姓名：黃明經

學歷：

美國匹茲堡大學博士

現職及經歷：

中央研究院生物醫學科學研究所副研究員（2000/04）

中央研究院生物醫學科學研究所助研究員

（1994/08-2000/03）

Biosym Technologies, Project leader (1991/06-1994/07)

Biosym Technologies, Postdoc (1989/06-1991/05)

著作名稱：

Leslie Y.Y. Chen, Szu-Hsien Lu, Edward S.C. Shih, and Ming-Jing Hwang (2002) "Single Nucleotide Polymorphism Mapping Using Genome-Wide Unique Sequences", Genome Research, 12: 1106-1111.

中文簡介：

互補，是生命的基礎（在DNA雙螺旋結構中，A總是跟T配成一對，G跟C則是另一對），亦是數位化的基礎（在0與1的世界裡，非0則1，非1則0）。這個或許並非偶然的關連，使得生命科學與資訊科學兩個原本不相干的學門得以經由基因體學的媒介迅速結合，孕育出當紅的生物資訊學，處理生命科學裡已經相當龐大且持續非常快速成長的資料。

在生物資訊學研究中，序列比對是最基本的演算，而BLAST則是近乎標準規格的序列比對演算方法，它主要的功能在於能夠提供快速精準的比對結果。BLAST在序列資料庫搜尋的表現亦是非常優秀，美國生物科技資訊中心（National Center for Biotechnology Information，通稱NCBI）更將這個方法建置為序列資料庫搜尋引擎，並將BLAST的各項延伸應用放置於網際網路上提供全世界做各項服務。然而，對於超大量的序列比

對，BLAST不盡理想的效率往往成為研究進展的瓶頸。舉例來說，人類基因體草圖完成後一個重要的後續工作是找出人與人之間在基因體上不同的地方，即變異(variation)。其中單核甘酸變異(Single Nucleotide Polymorphism, SNP)是許多種變異中最主要的。簡單地說，SNP所代表的就是我們所謂的「體質」。當我們把所有常見的SNP找出來後，理論上就可以透過統計以SNP將「體質」量化、分類與歸納，從而知道為什麼有人可以千杯不醉，有人卻一杯下肚就開始胡言亂語了。當然更重要的，我們期待能利用每個人的基因體型（即SNP或「體質」的型式）預測哪些人比較容易得哪些疾病及其對藥物會有何種反應等等。

這個道理雖然很簡單，困難的是，在人類基因體裏常見的SNP有好幾百萬（目前公衆的資料庫已記載超越八百萬筆的SNP序列），如何處理這龐大的資料並一一加以註解是生物資訊的一個挑戰。而註解SNP序列的首要工作便是將每一筆SNP序列找出它在基因體上的位置，這是有別於定序的定址工作，前者靠定序實驗後者則靠電腦計算。通常的作法是將每一筆SNP序列（約幾百個鹼基）與人類基因體（約30億鹼基）以BLAST比對，並以比對結果來決定該SNP在基因體上的位置，這樣的計算在一般的工作站上執行大概只要花上幾秒的時間就可完成一筆SNP的定址，但幾秒乘上幾百萬就等於整月甚至整年不眠不休的運算了。一般的實驗室是不太可能會有足夠的資源去做這樣的工作。為了解決這個問題，我們實驗室發展出一個定址的快速演算法，利用這個演算法，同樣的工作我們只利用四台個人電腦(1G Hz CPU+ 512MB RAM+ 40G HD)在不到一天的時間就全部做完了。而且我們能夠定址的SNP數量比NCBI多出5%，同時有>99%的定址是與NCBI一致的。

我們的秘訣在於完全不用序列比對，而是先找出所有在人類基因體上只出現過一次、15字元長的鹼基序列當作標記，這樣的標記我們稱之為Uni-Marker。如果SNP序列也有一些這樣的標記的話，我們就可以很快的知道這條序列是出自於人類基因體的哪一個位置。在一般情況下，標記所指出的位置會非常一致，但有時會因為SNP序列與基因體草圖不符，造成標記會有些許雜訊存在，這時可以「訊號/雜訊比」的概念來排除一些不正常的標記反應。由於使用了Uni-Marker，我們也稱這個方法叫做UM method，整個系統架構是以DNA序列在基因體的定址做為主軸，所以也可稱之為Genome Positioning System (GPS)。

這個的方法就像是在人類基因體上建立一個大型的索引表，然後靠著索引，快速的找到資料的位置；其實這樣的概念在資料庫的運作上早已行之有年，近年來發展出的基因體序列比對演算法、DNA序列資料庫搜尋系統與基因體組合系統等也都開始使用了類似的概念，在效能上也都優於使用BLAST的方法。當然，BLAST仍然有它過人之處，其延伸應用的範圍也非常廣泛，但在一些特定的問題方面，更適合的方法已不斷地被發展出來。

最近，我們更將 UniMarker 推廣成可以比對大基因體的演算法，我們發現在人與小鼠的基因體裡共有的獨特標記(UMs)，有些源自演化的共同祖先，有些則是經由突變產生。再度利用「訊號 / 雜訊比」的概念，我們成功地區分兩者，並由前者迅速且精確地推斷兩個基因體裡演化同源的區域。隨著全基因體序列數目的快速增加， UniMarker將可成為比較基因體學生物資訊研究的一利器。

評審簡評：

自從人類基因體序列解碼完成之後，將人類單核苷酸多形性 (Single nucleotide polymorphism, SNP) 序列在基因體上定位已成為找尋疾病基因的首要工作。傳統的方法是每次將一條約數百個核苷酸長度的DNA序列，利用電腦之幫助，與基因體DNA序列排列比對以找出該 DNA 序列在基因體的位置。此種方法雖然有很好的理論基礎與實用價值，但是非常的耗時費力。

本研究論文，作者黃明經先生根據在人類基因體中，含有15個核苷酸長度的特異序列(unique sequence) 可能只會出現一次，此種特異序列標記將之稱為單一標記（Uni Markers ，簡稱 UMs）。利用此單一標記可以找到單核苷酸多形性序列在基因體之位置。此種方法稱為單一標記法(Uni Markers Method)。

單一標記法只需利用個人電腦就能在極短時間內找到整個人類單核苷酸多形性資料庫中每筆序列在基因體上的位置。相對於傳統方法，不需 DNA 核苷酸序列的排列比對，即可於省下數百倍的時間內找到基因的位置。這是一項非常新穎的創意。對於未來基因體涵義的解密工作，會有很大的幫助。