

# 关于“蛋白质氨基酸的组合问题”的评注

韩 继 业

(中国科学院应用数学研究所, 北京 100080)

**摘要** 本文介绍 1992 年数学模型竞赛中一个离散数学问题的一个比较好的答案, 并且扼要地讨论了大规模离散数学问题的一些求解途径, 最后阐述了离散数学在理论上和实用上的重要意义。

关于蛋白质的氨基酸的可能组成, 问题是: 生命蛋白质是由一些氨基酸的不同组合构成的, 现只考虑 18 种主要的氨基酸, 它们的分子量分别为 57, 71, 87, 97, 99, 101, 103, 113, 114, 115, 128, 129, 131, 137, 147, 156, 163, 186. 令  $a_i$  表示上述分子量 ( $i \leq 18$ ), 给定某一蛋白质的分子量  $X$  ( $X \leq 1000$ ,  $X$  为正整数), 设计出数学模型以给出该蛋白质的所有可能组成. 即确定该蛋白质是由哪几种氨基酸组成以及每种氨基酸的个数.

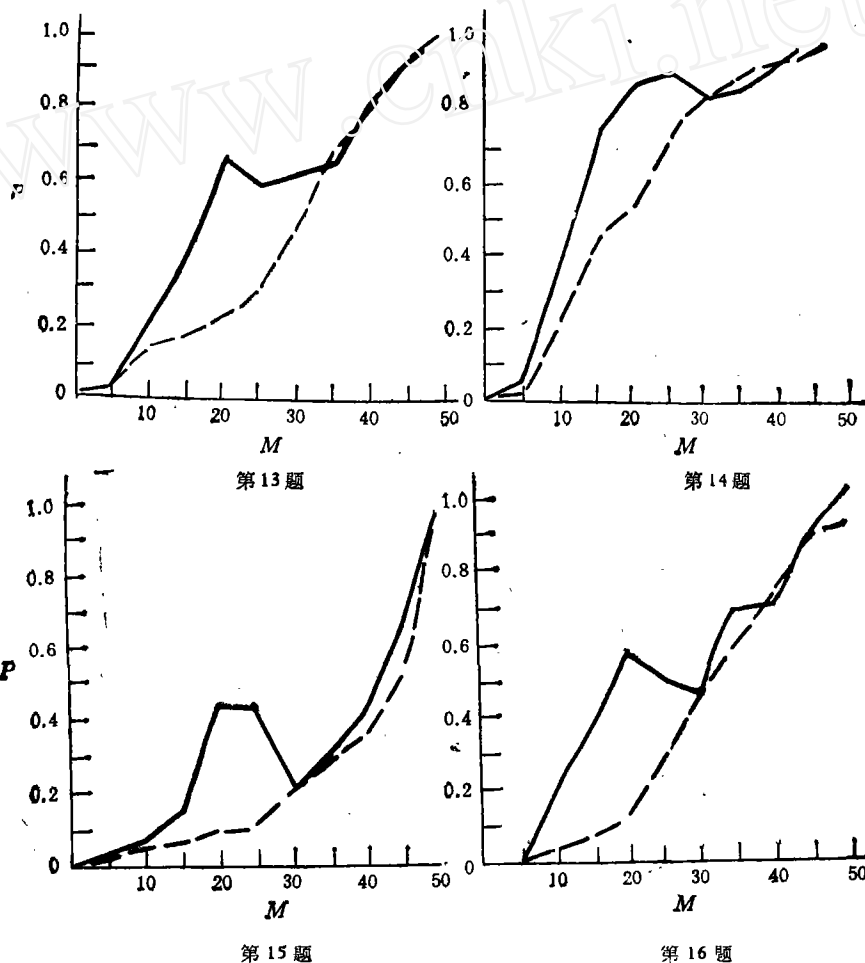
上述问题易被描述为线性不定方程式  $\sum_{i=1}^{18} a_i x_i = X$ , 这里  $x_i$  是第  $i$  种氨基酸的个数 ( $i = 1, \dots, 18$ ), 它只取非负整数. 该蛋白质的所有可能组成都包括在方程式  $\sum_{i=1}^{18} a_i x_i = X$  的所有非负整数解中. 从理论上说, 采用枚举法可以求出此方程式的全部非负整数解. 但实际上, 当  $X$  很大时, 方程式的全部解 (指非负整数解) 的个数是一很大数目 (如  $X = 1000$ , 解的个数为 28268), 并且解的个数随  $X$  的增大以指数关系增加. 一般的蛋白质的分子量大于 5000, 所以即使利用计算机也难以得出全部解. 同时大量的解无实际意义. 解决这类大规模离散数学问题, 需要建立若干补充的约束条件, 以尽可能多地消除无实际意义的解. 另一方面也需要从数学上研究离散数学问题的性质, 以探求更快的求解算法.

去年北京市大学生数学模型竞赛时, 有近半数的参赛小组选择了这一试题. 各小组在使用计算机的条件下都能较快地求出 28268 个解 ( $X = 1000$ ). 很多小组从数学上考虑了一些技巧, 以简化算法. 但如  $X$  变大, 例如  $X = 1500$ , 解的个数将急剧增大, 很多小组所考虑的简化计算的技巧将失去作用. 这时建立补充的约束条件显得非常重要. 中国人民大学的程龙, 张云军和赵蕊小组, 在建立约束条件方面被认为处理得比较好, 他们的答案获得了去年北京市数学模型竞赛的特等奖 (见本期文章“蛋白质氨基酸的组合问题”). 他们在比较多地查阅了有机化学有关文献后, 首先注意到: 生命蛋白质中氮的含量一般占总量的 15%—17%. 利用这一性质建立的约束条件对  $X = 1000$  可使解的个数从 28268 减为 10954. 即消除了一半以上的无实际意义的解. 其次, 他们注意到生命蛋白质中常见的氨基酸是由氮、碳、氢、氧和硫五种元素组成, 并且利用质谱实验法可以得到化合物的分子结构的信息及准确分子量和分子式. 利用这一情况, 建立了若干约束条件. 令  $d_j$  ( $j = 1, \dots, 5$ ) 表示氮等五种元素的原子总个数, 若  $d_j$  已知, 则有约束条件  $\sum_{i=1}^{18} c_{ij} x_i = d_j$ ,  $j = 1, \dots, 5$ , 其中  $c_{ij}$  是第  $i$  种氨基酸中所含的第  $j$  种元素的数目. 这

组约束条件消除了更多的不需要考虑的解,使解的个数减少到原来解的数目的廿分之一。他们还考虑了其他的约束条件。尽管有些约束条件的合理性需要讨论,但这种处理一些离散数学问题的途径是有效的。

离散数学问题,特别是离散最优化(或组合最优化)问题目前在国民经济方面、工程技术方面和军事方面等有广泛的应用背景。例如平板的最优激光钻孔、油田的最优勘探、血液银行的管理、基因密码、计算机切割的最优安排、交通工具的调度计划、板材的最优切割、偏好和选择判断的汇集、生产过程的调变安排、工厂的最优选址等等。都可归结为离散最优化问题。这门学科随着计算机的迅速发展而在理论、方法和实用上得到了巨大的发展,许多组合优化问题没有快速的计算机就不能被解决。1984年美国数学科学资金来源特别委员会在其报告“进一步繁荣美国数学”中提出:近几十年内国际上数学发展的趋向包括了“离散数学的作用将不断扩大”(趋向的其他内容是“对非线性问题的关注将进一步增长,概率分析的作用将不断扩大,大规模科学计算将进一步发展”等)。这种现象已经反映到国外数学模型课程的内容以及数模竞赛的题目方面,值得我们重视。

(上接第31页)



上面给出了七种进行偏差分析的方法,每种方法都是从各自的角度进行分析,所得结论也不尽相同,读者可根据自己的需要及兴趣利用其中的某些结论或进行综合分析。