第 31 卷第 1 期 数学的实践与认识 Vol. 31 No. 1
2001 年 1 月 MATHEMATICS IN PRACTICE AND THEORY Jan. 2001

DNA 序列中的结构与简化模型

孟大志

(北京工业大学, 北京 100022)

摘要: 本文简述 2000 年全国大学生数学建模竞赛A 题的科学研究背景, 以及题目的立意和设计. 进而对解答A 题的大学生们的出色方法进行介绍与评述

1 引 子

这是我第一次参与全国大学生数学建模竞赛, 深深地被这一十分有意义的赛事蒸蒸日上的发展所鼓舞, 为在赛事中涌现出来的青年学生们聪明才智和对科学强烈的热爱而惊喜, 为自己在这次参与中学到的和感受到的十分有益的影响而兴奋. 2000 年 7 月清华的唐云教授电话约我为竞赛出一道题, 出于个人兴趣, 也出于希望青年学生更关注在重大科学问题中运用数学和发展数学, 于是就在全世界被人类基因组计划的成果掀起的巨大热潮中, 找一个题目, 以期诱导有志青年投入这一二十一世纪的科学热点中. 我和领导建模比赛的全国组委会的一些教授们(叶其孝、姜启源、王强、唐云等) 共同讨论了这个题目, 反复修改和润色, 希望更适合中国大学生的实际. 但一直担心这样一个热点科学中引出的问题, 一个开放式问题的太大的自由度是否会为难青年学生. 结果出人意料, 特别是重点大学的参赛队, 十分热烈地选择A 题作为他们一显身手的考卷, 而且答出了同样出乎意料的水平. 然而在A 题的理解、解法及评判的一系列问题中, 仍有许多问题需要明确, 于是我应组委会之邀, 特写此文力窥全豹, 也对参与竞赛的师生们作一个交待.

2 A 题的背景

2000 年 6 月 26 日, "人类基因组计划"规定的禁发时间(EMBARGO) 北京时间 18: 00 刚过, 新华社、法新社、美联社、路透社……各国新闻发布机构以第一条消息发布了人类基因组草图绘就的重要消息. 美国总统克林顿在白宫举行的庆祝仪式上表示, 人类基因组草图是迄今"人类所绘制的最为奇妙的图谱"; 英国首相布莱尔说: "这是 21 世纪第一项伟大的科技成就……医学科学领域一场革命, 其意义远远超过抗生素的发现"; 日本首相森喜郎在声明中指出, 人类基因组草图绘制成功, 代表人类在破解自身构成方面向前迈出巨大的一步; ……许多国家的元首, 科技官员和著名科学家纷纷发表谈话, 赞扬人类基因组草图的完成, 评估这一伟大成果的意义. 直到 6 月 28 日, 中国主席江泽民在中央思想政治工作会议上也对人类基因组的意义作出评价并赞扬了中国科学家在其中的出色工作[1].

显然, 当 7 月份组委会提出建模赛题一事时, 顺应这一世纪科学大事, 在其中构造赛题, 将引导青年学子关注世界科技热点, 鼓励学生敢于投身到科学重大问题中去, 培养学生用数学为工具去解决科学技术问题的能力方面都具有了特殊的意义.

2003 年将完成人类基因组DNA 全序列的测序, 它将带给人类一本"自身的说明书", 这

对人类认识自己, 保护自身, 发展新的生物产业都将是意义重大的. 在许多科普读物中, 将人类基因组全序列这部"书"描绘成一座巨大金矿, 解读这部书就是从中发掘出无量的财富, 这种比喻一点儿也不过分. 生命科学称这一研究阶段为"后基因组时期"或"后基因组计划"(Post-Genome Project), 而将数学与计算机科学融入这一计划之中, 又常被人称为生物信息学(Bioinformation). 人类基因组研究中已经浮现出大量的数学问题, 已为世界上众多数学家关注[2]. 作为解读基因组这一庞大计划的一个十分重要而又基础的部分, 就研究基因组的结构, 而其中更基础的是DNA 序列的结构. "结构"这个词在这里的含义是十分广泛的, 也就是说, 作为由A、T、C、G 四个字符组成的一个有序字符串, 任何呈现规律性的特征都可以称为结构. 由于规律呈现范围不同, 我们又可以分为局部结构与整体结构, 或称小尺度结构与大尺度结构, 这些结构的揭示将大大有助于人们对于基因与基因组的解读. 这一点可以形象地比喻为一部 100 万页的书, 如果我们能够知道这部"天书"的篇、章、节的结构, 甚至段落、语句或词的结构都清楚了, 要读懂这部书的内容就变得容易了. 从这种意义上说, DNA 序列的结构的研究显然是生物信息学中重要的内容之一.

本届数学建模比赛的A 题是在这一世界科学发展的大背景下, 作为二十世纪最后一届比赛, 以翘首二十一世纪的姿态, 选择基因组研究为命题的学科领域, 以后基因组计划中生物信息的DNA 序列结构作为课题, 是顺应时代潮流的具有前瞻性的选题.

3　A 题的立意

在A 题设计之前, 立意就很明确: 源于科学实际, 解法充分开放.

本题取材于DNA 的结构的研究, 这里的结构指的是在DAN 序列中重复出现的有特征的片断, 这种重复出现形成了规律. 由于结构的含义是广泛的, 担心学生因此而无从下手, 我们特别举出三种结构为例, 其目的仅仅是为了说明, DNA 序列貌似随机地由A、T、C、G 四个字符组成, 但它之所以有"万能"的功能, 正是由于在随机的外衣下隐藏着大量的结构, 正是这种结构决定了功能. 因此, 在生物信息学中, 人们普遍相信这样一个信条: 序列——结构——功能. 这一信条引导人们成功地在DNA 序列中挖掘出许多与生物功能相关的自然规律. 在A 题中举出的三种结构是十分基础而且在科学界广泛为人们所接受的. 一种是四种碱基的丰度, 对于DNA 序列的不同的片段常常表现出碱基丰度的差别, 因此碱基的丰度往往成为区别不同序列片段的特征; 第二种是三联子对蛋白质的编码, 它首先由发现DNA 双螺旋结构的克里克和南非的分子生物学家西德尼·布伦纳确定的, 这种不重叠的三联子组成的编码区(Exon)与非编码区的交替出现形成了DNA 序列中一个重要的结构. 如果读者想了解这一方面的知识只要在互联网上搜索"Exon-Intron Structure", 你会得到供选读的大量文献; A 题举的第三个例子是所谓DNA 序列的长程相关性, 这一规律最早由C·K·Peng 等人在 1992 年Nature 上报导[3], 此后人们研究了各种DNA 长序列, 分别发现了DNA 序列在大尺度的范围内具有统计相关性, 然而这种相关性的细节及意义至今还是一个迷. A 题中举出这三种结构, 也为了说明在DNA 序列的结构中既有大尺度全局性的, 也有局部性的, 研究和发现DNA 序列中的这些规律均有重要意义.

正由于这种结构的多样性和一般性, 为求解A 题确定了解法的开放性. 虽然事实上许多试卷都把这一结构理解成为编码区与非编码区, 但这种局限性的理解并没有比一般性理解结构的试卷更好些. A 题定义结构的一般性, 有两方面的理由. 一方面希望在求解A 题

时对生物知识的依赖不要太多, 除了最基本的DNA 序列的背景外, 解题中并不需要有更多的基因组结构的知识(例如, 是否知道 Exon 与 Intron 并无大关系). 这样做是为了在"数学建模"这一基本的专业性质下平等. 第二个方面就是希望这种开放性, 可以使从初等到高等的许多数学模型化方法均能对A 题做出一定水平的解答. 而且也希望发现一些富有创造性的、十分有效的方法. 事实上, 本届比赛中也的确涌现出大量富有创意的方法, 实在令命题者兴奋不已.

解答方法的开放性, 是A 题的命题领域本身就决定了的. 事实上, 仅在编码区预测的文献中就有了许多不同的方法. 有通过核苷酸片段差异的区分方法[4], 同源比较算法[5], 隐马尔可夫模型(Hidden Markov Model, HMM). 这种方法将DNA 序列的形成看作随机过程, 而 HMM 可自动找出其隐藏的统计规律性[6]. 大家熟知的动态规划方法[7], 以及傅立叶分析[8], 线性判别分析(Linear Discriminant Analysis, LDA)[9]. 此外许多专门的方法用于DNA 的结构分析与寻找: 法则系统(rule-based system)[10], 语言系统(linguistic)[11], 决策树(decision tree)[12]. 这些方法对于从DNA 序列中找出编码序列均有很好效果, 有些准确率高达 90%. 有兴趣的读者可以在最近出版的《解码生命》[13]一书中查到有关评论.

A 题将DNA 结构的研究具体化为不同序列的分类, 这种分类对于寻找出序列的结构具有基础的价值. 它是寻找结构的一种简化而有效的变形, 这种具体化在帮助学生模型化是有益的. 然而这种具体化也给出题带来一定困难, 为了方便广大参赛队对这种分类方法的理解与数值实验, 我们设计了两套数据. 一套是人工构造的数据, 而另一套是来源于自然的DNA 数据库. 显然这两套数据既有联系又有明显的差别, 这种差别使得企图用比较简单的方法而不加区别地处理这两类数据将不会得到好的效果. 正如自然界给人类提出的问题不太可能恰好满足我们希望的数学条件一样, A 题也要求解题者具有立足于实际, 从有限而不完全的已知数据去探索更复杂的数据中的未知规律这样一种研究素质.

4 阅卷随想

在评阅试卷时, 老师们对年轻学子在A 题解法中表现出的热情、智慧、严谨和富予创造性都留下极深刻的印象. 作为命题人, 更对本科学生能在短短的三天中所做出的成果惊喜, 并在许多十分聪明的解法中学习到了新的东西. A 题的试卷几乎令所有阅卷老师叹服: 中国大学生年轻有为!

学生论文的立意大多在"特征提取→分类方法"这一模式, 这显然是最容易想到的, 大多数试卷也在这一立意之下, 选择好的方法而得到较好的结果. 特征的选择, 首先易于让人想到的是A、T、C、G 四个字符在字符串中出现的频率, 这在文献中常称为"单个碱基丰度". 单纯使用这一特征, 许多学生的文章对人工数据得到好的结果, 但对后面 182 个序列的分类却常常不太理想. 在优秀论文中浙江大学的一个队将这种特征提取后形成四维特征向量, 然后分别用欧氏距离、马氏距离分类法和 Fisher 判别模型, 对人工数据得到理想的分类, 对自然数据(182 个)也得到很高的分类正确率, 是这一类算法中较突出的卷例. 另有一些试卷在这一特征基础上考虑到字符的顺序, 将模型做得更复杂些. 更多的论文是用 4 个字符的字符串作为特征, 由于这时特征一下子增加了许多, 于是需要从其中评判挑选并排出特征的重要性顺序, 这种特征的提取往往可以得到较好的效果. 特别是对于自然序列, 大连理工大学的一个队通过概率统计方法首先对已知的人工序列集进行特征提取, 从而形成特征向

量较为全面地表达分类特征, 当然也出现了高维问题的计算复杂性, 他们得到了很好的分类效果　值得指出的是, 由于竞赛题一方面源于生物学实际问题, 同时又相对地独立于生物而形成适当抽象的"试题", 因此试题并不是基因组中某种结构的翻版　有些试卷过多地研究了生物学的来源, 而且将A 题仅局限于他们所想象的结构(例如 Exon 结构), 于是三联子编码成为分类的唯一特征, 而三联码的不重叠性又使他们在阅读框的起始位置前不知所措, 以至所产生的结果不理想

在分类方法上, 统计的方法(特别是聚类方法) 是最易于想到的, 许多试卷从而构造了好的方法　但是简单而不加修正地使用统计方法并不能得到好的结果　这是因为人工已知序列的样本数只有 20 个, 而且都很短, 待分类的自然数据样本数 182 且都长得多, 因此从小样本中得到的统计规律在处理大样本时效果显然不佳　这是众多用统计方法所得到结果不理想的一个直接原因　有些学生看到并指出了这一点, 而且有的试卷注意到人工数据与自然数据的生物学的差别而在分类自然序列时修改了分类方法而得到较好的结果, 显然概念的清楚与思维的灵活得到很好的统一.　用各种方式构造判别函数的方法以及神经网络的方法, 特别对于非线性系统的识别很有效　因此通过构造各种神经网络来进行分类, 更多的队得到很好的效果　例如大连理工大学的一个队, 用统计方法提取较好的特征又用BP 网络进行分类, 方法严谨, 考虑细致, 对自然序列的分类正确率高达 88%.　而科技大学的一个队通过对神经网络方法的逐层的改进, 又辅以统计方法, 产生了比较精细的网络算法, 也得到分类自然数据的正确率达 65% 的好效果

除了上述大量"正规方法"以外, 一些试卷有创意地提出了一些十分新颖的思想, 有些还取得了很好的效果　例如中国科技大学的一个队将序列看作信息流, 注意到字母出现的特征是熵的改变, 是十分新意的, 他们最终又将设计好的几个模型形成综合判别的目标函数, 也得到好的分类效果, 对自然数据分类正确性达 58%.　而北京大学的一个队将DNA 字符串看作一篇文章, 而利用了类似文本分类中的特征判别方法定义关键词标准, 进而使用优选法, 找出关键词的特征, 然后使用层次分类　他们的方法精细, 尽管分类最终效果并不十分理想, 仍不失为值得一读的好文章　由于篇幅有限, 有些文章虽然没有作为优秀论文刊出, 但是在其中仍然表现出学生丰富的想象力和创造精神　一篇十分有趣的文章是大连理工大学的另一个队, 这些学生既没有拘泥于"特征提取→分类"的模式, 也没有局限自己的思维于"概率统计"、"神经网络"、"判别函数"等"大路"方法　他们深入地分析了序列问题的生物来源, 又观察人工序列的数学结构和数值试验结果, 在一些DNA 序列几何表达文献的启发下, 提出了简捷的几何分类法, 得到了出色的分类结果　对自然数据分类的正确率高达 94%.　而且这种不依赖训练集的方法, 属于目前研究基因组结构的令人关注的方向

应当指出, 科研能力的表现是多方面的.　在试卷中, 我们注意到许多学生十分用心于科学文献的检索、阅读与借鉴　例如一些试卷研究了我国著名学者, 中科院院士张春霆教授的 Z 曲线方法[14], 并简化用于A 题分类(例如中国科技大学的另一个队), 也取得好的结果　此外, 特别值得指出的是香港城市大学的论文, 该文的思路清晰, 表述严谨, 图表数据完整, 行文流畅, 作为本科学生三天完成的科研论文值得赞赏!

综上所述, 作为A 题的命题人, 原先的担心与顾虑被事实扫得干干净净.　学生的聪明才智、扎实的数学功底和运用于实际问题的灵活性、创造性证明, 中国大学生完全可以适应更贴近科学研究实际, 更贴近工程技术实际, 更贴近社会经济生活实际的数学建模比赛问题

中国大学生在数学建模比赛的锻炼中必将大大提高应用数学的能力,在二十一世纪的人类新科技的发展中做出出色的成绩

参考文献:

[1] 子言,基因:讲述生命的故事.经济日报出版社,2000 年 7 月.

[2] Mathematics: Frontiers and Perspectives, AMS. Providence, 2000. M. Atiyah 前言.

[3] Peng C. K. Buldyvev, S. V. Goldberger, A. L. Havlin, S. Sxiortino, F. Simonso, M. And Stanley, H. E. Long-range correlation in nucleotide sequences, nature 356: 168~170.

[4] Claverie J M. Computational methods for the identification of genes in vertebrate genomic sequence hum Mol Genet, 1997, 6(10): 1735~1744.

[5] Green P, Lipman D, Hillier L, Waterston R, States D, Clavierie JM. Ancient conserved regions in new gene sequences and the protein databases. Science, 1993, 259: 1711~1716.

[6] Kroyh A, Mian I S, Hanssler D. A hidden Markov model that finds genes in E. coli DNA, Nucleic Acids Res, 1994, 22(22): 4768~4778.

[7] Gelfand M S, Roytberg M A. Prediction of the exon-intron structure by a dynamic programming approach. Biosystems, 1993, 30(1~3): 173~182.

[8] Tiwavi S. Ramachandran S, Bhattacharga A, Bhattacgarga S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. comput Appl Biosci, 1997, 13(3): 263~270.

[9] Fickett JM, Tung C S. Assessment of protein coding measures. Nucleic Acids Res, 1992, 20(24): 6441~6450.

[10] Guigo R, Knudsen S, Drake N, Smith T. Prediction of gene structure. J Mol Biol, 1992, 226(1): 141~157.

[11] Dong S, Searls D B. Gene structure prediction by linguistic methods. Genomics, 23(3): 540~551.

[12] Salzberg S. Locating protein coding regions in human DNA using a decision tree algorithm. J Comput Biol, 2(3): 473~485.

[13] 贺林主编 解码生命——人类基因组计划和后基因组计划.科学出版社,2000 年.

[14] ZHANG Chun-Ting, LIN Zhe-suai, YNA Ming, ZHANG Ren. A Novel Approach to Distinguish Between Intron-containing and Intronless Genes Based on Format of Z Curves. J theor Biol, 1998, 192: 467~473.

The Structure of DNA Sequence and Simple Model

MENG Da-zhi

(Department of Applied Mathematics, Beijing Polytechnic University, Beijing 100022)

Abstract: In this article the scientific research background of the problem A in the CUMCM-2000 as well as its intention and conception are simply stated. Moreover, some excellent methods proposed by the students participating in the contest for the answer to this problem are introduced and reviewed.