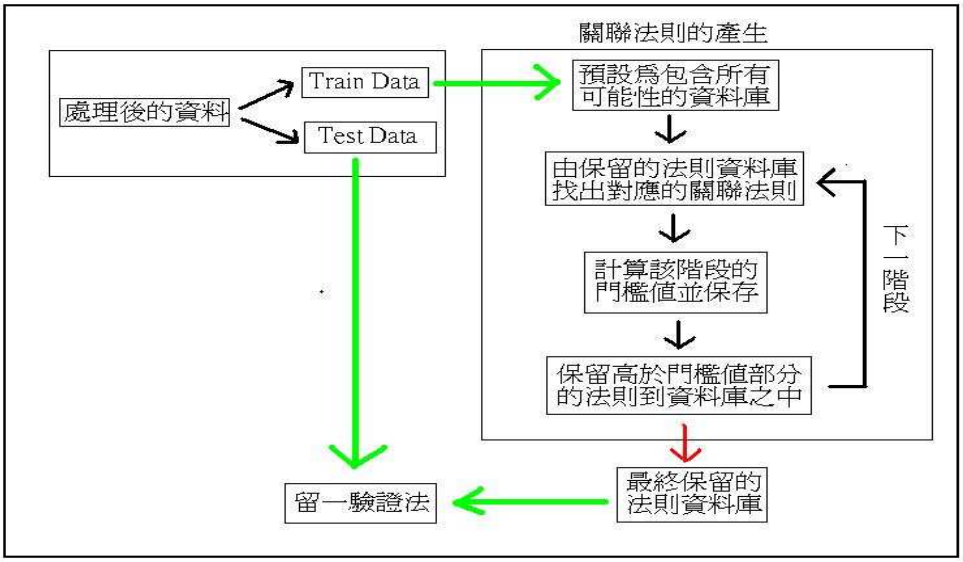


圖 2.基本運作演算法

對於那些符合的各階段信心值，我們提出的一個量測評估的方法來分類辨識蛋白質複合和非辨識蛋白質複合體。

此外，在討論蛋白質的辨識時，不同階段的信心值可能有不同程度的作用。因此，這些不同程度的信心值會給予不同程度的權重，以區分不同程度的重要性。

因此，在本文中，我們給予的權重值 1-1 信心值以 1 為基礎。而且，在 1-2 信心值給予 2 倍於 1-1 信心值的權重值。然後依此類推到其他階段的信心值。

下列為量測評估方法的圖示。

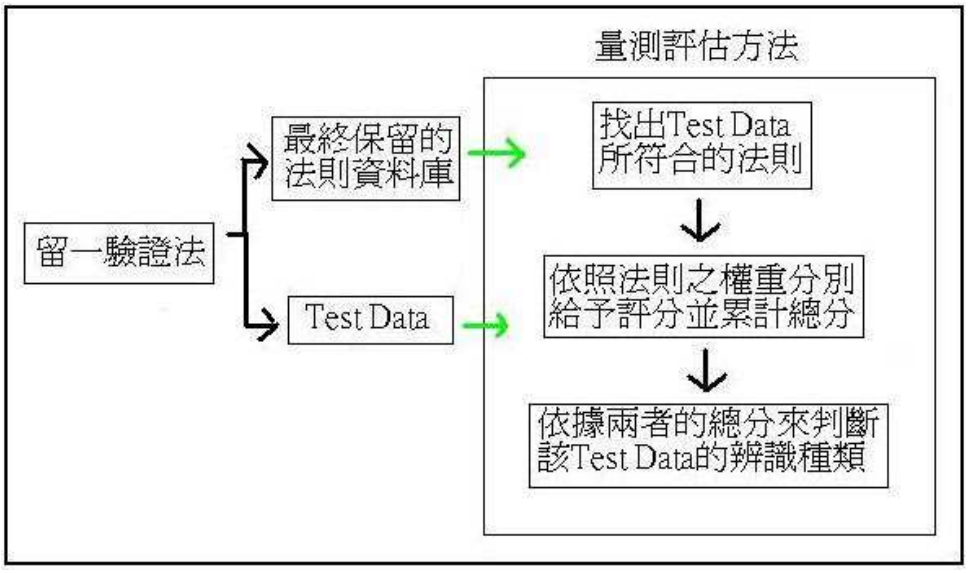


圖 3.量測評估方法

該量測評估的方法可以分為三個步驟。

第一步：根據最後 X-Y 信心值以及下推到基本 1-1 信心值，作為留一驗證輸入一筆預留的測試資料。

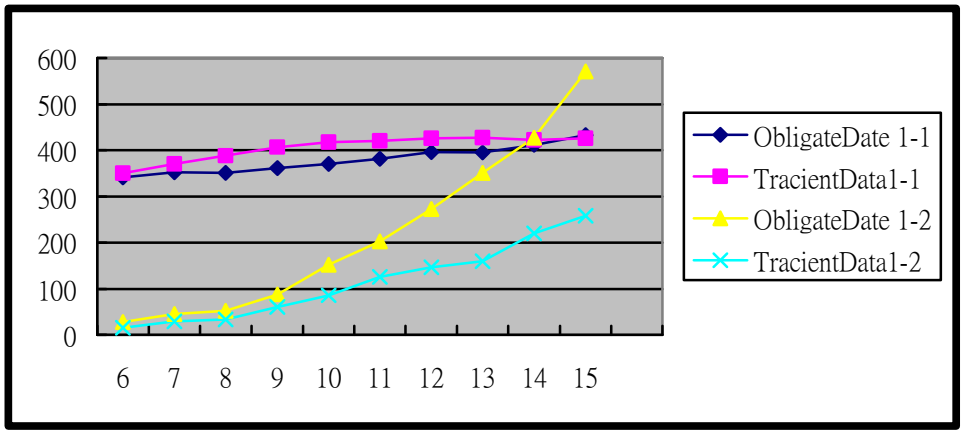
第二步：根據 X-Y 信心值的輸入資料值，找出符合該法則的信心值，再乘以根據該階段的權重值，與所以的符合法則進行累計作為輸出。

第三步：根據所有 X-Y 信心值累積的結果，較高的部分將被作為預測的種類的蛋白質複合體的交互作用的類別。

肆、結論(Conclusion)

根據圖 2 中所描述的基本運作演算法，在分別輸入不同預設距離所產生的 Transaction 資料，可以得到分別採用辨識蛋白質複合體以及非辨識蛋白質複合體所產生的關聯法則，如下表 2，其法則的數量為細部的實驗結果。

表 2. 不同預設距離下所產生的關聯法則的數量



仔細觀察表 2 可以發現到，在預設距離超過 10 原子單位以後，所產生的法則數量也開始逐步劇增，其後是否會影響到分類的準確率，仍待觀察。

表 3. 應用混用以及獨有之關聯法則的準確率

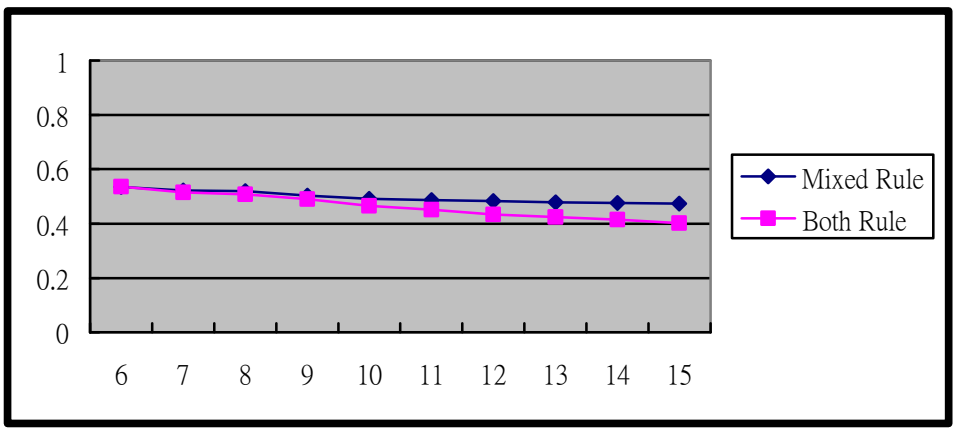
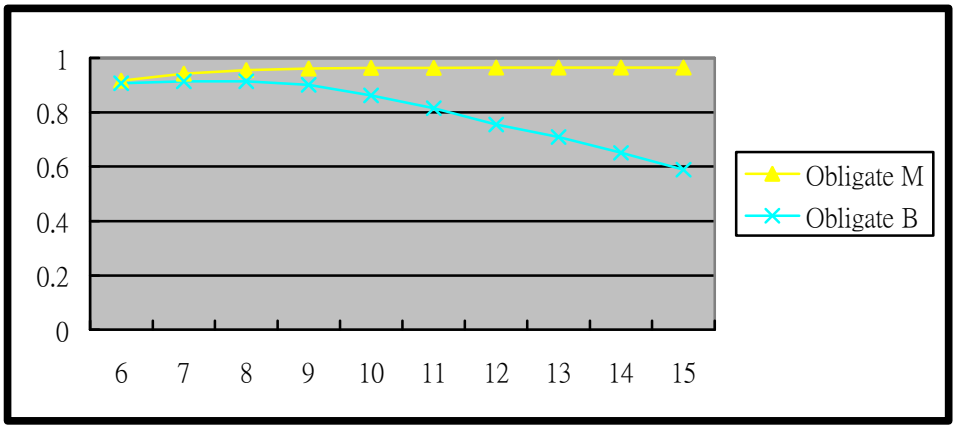


表 3 為應用表 2 所產生的關聯法則分別根據其所產生的類別進行結合，分別為兩者所獨有的法則以及兩者都具備的法則進行留一驗證法的結果。同樣的，可以發現到，在預設距離超過 10 原子單位以後，準確率已然是相當的差了。

表 4. 針對辨識蛋白質的準確率



然而，若是單純針對辨識蛋白質複合體的分類結果來看，同樣分別採用獨有法則以及兩者都具備的法則來進行分類，由上表4可以得知，雖然在預設距離超過10原子單位以後，混用法則的準確率有下降的趨勢，但整體而言，採用獨有法則的分類結果的準確率平均約落在95%附近。

參考文獻

1. J.R. Bock, D.A. Gough, “Predicting protein–protein interactions from primary structure,” Bioinformatics, Vol.17, NO.5, Pp.455-460, 2001
2. Y.R. Cho, A. Zhang, “Predicting protein function by frequent functional association pattern mining in protein interaction networks,” IEEE Transactions on Information Technology in Biomedicine, Vol.14, No.1, Pp.30-36, 2010
3. S. Dohkan, A. Koike, and T. Takagi, “Prediction of Protein–Protein Interactions using support vector machines,” Fourth IEEE Symposium on Bioinformatics and Bioengineering, Pp.576-583, 2004
4. C. He, H.T. Loh, “Pattern–oriented associative rule–based patent classification,” Expert Systems with Applications, Vol.37, No.3, Pp.2395-2404, 2010
5. M. Lan, C.L. Tan, J. Su, “Feature generation and representations for protein－protein interaction classification,” Journal of Biomedical Informatics, Vol.42, No.5, Pp.866-872, 2009
6. K.S. Leung, K.C. Wong, T.M. Chan, M.H. Wong, K.H. Lee, C.K. Lau, S.K.W. Tsui, “Discovering protein－DNA binding sequence patterns using association rule mining,” Nucleic Acids Research, Vol.38, No.19, Pp.6324-6337, 2010
7. B. Liu, Y. Ma, C.K. Wong, “Classification using association rules: weaknesses and enhancements,” Data Mining for Scientific Applications, Pp.591-602,2001

8. L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," Amino Acids, Vol.34, No.4, Pp.653-600, 2008
9. S.H. Park, J.A. Reyes, D.R. Gilbert, J.W. Kim, S.S. Kim, "Prediction of protein-protein interaction types using association rule based classification," BMC Bioinformatics, Vol.10, No.36, Pp.01-15
10. F.A. Thabtah, P.I. Cowling, "A greedy classification algorithm based on association rule," Applied Soft Computing, Vol.7, NO.3, Pp.1102-1111, 2007
11. B. Wang, P. Chen, D.S. Huang, J.J. Li, T.M. Lok, M.R. Lyu, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," Federation of European Biochemical Societies, Vol.580, No.2, Pp.380-384, 2006
12. L. Yu, Y. Guo, Z. Zhang, Y. Li, M. Li, G. Li, W. Xiong, Y. Zeng, "SecretP: A new method for predicting mammalian secreted proteins," ScienceDirect Peptides, Vol.31, No.4, Pp.574-574, 2010
13. M. Zhang, X. Gao, W. Lou, "A new crossover operator in genetic programming for object classification," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol.37, No.5, Pp.1332-1343, 2007
14. Y.P. Zhang, Y.C. Wang,L.N. Zhang, C.C. Xu, "Prediction of protein-protein interaction sites using covering algorithm,"The 5th International Conference on Computer Science and Education, Pp.334-339, 2010

Classifying Protein Interaction Type with Association Rule Mining

Huang-Cheng Kuo
Department of Computer Science and Information Engineering, National Chiayi University
hckuo@mail.ncyu.edu.tw

Che-Wei Yang
Department of Computer Science and Information Engineering, National Chiayi University
s0980402@mail.ncyu.edu.tw

Abstract

There are more and more studies in the area of classifying protein-protein recognition in recently years. By studying how protein-protein recognition works inside the protein complex, it is helpful when determining the significances of each amino acid on bioinformatics. Therefore, by using the properties of each amino acid ratio on protein-protein recognition complex and non-recognition complex, we can induce some rules that are meaningful on protein-protein recognition. In data mining, association rule mining is a well-known method for discovering the potential association relationships. In this paper, we present a classifying method with association rule on protein-protein recognition based on the contrast ratio of each amino acid in protein binding site. We successfully classify the recognition proteins with an average accuracy rate of 95% with leave-one out cross-validation.
Keyword：Protein、Interaction、Classification、Association Rule