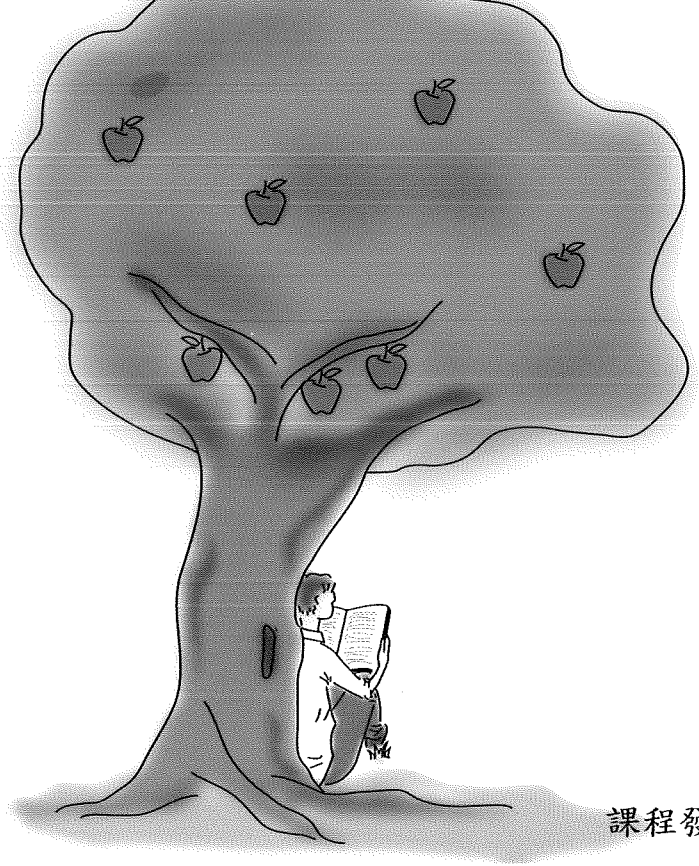
數學百子櫃系列(七)

數學的應用
基因及蛋白的分析



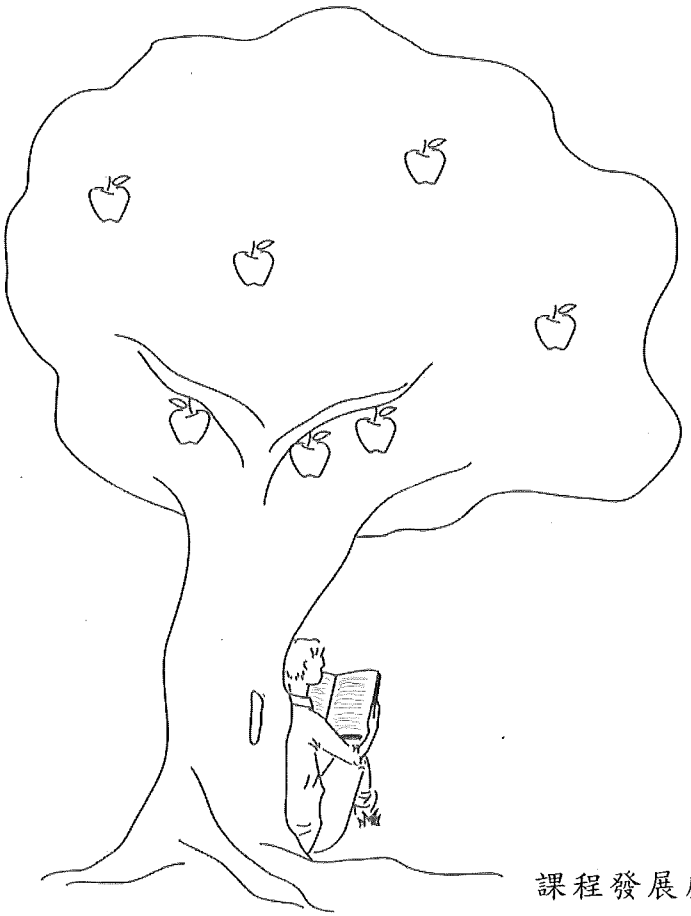
教育局
課程發展處數學教育組

數學百子櫃系列（七）

數學的應用

基因及蛋白的分析

作者 徐國榮教授



教育局
課程發展處數學教育組

ISBN 978-988-8019-00-7

目錄

前言

為配合香港數學教育的發展，並向教師提供更多的參考資料，課程發展處數學教育組於 2007 年開始蒐集和編撰一系列的文章，當中包括大學學者及資深教師的著作和講座資料，輯錄成《數學百子櫃系列》。本書《數學的應用：基因及蛋白的分析》是這個系列的其中一冊，當中輯錄了香港中文大學生物化學系徐國榮教授於 2007 年 1 月在「新高中數學課程知識增益系列–數學應用」研討會上演說的內容，其主題為基因的組成及如何運用數學研究基因突變及進行蛋白分析等，研討會內容精彩豐富，介紹現今課堂上較少討論到的應用。現將研討會的講章輯錄成書，供教師參考。本書內容由作者提供，並不反映教育局的立場。

本系列能夠出版，實在是各方教育工作者共同努力的成果。在此，謹向提供資料、撰寫文章的教師、學者，以及所有為本書勞心勞力的朋友，致以衷心的感謝。

如有任何意見或建議，歡迎致函：
九龍油麻地彌敦道 405 號九龍政府合署 4 樓
教育局課程發展處
總課程發展主任（數學）收
(傳真：3426 9265 電郵：ccdoma@edb.gov.hk)

教育局課程發展處
數學教育組

作者簡介

徐國榮教授在一九八五年畢業於香港中文大學化學系。在中學任教化學科六年後，他重返香港中文大學完成生物化學博士課程。在一九九七年，他獲聘為中大生物化學系助理教授，至二零零四年晉升為教授，現為微生物基因組和蛋白組中心及香港生物信息中心主任。徐教授的主要研究興趣是人類及致病微生物的遺傳基因，亦在國際期刊內發表七十多篇相關的論文。

I. 簡介

香港中文大學設有香港生物訊息中心。生物訊息主要由三個範疇組成；生物學、數學及電腦工程學。

細胞

人體大約有二百六十種細胞，而每個細胞並不一定像球體；有的像橢球體，有的像柱體。細胞的直徑平均為 30 微米 (μm)，每個人擁有大約 10^{13} 至 10^{14} 個細胞，那到底是怎樣知道的呢？首先假設這些細胞是球體，然後用幾十種細胞計算出這些細胞的平均直徑(以 d 表示)，而一個人的體積可用排水法得出。計算如下：

一個細胞的體積 $\approx \frac{\pi}{6}d^3 = \frac{\pi}{6}\times(30\times10^{-6})^3\text{m}^3$

一個人的體積 $\approx 0.5\text{m}^3$

因此，一個人擁有 $\approx 0.5 \div (\frac{\pi}{6}\times(30\times10^{-6})^3)$

$\approx 3.5 \times 10^{13}$ 個細胞。

一個細胞的分裂，從最初開始的一變二，然後二變四，四變八，八變十六.....不斷分裂下去。而一粒蝌蚪本身，就已經擁有一百萬個細胞。

由一個細胞，要進行多少次分裂，才能演變成一百萬個細胞呢？

即解方程

$2^n = 10^6$，其中 n 代表分裂的次數。

要解此方程，首先將方程左、右兩方也取對數

$$n\log 2 = 6\log 10$$

$$n = \frac{6}{\log 2}$$

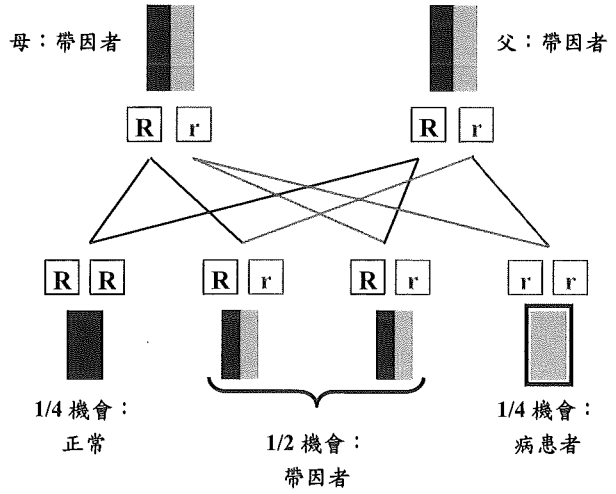
$$= 19.93 \text{（取至 2 個小數位）}$$

$$\approx 20\text{次}$$

若蝌蚪需約 5 小時作一分裂，此分裂過程大約需要一百一十個小時(約五天)。

染色體

人體有廿三對(即四十六條)染色體，最末的一對是性染色體；男性就是 XY，女性就是 XX。舉個例說，假設父母都是某種疾病的帶因者，而該疾病是隱性的，他們的孩子就有四份之一的機會得到此病，而孩子不是致病基因的攜帶者的機會仍會是四份之一，孩子成為此疾病的帶因者的機會是二份之一。



可是，這樣的算法是比較簡單的一個例子，只可以應用在第一至第二十二對的染色體。為什麼這樣計算，是過份簡化呢？因為某一個基因的突變並不一定令某功能完全失去。以葡萄糖六磷酸去氫酶缺乏症(簡稱 G6PD[1]缺乏症，是一種遺傳性代謝缺陷，輕則可以無任何症狀，重則不能接觸樟腦等化合物，否則便會引致溶血，甚至死亡。)為例，此病是跟隨著 X 染色體遺傳下來的，所以並不能如上述用簡單的算法來計算。香港男性患有 G6PD 缺乏症的機會約為十分之一[2]，而女性必須要兩條 X 染色體都有缺陷時才會患上 G6PD 缺乏症。由於兩條 X 染色體皆有缺陷的機會約為 $\left(\frac{1}{10}\right)^2 = \frac{1}{100}$，因此，患此病的男性一般較女性為多。

去氧核糖核酸(或稱脫氧核糖核酸)(DNA)

去氧核糖核酸是遺傳指令的分子，作用主要是儲存遺傳訊息。帶有遺傳訊息的 DNA 片段稱為基因，其他的 DNA 序列，有些在結構上有所作用，有些則參與調控遺傳訊息的表達。DNA 是四種去氧核糖核酸的聚合物，而這四種分子的結構十分相似，不同的部分稱為鹼基，遺傳密碼正是由四種不同的鹼基所組成：

- 腺嘌呤 (A)
- 鳥嘌呤 (G)
- 胞嘧啶 (C)

[1] glucose-6-phosphate dehydrogenase

[2] 編者按：衛生署 2006 年的最新數字為男性：4.5%, 女性：0.5%
http://www.dh.gov.hk/english/main/main_cgs/files/G6PD.pdf

- 胸腺嘧啶 (T)

在製造蛋白質的過程中，DNA 密碼轉錄為信使核糖核酸(messenger RNA)，再轉譯成為氨基酸。組成 RNA 的鹼基以尿嘧啶 (U)代替了胸腺嘧啶(T)。可是，到底是怎樣轉譯呢？轉譯過程是根據一類似矩陣名為密碼子表(見下表)。

密碼子表

第一位鹼基	第二位鹼基 U	第二位鹼基 C	第二位鹼基 A	第二位鹼基 G		第三位鹼基
U	UUU Phe[3]	UCU Ser	UAU Tyr	UGU Cys	U	
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C	
	UUA Leu	UCA Ser	UAA *Stop*	UGA *Stop*	A	
	UUG Leu	UCG Ser	UAG *Stop*	UGG Trp	G	
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U	
	CUC Leu	CCC Pro	CAC His	CGC Arg	C	
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A	
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G	
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U	
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C	
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A	
	AUG Met, *Start*	ACG Thr	AAG Lys	AGG Arg	G	
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U	
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C	
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A	
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G	

從上表，每三個密碼子代表一個胺基酸，所以會出現

$4\times4\times4 = 4^3 = 64$ 種組合，即是有六十四個密碼子。

譬如 UUU 代表一個胺基酸，名為苯丙氨酸。但是胺基酸只

[3] UUUphe 代表某一種名為 Phe 的胺基酸，由密碼子表可見各 Phenylalanine (Phe) 的胺基酸可由 UUU 或 UUC 密碼子形成。

有二十種，即密碼子與胺基酸是沒有一一對應的關係，有時候幾個密碼子會代表同一個胺基酸，例如一胺基酸叫做亮氨酸(Leu)，它可由幾個密碼子如 CUU，CUA 等代表(見表)。

II. 人類基因組

鹼基對

DNA 是由兩條互補的去氧核糖核酸聚合鏈組成。鹼基腺嘌呤 (A)與胸腺嘧啶(T)互補，而鹼基鳥嘌呤(G)則與胞嘧啶(C)互補。人體的細胞內有 3.2×10^9 鹼基對，假設要把人體的鹼基對印出來的話，到底需要多少張 A4 紙呢？以一張紙可印滿六十行，每行八十個字母，底面共印出 $80\times60\times2=9600\approx10^4$ 字母，所以共需要

$3.2\times10^9\div10^4 = 3.2\times10^5$ 張紙(即約需 30 萬張紙)。

基因

人體內大約有 26 000 至 39 000 個基因，為什麼基因的數目會有這麼大的差距？因為它們是以不同的理論及假設建構出對應的數學模型，再借助電腦計算出來的。例如有些研究組以鹼基的成份作起始點，另一些則比較側重鹼基的排列與組合。是故有一些數學模型，可以得出 26 000 個基因。倘若假設的數目少一點，則基因的數目便可能會多一些。整個基因組裡面，有 1.1-1.4%可轉譯為蛋白質，其餘 98.6-98.9%的作用是什麼呢？它們的主要作用是儲存生物訊息，我們可以運用數學在 DNA 裡面找出規律，然後再找出其功能，這程序稱為模式識別。

分子流行病學

流行病學，是運用統計學預測某種疾病何時發生，其中經常用的是統計學的多變量分析。假設某種疾病，已知發病的頻率，我們便會進行很多的調查，例如不斷去訪問病人，並詢問他們的生活習慣。當蒐集數據後，便應用統計學假說的 p-值，我們會找出哪一個數，哪樣東西和某種疾病的關係；我們也會計算其相關係數 r-值，最後作出預測，譬如某種疾病會否發生及於何時發生。

多年前，香港中文大學獲得資助，研究為何特別多亞洲人患有肝癌。六十年代中期，科學家才發現及分離出乙型肝炎病毒(HBV)。在香港約有 10%的人帶有此病毒。在約六十萬名乙型肝炎病毒帶病毒者中，約四分之一將會死於肝癌或肝硬化，他們通常會在 40 歲後發病。我們發現每一百名肝癌病患者中，約有 85 個是病毒帶菌者，所以此病毒與肝癌有著密切的關係。究竟有乙型肝炎病毒的人比沒有此病毒的人患肝癌的風險高多少倍呢？

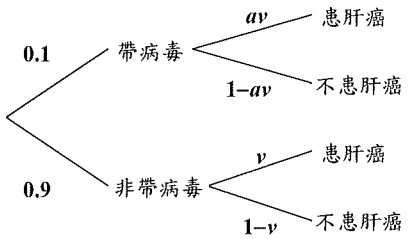
假設有此乙型肝炎病毒的人患肝癌的風險是沒有病毒的人的 a 倍，整體人口是 p，非病毒帶菌者患肝癌的機會率是 v，即

P(帶有乙形病毒) = 0.1

P(帶有乙形病毒|患肝癌) = 0.85

P(患肝癌|非病毒帶菌者) = v

用樹圖表示：



$$\frac{0.1p\times av}{0.9pv+0.1pav}=85\%$$

$$\frac{0.1av}{0.9v+0.1av}=0.85$$

$$0.1av=0.765v+0.085av$$

$$0.015av=0.765v$$

$$a=51$$

所以有乙型肝炎病毒而又變成肝癌患者是沒有攜帶此病毒而變成肝癌患者的 51 倍。我們亦可用類似方法統計吸煙者佔的人口比例及有肺癌的病患者裡面有多少是吸煙者，從而計算吸煙增加患肺癌的風險是多少。

腫瘤

大家有否發覺有些人經常吸煙，但是他們仍可活到八、九十歲也沒有患肺癌，為什麼？其實這是一個數學問題。煙草內有些化學物質會引起 DNA 突變而這過程是隨機的；它可以把鹼基 A 變成鹼基 T，鹼基 G 變成鹼基 C 等。在三萬幾個基因裡面，大約只有 1%是負責細胞分裂。當基因發生突變後，密碼子雖然有改變，但未必導致細胞分裂失控。例如，因為有六個密碼子(UUA,UUG,CUU,CUC,CUA,CUG)都是代表亮氨酸(Leu)，所以當

有突變時，氨基酸都可以沒有任何改變。

為什麼腫瘤初起時，我們是不容易察覺呢？回想細胞是由一個變兩個，兩個變四個，四個變八個如此類推。腫瘤細胞是以指數方式增長，一個失控的細胞長成 1mm 大小要數個月甚至半年的時間。當腫瘤成長到 1mm 後，它就很快長成 2mm 甚至 4mm，所以初起時，腫瘤是不易被察覺到。因此，我們仍會建議大家定期驗身。

一個細胞的 DNA 在複製時產生突變的機會十分微少，縱有突變也不一定落在負責細胞分裂的基因上。縱然突變落在負責細胞分裂的基因上，也不一定令到細胞分裂加快及失控。所以形成腫瘤純粹是一個機會率，你可能一生也不會發生。但是現代人的壽命增長了，發生腫瘤的機會自然地增大。

基因序列搜索器

BLAST是一個網上基因序列搜索器，它能夠把你手上的一基因序列(稱為目標序列)與一個很大的資料庫內的序列作比較並從中找出最相似的序列，這樣我們便能知道手中的序列是否代表一個新的基因，或這個序列是否帶有不尋常的基因突變。BLAST本身有不同的程式作不同搜索功能，例如BLASTp是用作搜尋蛋白質序列；BLASTn是用作搜尋核酸序列等。在BLAST程式裏，我們會輸入一組序列，BLAST便會找出很多類似的序列出來。在這麼多的序列裡，哪個才是我們需要找出來的？它首先會以較短的字長開始，預設字長是3。例如：有兩個序列分別為AGTTAG及 ACTTAG，BLAST會找出這兩序列相同的字長TTA，然後再將比對的結果向兩邊延伸直至找出最相似的序列。怎樣才是最相似？程式是根據一個計分矩陣計算，稱為

BLOSUM Matrix[4] (見下表)。BLOSUM 62是BLAST程式內預設的計分矩陣之一。

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	−1	④																		
T	−1	1	⑤																	
P	−3	−1	−1	⑦																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	⑤						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	⑥		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11

(對角線下: BLOSUM 62)

舉一例說明 BLOSUM 矩陣是怎樣找到最近似的序列。假設你輸入的蛋白質序列是 PNSTFAM(每一個字母代表序列中的一個氨基酸)，而 BLAST 輸出三個序列分別為序列 A :PASTFPM，序列 B: PNSYFAM 及序列 C:TNSVVAM。根據上表，

序列 A： PASTFPM

輸入的序列： PNSTFAM

因為序列 A 的第一個「P」與輸入的序列的第一個「P」一樣，此位置得 7 分，但序列 A 的第二個「A」與輸入的序列的

[4] Blocks of Amino Acid Substitution Matrix

第二個「N」不同，此位置得 −2 分，如此類推。所以序列 A 得 7−2+4+5+6−1+5 = 24 分。
同樣地，序列 B 得 30 分及序列 C 得 17 分，所以序列 B 得分最高，因此 B 便是要搜索的序列。

在序列比對時，我們有可能需要加插入間隔來找出相似的序列，但所得的分數則會按加入間隔的數目而扣減。

輸入序列　　：YSPALNKMFCQ.....

加入間隔序列　：YSPALNK　CQ.....

比對的序列　　：YSTELKKLYCQ.....

BLOSUM 矩陣內的數字是怎樣得出來？內裡涉及抽樣及概率。因為抽樣的方法不同，所以有些人用的 BLOSUM 矩陣會有些不同。舉一個例子，以便說明矩陣內的數字代表什麼意思。一個血紅蛋白，在很多動物身上都能輕易找出，但是牠們的序列可能會不一樣。例如在這些蛋白中的 P 變做 N 後，它仍然是血紅蛋白，因為變化後功能沒有改變，所以這個 P 變 N 的過程扣分會少一點。如果觀察一百種生物，P 很少會變做譬如 S，所以由 P 變 S 這個位置的數值會負得很大，因為發生的概率很低。因為大家所用的動物樣本不同，或使用的蛋白質不同，故此便產生了不同的 BLOSUM 矩陣。

III. 坐標幾何在生物訊息的應用

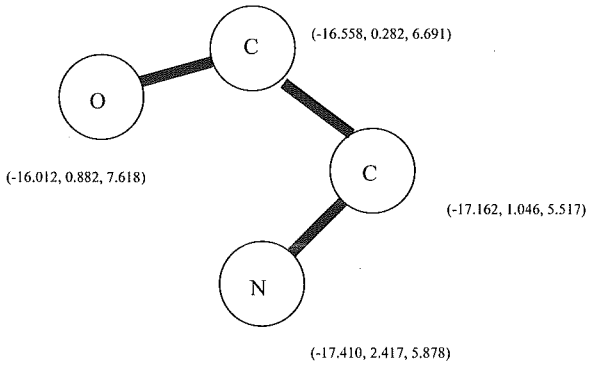
除了概率和矩陣外，坐標幾何亦在生物訊息科學扮演重要角色。蛋白質資料庫是三維生物分子結構數據的貯藏庫，它會顯示出氨基酸序列的原子三維坐標以幫助了解蛋白質的結構與

功能之間的關係。從蛋白質資料庫找出一蛋白質名叫氧化還原酶，其中一部分資料如下：

第五、六及七行便是代表個別原子三維坐標值。(見下表及示意圖) 當有了坐標，我們便可以運用數學公式計算兩個原子之間的距離，而決定它們有否譬如存在氫鍵，然後幫助理解分子的結構。

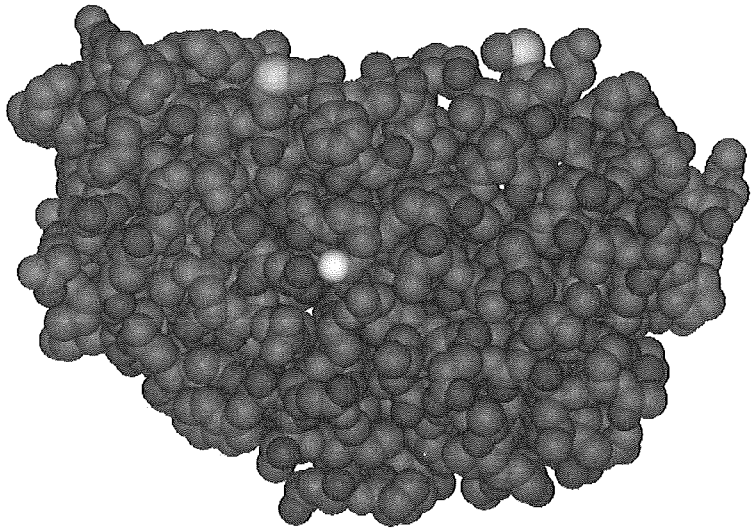
ATOM	1	N	MET A	0	-17.410	2.471	5.878	1.00 34.70	N
ATOM	2	CA	MET A	0	-17.162	1.046	5.517	1.00 35.11	C
ATOM	3	C	MET A	0	-16.558	0.282	6.691	1.00 32.97	C
ATOM	4	O	MET A	0	-16.012	0.882	7.618	1.00 32.19	O
ATOM	5	CB	MET A	0	-16.218	0.968	4.319	1.00 37.62	C
ATOM	6	CG	MET A	0	-14.878	1.638	4.550	1.00 41.52	C
ATOM	7	SD	MET A	0	-13.681	1.206	3.278	1.00 48.55	S

...　　(來源: http://www.rcsb.org/pdb/files/2atj.pdb)



目前已有不少電腦程式能將這些原子坐標轉化為三維結構，並且可以利用這些結構設計用以治病的藥物，用以醫治愛滋病的蛋白酶抑制劑便是一個較有名的例子。

HIV 蛋白酶抑制劑



參考資料：

[1] http://en.wikipedia.org

[2] http://liver.standard.edu/Edu/Edu_hepbinasians.php

[3] http://www.dh.gov.hk/english/main/main_cgs/files/G6PD.pdf

[4] http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

[5] http://www.rcsb.org/pdb/home/home.do

提供詞彙

乙型肝炎病毒	Hepatitis B virus, HBV
人類基因組	Human genome
分子流行病學	Molecular epidemiology
去氧核糖核酸	Deoxyribonucleic acid, DNA
生物訊息	Bioinformatics
目標序列	Target sequence
字長	Word size
多變量分析	Multivariate analysis
血紅蛋白	Haemoglobin
尿嘧啶	Uracil
物種	Species
亮氨酸	Leucine
信使核糖核酸	Messenger ribonucleic acid

染色體	Chromosome
突變	Mutation
胞嘧啶	Cytosine
苯丙氨酸	Phenylalanine
計分矩陣	Scoring matrix
香港生物訊息中心	Hong Kong Bioinformatics Centre
氧化還原酶	Oxidoreductase
氨基酸	Amino acid
胸腺嘧啶	Thymine
基因	Genes
基因序列搜索器	Genetic sequence alignment tools
密碼子表	Codon table
帶因者	Carrier
氫鍵	Hydrogen bond
蛋白質資料庫	Protein Data Bank, PDB

鳥嘌呤	Guanine
腺嘌呤	Adenine
模式識別	Pattern recognition
隱性	Recessive
轉錄	Transcribe
轉譯	Translate
鹼基對	Base pairs

數學百子櫃系列	作者
（一）漫談數學學與教－新高中數學課程必修部分	張家麟、黃毅英、韓藝詩
（二）漫談數學學與教－新高中數學課程延伸部分單元一	韓藝詩、黃毅英、張家麟
（三）漫談數學學與教－新高中數學課程延伸部分單元二	黃毅英、張家麟、韓藝詩
（四）談天說地話數學	梁子傑
（五）數學的應用：圖像處理－矩陣世紀	陳漢夫
（六）數學的應用：投資組合及市場效率	楊良河
（七）數學的應用：基因及蛋白的分析	徐國榮

教育局數學教育組編訂
政府物流服務署印
Prepared by the Mathematics Education Section,
the Education Bureau of the HKSAR
Printed by the Government Logistics Department

ISBN 978-988-8019-00-7