數學在分子生物學中不合理的有效性

作者：雷斯克 Arthur M. Lesk　譯者：周樹靜

作者簡介：雷斯克是美國普林斯頓大學的物理與物理化學博士，曾任職於英國劍橋大學臨床醫學院與劍橋分子生物實驗室，並在德國海德堡的分子生物實驗室肇建生物計算部門，如今他任職於賓州州立大學。雷斯克相信由數學而物理、化學、生物的知識層序。

重點摘要：

▶基因是生物的基本藍圖，存於 DNA 的鹼基序列中，再靠著「遺傳密碼」直接轉譯成蛋白質的胺基酸序列，蛋白質於是自發折疊成天然的三維結構。序列和空間結構都可用數學的語言描述。

▶分子生物學運用序列比對和結構疊合這兩項數學工具，試圖描述並分類序列或結構。希望以演化做為指導原則，描述並預測蛋白質的序列、結構、功能彼此之間的內在關係。

▶從胺基酸序列決定蛋白質的三維結構是大自然的演算法，但受制於生物學的特性，從物化原理預測蛋白質結構並不容易。這個熱門又重要的問題，應該從基本原理做起，或採用實用而有效的預測，考驗數學的應用能耐。

我的文章標題仿效威格納的知名文章〈數學在自然科學中不合理的有效性〉[1] 當然，其中的諷刺之處在物理學與分子生物學正好相反。在物理學，數學顯然是有效的，物理學家所立足的巨人肩膀許多是數學家，令人意外的是威格納竟然說這是不合理的。但在分子生物學裡，數學的合適角色並不明顯，與物理學相比，認為數學在生物學中有效才不合理的說法恐怕更有道理。當然，許多計算分子生物學的常用工具，例如在資料庫中搜尋與給定的探針序列（probe sequence）相近序列的工具，其基礎就是數學和電腦科學。但是對生命過程的終極理解是否能以數學語言來表示——就像物理定律以對稱概念為基礎，還是採用傳統的描述性、「軼聞式」的生物學語言，這一點依舊懸而未決。

為什麼懷疑數學在生物學的有效性是合理的？生命系統所觀察的性質，來自下列因素的組合

- 物理和化學定律
- 演化的機制
- 歷史性的偶然因素

我們很難區辨這些因素，它們彼此之間的創造性張力遍布於我們的研究。許多物理定律在描述自然世界時（包括生命系統），需要具體說明起始條件和最終條件之間的關係。但在生物學中，所有可能起始條件的組合複雜性造成困難，歷史偶然因素的巨大角色阻礙了研究，也讓我們更謹慎。就算物理和化學基本定律的簡單結果足以描述生命過程的細節，我們卻不見得可以發現這些過程，因為我們所能觀察的要複雜得多，因此排除了簡潔的理想化方法，而觀察到的特性又取決於起始條件的選擇，這些條件又來自龐大且分雜的各種可能性。在生物學裡，蘋果可不只是掉到頭上而已。

計算分子生物學在研究什麼

不過我們的研究對象，至少形式上可以嘗試著應用數學，包括

- 基因的 DNA 序列
- 蛋白質的胺基酸序列
- 蛋白質結構
- 蛋白質功能

讀者應該都聽過的**基因體計畫**（genome projects），希望能確定有機體 DNA 的完整序列——生物的藍圖。基因體的 DNA 序列包含生物出生、發育、成長、死亡所需的所有資訊。在 1996 年完成酵母菌基因組的定序後，我們對酵母細胞所知的已經和酵母細胞本身一樣，這樣說並不像初聽那麼倨傲，因為我們**真的**掌握了所有的資訊。無可否認，人類不像酵母菌在解釋這些資訊時那麼有效率，但是我們的確擁有完備的藍圖組合。不過藍圖只能**靜態**描述結構與蛋白質的活性，我們仍然需要擴大觀察，在有機體的時空架構裡，蛋白質的表現與功能的整體性。這些資料的匯集稱為「蛋白體計畫」（proteome project），在後基因組時代正逐漸取得重要性。

基因定序測量的進展速度很快，並且還在加速中。1998 年，隱桿線蟲（Caenorhabditis elegans）的 DNA 已經完成定序（9.7×10^7 鹼基），1999 年與 2000 年，果蠅（3.4×10^9 鹼基）和人類（3.4×10^9 鹼基）基因組也可能完成定序[1]，另外還有許多其他大大小小的生物。路易十五可以說：「朕死之後，管他巨浪滔天。」相較之下，諾亞不會這樣說，我們也不能不管。

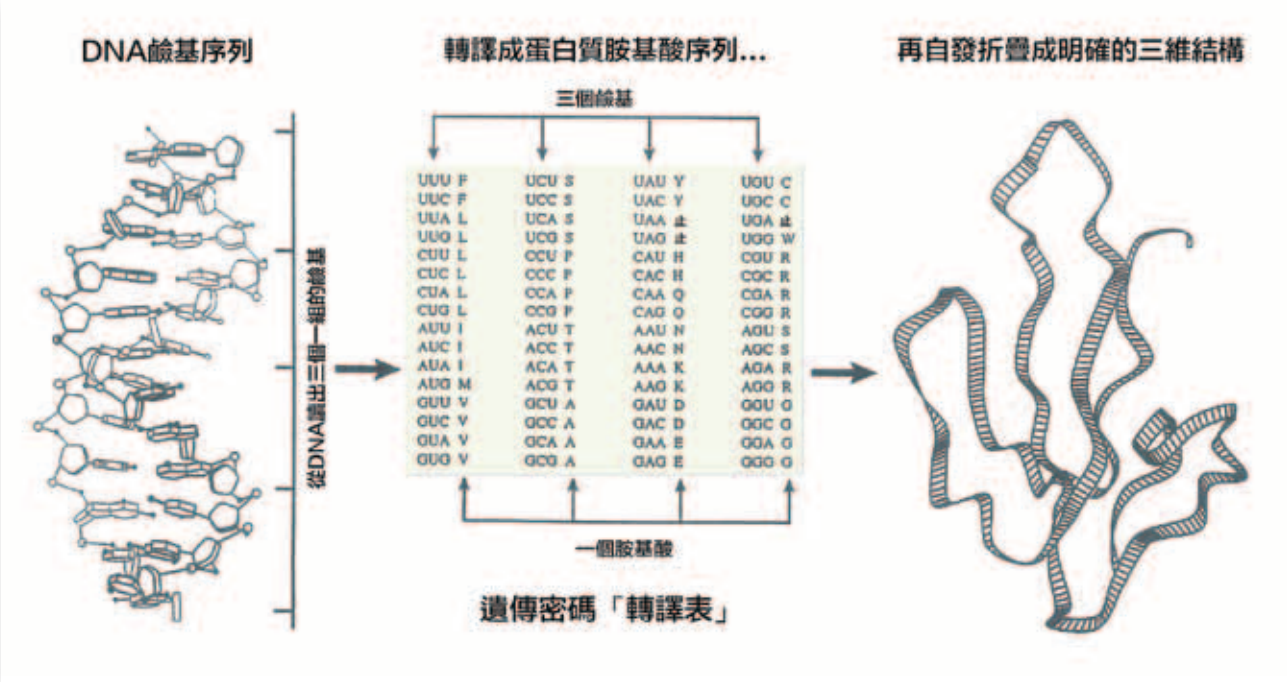
基因序列與我們研究的蛋白質結構牽連甚深又很重要。在分子層次，基因的 DNA 序列會轉譯成蛋白質的胺基酸序列，蛋白質的胺基酸序列再決定蛋白質的三維結構，蛋白質結構又決定了蛋白質的功能（圖一）。明確的三維結構是蛋白質功能的基礎，因為其中所需的相互作用，需要將分子的不同部位以精確的空間關係配置在一起。到最後，蛋白質功能又會回饋到基因序列，藉由天擇的演化完成整個循環。

在電腦中，DNA 序列可用字串表示，這是一維的物件。基因做為基因組序列的子字串，以近乎普適（universal）的密碼表轉譯成蛋白質的胺基酸序列。蛋白質胺基酸序列也可用一維的字串表示。然後蛋白質會自發的折疊成唯一且「自然」的三維結構（蛋白質會因加熱導致三維結構破壞而變性，但冷卻後它又會恢復原先的形式，就好像記憶金屬一樣。）這個蛋白質自發折疊的表現，正是讓大自然從一維基因序列得以一舉入我們居住的三維世界的關鍵。

計算分子生物學的目標

我們的目標是什麼？首先是描述序列之間與結構之間的相似與差異，並做分類。序列空間的拓樸是什麼？結構空間的拓樸呢？蛋白質功能的空間呢？這些空間彼此的映射又是什麼？我們希望能夠描述與預測，序列、結構、功能之間的內在關係，並以演化做為組織的原理。

[1] 譯註：果蠅 2000 年完成定序，人類基因組號稱 2003 年完成定序。



圖一　從基因可以讀出的資訊。**基因是生物的基本藍圖，包含在 DNA 的結構中**（左），左圖是 DNA 的雙螺旋，包含了兩條相互纏繞的鏈串，圖中各自用細線和粗線呈現。螺旋中的「樓梯效應」是由一組稱為「鹼基」的化學單位構成，鹼基一共有四種：A、T、G、C。眼尖的讀者可能注意到這些鹼基（梯板）的形狀不同，兩螺旋同層的鹼基相互作用，而且遵守嚴格的互補規則：A 和 T 對應；G 和 C 對應。於是，每一條鏈串都具有足夠的資訊，可以導引另一條的合成。邏輯上來說，所謂複製 DNA 就是將兩鏈解開，分開的鏈串再各自合成另一半。**基因中的鹼基序列，靠著稱為「遺傳密碼」的直接轉譯表，轉譯成蛋白質的胺基酸序列**（中）。基因的訊息使用 A、T、G、C 這四個字母書寫，蛋白質是包含一條化學殘基（residue）序列的聚合物，每個位置都是 20 種胺基酸其中之一所構成，記為 A、C、D、E、F、G、H、I、K、L、M、N、P、Q、R、S、T、V、W、Y。要確定這 20 個胺基酸，需要兩個以上的鹼基加上一個對演化很重要的多餘碼，事實上 DNA 序列是每三個鹼基一讀（另有三組三鹼基組保留給結束訊號「止」（End-of-file））。**蛋白質會自發折疊成天然有活性的三維結構**（右），這是大自然從一維基因序列躍入我們居住的三維世界的關鍵。圖例取自一種海蛇毒素，是用 X 光晶體學已決定的一種蛋白質結構。每個基因具有一個鹼基序列，先轉譯成蛋白質胺基酸序列，再決定其三維結構，最後是此蛋白質的功能。

要怎麼對付這些問題呢？布列納（Sydney Brenner）曾經感嘆說：「生物學的麻煩在於缺乏諧和振子（harmonic oscillator）。」他的意思是生物學不像物理學，就算**透過**理想化，還是避免不了複雜性。諧和振子是物理中的簡單問題，可以用很多方法精確求解，不但可以確切應用於某些問題，對其他問題也是有用的逼近，諧和振子在物理學是許多新方法的傳統測試台。事實上，在**計算**分子生物學中，也有兩個布列納意義的「諧和振子」：序列比對（sequence alignment）與結構疊合（structure superposition），這兩個可以精確和有效執行的操作，在分子生物學裡是眾多序列 / 結構關係的分析基礎。當然，現實世界的不諧和並不令人意外，但即便如此，從這些簡單的情況還是發展出許多有價值的工具。

但是，工具只能提供答案而非問題。這個領域的研究仍然得持續依賴人類科學家與資料的互動，並輔以數學和電腦方法。

BOX

傑出數學家與生理學家葛爾方德（I. M. Gelfand）激烈的否認他是數理生物學家，針對威格納原理：數學在物理學中不合理的有效性，他反過來說：**數學在生物科學不合理的無效性**

有些他的跟隨者稱之為威格納 / 葛爾方德原理。

序列與比對

基因和蛋白質序列都具有字串的形式。基因序列的字母是 A, T, G, C，依序表示腺嘌呤（Adenine）、胸腺嘧啶（Thymine）、鳥嘌呤（Guanine）、胞嘧啶（Cytosine）的核苷酸符號。蛋白質序列則有 20 個字母，分別表示 20 種標準胺基酸。

兩條字串的**比對**（alignment）就是決定兩字串字母的有意義對應[1]，例如底下兩字串：

gctgaac 和 ctataatc

兩種可能的比對是

```
gctga-a--c        gctg-aa-c
-ct-ataatc   和   -ctataatc
```

要如何判定這兩組或更多組比對，哪一組才是最佳的比對呢？我們可以為字串設計一種度量（metric），定義兩組字串的距離嗎？測量字串是否相近的測量方式包括：

（1）**漢明距**（Hamming distance）：兩等長字串的距離定義為相應兩字母不一致的位置總數。

（2）**列文史坦距**（Levenshtein distance）：對兩可能不等長的字串，從一字串透過「編輯操作」變成另一字串的最小操作次數，其中編輯操作包括刪除、插入，以及更換某一字母。一連串的編輯操作可以得到唯一的比對字串，但反之不真。

在分子生物學的情況，已知基因或蛋白質序列會發生插入與刪除，因此不能使用漢明距。而且也有證據顯示某些變化比其他的情況更常發生，因此就算列文史坦距也必須基於演化模型做推廣，對不同的編輯操作做差異性的加權。例如突變比想像中的保守，如果蛋白質中的胺基酸要更換，換成大小相近或物化性質相似胺基酸的情況，可能性大於不相似近的胺基酸。為了反映這個事實，本來只是對編輯操作離散計數的作法，就必須代之以讓序列中每一變化都指定某實數值「成本」的想法。

另外的證據顯示，序列中的空隙成本不像列文史坦模型，並不與空隙長度成正比。然而要如何適當選擇空隙長度的函數做為空隙加權的方式卻相當棘手。許多計算法[2]使用線性函數，以α表示空隙起始的固定成本，再用較小的參數β於空隙大小，使空隙成本形如$\alpha + \beta \times$(**空隙長度**-1)。已經有演算法可以計算由一字串轉成另一字串時，編輯操作成本和的最小情況，因此可以達成最佳的字串比對。

字串比對最佳化的形式敘述可以說明如下：給定兩個字串$A = a_1a_2\cdots a_n$和$B = b_1b_2\cdots b_m$，其中a_i、b_j都是字母集$\mathcal{A}$的元素。令$\mathcal{A}^+ = \mathcal{A} \cup \{\emptyset\}$。所謂編輯操作的序列指的是一個有序字對$(x, y)$所成的集合，其中$x, y \in \mathcal{A}^+$，個別的編輯操作包括：

(a_i, b_j)：表示用b_j取代a_i。

$(a_i, \emptyset)$：表示從A刪除a_i。

$(\emptyset, b_j)$：表示在B插入b_j。

成本函數是編輯操作的函數：

$d(a_i, b_j)$是突變的成本

$d(a_i, \emptyset)$或$d(\emptyset, b_j)$是刪除或插入的成本。

而A和B之間的最短加權距離是

$$D(A, B) = \min_{A \to B} \sum d(x, y)$$

其中$x, y \in \mathcal{A}^+$，而極小值是比較所有由A轉換成B的編輯操作序列而得到的。如果$d(x, y)$在$\mathcal{A}^+$上是度量，那麼$D(A, B)$在$\mathcal{A}^+$字母所形成的字串空間上也是度量。（以上問題的敘述預設空隙成本和空隙長度無關；更實際的空隙加權計算法則是推廣的情況。）

問題是我們希望找到$D(A, B)$以及符合這個條件的單個或更多序列比對，能在$O(mn)$時間內解決

[1] 譯註：在分子生物學中，alignment 有「比對」和「排序」兩種常用譯法，鑑於本文之數學模型，比對較易望文生義，故從之。

[2] 譯註：scheme 基本上是數值或計算數學對某問題的數值解法，通常是整合了許多計算法在內的多重組合，目前似乎常譯成很難望文生義的「格式」，暫譯成「計算法」。

此問題的算法久為人知 ❶，並且已經運用於很多問題如文本編輯、語音識別、鳥語分析等 [2]。將它介紹到分子生物學的是尼德曼（Saul Needleman）與溫煦（Christian Wunsch）的重要論文 [3]。

這個演算法有幾個值得關注的特色：

- 它得到的是**絕對最小值**（記得這是計算分子生物學兩大「諧和振子」的第一個），這個方法保證我們不會陷在局部極小值裡。
- 以上是好消息，壞消息是如果想解釋得到的結果並不簡單。雖然從最佳比對得到的編輯操作序列**或許**對應到實際的演化途徑，但卻無法證明。編輯操作距離越大，合理的演化途徑就會越多。不但最佳比對可能不唯一，而且還可能有很多得分很接近最佳比對的次佳選擇。例如，菲奇（Walter Fitch）和史密斯（Temple Smith）研究雞的α和β血紅蛋白基因時 [4]，就找到 17 組最佳比對，其中一組符合已知血紅蛋白結構的比對，但得分在最佳比對 5% 範圍內的超過 1000 組之多。

雙序列比對的問題

已知在蛋白質演化時，胺基酸序列分歧的速度遠比結構的分歧快。在許多情況中，我們察覺到兩個蛋白質結構間存在演化關係，但基因序列或蛋白質序列卻偵測不出什麼相近性。原因是這樣：儘管基因可以在 DNA 序列空間中探索前進，但是天擇卻會在結構的變化上踩煞車，這是為了維持蛋白質的

```
TYLWEFLLKLLQDR.EYCPRFIKWTNREKGVFKLV..DSKAVSRLWGMHKN.KPD
VQLWQFLLEILTD..CEHTDVIEWVG.TEGEFKLT..DPDRVARLWGEKKN.KPA
IQLWQFLLELLTD..KDARDCISWVG.DEGEFKLN..QPELVAQKWGQRKN.KPT
IQLWQFLLELLSD..SSNSSCITWEG.TNGEFKMT..DPDEVARRWGERKS.KPN
IQLWQFLLELLTD..KSCQSFISWTG.DGWEFKLS..DPDEVARRWGKRKN.KPK
IQLWQFLLELLQD..GARSSCIRWTG.NSREFQLC..DPKEVARLWGERKR.KPG
IQLWHFILELLQK..EEFRHVIAWQQGEYGEFVIK..DPDEVARLWGRRKC.KPQ
VTLWQFLLQLLRE..QGNGHIISWTSRDGGEFKLV..DAEEVARLWGLRKN.KTN
ITLWQFLLHLLLD..QKHEHLICWTS.NDGEFKLL..KAEEVAKLWGLRKN.KTN
LQLWQFLVALLDD..PTNAHFIAWTG.RGMEFKLI..EPEEVARLWGIQKN.RPA
IHLWQFLKELLASP.QVNGTAIKWIDRSKGIFKIE..DSVRVAKLWGRRKN.RPA
RLLWDFLQQLLNDRNQKYSDLIAWKCRDTGVFKIV..DPAGLAKLWGIQKN.HLS
RLLWDYVYQLLSD..SRYENFIRWEDKESKIFRIV..DPNGLARLWGNHKN.RTN
IRLYQFLLDLLRS..GDMKDSIWWVDKDKGTFQFSSKHKEALAHRWGIQKGNRKK
LRLYQFLLGLLTR..GDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKR
  L  fl  1L          i W       F          a  WG  K
```

圖二　是所謂 ETS 區域的蛋白質家族的部分序列的多重序列比對。每一列來自一蛋白質的胺基酸序列，這個字母序列的每個字母表示一個胺基酸。垂直欄位的字母是這些家族蛋白質在該位置的胺基酸，上下觀察可以看出一些模式。例如每一條序列的第三個都是白胺酸（L 即 leucine），這表示有某種結構或功能上的限制，阻礙演化在這個位置產生變異。表中最後一列的大寫子母表示所有位置都不變的殘基，只有一位置不同的則以小寫表示。注意各欄變化的分布並不均勻。如果保留殘基的週期性（3、4、8）提示這些蛋白質內有螺旋結構（這是真確的），其他還有些藏得更深的模式，需要計算分析才能確認。這些模式包括不同位置胺基酸分布的相關性，例如左起第四欄只有最後兩列是酪胺酸（Y 即 tyrosine），其他都是色胺酸（W 即 tryptophan）。此欄的變化和右起第四欄與第五欄有很高的相關性。一般相信（至少是希望）像這類序列表上不同位置變化模式的相關性，可以提供線索，指出三維結構中發生相互作用的區位（site）。不幸的是這些信號很微弱。

功能。遺傳密碼的冗餘特性，像是多個三鹼基翻譯成相同的胺基酸；許多單鹼基改變轉換的胺基酸仍保留類似的物化性質，都是為了緩解序列變化所導致的結構改變。

即使在序列層次可以偵測相近性，但對關係疏遠的蛋白質，雙序列的最佳比對結果卻經常是錯誤的，這是從做為最後裁決的結構比較而得知的。

然而，如果能找到許多相關的序列，那麼多重序列比對將比雙序列比對給出更有意義與準確的結果。為什麼多重比對能提升序列的資訊？因為它能顯現被保存的模式。個別位置變化的程度與特性，是序列的不同區域在結構或功能角色上的重要指標（見圖二）。例如，在一整個蛋白質家族中都能夠保存下來的殘基（residue）❷，經常與蛋白質功能相涉，或者至少經常與蛋白質結構有密切關係。相反的，經常發生插入或刪除的區域，則通常對應到結構的外圍部分。

（說起蛋白質的結構　一條序列忸忸怩怩
兩條序列嘻嘻暗笑　更多序列大聲咆哮）

不過如果序列訊息只能對蛋白質結構間接一瞥，為什麼不直接處理結構呢？因為已知的序列資料量遠遠超過結構的資料。目前有大約二十種生物的基因已經完全定序，給出完整的基因序列，但是其中只有極少數的基因，我們知道對應蛋白質的結構。

蛋白質結構的分析

分析蛋白質這麼複雜的分子結構，首要的問題就是結構呈現的方法。目前已經發展出許多電腦繪圖技術能簡化的呈現蛋白質。圖三以一個小蛋白質分子為例，顯示要詮釋完整細節、如實呈現的難處，以及一些利用程式簡化的圖形，讓人們得以見到這些結構。活躍的「家庭工業」（也就是很多素人）提出許多簡化呈現的方法，最後被納入各種繪圖套裝軟體。熟練的分子畫家利用它們，以精校程度的細節，從不同角度呈現分子的結構。這些圖形被上色並加上花俏的陰影效果後（考慮可見光的波長和分子的大小，這當然不真實），裝飾了許多期刊、海報，甚至 T 恤和馬克杯。

已知的蛋白質結構有一萬種❸，其空間模式有非常大的差異。針對羅塞福（Ernest Rutherford）的評價：「所有科學除了物理學以外，都是集郵。」，我現在的回答是蛋白質結構的研究結合了兩者的優點。我們既擁有壯觀的多樣性，但也相信存在基礎性的統一原理。

每個蛋白質有一條線狀（也就是不分岔）重複的聚合物主鏈，每隔固定間隔有一些不同的胺基酸側鏈接在上面。所以蛋白質看起來就像聖誕樹上的燈線，電線就是重複的主鏈，而五顏六色的燈光序列則是個別側鏈的序列。

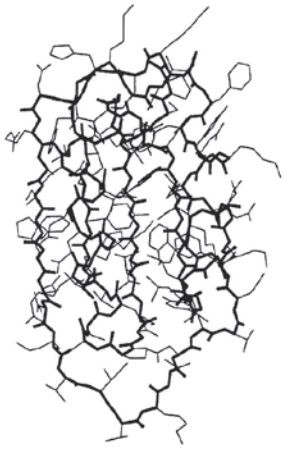
主鏈是一條由不同側鏈因相靠近時之相互作用而平衡穩定後所得的空間曲線。這條主鏈在圖三的中圖表現的最明顯。圖前有兩段**螺旋**（就像傳統的理髮店螺紋招牌），其螺旋軸幾乎是鉛直的，這是兩種局部區域標準結構之一，另一種標準結構是延伸的**摺板**（sheet）長串，圖三的蛋白質有四條摺板串，方向也幾乎都是鉛直的。這些摺板串是靠側面的交互作用來維持結構。圖三的底圖，螺旋和摺板串是用「示意圖」呈現的：螺旋是圓柱體，摺板串是大箭頭。至於圖三的上圖則是最具細節的結構圖，包括主鏈和側鏈，而其中粗細的對比顯示，就算是小型的蛋白質，簡化作圖對製作視覺可理解的圖形仍然很重要。

分析新結構的第一步是找出螺旋和摺板的區域，這是畫出圖三中圖與下圖所需的資訊。蛋白質中最

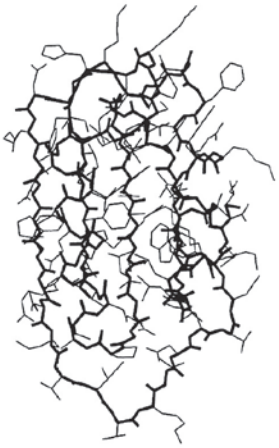
❶ 譯註：O記號表示計算法的漸近時間複雜度，如本例 $O(mn)$表示當 m 、n 很大時，所需要的計算時間大約和 mn 成正比。

❷ 譯註：殘基：大分子中的某一部位。當胺基酸形成巨大的蛋白質後，這些胺基酸就構成蛋白質的殘基。

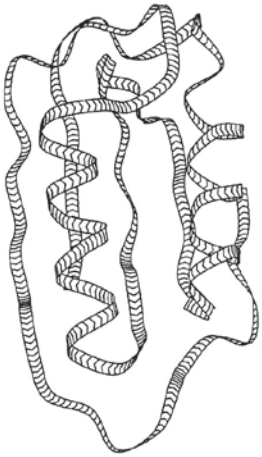
❸ 譯註：到 2013 年已經接近十萬種。



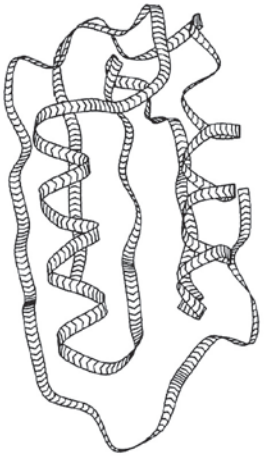
醯基磷酸酯酶



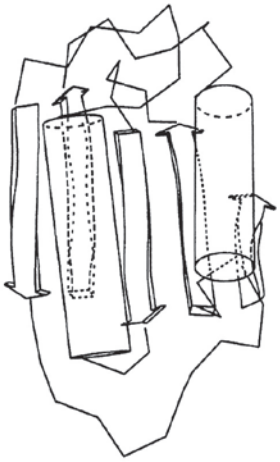
醯基磷酸酯酶



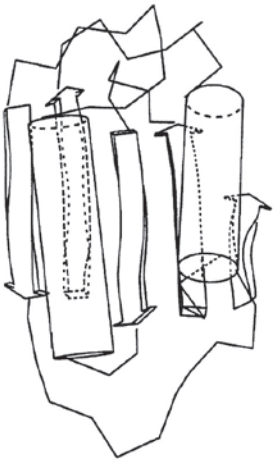
醯基磷酸酯酶



醯基磷酸酯酶



醯基磷酸酯酶



醯基磷酸酯酶

圖三　蛋白質具有相當複雜的結構，必須發展特定的工具來呈現。本圖以三種不同程度的簡化方式，呈現一種較小的蛋白質，稱為醯基磷酸酯酶（Acylphosphatase）。上圖：完備的骨骼模型（skeletal model），主鏈畫得比側鏈粗。中圖：鏈的路徑以光滑的內插曲線呈現，其中 V 型記號標記鏈的走向。下圖：在此略圖中，以圓柱表示螺旋，以箭頭表示褶板長串。圖中的立體是半透明的，並用虛線表示被遮住的部分。試著將相鄰兩圖疊合，可以用立體視覺觀看（不要看太久）。

常見的螺旋類型每個迴旋包含 3.6 個殘基。因此如果在胺基酸序列中顯出這種週期性的特點，就可能是螺旋區域。

結構的疊合

就像比對序列一樣，分析結構的基本問題是設計並計算結構相近度的測量方式。假設將結構用坐標來呈現。

$p_i = (x_i, y_i, z_i)), i = 1, \cdots, N$

與

$q_i = (x'_j, y'_j, z'_j)), j = 1, \cdots, M$

就和序列的情況一樣，這裡也有比對的問題。試比較在計算化學中出現的三個相關問題：

（1）測量對應已知的兩組原子的相近度：

$p_i \leftrightarrow q_j, i = 1, \cdots N$

這和序列問題的漢明距類似，這個情況可以精確有效的解決，這就是計算分子生物學的第二個「諧和振子」。

（2）測量原子對應未知，但分子結構給出對應的條件（尤指殘基依線性順序）的相近度：

$p_{i(k)} \leftrightarrow q_{j(k)}, k = 1, \cdots K \leq N, M$

且遵守以下限制條件

$k_1 > k_2 \Rightarrow i(k_1) > i(k_2), j(k_1) > j(k_2)$

這可想成對應列文史坦距或含空隙的序列比對。

（3）測量原子對應未知，也無對應的限制條件的相近度：

$p_{i(k)} \leftrightarrow q_{j(k)}, k = 1, \cdots K \leq N, M$

這個問題出現於下述重要情況：假設有兩個或多個分子具有類似的生物效應，例如共同的藥物活性。這通常表示這些結構共享一小區的原子組合，可以解釋這個生物效應，稱為**藥效團**（pharmacophore）。想找出藥效團，就要從這兩個或多個分子找出擁有類似結構的最大區域。

問題（2）和（3）都需要點對點做比對，只與坐標有關的比對方法稱為結構比對（structural alignment）。在結構比對裡，相對應的殘基被視為相同，因為它們在整體結構中佔據同樣的位置。人們必須思索如何抽取出最大的共同子結構，並以此作為比對的基礎（這就像字母 B 和 R 的最大共同子結構是 P），而在最大共同子結構之外的殘基則無法比對。這是雙序列比對無法偵測到的事實，因此是該方法的弱點。

對這三個問題最一般的解決方法，是奠基於問題（1）的解，也就是已知對應$p_i \leftrightarrow q_i$的情況。兩個**全等**的物件可以透過剛性平移與旋轉相互疊合，因此兩個相近的物件也能用旋轉和平移達成**逼近**（approximate）的疊合。如果考慮的物件是有序的點集合，一種相近度的度量方式是其最佳疊合的均方根（root-mean-square）標準差Δ：

$\Delta^2 = \min_{\mathbf{R},\mathbf{t}}(\sum_{i=1}^{N} \|\mathbf{R}p_i + \mathbf{t} - q_i\|)$

其中$\mathbf{R}$是恰當的旋轉矩陣，而$\mathbf{t}$則是平移向量。在最佳疊合時，這兩個點集合的平均位置（即俗稱的重心）會重合。尋找正確的相對方向的問題稱為「正交普洛克拉堤斯問題」（Orthogornal Procrustes problem），已知有以線性代數標準技巧為基礎的解法。[5]

解決最大共同子結構的問題提供了度量結構的基礎，它容許對局部與微弱相近度的偵測，並可導出蛋白質結構全體的分類樹。

最大共同子結構的計算方法奠基於兩種呈現結構的方式：

（1）坐標列表$p_i = (x_i.y_i, z_i), i = 1, \cdots, N$；

（2）距離矩陣$D(i, j) = |p_i - p_j|$。

採取距離矩陣的最大好處是它提供了和原點與方向無關的結構呈現方式，而且兩距離矩陣差的最大分量$\max_{i,j} |D_p(i, j) - D_q(i, j)|$，提供了測量兩個已比對點集合差異的方法。

坐標和距離矩陣在呈現點集合時近乎相等。從坐標資料計算距離矩陣簡單到不行，然而是否能從距離矩陣精確又直接的復原坐標資訊就不是那麼

明顯，不過已知這能用對角化（diagonalization）的方法解決 [6]，更確實的說法是，從距離矩陣可以得出原先的結構以及其鏡像異構物（enantiomorphs，例如左右兩手的手套互為鏡像異構物），不過這點含糊對分子生物學的應用並不嚴重。另外位置和方向的資訊當然也消失了。

計算（2）和（3）類型的最大共同子結構，最大的困難是各種比對可能方式的組合複雜度，就這個問題，以距離矩陣為基礎的演算法相對於坐標資料的方法要更有效率。另外以螺旋和褶板等結構要素為基礎的相關矩陣表現，也比光是運用坐標資料，可以得到蛋白質折疊模式更簡潔的呈現方式。抽取最大共同子矩陣的方法顯露具有共同折疊模式的最大子結構。這樣的矩陣表現也容許我們計數所有可能的蛋白質折疊模式。經驗估計所有自然蛋白質的折疊模式少於 1000 種。完全的計數讓我們可以檢視大自然的選擇，嘗試去分辨歷史的偶然，抑或是結構上的必然。

蛋白質的演化

蛋白質演化探討在相關物種裡，相對應的胺基酸序列和蛋白質結構如何不同。這是能提供許多資訊的研究，幫助理解序列和結構的關係。雖然胺基酸序列包含了所有形成蛋白質結構所需的資訊，但我們還不知道如何從序列推演出結構，姑且稱之為蛋白質折疊問題的「整體性形式」（integral form，「積分形式」），這個問題還未解決。另外在研究蛋白質演化時，我們可以觀察到序列的變化如何反映在結構的變化上，這個問題應該比較容易理解，稱為蛋白質折疊問題的「差異化形式」（differential form，「微分形式」）。

主題	蛋白質折疊	蛋白質演化
觀察	序列得出結構	序列變化得出結構變化
問題的形式	整體性形式	差異化形式
問題的狀態	未解	未解但應該比較簡單

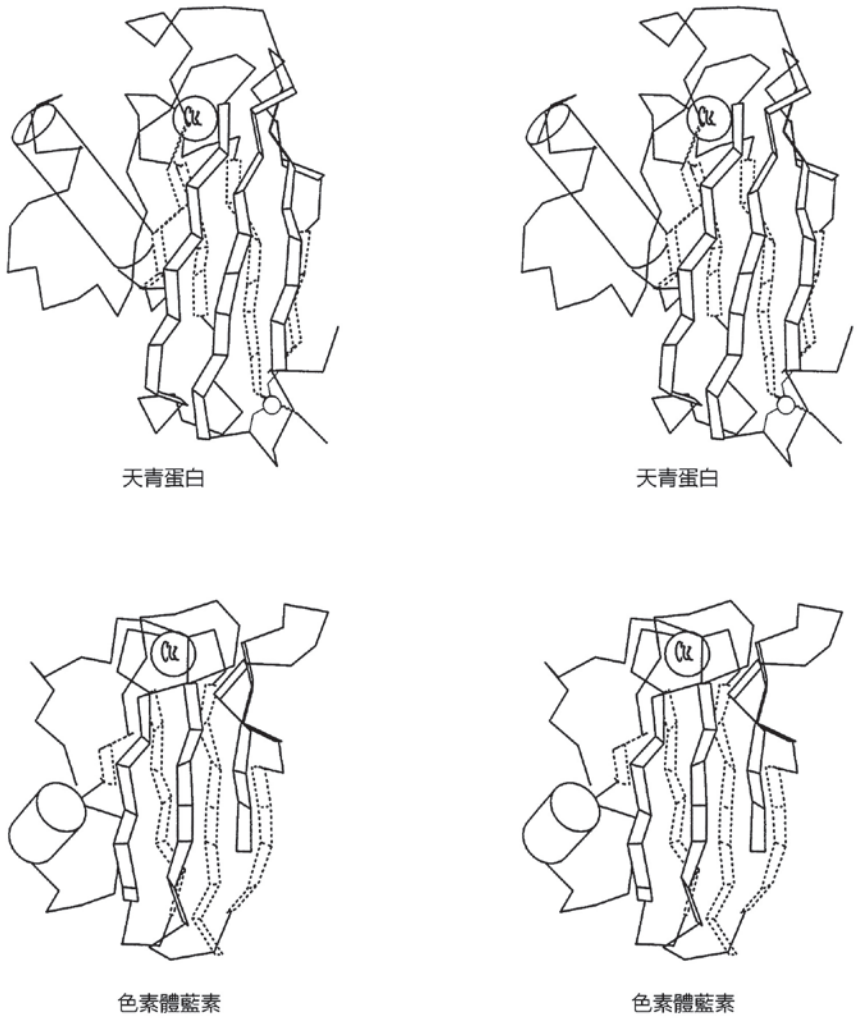
由簡單的論證知道，結構近乎序列的「連續」函數，至少對自然演化的序列和結構是如此。如果有種蛋白質，它的**任何**突變（胺基酸序列的任何變化）都導致不穩定的結構，那麼靠著大自然的演化過程根本無法到達這種蛋白質，因為它沒有穩定的前身（precursor），這表示來自大自然的結構是穩健的（robust）。大部分序列的小變化並不會改變結構（這對人造蛋白質結構並不適用）。

確實，自然的蛋白質如果序列相似，結構也相近。在人工合成胰島素上市之前，使用豬胰島素是治療人類糖尿病的有效臨床療法，即使豬和人類胰島素的胺基酸序列不盡相同。根據這種對相近性的信心，提供了一種由已知蛋白質結構預測相近蛋白質結構的方法，稱為**同源建模**（homology modeling）。不過隨著演化的進行，序列和結構終究會益發分歧。圖四顯示了兩個距離很遠的蛋白質：色素體藍素（plastocyanin）和天青蛋白（azurin），其中兩者右側區域都包含兩條面對面包在一起的褶板，這是保存下來的結構「核心」，至於左側細長的螺旋區域，則顯現完全不同的構形（conformation）。

蛋白質結構預測

大自然有一個演算法，可以單從蛋白質的胺基酸序列，就能明確得到三維結構，照理說我們可以發現這個演算法。如此一來，就能夠預測人類或其他基因組基因序列中與生俱來的蛋白質結構，並應用於實用的問題如藥物設計。但是預測蛋白質結構是困難的問題，人們已經嘗試過很多想法，其中頗多宣稱有進展。不過直到現在，除非先給定某個很接近的蛋白質，不然還沒有計算方法可以從胺基酸序列一致的預測出蛋白質結構，即使只考慮定性預測也做不到。

假設給定一個新蛋白質的胺基酸序列，請你預測它的結構，你大概能預測什麼？預測所能給出的最



圖四　在演化的過程裡，基因序列會累積突變，導致蛋白質序列和結構益發分歧。本圖呈現兩個有關的電子傳遞（electron-transport）蛋白質：白楊葉色素體藍素與細菌的天青蛋白。圖中右半邊的結構，包含實線或虛線像緞帶一樣的摺板區域，以及一處銅結合位（copper binding site），這些是在演化時保留不變的部分。至於左半邊的結構則有很根本的分歧。

完備資訊，莫過於該蛋白質模型的三維坐標，這是 3-D 預測。野心沒那麼大的預測，是指出螺旋與摺板在序列中發生的部位，此即 1-D 預測。介於這兩類預測之間，是超越 1-D 這種二級結構預測，但只給出折疊模式一般空間配置的定性資訊，這姑且稱為 2-D 預測。

預測蛋白質結構時你需要怎樣的資訊？最終目標是單純的「**從頭做起**」（ab initio），單單只運用到目標蛋白質的胺基酸序列，畢竟大自然就是這樣運作的，蛋白質在折疊前並沒有先上網搜尋資料庫。但是我們不妨試試，而且運用資料庫資訊，從已知結構去指認目標蛋白質的折疊方式也獲得一些成功。這個問題稱為**折疊辨識**（fold recognition）。當然這個方法成功的前提是，和目標蛋白質折疊方式相同的蛋白質結構已經在你的資料庫裡面。

談到你必須讓誰滿意？右列名單大略以難度遞減的次序排列。大部分科學家都接受「經費審核單位」是恰當的目標！

你必須讓誰滿意？
1. 結晶學家
2. 核磁共振光譜學家
3. 經費審核單位
4. 論文審閱人
5. 同事
6. 你的母親

說實在的，要如何說服別人你有一個可以成功預測蛋白質結構的方法？有兩種宣告原則上是不可測試的。一種是能夠預測已知的蛋白質結構，另一種是你預測的蛋白質結構，不但目前實驗所得的結構未知，而且可能很久以後還是不知道。我們得走在「已知」與「很久都不可知」兩者之間，協調結構預測與進行中的結構測定才行。

為了讓這項活動更有規矩，能夠鼓勵真正有進展的人，拒斥那些堅持他們「已經解出蛋白質結構預測問題」的人，莫耳特（John Moult）於是提出組織盲檢試驗（blind test）的想法：正用實驗解出蛋白質結構的科學家公開他們的胺基酸序列，但在公定截止日期前必須保持該結構的秘密，所有相信自己掌握蛋白質預測方法的人，在這個日期前要送入他們的預測，最後再將這些結構與實驗結果比較，通常是幾家歡樂萬家愁的局面。這個想法最後發展成兩年一次的 CASP（Critical Assessment of Structure Prediction，結構預測的關鍵評量）計畫。

預測結構的方法分成兩大類：歸納法和演繹法。歸納法直接使用序列和結構的資料庫。而演繹法是真正的「**從頭開始**」，就像裸身登上熱帶荒島一樣，試圖只用物理、化學和生物的一般原理預測蛋白質結構，卻不明顯參考已知的序列和結構。當然發展「從頭做起」的方法時，勢必會依賴已知序列和結構的研究知識。差別是從這些研究所得到的理解，將被提煉成一般法則，預測時不再從資料庫查詢特定的資訊。

「從頭做起」法又可分為兩類，我稱之為「自然型」（Nature）與「推調型」（Nudger）。自然型的路數尋求理解大自然的折疊過程，再依循或模仿它。推調型的方法允許任何能夠讓蛋白質鏈放到恰當構形的程序，就算這個程序並不自然，甚至違反物理原理也無所謂。

有證據顯示天擇不但形塑蛋白質最終的天然狀態，也作用在折疊的途徑上。因為蛋白質不只是演化成穩定有效的構形，它還必須從有許多隨機混合構形可能的非折疊狀態，在合理的時間內折疊到這個特定構形。基於在溶液中原子移動速率的簡單計算，發現如果要窮盡可能的構形，時間上會來不及到相差好個幾數量級（這有時稱為列文薩悖論（Levinthal's paradox））。雖然理論上可能，但沒有證據顯示折疊的途徑會影響最後的狀態。如果存在其他的折疊狀態，但途徑的演化促使其中之一發生，那我們這些預言者就**必須**選擇「自然型」而非「推調型」的方法。

結構預測的困難何在？我們認為自己理解讓天然蛋白質構形穩定的作用力，甚至可以清楚寫出構形的坐標能量函數，需要做的只是極小化。然而，重要的是要意識到——用熱力學的術語來說——蛋白質只是邊緣穩定的（marginally stable）。事實上，折疊蛋白質的構形能（conformational energy）與許多對照項的構形能只有很小的差別，這是數值分析的惡夢。

其中的困難是因為能量函數寫得不夠精確？還是這個函數太複雜以致於無法最佳化？一個測試的方式，是**從蛋白質的天然狀態出發**，嘗試極小化蛋白質的構形能量函數。在起點附近這樣的計算的確會收斂到最小能量構形，這表示這個能量函數至少在正確答案的附近是適合的（不令人意外，因為這個函數的定義，本來就為了重現已知天然狀態做過參數調整。）不過這還不夠。在最小值附近正確的函數，並不能在構形空間中提供完整的軌跡，讓程式能從任意起點開始找到整體最小值。

這裡有兩個問題。首先，許多穩定蛋白質的作用力是短距作用力，就算知道確切的能量函數，如果從隨機延伸鬆弛的構形開始，根本沒有長距力可以驅使系統轉變到正確的結構。其次，就算能夠折疊到緊密的狀態，坐標能量函數的地景（landscape）包含許多局部極小值，彼此隔著高能量障壁，這些極小值中許多都是天然態的候選者。實際的蛋白質

能克服這兩個問題，是結合了（1）大量的「平行處理」（parallel processing），所有的殘基同時探索它們局部的構形空間，而且（2）演化的折疊途徑引領整個系統走向正確答案。我們的電腦無法平行處理；我們的能量函數無法說明長距的折疊途徑；我們的演算法很難找到複雜、多變數、非線性函數的最小值（需要的時候，諧和振子跑哪去了？！）。

基於這些先驗方法的困難，導致我們發展先前提過的，以已知序列和結構為基礎的經驗方法。運用資料庫的預測方法有（1）同源建模法：從相近蛋白質的已知結構來預測目標結構；（2）折疊辨識法：評斷胺基酸序列與已知蛋白質折疊模式的相容度。這些方法的威力日益強大，部分但並非全部可以歸功於資料庫的增長。如果已知的序列和結構越多，新蛋白質當然越可能與已知的資料近似。相較之下，「從頭開始」法的改善就緩慢得多，最近一次 CASP 競賽之後，有個心不甘情不願的評論說，這個方法至少「不再保證失敗。」[8] 悲觀的人或許會預測，資料庫的增長終究意味著以資訊為基礎的方法，將為大部分的問題提供實用的解答，而那些對發展「從頭開始」法感興趣或支持的聲浪終將式微。如果最有趣的生物計算想法從此對計算生物學關上大門，這真是令人遺憾。

譯後記

作者後來去信編輯（見延伸閱讀），談起他的講題名稱本來想用「無效性」，但會議負責人不同意才改用「有效性」（「不就是一杯水是半空還是半滿的問題！」）結果看到文章上編輯加入葛爾方德的說法後，他半開玩笑說失去了和威格納、葛爾方德齊名的機會。然後他做了底下的反省：

數學在將生物觀察予以理性化時無疑是有效的，但是生物學缺乏物理學那種奇妙的簡潔，能以少許基本原理就很精確的定量預測許多觀測。生物學家面對非常多無法理解的觀察結果，沒有信心光靠發現正確的數學結構，就能以隱藏的規律性解釋所有現象。

問題在於歷史的偶然性扮演太重要的角色，有位知名物理學家對我的研究不以為然，他說：「你做的不是科學，只是考古學！」這對我不公平（對考古學家顯然也是），但是他的確說出在生物學中運用數學時實質而嚴厲的困難。

不過作者認為數學終將成功，才敲定文章的標題。對於數學在生物學是否有效的質疑，他引用知名脫口秀藝人楊曼的橋段：

有人問楊曼：「你的老婆如何？」

他回說：「跟什麼比？」

本文參考資料請見〈數理人文資料網頁〉http://yaucenter.nctu.edu.tw/periodical.php

本文出處

The Mathematical Intelligencer 22（2000）No.2, Springer. 作者改寫自 1998 年在牛頓數學科學研究院（Isaac Newton Institute for Mathematical Sciences）「基因組計畫脈絡下的生物分子功能與演化」會議的演講。

譯者簡介

周樹靜為臺灣數普譯者。

延伸閱讀

▶ Lesk, A., “Compare to What?”, Letters to the editor, *The Mathematical Intelligencer* 23（2001）no.1.

▶ Lesk, A., *Introduction to Bioinformatics*（2014）Oxford University Press. 雷斯克撰寫之生物資訊學教科書。他還有好幾本與數學不相關、談基因體學、蛋白質學的書。

▶ Lessick, Bob *Bioinformatics: Life Sciences on Your Computer Coursera* 網路課程，Johns Hopkins University。

▶〈生物資訊（Bioinformatics）專題（上）（下）〉網頁，《INVESTIGATOR 生物科學研發策進社群網站》，這是一批臺灣生命科學年輕人建立之網站。