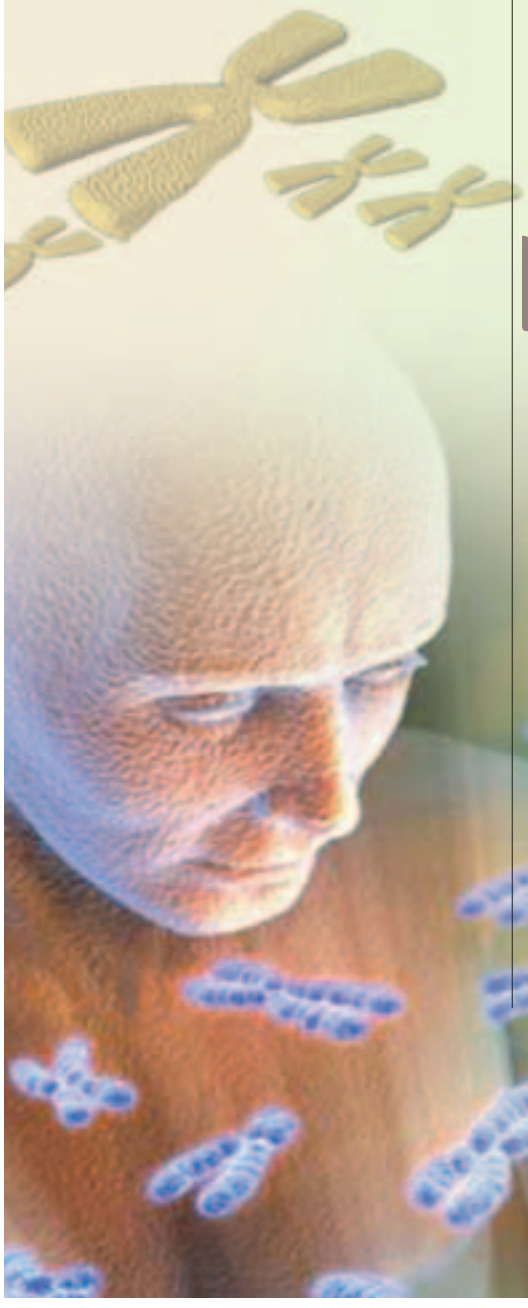
高密度片段的尋找

生物資訊學的問題重整

「生命科學」、「電腦」與「幾何」可以有怎樣的關聯呢？
「演算法」的技術很巧妙地，
把似乎沒有交集的三個領域串連在一起。

呂學一



「問題重整」（problem reduction）是資訊科學很重要的觀念與技巧，科學家遇到一個不熟悉的問題時，會把原始的問題轉換成比較熟悉的「形式」（formulation），進而藉由新形式相關領域中的工具解決原先的問題。在生物資訊研究中，這種「問題重整」的例子屢見不鮮。

「生命科學」、「電腦」與「幾何」可以有怎樣的關聯呢？「演算法」的技術很巧妙地把似乎沒有交集的三個領域串連在一起。

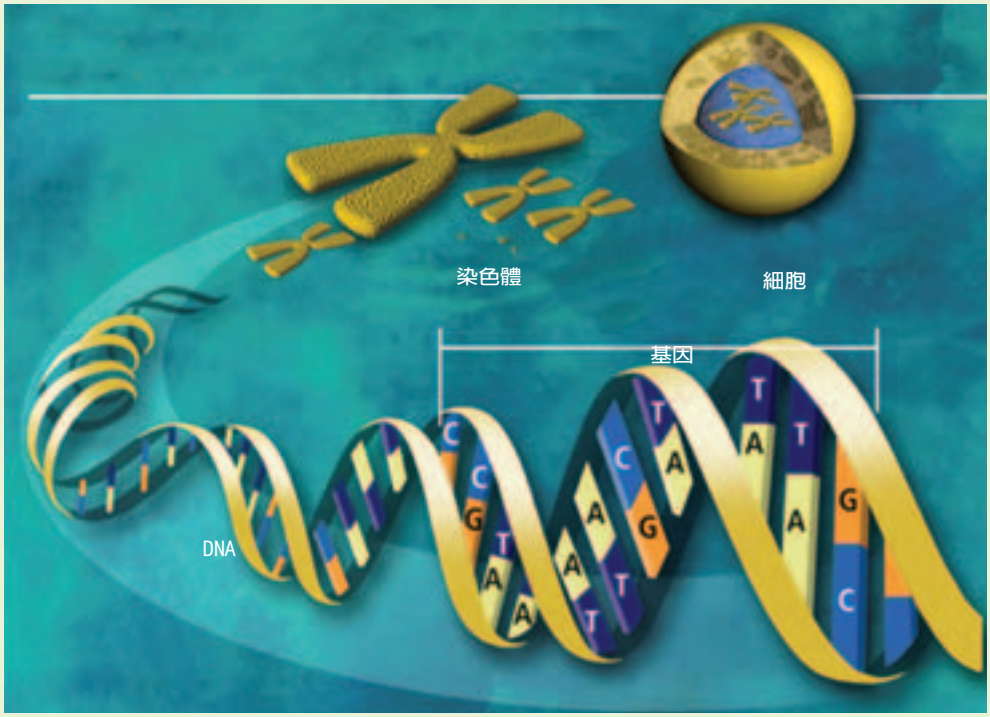
問題重整

問題重整是資訊科學很重要的觀念與技巧，科學家遇到一個不熟悉的問題時，會把原始的問題轉換成比較熟悉的形式，進而藉由新形式相關領域中的工具解決原先的問題。

生命科學

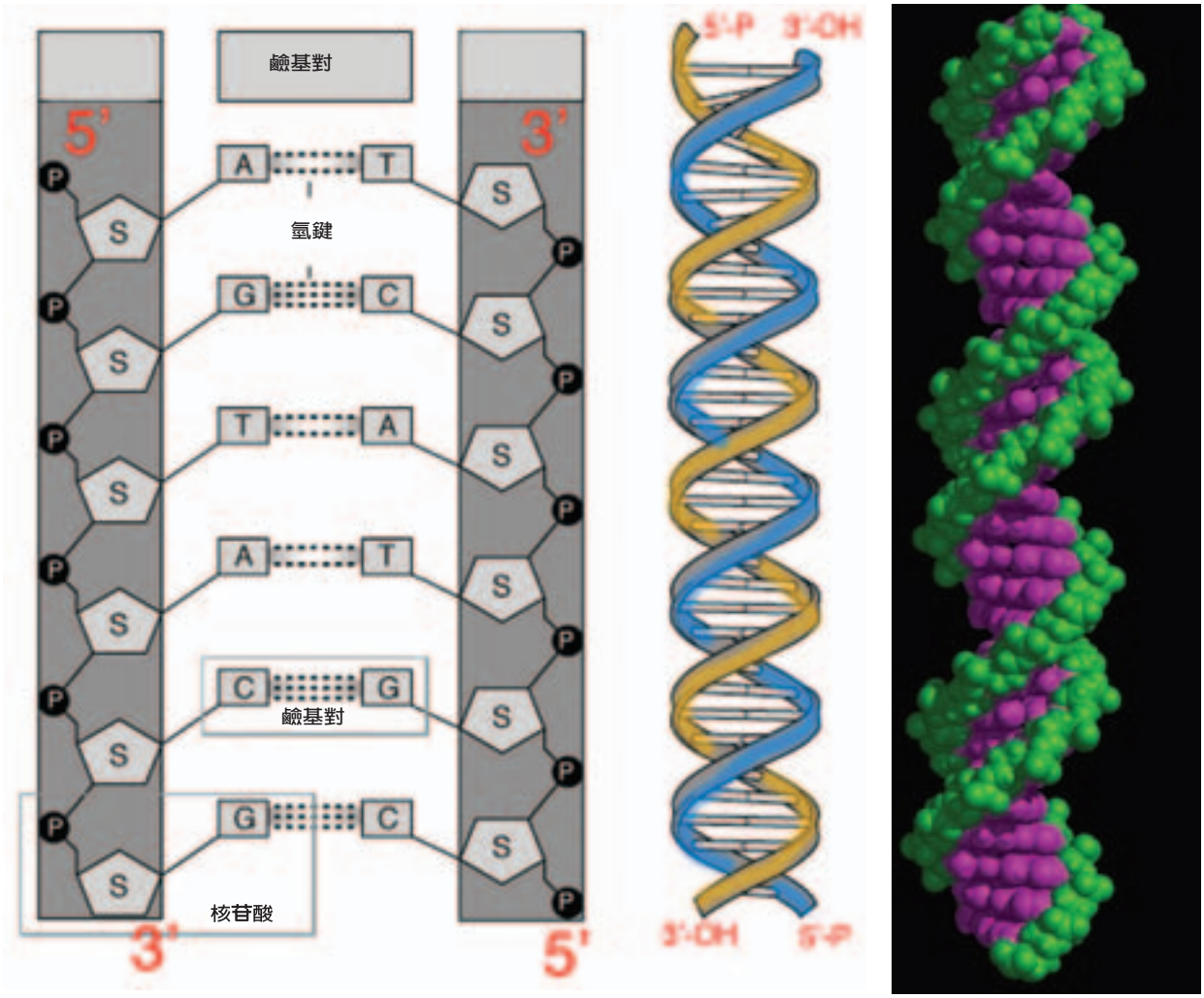
先從生命科學談起吧。

半個世紀以前，科學家對於去氧核醣核酸（DNA）的結構並沒有太多的認識，而顯微鏡可憐的放大倍數，也很難提供肉眼可見的影像。直到1949年查爾葛夫與維雪才確定DNA是由腺嘌呤（adenine，A）、鳥糞嘌呤（guanine，



http://www.paternityexperts.com/images/DNA-of-life.jpg

每個細胞有23對染色體，它們其實是由捲得很緊密的DNA所構成。DNA已被證明是遺傳的基本物質，它是由A、G、C及T四種鹼基組合而成的長鏈分子。所謂基因，就是指那些儲存蛋白質製造模具的DNA片段。



http://web.uconn.edu/gage/Media/229%20Pictures/1-DNA%20structure.JPG

華生與克里克所發現的雙螺旋結構深深地影響了過去這半個世紀的生命科學

G）、胞嘧啶（cytosine，C）、胸腺嘧啶（thymine，T）4種成分所組成。稍後薩門霍夫加入這個團隊，開始對DNA的4種成分進行定量的分析。

當時普遍猜測A、G、T、C在DNA裡的比率應該相去不遠，但3人獲得的實驗數據卻完全不是這麼回事，他們發現在不同DNA當中，A、G、T、C的比率並不相同。最有趣的是，不管4種成分的比率如何變化，A與T的數量總是非常相近，G與C的數量也幾乎相同。當時查爾葛夫甚至在論文中寫下：「我們的定量分析觀察到一個令人驚訝，但或許是毫無意義的規律性。」

幾年之後全世界才恍然大悟，原來查爾葛夫等人所觀察到的規律性，有非比尋常的重大意義。1953年華生與克里克提出DNA的「雙螺旋結構」，結構中互相纏繞的兩道DNA序列裡，A總是黏著T，而G總是與C爲伍。那篇短短兩頁不過900字的論文，深深影響過去這半個世紀生命科學的研究發展。爲此華生與克里克在1962年與威爾金斯獲得諾貝爾生理與醫學獎的殊榮。

雖然查爾葛夫等人所觀察到的規律性已經有了清楚的解釋，但是A－T與G－C的數量在不同DNA序列爲什麼會有明顯的差異，至今科學家還是有兩派意見，誰也不服誰。倒是有一些研究顯示，在G－C密度比較高的片段當中，通常會有比較豐富的生物意義。於是這裡衍生出一個電腦演算法的問題，就是怎麼在一個

http://www.time.com/time/time100/scientist/profile/watsoncrick.html

解開DNA結構之謎的華生（左）和克里克（右）

April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

http://www.lecb.ncifcrf.gov/~toms/icons/Watson.Crick.Nature.jpg

1953年華生與克里克發表於《自然》期刊有關DNA雙螺旋結構的論文

DNA序列中找到一個G－C密度最高的片段。

電腦

以上這個題目讓我們想到「電腦」。電腦的快速計算能力，這幾年成了「生命科學」相關研究的一具強力噴射引擎，使得DNA定序的進展一日千里。過去生物學家必須埋頭苦幹好幾年才能完成的實驗，如今靠著電腦的幫忙，可以在短短幾天之內完成。

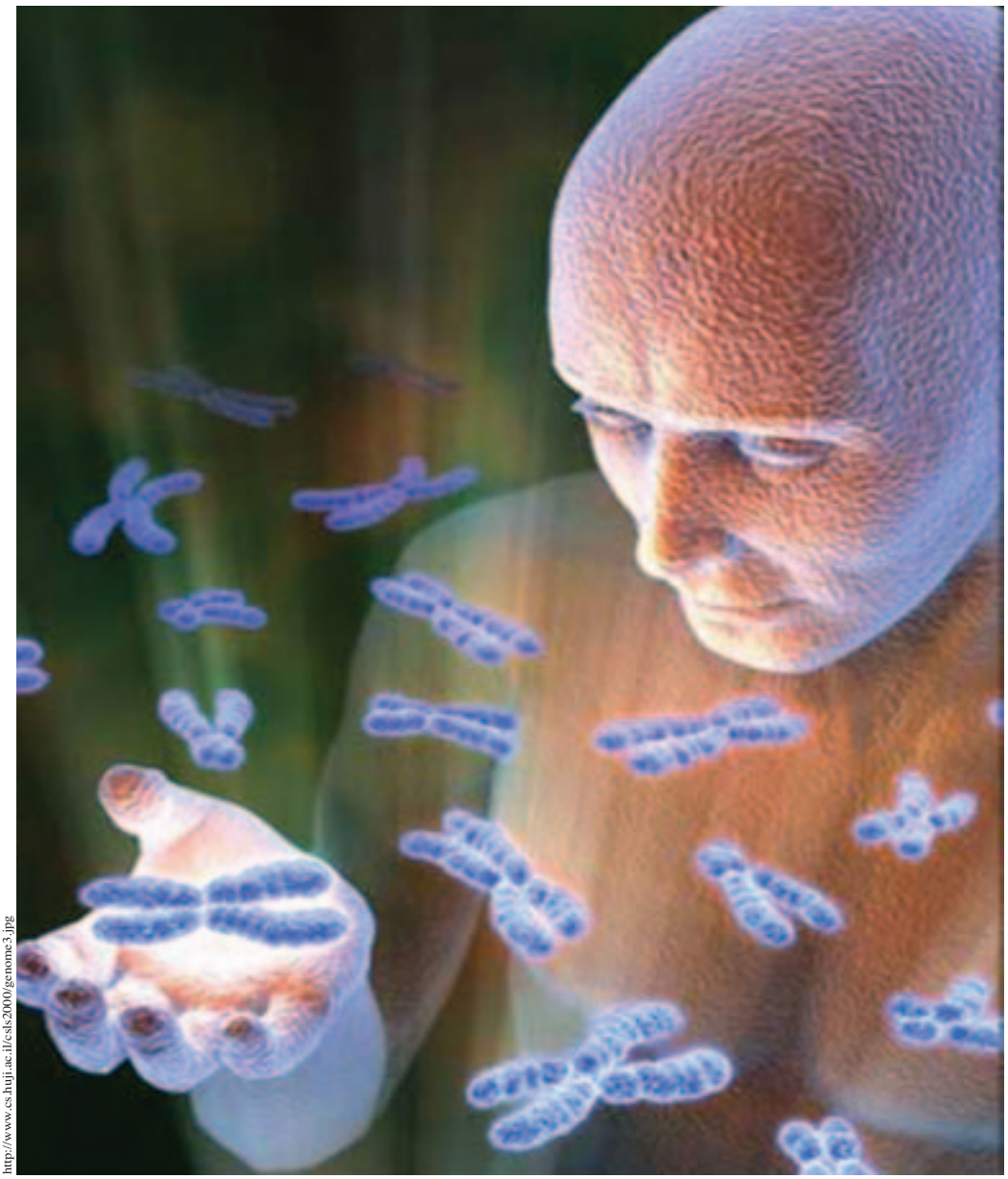
不過要讓電腦幫忙，得靠程式設計人員撰寫程式。電腦跑得快不快，跟程式寫得好不好大有關係，而一個程式寫得好不好，又跟程式背後那個解決問題的想法，也就是「演算法」（algorithm）有絕對的關聯。

比較演算法孰優孰劣有一套粗略但相當客觀的標準，就是演算法解決問題時所需要的運算時間，跟

所輸入資料的長度之間是怎樣的關係。如果是成線性的關係，這個演算法就是最佳的解題法，如果是平方的關係，這個演算法就沒有那麼受人青睞，萬一是立方的關係，這個演算法就不切實際了。

以上面提到的例子來說：我們手中有一個長度是n的DNA序列，而想要寫個程式找出這個序列當中長度不短過m，而且G－C的密度最高的片段。這個程式所需要的運算時間如果與m×n成正比（即所謂成平方關係），程式背後的演算法就還有待改進。如果程式所需要的運算時間與n成正比（即所謂成線性關係），這個程式背後的演算法便是最佳的解題方法。

這個「G－C密度最高片段」的問題，很自然地可以重整成一個「數列」的問題：輸入一個長度是n的數列，其中每個數字非0即1（A或T用0代表，G或C就用1代替），要求輸出該數列一個不短過m的片段，使得這片段的平均值爲最高。而所謂平均值就是，這個片段當中數字的和除以片段的長度。



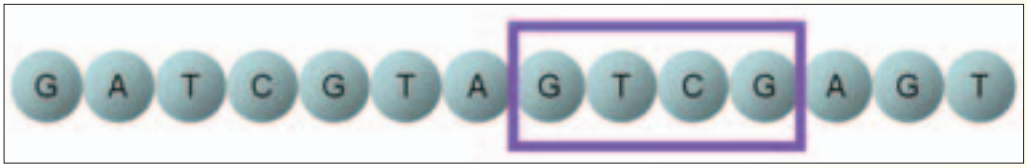
http://www.cs.huji.ac.il/csls2000/genome3.jpg

電腦提供了生物學家分析染色體所含龐大基因資訊的能力

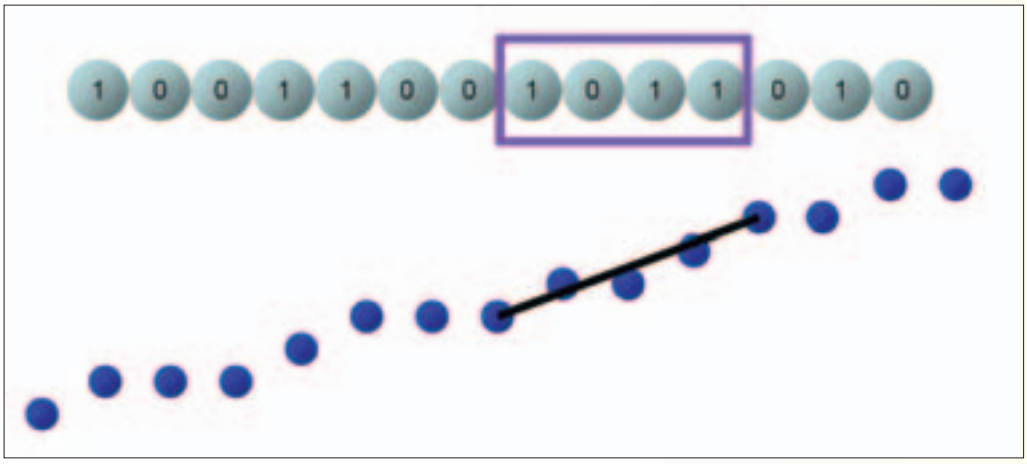
這個數列上的問題，很容易就有一個平方時間的演算法，道理很簡單，因爲長度是n的數列最多只有n^2個片段，只須挑出這個片段當中長度大於或等於m，且平均值爲最高的一個片段即可。簡簡單單就讓平方時間的演算法完成任務。不過如果想讓執行時間跟數列長度的關聯降低到線性關係，數列演算法好像沒有現成的工具可以直接套用，這時候就需要「幾何」來幫忙了。

幾何

幾何學是一



為了找出長度大於等於 3，且G－C密度最高的片段，什麼樣的演算法才是最佳的解題方法？



為了找出在平面上橫座標距離不小於 3，且可拉出斜率是最高的兩點，什麼樣的演算法才是最佳的解題方法？

門非常古老的學問，上至天文，下至地理，莫不與幾何密切相關。過去這 20 年在「計算幾何」這個領域當中，有許多問題被研究得非常透徹。舉例來說，如果給定平面上 n 個點，如何快速從這 n 個點當中挑出兩個點，使得通過這兩點的那條直線的斜率爲最大，這就是曾被深入研究過的「斜率選擇」問題。

其實上述的最高均值片段的問題，正可以「重整」成一個特別的「斜率選擇」問題。乍聽之下這兩個問題似乎毫不相干，但是底下這個轉換，說穿了一點也不稀奇。

把數列當中的 n 個數字分別對應到平面上面的 n 個點：第 i 個數字的 x 座標就是 i。至於 y 座標，爲了方便起見，我們想像有個第 0 點在平面的原點（0,0）上，也就是第 0 點的 y 座標是 0。此後第 i 個點的 y 座標，就是第 $i-1$ 個點的 y 座標加上第 i 個數字。有了這 $n+1$ 個在平面上的點，尋找平均值最高的數列片段，就變成尋找兩個 x 座標相差大於或等於 m 的兩個點，使得通過這兩個點的直線有最大的斜率。

由於「計算幾何」領域當中，有諸多現成的演算法工具可以處理各式各樣「斜率選擇」問題的變形，所以稍爲再費點心思，線性時間的演算法就唾手可得。

換個角度看

科學家在著手研究的時候，有高「智商」（IQ，intelligence quotient）固然吃香，但是高 CQ 或許更要緊。甚麼是 CQ 呢？就是「創意商數」（creativity quotient）。當年愛因斯坦用微分幾何發展出相對論，近年來不管是在生命科學研究當中引進電腦演算法，或是利用物理上「量子」的性質來解決數學上「因數分解」的大難題，背後全都是「問題重整」的創意。

您手中有甚麼懸宕已久的難題嗎？換個角度來看看吧，做個「問題重整」或許就會「山窮水盡疑無路，柳暗花明又一村」呢！ □

呂學一
台灣大學資訊工程學系