

2000 年全国大学生数学建模竞赛

A 题 DNA 序列分类

2000 年 6 月, 人类基因组计划中 DNA 全序列草图完成, 预计 2001 年可以完成精确的全序列图, 此后人类将拥有一本记录着自身生老病死及遗传进化的全部信息的“天书”。这本大自然写成的“天书”是由 4 个字符 A, T, C, G 按一定顺序排成的长约 30 亿的序列, 其中没有“断句”也没有标点符号, 除了这 4 个字符表示 4 种碱基以外, 人们对它包含的“内容”知之甚少, 难以读懂。破译这部世界上最巨量信息的“天书”是二十一世纪最重要的任务之一。在这个目标中, 研究 DNA 全序列具有什么结构, 由这 4 个字符排成的看似随机的序列中隐藏着什么规律, 又是解读这部天书的基础, 是生物信息学 (Bioinformatics) 最重要的课题之一。

虽然人类对这部“天书”知之甚少, 但也发现了 DNA 序列中的一些规律性和结构。例如, 在全序列中有一些是用于编码蛋白质的序列片段, 即由这 4 个字符组成的 64 种不同的 3 字符串, 其中大多数用于编码构成蛋白质的 20 种氨基酸。又例如, 在不用于编码蛋白质的序列片段中, A 和 T 的含量特别多些, 于是以某些碱基特别丰富作为特征去研究 DNA 序列的结构也取得了一些结果。此外, 利用统计的方法还发现序列的某些片段之间具有相关性, 等等。这些发现让人们相信, DNA 序列中存在着局部的和全局性的结构, 充分发掘序列的结构对理解 DNA 全序列是十分有意义的。目前在这项研究中最普通的思想是省略序列的某些细节, 突出特征, 然后将其表示成适当的数学对象。这种被称为粗粒化和模型化的方法往往有助于研究规律性和结构。

作为研究 DNA 序列的结构的尝试, 提出以下对序列集合进行分类的问题:

1) 下面有 20 个已知类别的人工制造的序列 (见反面), 其中序列标号 1—10 为 A 类, 11-20 为 B 类。请从中提取特征, 构造分类方法, 并用这些已知类别的序列, 衡量你的方法是否足够好。然后用你认为满意的方法, 对另外 20 个未标明类别的人工序列 (标号 21—40) 进行分类, 把结果用序号 (按从小到大的顺序) 标明它们的类别 (无法分类的不写入):

A 类	B 类
-----	-----

请详细描述你的方法, 给出计算程序。如果你部分地使用了现成的分类方法, 也要将方法名称准确注明。

这 40 个序列也放在如下地址的网页上, 用数据文件 Art-model-data 标识, 供下载:

- 网易网址: www.163.com 教育频道 在线试题;
- 教育网: www.cbi.pku.edu.cn News mcm2000
- 教育网: www.csiam.edu.cn/mcm

2) 在同样网址的数据文件 Nat-model-data 中给出了 182 个自然 DNA 序列, 它们都较长。

用你的分类方法对它们进行分类, 像 1) 一样地给出分类结果。

提示: 衡量分类方法优劣的标准是分类的正确率, 构造分类方法有许多途径, 例如提取序列的某些特征, 给出它们的数学表示: 几何空间或向量空间的元素等, 然后再选择或构造适合这种数学表示的分类方法; 又例如构造概率统计模型, 然后用统计方法分类等。

Art-model-data

```
1.aggcacggaaaaacgggaataacggaggaggacttggcacggcattacacggaggacgaggtaaaggaggcttg
tctacggccggaaagtgaaggggatatgaccgcttg
2.cggaggacaaaacgggatggcgggtattggaggtggcggactgttcggggaattattcggtttaaacgggacaagg
aaggcggctggaacaaccggacggtggcagcaaagga
3.gggacggatacggattctggccacggacggaaaggaggacacggcggacatacacggcggcaacggacggaacg
gaggaaggagggcggcaatcggtacggaggcggcgga
4.atggataacggaacaaccagacaaacttcggtagaaatacagaagcttagatgcatatgtttttaataaa
atgttattattatggtatcataaaaaaaggttgcga
5.cggctggcggacaacggactggcggattccaaaaacggaggaggcggacggaggctacaccaccgttccggcgg
aaaggcggagggtggcaggaggctcattacggggag
6.atggaaaatttccgaaaggcggcaggcaggaggcaaaggcggaaaggaaggaacggcggatattcggaagt
ggatattaggagggcggaaataaaggaacggcggcaca
7.atgggattattgaatggcggaggaagatccggaataaataatggcggaaagaactgttttcggaatggaaaa
aggactaggaatcggcggcaggaaggatattggaggcg
8.atggccgatcggcttaggctggaaggaacaaataggcggaaattaaggaaggcgttctcgcttttcgacaaggag
cgggacataggaggcggattaggaacggttatgagg
9.atggcggaaaaaggaatgttggcatcggcgggctccggcaactggaggttcggccatggaggcgaaaatcgt
ggcggcggcagcgtggcgggagttgaggagcgcg
10.tggccgcggaggggcccgtcgggcgggatttctacaagggtcctgttaaggaggtggcatccaggcgtcg
cacgctcggcggcaggaggcacgcgggaaaaaacg
11.gttagatttaacgtttttatggaattatggaattataaaatttaaaatttatatttttaggtaagtaac
caacgtttttattacttttaaaattaaatattttt
12.gtttaactttatcatttaatttaggttttaatttaaaatttaatttaggtaagatgaatttggtttttt
taaggtagttattatcatcgttaaggaaagttaa
13.gtattacaggcagaccttatttaggtattattatttttgatttttttttttttttaagttaaccg
aattatttctttaaagcgttacttaatgtcaatgc
14.gttagtcttttttagattaaatttagattatgcagttttttacataagaaaatttttttcggagttca
tattctaactgtctttataaatcttagagatatta
15.gtattatattttttattttatttttagaataaatttgaggtatgtgttaaaaaaattttttttt
tttttttttttttttaaaattataaatttaa
16.gttatttttaaaatttaatttttaaaatacaaaatttttactttctaaaattggtctctggatcgataa
tgtaaacttattgaatctatagaattacattatgat
17.gtatgtctatttcacggaagaatgcaccactatgatattgaaattatctatggctaaaaaccctcagtaaaa
tcaatcccataacccttaaaaaacggcggcctatccc
18.gttaattatttattccttacgggcaattaattattattacggtttatttacaatttttttttctgctca
tagagaaacttacttacaacacgttattttacatactt
19.gttacattatttattattatccggtatcgataattttttacctctttttcgtgagttttattcttacttt
tttctctttatattaggtctcatttaatatcttaa
20.gttttaactctcttacttttttttactctctacatttcatcttctaaaactgttgatttaaaactttt
gttctttaaggatttttttacttactctctgttat
21.tttagctcagtcagctagctagtttacaatttcgacaccagtttcgaccatctaaatttcgatccgtacc
gtaatttagcttagatttgatttaaaaggatttagattga
22.tttagtacgttagctcagccaagaacgatgtttaccgtaacgtqacgtaccgtaccgtaccgtaccgatt
ccggaagccgattaaggaccgatcgaaaggg
23.cggcgggatttaggccgacggggaccgggattcgggacccgaggaaattcccggattaaggtttagcttccc
gggatttagggcccgatggctgggaccc24.tttagctagctactttagctatttttagtagctagccagcctt
aaggctagcttagctagcattgttctttattgggaccaagtgcactttttagcttagtttgaccgt
25.gaccaaaggtggccttagggaccgatgcttttagctgcagctggaccagttcccagggtattaggcaaaag
ctgacgggcaattgcaatttaggcttaggcca
26.gatttacttttagcatttttagctgacgttagcaagcattagcttttagccaatttcgactttgccagtttcgca
gctcagtttaacgcgggatcttttagcttcaagcttttac
27.ggattcggatttaccggggattggcggaaacgggaccttaggtcgggaccattaggagtaaatgccaaagg
acgctggttagccagtcctgtaaggcttag
28.tccttagatttcagttactatatttgacttacagcttttagatttccttacgattttgacttaaaatttag
```

acgttagggcttatcagttatggattaatttagcttattttcga
29. ggccaattccggttaggaaggtgatggcccgggggttcccgggaggatttaggctgacgggcccggccatttcgg
tttagggaggccgggacgcgtagggc30. cgctaagcagctcaagctcagtcagtcacgtttgccaagtcagta
atttgccaaagttaaccgtagctgacgctgaacgctaaacagtattagctgatgactcgta
31. ttaaggacttaggctttagcagttactttagtttagttccaagctacgtttacgggaccagatgctagctagc
aatttattatccgtattaggcttaccgtaggttagcgt
32. gctaccgggcagctttaaactagctaccggtta
gtttggcccagccttgcggtgttcggattaattcgttgcagtcgctcrttgggttagtcattcccaaaagg
33. cagttagctgaatcgtttagccatttgacgtaaacatgattttacgtacgtaaatTTtagccctgacgttag
ctaggaatttatgctgacgtagcgatcgactttagcac
34. cggttagggcaaaggttgatttcgaccagggggaaagcccgggacccgaaccagggttagcgttaggt
gacgctaggcttaggttgaaccggaaa
35. gcggaagggcgttaggttgggatgcttagccgtaggctagctttcgcacagatcgattcgaccacaggataa
aagttaagggaccggttaagtcgcgtagcc
36. ctagctacgaacgcttttagcgcccccgggagtagtcggttaccgtagtatagcagtcgagtcgcaattcgc
aaaagtcccagctttagccccagagtcgacg
37. gggatgctgacgctggttagctttaggcttagcgtagctttagggccccagctcgcaggaaatgcccaaagga
ggcccaccgggtagatgccasagtcaccgt
38. aacttttagggcatttccagttttacgggttattttccagttaaactttgcaccattttacgtgttacgatt
tacgtataatttgacctattttggacactttagttgggttac
39. ttagggccaagtcccaggcaaggaattctgatccaagtccaatcacgtacagtcacaagtcaccgtttgcagc
taccgtttaccgtacgttgcaagtcaaatccat
40. ccattagggtttattacgtttattttttcccagaccttaggtttaccgtacttttaacggtttacctt
tgaaattttggactagcttaccctggatttaacggccagttt