蛋白质氨基酸的组合问题

程 龙　张云军　赵 蕊

教练：胡云芳　龙永红

（中国人民大学经济信息管理系、财政金融系，北京　100872）

摘要　试题B要求参赛者给出模型测定，给定分子量的某一蛋白质的氨基酸组成，这是一个组合问题．文章首先给出了一般的多元线性方程模型，测试结果表明当 $X = 1000$ 时，解的个数为28268个．而实际蛋白质的分子量均在5000以上，因此文章对一般模型加入补充信息和约束条件，给出模型A、B、C和D．考虑到不拥有微机的情况，加强了补充信息和约束条件，给出了模型E和F．文章还对每一个模型都选取了一组或多组数据进行测试，并对测试结果，主要是解的个数与运行时间作了分析。

从整体结构上，文章划分为三部分．第一部分是建立模型前的准备，包括问题重述，问题分析，假设条件和符号约定；第二部分是文章的主体，详细阐述了最一般模型及改进模型A至F的建立，数据测试和结果分析；第三部分是建立模型的善后工作，包括对模型进一步推广和改进的设想，模型误差分析和优缺点分析．

一、问题的提出

生命蛋白质是由若干种氨基酸的不同组合构成的．各种氨基酸的已知分子量 $a[i]$ $(i = 1,2,\cdots n)$ 分别如下：

$n = 18$

$a[1:18] = 57,71,87,97,99,101,103,113,114,115,128,129,131,137,147,156,163,186.$

给定某一蛋白质的分子量 X（$X \leqslant 1000$ 且 X 为正整数）设计数学模型给出该蛋白质的所有可能的组成．即确定该蛋白质是哪几种氨基酸组成以及每种氨基酸的数目．

二、问题的分析

根据给定的分子量 X 及 a_i 测定蛋白质的组成，实际是求多元线性方程：

$$\sum_{i=1}^{18} a_i x_i = X$$

的所有整数解的问题．一般采用枚举法求解，即将所有可能的组合代入方程试验，等式成立即为解．在本问题中，所有可能的组合共有 $\prod_{i=1}^{18} ([X/a_i] + 1)$ 种．因此对于所有的组合，一方面计算量大、耗费时间长（对于计算机尚且如此，在没有微机的情况下更是无法

想象的);另一方面,给出的解个数过多反而失去了解的意义. 考虑到这一点,模型的设计和改进围绕着**减少运算时间**和**缩小解的范围**的思路展开,根据实际化学试验研究中采取的办法,对一般模型加入辅助信息和约束条件. 对实现模型的程序的改进则从改良算法和加入合理判断条件出发.

三、模 型 假 设

1. 给定的蛋白质的分子量 X 和氨基酸已知分子量 a_i 是准确的,没有测试误差;

2. 假设所有被测定的蛋白质均由给定分子量的这几种氨基酸构成,而不含有其它种类的氨基酸. 实际中,构成生命蛋白质的主要氨基酸有 20 种[1-6]. 其中两对氨基酸的分子量相等(见附录 C);

3. 假设蛋白质分子式构成过程中,各个氨基酸分子之间相互结合的方式不影响蛋白质的分子量. 通过计算可知,给定的已知分子量均是氨基酸分子失去 1 分子水后的分子量. 因而在此假设条件下,给定的蛋白质分子量 X 只是几个已知分子量之和而不考虑其它因素;

4. 假设被测定的蛋白质所含氨基酸的个数 $\geqslant 2$,即 $X \geqslant 114$;

5. 假设氨基酸分子结合过程中是任意排列组合的,不存在互斥或互补现象,即任何两种氨基酸都可以同时存在于同一个蛋白质中,没有任何一种氨基酸的存在是以其它氨基酸的存在为前提的. 实际中这一假设是成立的[1-6].

6. 假设在蛋白质中,每种氨基酸存在的概率是相等的,不存在某种必须存在的氨基酸;

7. 假设该实验室拥有测定物质化学性质的仪器.

使用符号说明

a_i 第 i 种氨基酸的已知分子量;

x_i 被测定的蛋白质所含第 i 种氨基酸的数目;

c_{ij} 第 i 种氨基酸所含第 j 种元素的数目;

d_j 被测定的蛋白质中第 j 种元素的数目;

其中 $j=1$ C 元素; $j=2$ N 元素;

$j=3$ O 元素; $j=4$ H 元素;

X 被测定的蛋白质的分子量.

四、最一般的模型

在没有任何其它补充信息和约束条件的情况下,最一般的模型可以表示为

$$\begin{cases} \sum_{i=1}^{18} a_i x_i = X; \\ x_i \text{ 是非负整数 } (i = 1,2,\cdots 18), \end{cases}$$

该模型的解(及解的个数)是由附录 A 的程序给出的. 此程序采用了深度优先算法[7],遍

历了整个解空间，由于采用了分枝限界，其实际最坏的时间效率也是远小于$\prod_{i=1}^{18}([X/a_i]+1)$的．下面的表1是该模型的试验数据．可以看出，当分子量每增加100时，解的个数和运行时间大约增为原来的3倍．

当 $X=1000$ 时，解的个数已达28268个．因此在实际应用中该模型已无多大可行性．为此必须对模型作某些方面的改进，排除无效解，减少解的个数．

在化学中，我们知道，生命蛋白质氮的含量约占总量的16%左右（其波动范围为15%—17%）[6]．蛋白质含量测定的凯式定氮法[5]就是利用了这个性质。在附录A的程序中，我们给出了考虑含氮量的模型（而且下面的几个模型B、C、D也考虑了这种情况）．

在表1中，已给出了考虑含氮量时的解的个数和运行时间的数据．可以看出，经过这种改进，效果一般比以前好得多．

表 1

蛋白质分子量 X	未考虑含氮量的模型		考虑含氮量的模型	
	解的个数	运行时间(秒)	解的个数	运行时间(秒)
200	4	<1	0	<1
300	14	1	0	<1
400	45	2	0	<1
500	158	5	115	3
600	522	15	0	1
700	1508	43	763	23
800	4291	125	0	9
900	11249	321	4301	133
1000	28268	810	10954	335
1001			10177	329

模型A

已知蛋白质的分子式．

根据有关质谱实验在有机化学中的应用方面的材料[2]可知质谱法可以“得到有关分子结构的信息以及化合物的准确分子量和分子式”，因此在模型A中加入如下的假设：

假设 $8a$　假设蛋白质的分子式是已知的．根据有关资料[1-6]，生命蛋白质中常见的氨基酸是由C、N、O、H、S五种元素组成的(见附录C)．已知蛋白质的分子式，即已知各种元素原子的总数目 $d_j(j=1,2,3,4)$．(由于把 S 作特殊处理，d_j 只有4种．)

模型A可以表示为：

$$\begin{cases}\sum_{i=1}^{18} a_i x_i = X; \\ \sum_{i=1}^{18} c_{ij} x_i = d_j; \quad (j=1,2,3,4) \\ x_i \text{ 为非负整数，} i=1,2,\cdots\cdots,18.\end{cases}$$

对该模型有两点说明：

1. 常见的 20 种氨基酸中，有两对的分子量相等. 其中亮氨酸与异亮氨酸为同分异构，分子量与分子式均相同，因而不会影响该模型的计算. 而另一对谷酰氨酸与赖氨酸仅是分子量相同，分子式不同. 因此在模型中，把含硫的两种氨基酸作特殊处理后，还剩下 16 种分子量不同，然后加入一个变量，用以区分谷酰氨酸与赖氨酸. 最后将结果合并.

2. 常见的这 20 种氨基酸中，只有两种氨基酸，即半胱氨酸与蛋氨酸含有 S 元素. 因此在蛋白质分子式中含有 S 元素时，可以通过简单的计算(以及化学试验)，确定含 S 的氨基酸的种类和数目. 我们的模型即假设对含 S 的情况已作过特殊处理.

当然，这样的模型可表为

$$\begin{cases} \sum_{i=1}^{17} a_i x_i = X, \\ \sum_{i=1}^{17} c_{ij} x_i = d_j \quad (j = 1, 2, 3, 4) \\ x_i \text{ 为非负整数 } (i = 1, 2, \cdots, 17). \end{cases}$$

在下面的表 2 中我们给出了一些试验数据.

表 2

蛋白质分子量 X	解的个数	运行时间(秒)
369	3	<1
569	2	1
671	24	6
982	618	118
1133	1195	239

与前面的最一般模型的解的情况作比较，可以看出，解的个数约为最一般模型的 1/20，运行的时间也大大地缩短了.

当然，由于氮的含量(对应其原子个数)事先已知道，所以不必再讨论含氮量的情况. (但我们的数据并不是来源于实际的蛋白质，所以可能含氮量是不符合前面所提到的性质的.)

模型 B

已知蛋白质中某些氨基酸是存在的.

在实际的蛋白质一级结构测定[3]中，通常可以对蛋白质经过充分水解后所得到的氨基酸混合液作离子交换层析、纸层析或薄层层析，定性研究的结果可以确定该蛋白质所含的全部或部分氨基酸种类.

在本模型中，给出如下的假设:

假设 8b 已知被测定的蛋白质中肯定含有其中的 k 种氨基酸，其分子量为 $\bar{a}_j (j = 1, 2, \cdots, k)$. 很显然对应的 $\bar{x}_j \geqslant 1 (j = 1, 2, \cdots, k)$.

因此，可假设 $X' = X - \sum_{j=1}^{k} \bar{a}_j$，即 X 中先扣除已知存在的 k 种氨基酸的分子量(都先减去一份)，现在的模型实际上已同最一般的模型.

$$x_i' = \begin{cases} x_i - 1 & (i \text{ 对应的氨基酸是已知存在的}); \\ x_i & (\text{其他的 } i). \end{cases}$$

则模型表为

$$\begin{cases} \sum_{i=1}^{18} a_i x_i' = X'; \\ x_i' \text{ 为非负整数 } (i = 1,2,\cdots,18). \end{cases}$$

我们注意到，在最一般的模型中，解的个数和运行时间是分子量 X 的单增函数(一般如此)．所以减少 X 就可减少解的个数和运行时间．

下面的表 3 给出了一些试验数据．含氮量的情况也作了类似最一般模型的处理．

表 3

蛋白质分子量 X	未考虑含氮量的模型		考虑含氮量的模型	
	解的个数	运行时间(秒)	解的个数	运行时间(秒)
300	0	<1	0	<1
400	3	<1	0	<1
500	12	1	8	1
600	32	1	0	<1
700	139	4	67	2
800	420	12	0	1
900	1287	37	459	14
1000	3631	104	1570	48
1001	3741	107	1219	38
1200			8821	276

表中的解是针对已知存在分子量为 57，71，87 三种氨基酸的．

模型 C

已知蛋白质中只含有某几种氨基酸．

在比较成功的氨基酸定性分析中，可以得到被测定的蛋白质完全水解生成的氨基酸的全部种类，从而可给出如下的假设：

假设 8c　假设某蛋白质由且仅由 k 种已知的氨基酸(或 k 种不同的分子量对应的氨基酸)构成．

只要 $k < 18$，就可以减少变量个数，从而提高求解速度，减少解的个数，使解限制在一定的范围之内．而且我们知道已知的氨基酸是肯定存在的，即对应的 $x_i \geqslant 1$，这样我们可以令 $X' = X - \sum_{i=1}^{k} a_i'$

(a_i' 是存在的 k 种氨基酸的分子量，$i = 1,2,\cdots,k$)，又令 $x_i'' = x_i' - 1$

(x_i' 对应 k 种氨基酸，$i = 1,2,\cdots,k$)．

模型可表为

$$\begin{cases} \sum a_i' x_i'' = X' \\ x_i'' \text{ 为非负整数 } (i = 1,2,\cdots,k). \end{cases}$$

下面的表 4 给出了一些试验数据. 很显然,由于 X 的量的减少,运行时间和解的个数也会减少.

对考虑含氮量的情况也作了测试.

表 4a

蛋白质分子量 X	解的个数	运行时间(秒)
500	0	<1
600	0	<1
700	0	<1
800	1	<1
900	2	<1
1000	4	<1
1001	3	<1
1100	5	<1
1200	10	<1
2000	129	3

不考虑含氮量的模型,假设由且仅由57、71、87、97、99 五种构成

表 4b

蛋白质分子量 X	解的个数	运行时间(秒)
300	0	<1
400	0	<1
500	0	<1
600	0	<1
700	0	<1
800	0	<1
900	0	<1
1000	0	<1

考虑含氮量的模型,假设由且仅由 57、71、87 三种构成

表 4c

蛋白质分子量 X	已知氨基酸的分子量	未考虑含氮量的模型		考虑含氮量的模型	
		解的个数	运行时间(秒)	解的个数	运行时间(秒)
612	57 115 163	1	<1	0	<1
579	71 87 103 128	1	<1	1	<1
697	57 101 128 137	1	<1	0	<1
1439	97 103 129 163	2	<1	0	<1
887	99 115 147 156	1	<1	1	<1
2069	57 87 101 114 128 156	132	3	60	1
2035	71 99 113 131 163	29	<1	2	<1
3047	57 71 89 114 128 147	1604	31	614	18

模型 D

18 种已知氨基酸分子量的平均值为 118.5,因而平均来看对于 $X \leqslant 1000$ 的蛋白质来说其所含氨基酸的分子数在 8—9 之间,为简化起见,我们不妨设每种氨基酸分子的数目仅为 0 或 1. 因而模型表示为

$$\begin{cases}\sum_{i=1}^{18} a_i x_i = X; \\ x_i = 0 \text{ 或 } 1; \end{cases}$$

运行结果如下表：

表 5

蛋白质分子量 X	未考虑含氮量的模型		考虑含氮量的模型	
	解的个数	运行时间(秒)	解的个数	运行时间(秒)
200	4	<1	0	<1
300	8	<1	0	<1
400	21	1	0	<1
500	53	2	40	1
600	87	2	0	1
700	171	5	101	3
800	226	6	0	1
900	331	10	146	5
1000	393	12	202	7
1001	379	11	148	8
1100	363	12		
1200	392	13	166	6
2000	0	6		
3000	0	5		

模型 E

若实验室不拥有微机，但可能拥有较先进的化学分析设备。设实验室可对完全水解后的氨基酸混合液作定性的分析[1-3]，并可以通过质谱仪测得蛋白质的分子式[2]。因而若设构成被测蛋白质的氨基酸分别为第 $i_1,\cdots,i_k$ 种，则模型可以进一步简化为：

$$\begin{cases}\sum_{l=1}^{k} a_{i_l} x_{i_l} = X \\ \sum_{l=1}^{k} c_{i_l j} x_{i_l} = d_j \quad (j = 1,\cdots,4) \\ x_{i_l} \text{ 为正整数 } (l = 1,\cdots,k). \end{cases}$$

当 k 的取值不大(如 $k \leqslant 8$)的情况下，可先求出线性方程组的通解，然后再找出其整数解。然而当 k 的取值较大时，对手工计算来说，该模型就不太可行了。

模型 F

进一步假设实验室拥有先进的氨基酸自动分析仪，可对完全水解后的氨基酸混合液作定性和定量分析，得出被测蛋白质所含氨基酸的种类及各种氨基酸之间的比例关系为：$b_{i_1}:b_{i_2}\cdots b_{i_k}$[1]，因而模型可表述为：

$$\begin{cases} X = \sum_{l=1}^{k} a_{i_l} x_{i_l} = r \sum_{l=1}^{k} a_{i_l} b_{i_l} \\ \text{其中 } x_{i_l} = r \cdot b_{i_l}, \ (l = 1,\cdots,k); \end{cases}$$

$$\therefore \quad r = X/\sum_{i=1}^{k} a_{i_l} b_{i_l}$$

经过上述简单的运算便能得出问题的解，并且解是唯一的，可见氨基酸自动分析仪对解决本问题是比较方便的.

模型的改进方向

从上述各模型可以看出：变量众多是解决该问题的困难所在，因而寻找有效的减少变量个数的方法是模型进一步改进的重要方向. 除了作上述的一些改进外，我们还可以从所给氨基酸分子量的内部联系出发，得到它们之间的一些关系，如

$$71 = 57 + 14 \qquad 99 = 57 + 42 = 57 + 3 \times 14$$

$$113 = 57 + 56 = 57 + 4 \times 14$$

$$\vdots$$

类似的分解可以使变量的个数大大减少，从而也大大减少计算量(特别是人工求解). 当然，如此求出解后再进行组合得原问题的解也是较复杂的.

在实际求解过程中我们发现： 只进行纯粹的分解而得到的解并非都符合实际情况. 由此，我们从实际应用的角度出发，充分利用可能得到的信息，对一般模型作了一系列的改进. 除此之外，我们还可以利用其它一些信息，如(1)质谱分析仪可以测定分子的结构，据此我们可以分析单键和双键的数目，也可以根据羟基、苯环等的性质测定其数目；(2)根据R基的不同可以将氨基酸分为极性和非极性或者中性、酸性、碱性，通过酸碱中和滴定、电泳分离等方式测试出蛋白质中每类氨基酸的含量，从而将18个变量的模型分解为几组变量较少的模型[2,4]. 然后再进行与上述模型相似的计算. 总之，如果能充分利用现实中得到的有用信息，将其作为约束条件纳入模型中去进行综合考虑，我们相信其效果将会更好.

模型的误差分析

1、X的测定误差是影响结果正确的一个重要因素. 如果X的测量值与真实值相差1，其结果将会有很大的变化.

2、a_i的测量误差对模型的结果也会有一定的影响.

3. 在生命蛋白质含氮量的约束条件中，关于含氮量的范围在不同的资料中有点不同，有为15%—17%，亦有为15%—17.6%，但确实说明有些规律存在，我们取了15%—17%可能会引起误差.

模型的优点

我们给出的一系列模型，特别是“最一般的模型”适用范围较广，这主要表现在：

1. 无论X增大或者氨基酸的种类增多模型总是有效的，并可以给出所有可能的解. 同时由于组成生命蛋白质最主要的氨基酸只有20种，分子量只有18种，因而我们的模型对于分析蛋白质组成这一问题更有实际意义.

2. 考虑到不同实验室的设备条件和获取以上信息的能力不同，我们给出了模型A—C、E、F以满足不同的实际情况的需要.

3. 我们建立这些模型的方法和思想对其它类似问题也很适用，象多糖等类似高分子化合物的组成分析，我们只需改变模型中的某些参数就可作类似分析。

模型的缺点

1. 我们模型的缺点仍然在于如何解决模型给出的解数目太多的问题。例如当 $X=1000$ 时，最一般的模型给出了 28268 个解，改进的模型中最多可以将其减少到几个，然而一般来说蛋白质的分子量都在 5000 以上，那么解的个数(即使是改进的模型)将仍然是很可观的。

2. 在改进的模型中，由于约束条件对试验数据的要求较严格，因而我们构造的某些测试数据可能是不现实的，从而某些模型中得到了不太理想的结果。

3. 一些模型所加入的约束条件可能也有不太现实的，如模型 D 中假设 $x_i=0$ 或 1 就可能与现实不太相符，但如果将 0—1 约束改为下限和上限的约束可能就比较实际了。

4. 蛋白质含氮量的约束条件只是基于一般情况，而没有考虑例外情况。

附录 C　20 种常见氨基酸的名称，分子式和分子量

分子量	名称	分子式	元素数目 C	N	O	H
57	甘氨酸	NH_2CH_2COOH	2	1	1	3
71	丙氨酸	$CH_3NH_2CHCOOH$	3	1	1	5
87	丝氨酸	$HOCH_2NH_2CHCOOH$	3	1	2	5
97	脯氨酸	$CH_2CH_2CH_2NHCHCOOH$	5	1	1	7
99	缬氨酸	$CH_3CH_3CHNH_2CHCOOH$	5	1	1	9
101	苏氨酸	$CH_3CHCHNH_2CHCOOH$	4	1	2	7
103	半胱氨酸	$HSCH_2NH_2CHCOOH$	3	1	1	5
113	亮氨酸	$CH_3CH_3CHCH_2NH_2CHCOOH$	6	1	1	11
113	异亮氨酸	$CH_3CH_3CH_2CHNH_2CHCOOH$	6	1	1	11
114	天冬酰胺	$NH_2COCH_2NH_2CHCOOH$	4	2	2	6
115	天冬氨酸	$OHOOCH_2NH_2CHCOOH$	4	1	3	5
128	谷酰胺	$NH_2COCH_2CH_2NH_2CHCOOH$	5	2	2	8
128	赖氨酸	$NH_2CH_2CH_2CH_2NH_2CHCOOH$	6	2	1	12
129	谷氨酸	$OHCOCH_2CH_2NH_2CHCOOH$	5	1	3	7
131	蛋氨酸	$CH_3SCH_2CH_2NH_2CHCOOH$	5	1	1	9
137	组氨酸	$CHNCHNHCCH_2NH_2CHCOOH$	6	3	1	7
147	苯丙氨酸	$C_6H_5CH_2NH_2CHCOOH$	9	1	1	9
156	精氨酸	$NH_2NHCNHCH_2CH_2CH_2NH_2CHCOOH$	6	4	1	12
163	酪氨酸	$OHC_6H_4CH_2NH_2CHCOOH$	9	1	2	9
186	色氨酸	$C_8H_6NHCHCCH_2NH_2CHCOOH$	11	2	1	10

注．表中给出的分子量及元素数目都是原氨基酸除去 1 分子水后的值。

参　考　文　献

[1] R.M. 罗伯茨等，近代实验有机化学导论，上海科学技术出版社。
[2] 基础有机化学，人民教育出版社。
[3] L.F. 费塞尔，K.L 威廉森，有机实验，高等教育出版社。
[4] C.D. 古奇，D.F. 帕斯托，有机化学基础，高等教育出版社。
[5] 黄梅丽，江小梅，食品化学，中国人民大学出版社。
[6] 华东华北区粮食学校编写组，有机化学，江西人民出版社。
[7] 邹海明，余祥宣，计算机算法基础，华中工学院出版社。