中文
成大研發快訊　第二十八卷 第二期 - 2014年十二月十九日
[ http://research.ncku.edu.tw/re/articles/c/20141219/4.html ]

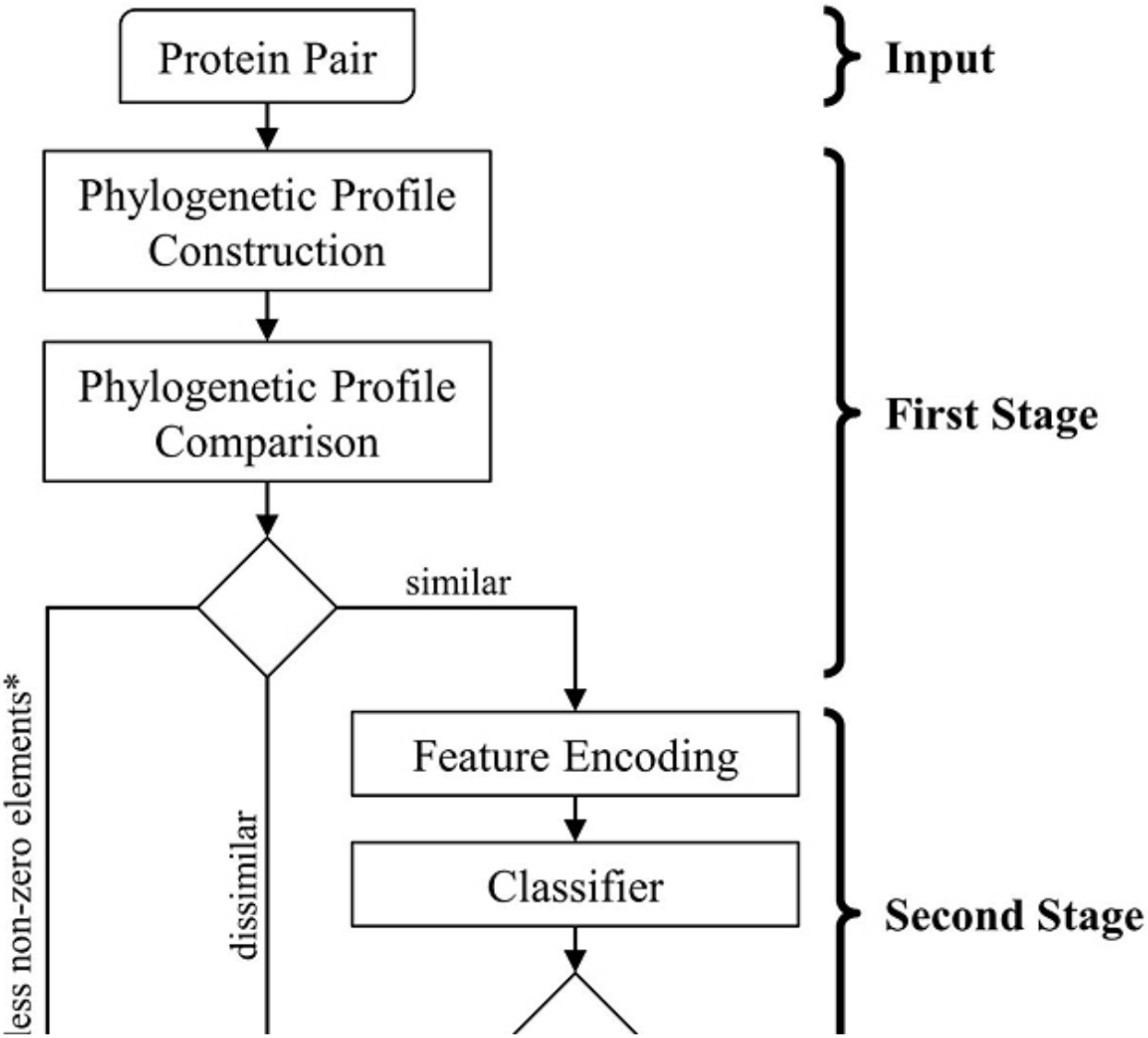
# 結合系統發育譜與機器學習技術來預測功能相關蛋白

林子文, 吳建緯, 張天豪*
國立成功大學電機工程學系
darby@mail.ncku.edu.tw

Optics Express, Vol. 18, No. 1, 165-172 (2010)

了解蛋白質的功能並聯結功能相關蛋白是系統生物學中很重要的課題。隨著基因體技術帶來了大量的資料，計算方法在其中所扮演的角色也越來越重要。系統發育譜(phylogenetic profiling)是利用共同演化特徵來預測功能相關蛋白的傳統計算方法，其缺點是只適用於原核生物，而無法用來分析老鼠、人類等真核生物。本研究提出了一套兩階段架構，使用機器學習(machine learning)技術來改善這個問題。

圖 1為本研究所提出的兩階段架構。其中第一階段的系統發育譜會濾除演化相似度較低的蛋白質對，減少第二階段機器學習所需處理的資料量。本方法在第一階段中的最大特點是一個 non-zero 過濾器(以星號表示)，它可以有效辨別系統發育譜色的可靠程度。第二階段則利用機器學習建構功能相關蛋白的數學模型，提供最終預測的結果。



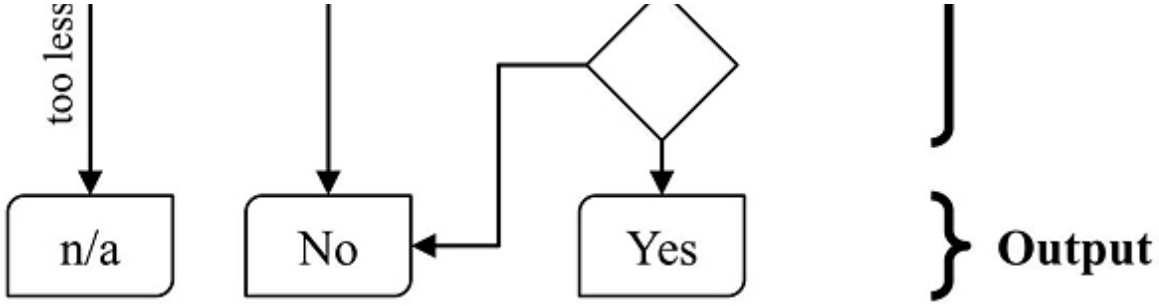



圖1本研究所提出用來預測功能相關蛋白的二階段架構

為了深入探討了non-zero過濾器以及建構系統發育譜所需要參考物種集所造成的影響，本研究分析了不同 non-zero 過濾器在不同大小、演化距離的參考物種集下，功能相關蛋白的預測效能。圖2a是使用829種原核生物的參考物種集，圖2b則為使用132種真核生物的參考物種集。在圖2a中高準度(只預測少量結果)的區域中，最佳的 non-zero 值相對較大，這說明了參考物種集的大小的確會造成影響。另一方面，圖2b中最佳的 non-zero 值則非常穩定。而且使用真核生物可以在前50名預測中達到超過90%的準確度，在前100名也還有超過80%的準確度。這些結果說明演化距離的影響又比數量更為重要。這些效能遠優於現有的方法，除此之外，這些分析也結也提供未來相關應用在選擇 non-zero 過濾器時一個很重要的參考依據。

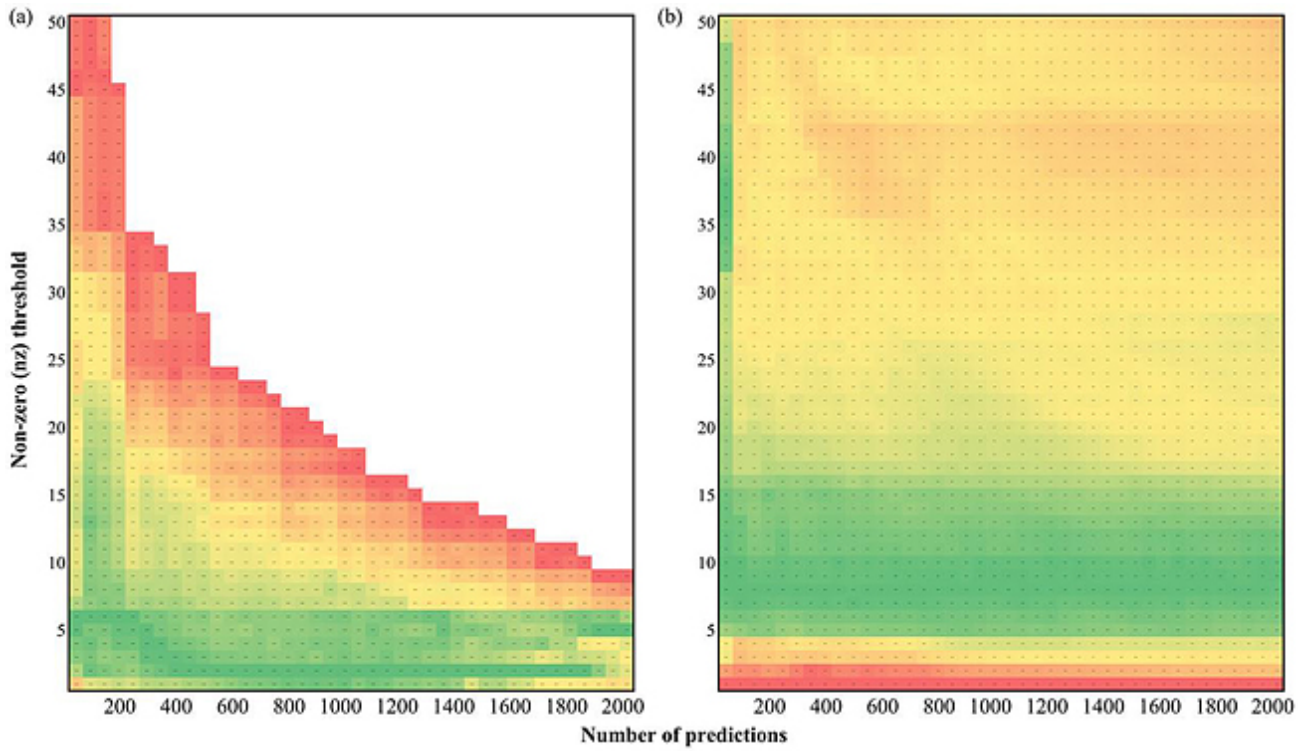



圖2預測數量(橫軸)、non-zero 值(縱軸)與預測效能(顏色)的關係圖。(a)使用829種原核生物作為參考物種集。(b)使用132種真核生物作為參考物種集。

本研究所提出的二階段架構有良好的效能並且保留了兩種類型技術各自的好處：(i)系統發育譜預測高名次的極高預測效能以及(ii)機器學習的穩定效能(在所有名次下)，這樣穩定高效的表現在實際應用時是非常重要的特性。除此之外，因為本研究所提出的 non-zero 過濾器，使得傳統系統發育譜突破只能應用於原核生物的瓶頸。這些研發成果除了幫助功能相關蛋白的研究，也是未來整合多類型技術的方法一個很重要的基礎。