摘　要

本文以生物序列的比较分析为背景，提出了一些新的图形表示，为生物序列的分类、分析、比较和储存等研究提供新的方法．另外，还展示了这些表示法在生物序列的相似性分析和构建进化树等问题上的具体应用．本文主要研究内容可以概括如下：

1. 将 DNA 序列和氨基酸序列转化为 2-D 图形表示．DNA 序列和氨基酸序列转化的二维图形类似于分子结构图，由此我们借助化学计量学方法计算了在经过转换所得图的基础上衍生出图的不变量 (数学不变量) — Balaban 指数和信息分布指数以及图对应的图论距离矩阵的平均频带宽度．并利用这些拓扑指数作为 DNA 序列和氨基酸序列的不变量分析了 9 个物种的 β- 球蛋白 (globin) 基因的第一个外显子 DNA 序列和 6 种 yar029w 等氨基酸序列的相似性和非相似性．

2. 用 1-D 随机游动来描述 DNA 序列，得到了 DNA 序列对应的两个随机序列 $\{Y_m\}$ 和 $\{X_n\}$，进而验证了两个随机序列 $\{Y_m\}$ 和 $\{X_n\}$ 都具有马尔可夫性，同时也得到了 DNA 序列的 1-D 游动曲线表示．基于 DNA 序列的图形表示以及马尔可夫链的转移概率分布、信息熵和随机序列的数字特征 (均协方差) 得到了 DNA 序列的一些新的数学不变量，进而利用这些数学不变量来比较了 9 个不同物种的 β- 球蛋白基因的第一个外显子的 DNA 序列的相似性．

3. 现有的方法一般是基于多个序列的比对来构建物种进化树，我们提出了一种新的方法：在 DNA 序列的三维图形表示的基础上，利用图的不变量给出了序列之间的距离度量，进而定义了物种进化距离，并利用基于距离法的 NJ 算法构建了生物系统进化树．选取 30 个物种线粒体 DNA 序列为材料，得到的结果与传统的根据物种形态和其他方法构建的系统进化树基本一致．

4. 在复平面上用二维随机游动来描述了 RNA 二级结构序列，得到了对应的随机游动曲线和随机复数字序列．在 6-D 空间中定义了使核苷酸集与点集之间一一对应的函数，进而利用这个函数在 6-D 空间中得到了 RNA 二级结构的 6-D 表示，然后基于 6-D 表示把它转化为矩阵表示和特征向量表示．并利用 RNA 二级结构对应的随机复数字序列的数字特征：模和相位，以及矩阵不变量：矩阵的最大特征值，特征向量来表征序列并且分析了 AIMV-3 等 9 种病毒的 RNA 二级结构序列的相似性．

5. 给出了把 RNA 二级结构序列映射为“波谱线”和“ Z 型曲线”表示的三个递归公式．利用这三个递归公式给出了 RNA 二级结构序列的 1-D 、 2-D 和 3-D 图形表示，进一步利用 1-D 图形表示给出了关于 RNA 二级结构序列频谱分析的方法．

6. 在 DNA 三联体密码子表示的基础上，在半复平面上给出了蛋白质序列的非退化的 2-D 图形表示，同时利用复向量的主要特征一模和相位，给出了蛋白质序列的一种数值刻划．进一步在 3-D 空间里，把 20 种氨基酸分别分配给正 12 面体的 20 个顶点，根据正 12 面体的对称性得到了 20 种氨基酸的 3-D 表示，进而得到了蛋白质序列的 3-D 图形表示和对应的数字序列，并利用图的不变量和数字序列的特征比较了 9 种动物的神经元

基因序列的相似性并构建了一组细胞色素 C 蛋白质的序列进化树.

关键词: DNA 序列; 蛋白质; RNA 二级结构; 特征数值; 图形表示; 距离矩阵; 最大特征值; 序列不变量; 进化树

Abstract

This dissertation mainly studied some new graphical representations of biological sequences based on biological background and structures of biological sequences, provided new method for classifing, analyzing, comparing and storing of biological sequences, etc. and discussed concrete applications of these representation methods to analysis of similarity constructions of evolutionary tree problems of biological sequences, etc. The main results, obtained in this dissertation, may be summarized as follows:

1. The DNA sequences and amino acid sequences have been translated into 2-D graphical representations. The 2-D graphical representations of DNA sequences and amino acid sequences are similar to the molecular structure graphs. Therefore we make use of chemistry metrology method to compute invariants of graphs—Balaban index, distribution index and the average bandwidths of corresponding distance matrix and consider them as a set of invariants for the DNA primary sequences and amino acid sequences. Similarity and dissimilarity analysis based on invariants of DNA primary sequences and amino acid sequences are given for the first exon genes of β-globin of nine species: human, goat, gallus, opossum, lemur, mouse, rabbit , rat, gorilla and six yar029w etc.

2. We describe the DNA primary sequence as a random walk. With the description, two random sequences $\{Y_m\}$ and $\{X_n\}$ correspond to a DNA sequence, and graphical representations of DNA sequences are given as well. We further prove that two random sequences $\{Y_m\}$ and $\{X_n\}$ have the quality of Markov chains. Based on the graphical representations of DNA, transition probability distributions, correlations and numerical characterizations of random sequences are given. We introduce some new invariants for the DNA primary sequences also. Then using these invariants, we compared primary sequences for exon-1 of $\beta-$globin genes that belong to nine species for analyzing the similarity and dissimilarity.

3. Construction of phylogenetic trees is key means in molecular evolutionary studies. We propose a new method for phylogenetic analysis, based on graphic representations of DNA sequences. Utilizing the invariants of graphs, we give the distance measure of DNA sequences and define the distance between species. We have chosen mitochondrial DNA sequences of 30 species and constructed their phylogenetic trees successfully. The method does not require sequence alignment and is totally automatic.

4. The sequences of RNA secondary structure on the complex plane are described as 2-D random walks. A random walk curve and a random complex numerical sequence are obtained. We define a function between the nucleotide sets and the point sets in the 6-D space. Therefore, we get the 6-dimensional representation of RNA secondary structure in the 6-D space by this function. Furthermore, we transform the representations into matrices and characteristic vectors.

We analyze the similarity of the RNA secondary structures of AIMV-3 and the other 8 kinds of viruses by using the numerical representation of random complex numerical sequence: module, phase, and the matrix invariant−the leading eigenvalues of the matrix and the distances between the characteristic vectors, which describe the sequences.

5. The RNA secondary structure sequences are translated into "Spectrum−like" and "Zigzag Curve" representations, from which we get three recursive formula, and obtain 1-D, 2-D and 3-D graphical representations of RNA secondary structure sequences by the three recursive formula. Furthermore using the 1-D graphical representation, we propose frequency−domain analysis method of RNA secondary structure sequences.

6. We give a new 2-D graphical representation of protein sequences based on nucleotide triplet codons in the half complex plane, which has no degeneracy. Meanwhile using main characterization of complex vector: module and phase, we give a kind of numerical description of protein sequences. Also in the 3-D space, we assign the 20 amino acids to 20 vertices of the dodecahedron. By the symmetry of the dodecahedron we obtain 3-D representation of 20 amino acids, and 3-D graphical representation and the corresponding numerical sequence of protein sequences. And similarity and dissimilarity analysis based on the invariants of graphs and characteristics of numerical sequences are given for nine RNA secondary structures of RNA-3 of virus. We construct sequence phylogenetic tree of a group of cytochromes C protein.

Keywords: DNA sequences; protein; RNA secondary structure; numerical characterization; graphical representation; distance matrix; leading eigenvalue; sequence invariant; phylogenetic tree

1 绪论

本章介绍了生物序列研究的背景、理论意义及应用价值、生物序列的图形表示研究概况．以生物序列的比较分析为背景，介绍了图形表示在生物信息学和计算分子生物学中的广泛应用．同时列出本文取得的主要结果．

1.1 生物序列研究的背景、理论意义及应用价值

随着人类基因组测序计划的完成，人们的研究重点由测序转向功能基因组的研究．同样，生物信息学也经历了由最初主要将基因组测序计划完成的序列数据通过数据库进行存储，到有效利用包括生物大分子的三维结构、代谢途径和基因表达等各类数据的发展过程．现在和将来，科学家们将着重于研究 DNA 序列信息，蛋白质结构信息，以及它们之间的相互作用．破译每一水平的生物信息提出了与基因或蛋白质有关的统计和组合数学问题．生物信息的急剧增长也带来了对计算机科学的挑战．为此，计算分子生物学和生物信息学便应运而生．

生物信息学大量地在生物学中引入了数学模型，它标志着生物学已经从实验科学向理论学科转变．对于生物学本身而言，这就是一次从量变到质变的飞跃．在生物信息学形成以前，一切的生物学理论的发展都是通过大量的实验证据所得到的经典理论，然而生物信息学的加入之后，生物学理论的研究用于指导、验证试验生物学．这将会使得试验生物学的目的更加明确，并且也将会大大缩短试验周期．

生物信息学的产生将生物学、信息学、数学、计算机科学、物理学等多门学科有机的整合为一个新兴学科，这一学科领域的建设必然会推动上述诸多学科的进一步发展．与此同时，在生物信息学建设的过程中，又以此为基础萌生出一系列分支科学，如 DNA 计算等．所有的这一切，其最直接的意义便是给各个领域带来了无限的商机，孕育了一个美好的市场．另一方面，伴随生物信息学的发展，人类必将揭示更多的生命活动本质规律，其中当然会有很多是与人类自身健康、疾病、衰老等相关的生物信息，而它们的发展必然导致新药物的设计与研发周期大幅度变短以及基因治疗的最终实现，从而彻底地改变人类自身的命运，这无疑是人类文明的又一次飞跃．当然，在这一过程中产生的巨大经济效益是现在无法估量的．

生物信息学主要研究生物信息的采集、处理、存储、传播、分析和解释等方面内容的一门学科．它利用生物学、计算机科学和信息技术综合分析大量而复杂的生物数据，揭示其所蕴涵的生物学意义．具体地说，生物信息学是把基因组 DNA 序列信息分析作为源头，在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测，最后依据特定蛋白质的功能进行必要的药物设计．

计算分子生物学不仅是生物信息学的前身，更是生物信息学的核心部分．可以说，对生物信息学的研究中数学技术发挥着重要作用．随着生物信息学算法的不断完善，已能进行生物序列家族或同源性分析；进行生物序列的聚类，建立进化树并确定生物序列间的进化关系；进行代谢途径相关基因的同源性分析，以及获取其它生物代谢途径的相关信息等．其中生物序列的比较是生物信息学中最基本的问题，因为对于 DNA 序列，即使我们考虑他的一个很短的片断，我们也不可能直接得出它表示的对象所具有的全部信息，然而如果我们比较不同的生物序列就有可能得到某些重要信息．然而这个问题非常复杂，至今还有许多未解决的问题．总之，对生物序列进行分析和比较是生物信息学的最基本也是最重要的课题之一，同时对生命科学的研究具有深远的意义 [130-139] ．

1.2 生物序列的图形表示研究概况

生物序列一般是指 DNA 、 RNA 序列或蛋白质序列．而 DNA 、 RNA 和蛋白质序列都是由较小的单元组成的无分枝的线性聚合体大分子．对于 DNA ，这些单元是 A(腺嘌呤) 、 C(胞嘧啶) 、 G(鸟嘌呤) 和 T(胸腺嘧啶) 这 4 种核苷酸残基；对于 RNA ，这些单元是 A, C, G 和 U(尿嘧啶) 这 4 种核苷酸残基；对于蛋白质这些单元是 20 种氨基酸，即 A(丙氨酸) 、 C(半胱氨酸) 、 D(天冬氨酸) 、 E(谷氨酸) 、 F(苯丙氨酸) 、 G(甘氨酸) 、 H(组氨酸) 、 I(异亮氨酸) 、 K(赖氨酸) 、 L(亮氨酸) 、 M(甲硫氨酸) 、 N(天冬酰氨酸) 、 P(脯氨酸) 、 Q(谷氨酰胺酸) 、 R(精氨酸) 、 S(丝氨酸) 、 T(苏氨酸) 、 V(缬氨酸) 、 W(色氨酸) 和 Y(酪氨酸) ．这样，一个 DNA(RNA) 序列可以看作是在一个有 4 个字母的字母表 $\mathcal{N}$={A, C, G, T(U)} 上的字 (word) ，同样，蛋白质序列也可以看作是一个在 20 个字母表 $\mathcal{M}$={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} 上的字 (word) ．而 RNA(DNA) 二级结构是由自由基 (free base) 和基对 A−U(A−T) 和 C−G 组成的，在一定程度上， RNA(DNA) 二级结构经过处理后都可以转化为线性序列．

生物信息学的理论分析已成为生物信息学又一种主要的研究手段，是生物学家获取信息的另一途径和生物信息学显示其价值的窗口，也是大的基因组研究中必不可少的．如前所述，生物序列的传统表示是由字母来表示的，这种表示具有自身的优点，但是随着计算机技术的发展和可视化要求的提高，它固有的缺点也随之暴露出来．在生物序列的研究分析中，对生物序列的有效表示，发挥很大的作用．自从 1983 年由 E. Hamori 和 J. Ruskin[1] 提出了 DNA 序列图形表示的思想 — 将 DNA 序列表示为一条平面或空间中的曲线，把 DNA 序列的研究带进了一个新的研究领域．自此国内外不少化学专家如 M. Randic, A. Nandy 以及国内郭晓峰、廖波和王天明等人提出了生物序列的众多的不同维的图形表示 [7-16] ， [17−23], [25−27], [28−31], [32, 43, 44, 66, 77], [94−121], [123−125] ． M. Randic 等人还基于他们的图形表示，将 DNA 序列转化为矩阵等数学表示，进一步用矩阵不变量来研究 DNA 序列，取得了很好的结果 [7−31, 120] ．生物序列的图形表示主要应用在序列相似性分析和基因识别等方面 [28−31, 33, 100−105] ．我国著名理论物理专家张春霆院士也提出了一种 DNA 序列的几何图形表示 —Z 曲线， Z 曲线是表示 DNA 序列的一个等价的三维空间曲线 [49] ．通过对 Z 曲线的研究来对基因组序列进行研究是一种几何学的途径．这种新颖的学术观点为引进更多的数学工具来分析生物序列提供了广阔

的前景. 天津大学生物信息中心用这种思路研究了真核和原核基因组中若干重要问题, 这样的思路是切实可行的. 原则上说, 基因组中的许多问题都可以通过这种途径加以解决, 这种独树一帜别开生面的研究思路已经得到国内外学术界的普遍好评和认可, 越来越多的学者, 加入到对 Z 曲线研究的行列中来 [46–56] . 可以预见, 用几何学方法研究基因组将会有一个广阔的发展空间.

DNA 是携带生物遗传信息的主要大分子, 但 RNA 是大部分病毒的遗传物质, 并且 RNA 还参与蛋白质的合成, 与细胞分化, 代谢, 记忆的储存等有重要关系. 正是由于 RNA 具有的这些特殊属性, 目前越来越多的人开始关注 RNA , 最近, 廖波和王天明鉴于现有比较 RNA 二级结构相似性的算法受不带假结的限制, 首次提出用几何图形表示 RNA 二级结构 [101–105] . 根据 RNA 二级结构组成和核苷酸 A, C, G, U 的化学结构分类, 他们给出了 RNA 二级结构的一种 3-D 图形表示和 6-D 图形表示法, 并利用这些表示的数据特征来比较 RNA 二级结构的相似性.

以上这些表示都还有各自的缺陷, 主要表现在以下几点: (1) 有退化现象; (2) 对完整序列而言, 使用的数学不变量计算太复杂, 有的甚至还没有算法解决; (3) 缺乏表征生物序列特征的更多的灵敏度足够好的数学不变量. 另有关生物序列 (DNA 序列、 RNA 序列和蛋白质序列) 的图形表示的应用研究还很少.

1.3 本文的主要工作

计算分子生物学的研究对象是与基因和蛋白序列有关的组合和计算问题. 计算分子生物学的主要课题有: 序列组合, 序列分析, 生物信息资料库, 基因认定, 种族树的构建以及结构预测等. 从计算理论的角度来讲, 它们都是难处理的; 换句话讲, 我们并不知道是否存在有效的算法去解决这些问题. 目前的研究集中在设计好的近似算法或概率算法; 这些算法虽然并不能对有关问题的每一个实例都能求出好的解, 但对大多数实例却行之有效. 本文就针对某些算法的缺陷性, 我们考虑用其他方法来试图解决问题. 比如我们更进一步的用新的几何图形表示的方法 (理论分析方法) 来比较生物序列的相似性等.

本文主要给出了生物序列的一些新的图形表示, 并利用生物序列的图形表示寻求新的特征数值, 利用这些特征数值来比较和分析了生物序列 (主要是相似性) , 进而将相似性转化为距离记分构建物种进化树、构造蛋白质序列进化树等方面做了一些研究和探讨. 本文的主要内容如下:

在第二章, 将 DNA 序列和氨基酸序列转化为 2-D 图形表示, DNA 序列和氨基酸序列转化的二维图形类似于分子结构图, 由此我们借助化学计量学方法计算了在经转换所得图的基础上衍生出图的不变量 (数学不变量)—Balaban 指数 [24] 和信息分布指数以及图对应的图论距离矩阵的平均频带宽度. 并利用这些拓扑指数作为 DNA 序列的不变量分析了 human, goat, gallus, opossum, lemur, mouse, rabbit, rat, gorilla 等 9 个物种的 β- 球蛋白 (globin) 基因的第一个外显子 DNA 序列和 6 种 yar029w 等氨基酸序列的相似性和非相似性.

在第三章里, 用 1-D 随机游动来描述 DNA 序列, 得到了 DNA 序列对应的两个随机序列 $\{Y_m\}$ 和 $\{X_n\}$, 进而验证了两个随机序列 $\{Y_m\}$ 和 $\{X_n\}$ 都是马尔可夫链, 同时也

得到了 DNA 序列的 1-D 游动曲线表示。基于 DNA 序列的图形表示以及马尔可夫链的转移概率分布、信息熵和随机序列的数字特征得到了 DNA 序列的一些新的数学不变量，进而利用这些数学不变量来比较了 9 个不同物种的 β- 球蛋白基因的第一个外显子的 DNA 序列的相似性.

现有的方法一般是基于多个序列比对的最大节约法 (maximum parsimony, MP) 、最大似然法 (maximum likelihood, ML) 和距离法。一种称为贝叶斯推断的统计学方法也开始使用. 在第四章，我们基于距离法的 NJ 算法提出了一种新的方法：在张春霆、廖波和王天明等提出的 DNA 序列 3-D 图形表示的基础上 [49,30] ，利用图的不变量给出了序列之间的距离度量，进而定义了物种进化距离，把它应用到基于 DNA 序列分析的生物系统进化树构建的研究中. 选取人类等 30 个物种线粒体 DNA 序列为材料，得到的结果与传统的根据物种形态和其他方法构建的系统进化树基本一致.

在第五章，在廖波提出的 RNA 二级结构特征序列 [102] 和张春霆的 Z 曲线表示 [49] 的基础上，根据 RNA 二级结构中自由基和基对的化学结构分类，在复平面上用二维随机游动来描述了 RNA 二级结构序列，得到了对应的随机游动曲线和随机复数字序列. 在 6-D 空间中定义了使核苷酸集与点集之间一一对应的函数，进而利用这个函数在 6-D 空间中得到了 RNA 二级结构的 6-D 表示，然后基于 6-D 表示把它转化为矩阵 Q 表示和特征向量表示. 并利用 RNA 二级结构对应的随机复数字序列的数字特征：模和相位，以及 Q 矩阵不变量：矩阵的最大特征值，特征向量作为序列不变量来分析了 AIMV-3 等 9 种病毒的 RNA 二级结构序列的相似性.

在第六章，结合 Zupan 和 Randic 提出的把 DNA 序列映射为“波谱线”和“ Z 型”的 1-D 、 2-D 和 3-D 图形表示的运算法则 [94] ，给出了把 RNA 二级结构的特征序列映射为“波谱线”和“ Z 型曲线”表示的三个递归公式：

$$R(x_{i+1}) = \frac{R(x_i) + S(x_{s_{i+1}})}{d};$$

$$R(x_{i+1}, y_{i+1}) = \frac{R(x_i, y_i) + S(x_{s_{i+1}}, y_{s_{i+1}})}{d};$$

$$R(x_{i+1}, y_{i+1}, z_{i+1}) = \frac{R(x_i, y_i, z_i) + S(x_{s_{i+1}}, y_{s_{i+1}}, z_{s_{i+1}})}{d},$$

其中， d 为任意非零实数. 利用这三个递归公式同样给出了 RNA 二级结构序列的 1-D 、 2-D 和 3-D 图形表示，进一步利用 1-D 图形表示给出了关于 RNA 二级结构序列频谱分析的方法.

在第七章，在 DNA 三联体密码子表示的基础上，在半复平面上给出了蛋白质序列的非退化的 2-D 图形表示，同时利用复向量的主要特征一模和相位，给出了蛋白质序列的一种数值刻划. 还有在 3-D 空间里，把 20 种氨基酸分别分配给正 12 面体的 20 个顶点，根据正 12 面体的对称性得到了 20 种氨基酸的 3-D 表示，进而得到了蛋白质序列的 3-D 图形表示和对应的数字序列，并利用图的不变量和数字序列的特征：自相关系数和自协方差系数来比较了 9 种动物的神经元基因序列的相似性以及构建了一组细胞色素 C 蛋白质的序列进化树.

本文的主要内容是作者近期获得的一些结果，我们希望由此能在将来的工作中探讨数学与生物两大学科的汇合点．由这些内容的讨论看到，许多数学理论与工具可在分析生物序列领域内应用，这些问题数据丰富，背景明确，将成为研究生物信息学的有力工具．但由于生物序列结构的复杂性，许多问题远远没有解决，因此继续深入研究的发展空间巨大．

2 拓扑指数的应用：生物序列的比较方法

本章在生物序列的二维图形表示的基础上，利用 Balaban 指数和信息分布指数以及矩阵不变量—距离矩阵主对角线以外的次对角线之和的平均值，比较了生物序列的相似性。我们以包括人类等 9 个物种的 β- 球蛋白 (globin) 基因的第一个外显子 DNA 序列和 yar029w 等 6 种蛋白质序列为例来说明该方法的应用。

2.1 引言

在计算分子生物学中序列比较是最重要和最常用的原始操作，是许多其他更复杂操作的基础．粗略地讲，这一操作包括发现序列的类同与序列的不同两方面．最常见的比较是蛋白质序列之间和核酸序列之间的两两比较．生物序列的相似性分析是通过生物序列的比较来实现的，但又不同于符号序列的序列比对，其理论基础是进化学说，如果两个序列之间具有足够的相似性，就推测二者可能有共同的进化祖先，经过序列内残基的替换、残基或序列片段的缺失、以及序列重组等遗传变异过程演化而来．注意，序列相似和序列同源是不同的概念，序列之间的相似程度是可以量化的参数，而序列是否同源需要有进化事实的验证．序列之间的相似程度是数量上的多或少的判断，而序列的同源性判断是质的判断，序列之间要么同源要么不同源．

在生物信息学中，序列的比较是通过将两个或多个核酸序列或蛋白质序列进行比对．通过比对未知序列与已知序列 (尤其是功能和结构已知的序列) 之间的相似性得到它们的同源性来预测未知序列的功能．序列比较的常用方法有：动态规划算法，压缩矩阵方法，图形表示的数值刻划方法．所有这些方法只考虑了序列的组成（由四种核苷酸组成的字符串）以及每个基的位置．然而， DNA 序列的表示，储存，比较都应当体现每个基的自身的化学性质和化学结构，传统的动态规划算法就存在这方面的缺陷。

序列比对的基本问题是比较两个或两个以上符号序列的相似性或不相似性．序列比对是生物信息学的基础，非常重要．两个序列的比对有较成熟的动态规划算法．有时两个序列总体并不很相似，但某些局部片断相似性很高． Smith−Waterman 算法 [37] 是解决局部比对的好算法，缺点是速度较慢． Aligment 算法自 Smith−Waterman 提出以来，已经有数十种 Aligment[137] 算法被提出和应用．然而应用动态规划算法的最大困难之一是罚分参数的选择．在某些情况下，核酸或氨基酸的权重或插入删除函数的微小变化在所到的比对中产生很大的改变．在另一些情况下，比对对于算法参数的改变是非常稳健的．没有一组“正确”参数：对一对序列来说，能够找到对一种统计特性的有意义匹配参数，对另一种类型匹配没有用．所以，对一大组参数值考虑序列的比较是有意义的．理

想情况是要对所有可能数值计算最优比对．这样一来需要大量的计算。

压缩矩阵最早是由 Randic 等人提出来的 [14]．它来源于计算化学中化学指标计算．他的基本思想是先构造一个适当的矩阵来表示一个序列，这样序列之间的比较就转化为矩阵之间的比较，而且如果矩阵是数值矩阵就可以选择一个适当的不变量进而把矩阵之间的比较转化为比较这些不变量．这使得把复杂的问题简单化了．

利用压缩矩阵方法来比较生物序列，不同于比对方法去直接比较生物序列，而是去考虑这些生物序列的不变量．这些不变量是从生物序列对应的矩阵中提取出来，即把初始的生物序列转化为数值序列，而这些数值序列的长度可以依靠被选择的不变量的性质并按照自己不同的需要进行修改．另外一个优点在于不变量的刻划非常简单，两个生物序列的比较被转换成了生物序列对应的数学对象的比较．然而他所付出的代价是在用不变量来刻划和比较生物序列时同时会伴随着某些结构方面的信息的丢失．所以如何能找到一些更适当的参数来刻划生物序列的特征，进而比较和分析生物序列是值得进一步研究的课题．

由于序列比对的动态规划方法和压缩矩阵方法有如上所述的一些缺陷，使得很多人试图寻找其他的方法来比较生物序列．最近 Randic[7−9, 12, 14, 16]，Nandy[18−23]，等人提出了一种新的方法来进行序列的比较，这就是所谓的压缩矩阵的不变量方法．也有不少学者给出了一些图形表示方法．如，张春霆院士 [49] 给出了一种 3-D 图形表示— Z 曲线，廖波和王天明提出了和他们不同的几何图形表示法 [28-31]．虽然这些表示不一样，但他们有个共同的思想就是：将生物序列转化为图形 (曲线)，利用图形构造矩阵，再利用矩阵不变量 (如最大特征值，次对角线上所有元素和的平均值，最大 (小) 行和，矩阵的迹等) 来比较生物序列的相似性．贺平安和王军 [39] 给出了 DNA 的一种 0, 1 特征序列表示法．这些方法考虑了每个基的自身的化学性质和化学结构，且这些方法的直观性和实用性，受到了计算分子生物学家的高度重视．

拓扑指数在分子相似性比较中一直是一个非常活跃的领域．用 A.Nandy[17] 的方法把生物序列转化的二维图形类似于分子结构图，由此可借助化学计量学方法在转换所得图的基础上衍生出诸多图的不变量 (即参数)，因此，本章中我们把图的不变量—拓扑指数作为生物序列的特征数值应用于 DNA 序列和蛋白质序列的相似性比较．以下我们都以 9 个不同物种的 β- 球蛋白基因的第一个外显子和 yar029w 等 6 种蛋白质作为研究对象，见表 2.1 、表 2.2 .

相似性的计算通常有如下三种方法： (1) 计算向量终点之间的欧氏距离，如果两个向量 (它们表示序列) 的终点距离较小就认为这两个比较相似． (2) 计算两向量之间的夹角，如果两个向量所成角比较小就认为这两个序列比较相似． (3) 计算两向量夹角的余弦值，如果两个向量所成角的余弦比较大就认为这两个序列比较相似．

表 2.1: 9 个不同物种的 β- 球蛋白基因的第一个外显子碱基序列

Species	Coding sequence
human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG
	GGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAG
	GTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGG
	TCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGG
	GGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemmur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGG
	GGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTG
	GGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGG
	GGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGG
	GGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG
	GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG

表 2.2: 6 种蛋白质的氨基酸序列

Code name	Amino acid sequence
yar029w 74	MNKYLFDHKIWSTPYYFYCEEDCHRLSFIEGRTFEKPTSNAEENVQETE
	AGESFTLNPGE
yarl02w a65	MYLLRGKVWKNKKYFFDGADYQAFFTGSFRNFLYSRTFMLCGLLNVGL
	WKYIQEAQLSSN
ybll08c-a42	MLTGIAPDQVTRMITGVPWYSSRLKPAISSALSKDGIYTIAN
ycl057c-a97	MSEQAQTQQPAKSTPSKDSNKNGSSVSTILDTKWDIVLSNMLVKTAMGF
	GVGVFTSVLFFKRRAFPVWLGIGFGVGRGYAEGDAIFRSSAGLRSSKV
yar020c 55	MVKLTSIAAGVAAIAAGASAAATTTLSQSDERVNLVELGVYVSDIRAHLA
	EYYSF
yar070c 99	MYKITTIYLWQKSYLSFFIGIDNLDCFTLIRFFQCRLQNKLGLQDILDFFC
	NLCGHSMVRTCNMVEAAQKQNRITFGSIYVKLHPLVKLCTGIVWAPRV

2.2 拓扑指数在生物序列相似性比较中的应用

2.2.1 拓扑指数

图形表示是由物种的 DNA 序列决定的，且能给出 DNA 序列的一个直观的刻画，但不是量化的数值. 因为 DNA 序列与图之间存在着对应关系，图的不变量因为能够很好的表征图的结构而得到广泛的应用 [25, 26, 29, 31]，因此图的不变量可以作为 DNA 序列的特征数值，即图的不变量能够表征 DNA 序列的特征. 得到图的不变量的最为简单的方法为拓扑指数法，到目前为止已有上百种拓扑指数，其中 Balaban 指数 [24]、信息分布指数和分子连接性指数 [41] 等因可以较好的表征图的结构而被广泛应用.

2.2.1.1 DNA 序列的图形表示

建立平面直角坐标系 oxy，x,y 轴的正负四个方向分别表示四种碱基 A, T, G, C. 从左到右观察 DNA 序列的碱基，画出 DNA 序列图 [17]. 即我们分别置 A, T, G 和 C 于 $-$x 轴，+x 轴，$-$y 轴和 +y 轴，并假设 A, G, T 和 C 分别为沿坐标轴的 4 个方向移动的矢量，其中 A, T 分别沿 x 轴的负和正的方向移动，而 G, C 分别沿 y 轴的负和正的方向移动，并且每个碱基仅移动一个单位，即 A$(-1,0)$; T$(1,0)$; G$(0,-1)$; C$(0,1)$，如取 human 的 β- 球蛋白基因第一个外显子碱基序列的前 20 个碱基 ATGGTGCACCTGACTCCTGA，则所得的坐标为 $(-1,0)$, $(0,0)$, $(0,-1)$, ..., $(2,2)$, $(2,1)$, $(1,1)$，与之对应的图形，示于图 2.1(a).

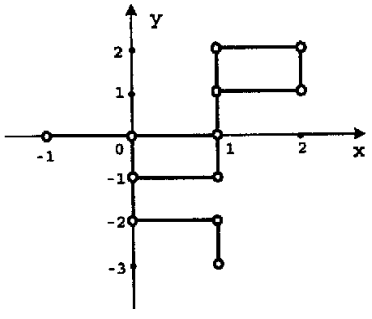
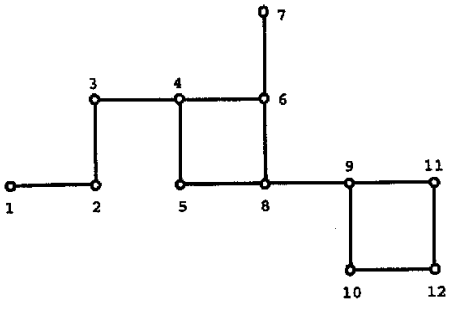


图 2.1 (a): human β- 球蛋白基因的前 20 个碱基序列对应的平面图



(b): human β- 球蛋白基因的前 20 个碱基序列对应的连通标号图

在图 2.1(a) 中我们分别置 A, T, G 和 C 于 $-$x 轴，+x 轴，$-$y 轴和 +y 轴，我们得到了一条特征曲线，同样如果我们分别置 A, G, T 和 C 于 $-$x 轴，+x 轴，$-$y 轴和 +y 轴，我们也可以得到一条不同的特征曲线，这就意味着这种特征曲线并不唯一，事实上它跟 A, G, T, C 的排列顺序有关. 因此对每一条 DNA 序列并不只有唯一矩阵表示. 而 A, G, T 和 C 的排列有 24 种方法，但并不意味着有 24 个图形表示. 我们分别置 A, G, T, C 于 $(-1,0)$, $(0,-1)$, $(1,0)$, $(0,1)$，顺时针旋转 90 度将顺序 AGTC 变成顺序 GTCA，但特征曲线并不改变，因为他们之间的距离并不改变，因此 AGTC, GTCA, CAGT, TCAG 对应相同的图形表示；再有我们交换 A 和 T 或 G 和 C，特征曲线也不改变，也就是说形式 ACTG 与形式 AGTC 或形式 TGAC 和形式 AGTC 具有同样的图形表示，因此我们

可得 AGTC, CAGT, TCAG, GTCA, ACTG, GACT, TGAC, CTGA 具有同样的图形表示，故每条 DNA 序列最多有 3 个图形表示. 我们改变 AGTC 的顺序为 ATCG 和 ATGC，显然 A 和 T 之间的距离改变，且将顺序 AGCT 变为 ATCG 会改变 A 和 G 之间的距离，因此形式 AGTC, ATCG 和 ATGC 具有不同的矩阵表示，因此每条 DNA 至少有 3 个矩阵表示，综上所述，我们可以断定每条 DNA 序列有且只有 3 个图形表示. 在文献 [57] 已证明：一条 DNA 序列的 3 个不同图形表示恰好对应于 A, G, T 和 C 的三种分类方法 ((1) 嘌呤 R={A,G}，嘧啶 Y={C,T}； (2) 酮基 M={A, C}，氨基 K={G, T}； (3) 弱氢键 W={A, T}，强氢键 S={C, G}) 来把 A, G, T 和 C 分配于坐标轴上. 因此对应于这三种分类方法我们可以得到三种不同的图形表示. 本文我们分别取 (1) A$(-1,0)$; G$(1,0)$; C$(0,1)$; T$(0,-1)$, (2) A$(-1,0)$; C$(0,1)$; G$(0,-1)$, T$(1,0)$, (3) A$(-1,0)$; T$(0,-1)$; C$(1,0)$; G$(0,1)$. 这样共得到了 9 个不同物种对应的 27 个二维图形 (略).

把 DAN 序列转化的图形看成一个边长为 1 的连通标号图，如图 2.1(b) 是 human 的 β- 球蛋白基因第一个外显子碱基序列的前 20 个碱基对应的连通标号图.

2.2.1.2　蛋白质序列的图形表示

蛋白质与 DNA, RNA 一样，也是线性大分子. 它是字符集 $\mathcal{M}$={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} 上的字符串. 蛋白质序列的比较不仅仅是字符串本身的比较，还要考虑他们的化学结构和化学性质. 我们根据 20 种氨基酸的化学性质以及蛋白质二级结构的相关性质把它们分成四类：亲水，极性，带电荷 X=HPC={D, N, S, H, T, C}；疏水，非极性 Z=HA={Y, F, V, I, W, M, L}；非极性且体积小 B=AS={G, P}；其他 J=O={R, K, E, A, Q}[34].

和 DNA 序列的图形表示一样建立平面直角坐标系 oxy，x,y 轴的正负四个方向分别表示 X, Z, B, J. 从左到右扫描蛋白质序列，根据每次观察到的氨基酸画出蛋白质序列图. 即我们分别置 X, B, Z 和 J 于 $-$x 轴，+x 轴，$-$y 轴和 +y 轴，并假设 X, B, Z 和 J 分别为沿坐标轴的 4 个方向移动的矢量，其中 X, B 分别沿 x 轴的负和正的方向移动，而 Z, J 分别沿 y 轴的负和正的方向移动，并且每个氨基酸仅移动一个单位，即 X$(-1,0)$; B$(1,0)$; Z$(0,-1)$; J$(0,1)$，如取蛋白质 yar029w 74 的氨基酸序列的前 15 个字符 MNKYLFYDHKIWSTP，则所得的坐标为 $(-1,0)$, $(-1,-1)$, $(-1,0)$, ..., $(-5,-4)$, $(-4,-4)$, $(-4,-5)$，与之对应的图形，示于图 2.2(a).

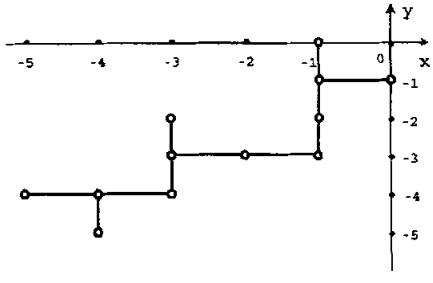
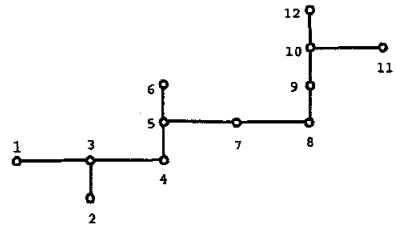


图 2.2 (a): 蛋白质 yar029w 74 氨基酸序列的前 15 个字符序列对应的平面图



(b): 蛋白质 yar029w 74 氨基酸序列的前 15 个字符序列对应的连通标号图

根据以上蛋白质序列中氨基酸的分类我们同样可以得到三种不同的图形表示。我们分别取 (1) X(−1,0); Z(1,0); J(0,1); B(0,−1), (2) X(−1,0); J(0,1); Z(0,−1); B(1,0), (3) X(−1,0); B(0,−1); J(1,0); Z(0,1)。这样共得到 6 种蛋白质对应的 18 个二维图形 (略)。把蛋白质序列转化的图形看成一个边长为 1 的连通标号图，如图 2.2(b) 是蛋白质 yar029w 74 氨基酸序列的前 15 个字符对应的连通标号图。

2.2.1.3 Balaban 指数和信息分布指数

为了阐明这种图形表示的数值特征，我们将这种图形表示转化为另一数学目标—矩阵。我们构造距离矩阵 $D=(d_{ij})$，其中元素 d_{ij} 为图中第 i 个顶点到第 j 个顶点的最短路所经过的边数，而每两个相邻顶点之间的边长取为单位长度，如图 2.1(b) 对应的矩阵示于表 2.3。

表 2.3: 图 2.1(b) 对应的 12 × 12 图论距离矩阵

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	2	3	4	4	5	5	6	7	7	8
2		0	1	2	3	3	4	4	5	6	6	7
3			0	1	2	2	3	3	4	5	5	6
4				0	1	1	2	2	3	4	4	5
5					0	2	3	1	2	3	3	4
6						0	1	1	2	3	3	4
7							0	2	3	4	4	5
8								0	1	2	2	3
9									0	1	1	2
10										0	2	1
11											0	1
12												0

利用构造的距离矩阵我们可以得到对应的 Balaban 指数和信息分布指数。

Balaban 指数 J

$$\mathrm{J}=\frac{q}{\mu+1}\sum_{i=1}^{n-1}(s_i\times s_{i+1})^{-\frac{1}{2}} \tag{2.1}$$

其中，μ 为图中的环数，$\mu=q-n+1$，n 为顶点数，q 为边的个数，s_i 为图对应的距离矩阵的第 i 行的行和。

根据等式 (2.1) 我们就得到了表 2.1 中 9 个不同物种的 DNA 序列和表 2.2 中 6 种蛋白质序列的 Balaban 指数见表 2.4、表 2.5。

信息分布指数 I

$$\mathrm{I}=\sum_{i=1}^{n}\frac{N_i\times d_i}{W}\log_2(\frac{N_i\times d_i}{W}) \tag{2.2}$$

其中，d_i 表示不同的距离，N_i 是与 d_i 相应的计数，W 是 Wiener 指数，$W=\sum_{i=1}^{n} N_i \times d_i$.

根据等式 (2.2) 我们计算了表 2.1 中 9 个不同物种的 DNA 序列和表 2.2 中 6 种蛋白质序列的信息分布指数见表 2.6 、表 2.7 .

表 2.4: 表 2.1 中 9 个物种的 DNA 序列的 Balaban 指数

type	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Corilla
AGCT	0.8168	2.2593	0.9119	0.7650	2.0923	0.7711	1.5480	0.9793	0.8110
ACGT	0.5353	0.5951	0.6556	0.7326	0.7420	0.7896	0.6209	0.5407	0.5236
ATCG	1.5900	1.8383	0.8960	0.9614	0.9893	1.1597	1.4004	0.8297	2.18095

表 2.5: 表 2.2 中 6 种蛋白质序列的 Balaban 指数

type	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
XZJB	3.5322	1.3897	1.1281	0.6487	3.1937	0.8756
JZBX	0.8495	1.3452	2.6185	1.7862	1.6797	3.5867
XBJZ	1.4629	1.2214	1.1071	3.2963	1.9111	1.6854

表 2.6: 表 2.1 中 9 个物种的 DNA 序列的信息分布指数

type	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
AGCT	5.0165	4.8318	3.5819	4.0051	5.2604	4.0848	5.0628	4.5639	5.1088
ACGT	4.8165	5.3382	4.0339	5.0865	5.0802	4.7252	5.0613	4.2418	4.9533
ATCG	4.9216	4.8818	3.8446	4.4850	4.9556	4.6583	4.9488	4.5054	4.9451

表 2.7: 表 2.2 中 6 种蛋白质序列的信息分布指数

type	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
XZJB	4.7759	4.7687	2.8936	5.0100	4.9578	4.7877
JZBX	4.8269	4.2810	3.2890	5.8143	4.9687	5.6809
XBJZ	4.3853	5.1816	3.4761	5.4974	4.6450	6.0408

2.2.2　相似性比较

为了比较生物序列的相似性，我们构造两种三维向量，Balaban 指数向量和信息分布指数向量，它们的坐标分别为每种生物序列的三种图形表示的 Balaban 指数和信息分布指数，通过计算各向量终点之间的欧氏距离和向量之间的夹角来比较相似性．我们约定：距离和夹角越小，相似性也就越大；反之，相似性就越小．由表 2.4 、表 2.5 、表 2.6 和表 2.7 的数据作为分量构造三维向量，显然这三维向量与 DNA 序列、蛋白质序列之间又建立了对应关系．进而得到了 9 个不同物种的 DNA 序列和 6 种蛋白质序列的相似性比较的数据，见表 2.8 、表 2.9 、表 2.10 、表 2.11 、表 2.12 、表 2.13 、表 2.14 、表 2.15 .

表 2.8：表 2.1 中 9 个物种的 DNA 序列对应的三维向量的两两向量之间的夹角弧度矩阵 (基于 Balaban 指数)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.4103	0.3477	0.3054	0.6245	0.2472	0.3456	0.3851	0.1341
Goat		0	0.2862	0.3910	0.2574	0.4209	0.1012	0.1981	0.5200
Opossum			0	0.1244	0.3494	0.1812	0.1886	0.1005	0.4816
Gallus				0	0.4728	0.0769	0.2898	0.2230	0.4303
Lemur					0	0.5354	0.2813	0.2603	0.7496
Mouse						0	0.3214	0.2784	0.3640
Rabbit							0	0.1138	0.4684
Rat								0	0.5175
Gorilla									0

表 2.9：表 2.1 中 9 个物种的 DNA 序列对应的三维向量在欧式距离上的相似性矩阵 (基于 Balaban 指数)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	1.4649	0.7107	0.6609	1.4249	0.5019	0.7602	0.7775	0.5911
Goat		0	1.6453	1.7380	0.8777	1.6471	0.8357	1.6305	1.4900
Opossum			0	0.1783	1.1872	0.3276	0.8126	0.1488	1.2957
Gallus				0	1.3276	0.2064	0.9046	0.3164	1.2382
Lemur					0	1.3330	0.6928	1.1423	1.7634
Mouse						0	0.8306	0.4628	1.0561
Rabbit							0	0.8097	1.0779
Rat								0	1.3618
Gorilla									0

表 2.10：表 2.1 中 9 个物种的 DNA 序列对应的三维向量的两两向量之间的夹角弧度矩阵 (基于信息分布指数)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.0610	0.0650	0.1141	0.0205	0.0792	0.0199	0.0167	0.001
Goat		0	0.0246	0.0575	0.0556	0.0487	0.0421	0.0769	0.0550
Opossum			0	0.0502	0.0673	0.0242	0.0507	0.0782	0.0623
Gallus				0	0.1125	0.0511	0.0974	0.1280	0.1100
Lemur					0	0.0862	0.0171	0.0343	0.0119
Mouse						0	0.0691	0.0892	0.0789
Rabbit							0	0.0365	0.0128
Rat									0.0255
Gorilla									0

表 2.11: 表 2.1 中 DNA 序列对应的三维向量在欧式距离上的相似性矩阵 (基于信息分布指数)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.5549	1.9572	1.1342	0.3608	0.9725	0.2506	0.8416	0.1667
Goat		0	2.0831	0.9509	0.5057	0.9918	0.3666	1.1898	0.4784
Opossum			0	1.3028	2.2686	1.1802	2.1135	1.2018	2.0947
Gallus				0	1.3407	0.4086	1.1549	1.0130	1.2032
Lemur					0	1.2635	0.1989	1.1793	0.1980
Mouse						0	1.0739	0.6976	1.0876
Rabbit							0	1.0569	0.1176
Rat								0	0.9982
Gorilla									0

表 2.12: 表 2.2 中 6 种蛋白质序列对应的三维向量的两两向量之间的夹角弧度矩阵 (基于 Balaban 指数)

Name	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
yar029w 74	0	0.5055	0.8585	0.9530	0.2478	0.9987
yarl02w a65		0	0.4007	0.5694	0.2577	0.5105
ybll08c-a42			0	0.6787	0.6265	0.1646
ycl057c-a97				0	0.7496	0.6242
yar020c 55					0	0.7574
yar070c 99						0

表 2.13: 表 2.2 中 6 种蛋白质序列对应的三维向量在欧式距离上的相似性矩阵（基于 Balaban 指数）

Name	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
yar029w 74	0	2.2123	3.0059	3.5431	1.0023	3.8209
yarl02w a65		0	1.3049	2.2469	1.9601	2.3460
ybll08c-a42			0	2.3906	2.4072	1.1557
ycl057c-a97				0	2.8995	2.4266
yar020c 55					0	3.0102
yar070c 99						0

表 2.14：表 2.2 中 6 种蛋白质序列对应的三维向量的两两向量之间的夹角弧度矩阵 (基于信息分布指数)

Name	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
yar029w 74	0	0.1168	0.1083	0.0745	0.0117	0.1269
yarl02w a65		0	0.0923	0.1173	0.1051	0.1079
ybll08c-a42			0	0.0499	0.1000	0.0204
ycl057c-a97				0	0.0706	0.0617
yar020c 55					0	0.1191
yar070c 99						0

表 2.15：表 2.2 中 6 种蛋白质序列对应的三维向量在欧式距离上的相似性矩阵 (基于信息分布指数)

Name	yar029w 74	yarl02w a65	ybll08c-a42	ycl057c-a97	yar020c 55	yar070c 99
yar029w 74	0	0.9655	2.5952	1.5055	0.3473	1.8628
yarl02w a65		0	2.7219	1.5840	0.8925	1.6427
ybll08c-a42			0	3.8655	2.9067	3.9858
ycl057c-a97				0	1.2018	0.6021
yar020c 55					0	1.5762
yar070c 99						0

由利用 Balaban 指数得到的相似性的数据表 2.8 、表 2.9 可知， Human 和 Gorilla 的 DNA 序列对应的三维向量之间的距离及夹角最小，即相似性最大，与实际相符. 还有 Goat 和其它物种的 DNA 序列对应的三维向量之间的距离及夹角都比较大，也基本符合已有的资料. 但是由于在 DNA 序列转化为图形时，出现碱基对应的顶点重叠的现象，这样丢失了一些信息，从而影响所得结果. 例如 Gallus 与 Opossum, Mouse, Rat 的 DNA 序列对应的三维向量之间的距离及夹角都比较小，使得它们的 DNA 序列比较相似，这与实际不太符合. 由利用信息分布指数得到的相似性的数据表 2.10 、表 2.11 可知， Human 和 Gorilla 的 DNA 序列对应的三维向量之间的距离及夹角最小，即相似性最大， Mouse 和 Rat 的 DNA 序列对应的三维向量之间的距离较小，所以也比较相似. 非哺乳动物 Gallus 和其它物种的 DNA 序列对应的三维向量之间的距离及夹角都比较大，因此相似性小， Opossum 和其它物种的 DNA 序列对应的三维向量之间的距离及夹角都比较大，因此相似性小，与实际相符. 再有，从表 2.12 、表 2.13 、表 2.14 、表 2.15 中都能看出蛋白质 yar029w 74 与蛋白质 yar020c 55 相似性最大，而在表 2.12 、表 2.13 中 yar070c 99 与 ybll08c a42 相似，在表 2.14 、表 2.15 中 yar070c 99 与 ycl057c a97 相似，这都比较符合实际已有的结果. 由于以上同样的问题，也存在一些不与实际相符的现象. 这是该种方法有待改进之处. 然而本文所提出的由生物序列转化为图的拓扑指数法有比较过程简单，速度快等优点. 由计算的结果来看，不论对 DNA 序列还是对蛋白质序列，信息分布指数所得结果较好. 并且表明该方法对于研究不同物种的 DNA 序列的进化和对蛋白质家族分析较有价值.

2.3　DNA 序列的特征数值

求序列的特征数值的一般方法是：首先把序列转换成图形表示，由图形表示提取特征矩阵 [10]: (1) 欧式矩阵 E: 其中 (i, j) 元由图形表示的曲线上两个基对应的点之间的欧式距离得到． E 为一对称矩阵，且主对角线元为零． (2) M/M 矩阵：其中 (i, j) 元由图形表示的曲线上两个基对应的点之间的欧式距离与它们之间存在的单位线段数之比 (即 $|j-i|$) 得到． M/M 为一对称矩阵，且主对角线元为零． (3) L/L (D/D) 矩阵：其中 (i,j) 元由图形表示的曲线上两个基对应的点之间的欧式距离与它们之间的图论距离 (曲线上两点间的单位线段长的和) 之比得到． L/L (D/D) 为一对称矩阵，且主对角线元为零，除此之外其它元都小于等于 1(大于 0) ． (4) 图论距离矩阵 D: 其中元素为图中第 i 个顶点到第 j 个顶点的最短路径所经过的边数，而每两个相邻顶点之间的边长取为单位长度． (5) 高阶矩阵： (a) 矩阵：亦为一个对称矩阵，其中 (i, j) 元由 D/D (L/L) 矩阵中每个元取 k 次幂得到． (b) 矩阵：为一 0, 1 矩阵，即对矩阵中每个元对 k 取极限得到．然后，从特征矩阵提取矩阵不变量 (不变量即独立于四个字符的一个量)： (1) 矩阵的最大特征值．这一不变量具有很大的优越性，已证实能比较好的反映序列的信息，但当序列长度较大时，矩阵特征值计算量很大，这就需要寻求其它的量来代替． (2) 矩阵的行列式值、矩阵的迹、矩阵的行 (列) 均值． (3) 平均带宽：对角线附近对称的一带状区域元素的平均值．最后把矩阵不变量作为序列的特征数值，对相应的序列进行相似性比较．

2.3.1　DNA 序列的图形表示

在这一节我们同样应用 A. Nandy 提出的二维图形表示，把 DNA 序列转化为图形表示．建立平面直角坐标系 oxy ，假定 4 种碱基 A, G, C, T 分别为沿坐标轴的 4 个方向移动的 2 维向量见文献 [17] ，则 DNA 序列就被转化为 oxy 平面内的 2 维图形．取人 (human) 和茶花鸡 (gallus) 的 β- 球蛋白基因的第一个外显子的前 15 个碱基组成的序列 ATGGTGCACCTGACT； ATGGTGCACTGGACT ，设 G, A 分别沿 x 轴的正和负的方向移动，而 C, T 分别沿 y 轴的正和负的方向移动，并且每个碱基仅移动一个单位，即 A$(-1,0)$, G$(1,0)$, T$(0,-1)$, C$(0,1)$ ，于是以上两个序列转化的 2 维图形示于图 2.3 ．

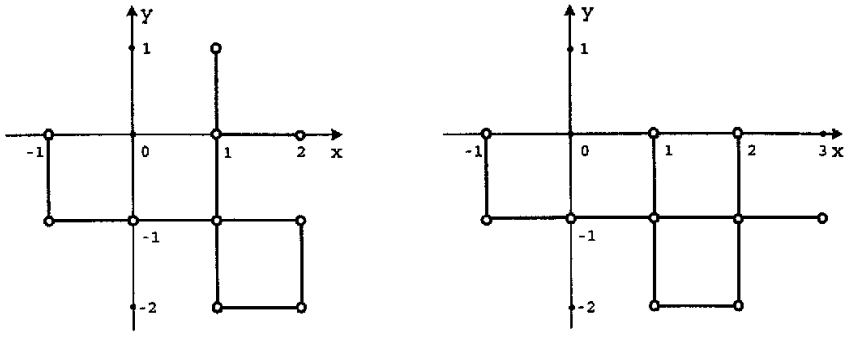


图 2.3　表 2.1 中的 human 和 gallus 的前 15 个 $\beta-$ 球蛋白基因的碱基序列对应的图形

根据上面已得到的 DNA 序列对应的 2 维图形来计算与图对应的图论距离矩阵。DNA 序列转化的图是一个连通图，给每个顶点标号如给图 2.3 的顶点标号示于图 2.4，则与之对应的图论距离矩阵是一个对称矩阵 $F=F^T$，其元素 (i, j) 是这样定义的：图中第 i 个顶点到第 j 个顶点的最短路所经过的边数之和，而每两个相邻顶点之间的边长取为单位长度，则图 2.4 对应的图论距离矩阵见表 2.16 .

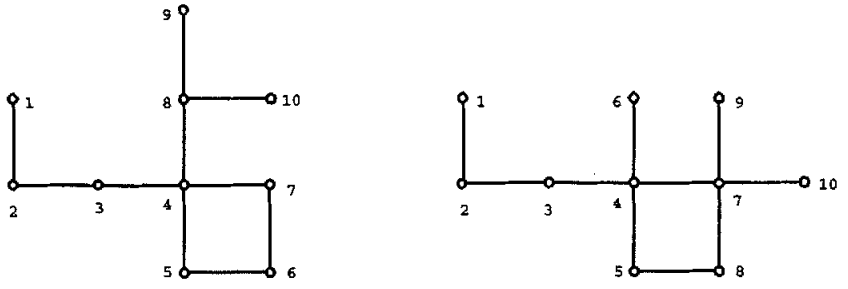


图 2.4　图 2.3 的顶点标号连通图

2.3.2　矩阵不变量

由前述可知，图与图论距离矩阵之间存在着一一对应关系，因此得到了矩阵的不变量就能够得到与其对应的图的不变量，矩阵的不变量有很多种，如矩阵的特征值、矩阵的最大 (最小) 行和、矩阵的行列式值等等. 在本节中我们取矩阵的主对角线以上 5 个次对角线元素之和的平均值，那么它也是矩阵的一个不变量. 这样我们得到了表 2.1 中前 8 个不同物种的 DNA 序列的特征数值，即每个 DNA 序列对应一个 5 维数组示于表 2.17 .

表 2.16: 图 2.4 对应的 10×10 图论距离矩阵

human	1	2	3	4	5	6	7	8	9	10	gallus	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	4	4	5	5	1	0	1	2	3	4	4	4	5	5	5
2		0	1	2	3	4	3	3	4	4	2		0	1	2	3	3	3	4	4	4
3			0	1	2	3	2	2	3	3	3			0	1	2	2	2	3	3	3
4				0	1	2	1	1	2	2	4				0	1	1	1	2	2	2
5					0	1	2	2	3	3	5					0	2	2	1	3	3
6						0	1	3	4	4	6						0	2	3	3	3
7							0	2	3	3	7							0	1	1	2
8								0	1	1	8								0	2	2
9									0	2	9									0	2
10										0	10										0

表 2.17: 表 2.1 中前 8 个不同物种的 DNA 序列对应的图论距离矩阵的主对角线以外的 5 个次对角线之和的平均值

Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat
1.2200	1.3000	1.3429	1.2727	1.3125	1.2826	1.2692	1.2766
2.0816	2.1633	2.4118	2.1395	2.0851	2.0444	2.1176	2.1739
2.9717	2.8758	3.0606	2.7619	2.8913	2.7955	2.8900	3.000
3.3191	3.6170	3.7419	3.1707	3.5556	3.5116	3.4018	3.7273
3.8696	4.3261	3.9032	3.5000	4.2727	4.2143	4.2708	4.4489

2.3.3　相似性分析

利用以上得到的 DNA 序列对应的特征数值对 DNA 序列进行相似性比较. 取每个 DNA 序列对应的特征数值 (5-D 数组) 作为在 5-D 空间向量的终点坐标，这样表 2.1 种前 8 个不同物种的 DNA 序列与 8 个 5-D 向量之间建立了一一对应关系. 设

$$V_1 = (v_{11},\ v_{12},\ v_{13},\ v_{14},\ v_{15}),\quad V_2 = (v_{21},\ v_{22},\ v_{23},\ v_{24},\ v_{25})$$

分别是两个序列 a 、 b 所对应的 5-D 向量. 通常地说，如果两个向量所指方向越相似则我们就认为这两个向量越相似. 对于这个假定，我们有两种方法计算：

(1) $d(V_1, V_2) = \sqrt{\sum_{i=1}^{5}(v_{1i} - v_{2i})^2}$ ， d 较小，就认为这两个序列比较相似；

(2) $\sin(V_1, V_2)$(或 $\cos(V_1, V_2)$) ，即两个向量所成角的正弦 (或余弦) ，如果两个向量所成角的正弦较小，就认为这两个向量比较相似.

由此我们得到了表 2.1 中前 8 个不同物种的 DNA 序列的相似性比较的数据，见表 2.18 、表 2.19 .

表 2.18: 表 2.1 中前 8 个不同物种的 DNA 序列对应的 5 维向量在欧式距离上的相似性比较矩阵

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat
Human	0	0.5666	0.0135	0.4070	0.4867	0.4015	0.4698	0.7776
Goat		0	0.5340	0.9491	0.1136	0.2193	0.0836	0.2237
Opossum			0	0.8106	0.5546	0.5989	0.5313	0.6383
Gallus				0	0.8755	0.7979	0.8865	1.1596
Lemur					0	0.1307	0.0661	0.3114
Mouse						0	0.1493	0.4246
Rabbit							0	0.2887
Rat								0

表 2.19: 表 2.1 中前 8 个不同物种的 DNA 序列对应的 5 维向量的两两向量之间的夹角弧度矩阵

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat
Human	0	0.0307	0.0531	0.0761	0.0322	0.0377	0.0308	0.0383
Goat		0	0.0784	0.0767	0.0102	0.0094	0.0053	0.0013
Opossum			0	0.0297	0.0814	0.0849	0.0781	0.0862
Gallus				0	0.0779	0.0083	0.0774	0.0856
Lemur					0	0.0079	0.0096	0.0143
Mouse						0	0.0097	0.4246
Rabbit							0	0.0099
Rat								0

由表 2.18 、表 2.19 可见，由于自身的 DNA 序列是完全相同，故主对角线上各元素欧式距离和夹角均为零，比较不同的两物种之间的相似性发现，非哺乳动物茶花鸡 (Gallus) 与其他人类 (Human) 等 7 个哺乳动物的 DNA 序列的相似性很小，还有 Opossum 与其它 7 个物种的 DNA 序列的相似性很小，这与图形的特征是一致的，也与实际相符．在 DNA 序列转化为图的形式时由于有些碱基重叠且不能表达连接的先后顺序，从而转化的图形有回路，这样就会丢掉一些信息从而影响所得结果，如 Mouse 和 Rat 同为鼠其 DNA 序列应有最大的相似性，可是在相似性数据里它们的数值不是最小，这与实际不符．虽然如此，本节提出的利用 DNA 序列的特征数值来比较 DNA 序列的相似性，比较过程简单、速度快并且对非哺乳动物与哺乳动物的 DNA 序列的相似性比较有研究价值．正如在文献 [10-14, 117, 118] 中指出的，不应该期望一个或几个不变量能够满足有限的甚至更短的 DNA 序列的特征，应该积极寻找发现其它新颖的更多的不变量，并且用它们来捕获 DNA 序列的主要特征，在 DNA 序列的比较和类比中有发现作用．

2.4 小结

近几年有很多人给出了不少方法比较人等生物的 β- 基因的第一个外显子 DNA 序列的相似性，如 M.Randic 分别构造了 12-D 向量 (向量元素为正规化最大特征值)[12] ， 16-D 向量 (向量元素为所有可能的有序相邻基对出现的的频率)[9] ， 64-D 向量 (向量元素为有序三联体出现的频率)[27] 和 n-D(n=5, 10 和 15) 向量 [10, 13] ，通过计算向量终点之间的欧氏距离来比较序列的相似性．贺平安和王军 [39, 40] 构造了 12 个矩阵，利用矩阵最大特征值来比较相似性．廖波和王天明 [28–31] 分别给出了生物序列 2D ， 3D 和 4D 表示法以及三联体的 6D 表示法，利用正规化最大特征值和序列不变量分别构造了 3-D, 5-D, 6-D, 10-D, 15-D 和 24-D 向量．他们所采用的方法大部分构造的向量元素为序列对应矩阵的正规化最大特征值，当序列长度较大时矩阵特征值计算量很大，这就需要寻求其它的量来代替．在本章我们基于生物序列的图形表示给出了能刻划生物序列特征的几种拓扑指数，并利用这些拓扑指数对生物序列进行了相似性比较，得到了与目前所有结果基本一致的结果，并且求得这些拓扑指数的算法简洁快速．

3 用随机游动描述 DNA 序列

在本章中，我们用随机游动来描述 DNA 序列，将 DNA 序列转化为一数字序列，从中提取特征，得到能够比较 DNA 序列相似度的方法，并用这个方法来比较了 9 个不同物种的 β- 球蛋白基因的第一个外显子的 DNA 序列的相似性。

3.1 引言

近年来在多种学科领域对 DNA 碱基序列的研究中发展起来的定量分析符号序列的方法、将碱基与数字对应起来的规则和进行统计分析的方法，被广泛应用．DNA 分子包涵了丰富的化学信息和生物信息，对于 DNA 序列的统计分析显得非常重要．将 DNA 序列表达成数字序列通常有从 1-D 到 n-D 不同维数空间的映射方式，其相应的统计方法：均方根涨落、熵近似方法、傅立叶变换和小波变换等，各种方法从多个角度、多个层次来分析揭示了 DNA 序列的结构规律． 1992 年， Peng 等 [58] 利用 " DNA Walk " 研究了核酸序列中的长程关联性质，并介绍了定量表示关联的方法． Dodin 等 [59] 运用傅立叶分析和小波变换作为 DNA 序列的可视化工具． Tsonis 等 [60] 用连续小波变换法探索 DNA 序列的局部结构，并通过实验考察讨论了基因进化的重要意义．罗辽复等 [61, 62] 也早就开展了 DNA 分子分形结构的研究，并计算了一些有机体的 DNA 的分维数值，他们还发现分维与进化水平保持很好的相关性．

DNA 序列比较的深度和复杂性无疑会使人们常用的比较方法达到一个新的高度．为了给出 DNA 序列的一个直观的刻划并比较，科学家们做了大量的工作．例如， Hamori 和 Ruskin[1-3] ， 张春霆 [49, 50, 55, 56] 由 G- 曲线和 H- 曲线、 Z 曲线表示 DNA 序列． M. A. Gate[4, 5] 和 Mogenthaler[6] ， Nandy[17–23] 由二维图形表示了 DNA 序列．近些年 Randic 等 [8, 13, 16] 将二维图形表示一般化到三维图形表示、四维图形表示等．还有罗辽复的随机游动 [35] 以及其他方法表示 DNA 序列 [39–44] ．这些方法为每个基因序列提供了一种简单的直观表示，使直接分析 DNA 序列中的碱基的分布及其关联，进而分析序列之间的相似性成为可能。

目前对 DNA 序列的理论分析中最普遍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有助于研究规律性和结构，以便对序列进行比较分析．

在本章中，用随机游动来描述 DNA 序列，将 DNA 序列转化为一数字序列，从中提取特征，得到能够比较 DNA 序列相似性的方法，并用这个方法来比较了 9 个不同物种的 β- 球蛋白基因的第一个外显子的 DNA 序列的相似性．

3.2 随机游动与 DNA 序列

要分析 DNA 序列就要将它通过某种规则映射成一个或多个离散的时间序列，表示为实数或复数值，这种方法称为 DNA 游动 (DNA Walk)，名称来源于统计学上随机游动的概念. 根据映射的度量空间的不同，可以得到不同维数空间的 DNA 游动. 其中一维映射方式，一维 DNA 游动有很多方法，使用最多的是碱基对应方法 [58, 63–65].

DNA 序列是字符集 $\mathcal{N}$={A, C, G, T } 上的一个词. 首先给 DNA 序列编号，然后把 DNA 序列看成一个具有 t 个元素的有限全序集合，它等同于 $[t] := \{1, 2, \cdots, t\}$. 从四种核苷酸基的化学结构入手，将它们分类，记 R 为嘌呤， Y 为嘧啶；即 R={A, G} 和 Y={C, T}. 类似的可以将这四个核苷酸基分为酮基和氨基两类：即 M={A, C} 和 K={G, T}. 从 DNA 双螺旋结构的构成还可以把四个核苷酸基分为弱氢键和强氢键两组：即 W={A, T} 和 S={C, G}[49]. 我们根据上面的分类结合文献 [58] 给出的一维映射：

$$\mathrm{Y}_m^{\mathrm{RY}} = \begin{cases} +1, & \text{如果 } m \in \mathrm{R} \\ -1, & \text{如果 } m \in \mathrm{Y} \end{cases}, \quad m \in [t] \tag{3.1}$$

$$\mathrm{Y}_m^{\mathrm{MK}} = \begin{cases} +1, & \text{如果 } m \in \mathrm{M} \\ -1, & \text{如果 } m \in \mathrm{K} \end{cases}, \quad m \in [t] \tag{3.2}$$

$$\mathrm{Y}_m^{\mathrm{WS}} = \begin{cases} +1, & \text{如果 } m \in \mathrm{W} \\ -1, & \text{如果 } m \in \mathrm{S} \end{cases}, \quad m \in [t] \tag{3.3}$$

如果 m 表示相继的时刻 $1, 2, \cdots, t$. 则称序列 Y_m^u， u=RY, MK, WS，为在 y 轴上的一个动点沿正负方向跳跃 1 个单位的随机游动. 又令： $\mathrm{X}_0 = 0$， $\mathrm{X}_n^u = \mathrm{X}_0 + \sum_{m=1}^{n} \mathrm{Y}_m^u$，$u$=RY, MK, WS， $0 \le n \le t$，则 X_n^u 表示动点（开始从 X_0 出发）第 n 时刻的位置，称 $\{\mathrm{X}_n^u, n \ge 0\}$ 为随机游动，称 $\{\mathrm{X}_n^u\}$ 的可能取值全体为状态空间，记为 S. 由此我们得到下面的结论：

命题：序列 $\{\mathrm{Y}_m^u\}$ 和 $\{\mathrm{X}_n^u\}$ 都是齐次马尔可夫链.

证明：根据 Y_m^u 的定义可知 $\mathrm{Y}_m^u (m = 1, 2, \cdots, t)$ 是独立分布的随机变量，显然是个马尔可夫链. 下证 $\{\mathrm{X}_n^u\}$ 也是马尔可夫链. 因为：

$$\begin{aligned} p_{i,j}^{(k)}(n) &= \mathrm{P}\{\mathrm{X}_{n+k} = j | \mathrm{X}_n = i\} \\ &= \mathrm{P}\{\mathrm{X}_n + \sum_{l=n+1}^{n+k} \mathrm{Y}_l = j | \mathrm{X}_n = \sum_{l=1}^{n} \mathrm{Y}_l = i\} \\ &= \mathrm{P}\{\sum_{l=n+1}^{n+k} \mathrm{Y}_l = j - i\} \\ &= \mathrm{P}\{\sum_{l=1}^{k} \mathrm{Y}_l = j - i\}, \end{aligned}$$

（由 $\{Y_m^u\}$ 的相互独立性知道 $P\{\sum_{l=n+1}^{n+k} Y_l = j-i\}$ 与 n 无关）

$$P\{X_0 + \sum_{l=1}^{k} Y_l = j | X_n = i\} = P\{X_k = j | X_0 = i\} = p_{i,j}^{(k)}(0) = p_{i,j}^{(k)},$$

即将来的状态仅依赖现在所在位置，而与以前所在位置无关，且与时刻无关. 所以 $\{X_n^u\}$ 是一个齐次马尔可夫链.

DNA 随机游动曲线定义为：

$$f(n) = X_n^u = X_0 + \sum_{m=1}^{n} Y_m^u, \quad X_0 = 0, \quad n \in [t], \quad u = RY, MK, WS$$

这样三种分类方式得到的曲线不同，做相似性分析时所研究的内容也不同. 如第一种编码方式探讨的是嘌呤与嘧啶的分布关系. 正因为如此，对每一条序列我们都分别计算它在每一种编码方式下得到的曲线的参数（不变量），并将其视为表征其 DNA 序列的一维特征. 为了下文叙述的方便，对这三种编码方式所得的曲线，我们分别称之为 DNA 序列的 RY−walk 曲线、 MK−walk 曲线以及 WS−walk 曲线. 我们取 9 种不同物种的 β-球蛋白基因的第一个外显子的 DNA 序列为研究对象，见表 2.1 . 例如我们取表 2.1 中人的 β- 球蛋白基因的第一个外显子的碱基序列，把它看成一个具有 92 个元素的有限全序集合，即 $[t]:=\{1,2,\cdots,91,92\}$. 然后我们根据 DNA 随机游动曲线定义画出其 RY−walk 曲线、 MK−walk 曲线以及 WS−walk 曲线见图 3.1—3.3 .

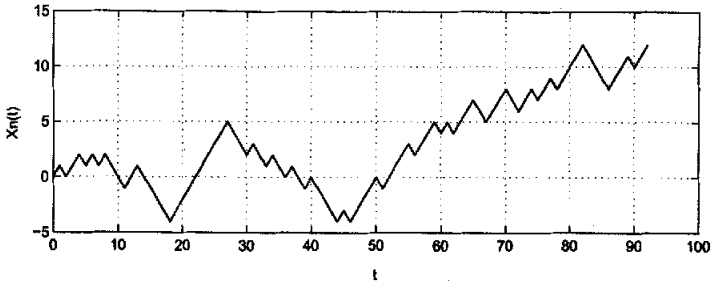


图 3.1　人的 β-Globin 基因的第一个外显子的碱基序列的 RY−walk 曲线

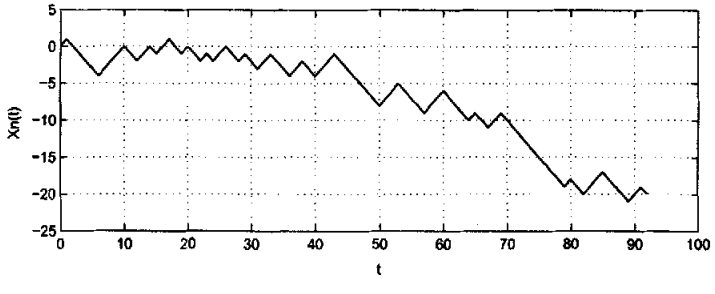


图 3.2　人的 β−Globin 基因的第一个外显子的碱基序列的 MK−walk 曲线

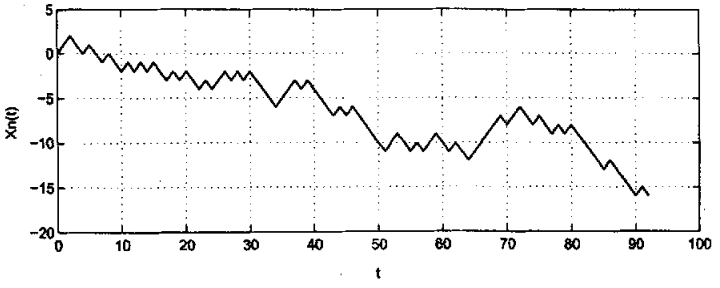


图 3.3 人的 β-Globin 基因的第一个外显子的碱基序列的 WS-walk 曲线

我们用同样的方法可以画出表 2.1 中其它物种的 β- 球蛋白基因的第一个外显子的 DNA 序列的三种 walk 曲线 (略)。

根据 DNA 随机游动定义我们能够得到表 2.1 中 9 种不同物种的 β- 球蛋白基因的第一个外显子的 DNA 序列对应的随机序列。为了下面的叙述方便，对三种分类方式所得的随机序列我们分别称之为 DNA 序列的 RY 、 MK 以及 WS 随机序列。如表 2.1 中 Human 的 β-Globin 基因的第一个外显子的碱基对应的的三种随机序列分别为：

$$
\begin{aligned}
Y_m^{RY} =& \{1,-1,1,1,-1,1,-1,1,-1,-1,-1,1,1,-1,-1,-1,-1,-1,1,1,1,1,1,1,1,1,1,-1,-1,-1,1,-1,-1,1,\\
&-1,-1,1,-1,-1,1,-1,-1,-1,-1,1,-1,1,1,1,1,1,-1,1,1,1,1,-1,1,1,1,1,-1,1,-1,1,1,1,-1,-1,1,\\
&1,1,-1,-1,1,1,-1,1,1,-1,1,1,1,1,1,-1,-1,-1,-1,1,1,1,-1,1,\};\\
Y_m^{MK} =& \{1,-1,-1,-1,-1,-1,1,1,1,1,-1,-1,1,1,-1,1,1,-1,-1,1,-1,-1,1,-1,1,1,-1,-1,1,-1,-1,1,1,\\
&-1,-1,-1,1,1,-1,-1,1,-1,-1,-1,-1,1,1,1,-1,-1,-1,-1,1,-1,-1,1,1,-1,-1,-1,-1,-1,-1,\\
&-1,-1,-1,-1,1,-1,-1,1,1,1,-1,-1,-1,-1,1,1,-1,1,1,-1,-1,-1,-1,-1,-1,-1,1,1,1,1,\};\\
Y_m^{WS} =& \{1,1,-1,-1,1,-1,-1,1,-1,-1,1,-1,1,-1,-1,1,-1,1,-1,-1,1,-1,1,1,-1,1,-1,1,-1,-1,\\
&-1,-1,1,1,1,-1,1,-1,-1,-1,-1,1,-1,1,-1,-1,-1,1,1,1,-1,-1,1,-1,1,1,-1,-1,1,-1,\\
&-1,1,1,1,1,1,-1,1,1,-1,-1,1,-1,-1,1,-1,1,-1,-1,-1,-1,1,1,-1,-1,-1,1,-1,-1,-1,1,-1,\};\\
X_n^{RY}(t) =& \{0,1,0,1,2,1,2,1,2,1,0,-1,0,1,0,-1,-2,-3,-4,-3,-2,-1,0,1,2,3,4,5,4,3,2,3,2,1,2,1,0,1,0,-1,\\
&0,-1,-2,-3,-4,-3,-4,-3,-2,-1,0,-1,0,1,2,3,2,3,4,5,4,5,4,5,6,7,6,5,6,7,8,7,6,7,8,7,8,9,8,\\
&9,10,11,12,11,10,9,8,9,10,11,10,11,12\};\\
X_n^{MK}(t) =& \{0,1,0,-1,-2,-3,-4,-3,-2,-1,0,-1,-2,-1,0,-1,0,1,0,-1,0,-1,-2,-1,-2,-1,0,-1,-2,-1,\\
&-2,-3,-2,-1,-2,-3,-4,-3,-2,-3,-4,-3,-2,-1,-2,-3,-4,-5,-6,-7,-8,-7,-6,-5,-6,-7,\\
&-8,-9,-8,-7,-6,-7,-8,-9,-10,-9,-10,-11,-10,-9,-10,-11,-12,-13,-14,-15,-16,-17,\\
&-18,-19,-18,-19,-20,-19,-18,-17,-18,-19,-20,-21,-20,-19,-20\};\\
X_n^{WS}(t) =& \{0,1,2,1,0,1,0,-1,0,-1,-2,-1,-2,-1,-2,-1,-2,-3,-2,-3,-2,-3,-4,-3,-4,-3,-2,-3,-2,\\
&-3,-2,-3,-4,-5,-6,-5,-4,-3,-4,-3,-4,-5,-6,-7,-6,-7,-6,-7,-8,-9,-10,-11,-10,-9,\\
&-10,-11,-10,-11,-10,-9,-10,-11,-10,-11,-12,-11,-10,-9,-8,-7,-8,-7,-6,-7,-8,-7,\\
&-8,-9,-8,-9,-8,-9,-10,-11,-12,-13,-12,-13,-14,-15,-16,-15,-16\};
\end{aligned}
$$

这样我们得到了 DNA 序列与随机数字序列之间、 DNA 序列与图形之间的对应关系，即可以用随机数字序列和图形来描述 DNA 序列。

3.3　DNA 序列的特征数值

因为 DNA 序列可看成一个由四个消息 (A, C, G, T) 组成的信息整体，同样也可以把对应的随机数字序列 Y_m^u 看成一个由两个消息 $(-1,+1)$ 组成的信息整体，由于这两个消息出现次数具有随机性，可采用平均信息量这个整体特征参数来描述 Y_m^u 或 DNA 序列. 即用下式引入的信息熵 [36] 来定量表示 DNA 序列.

$$H(p) = -\sum_i p_i \log p_i \quad (or \ \ H(p) = -\frac{1}{L}\sum_i p_i \log p_i) \tag{3.4}$$

式 (3.4) 中 $i \in \{-1,1\}$，p_i 为不同状态的转移概率，L 表示序列的长度. 从而我们可以用状态转移概率和信息熵来表征 DNA 序列.

例如设 Y_t^{RY} 为第 $t(t=1,2,\cdots,91,92)$ 个时刻的状态，状态空间为 S=$\{1,-1\}$，92 次状态 $\{Y_t^{RY}, t \geq 1\}$, 转移的情况是:

$-1 \longrightarrow -1$，19 次；$-1 \longrightarrow 1$, 21 次；$1 \longrightarrow -1$, 21 次；$1 \longrightarrow 1$, 30 次.

因此，一步转移的概率可用频率近似的表示为:

$$P_{-1,-1}^{RY} = P\{Y_{t+1} = -1 | Y_t = -1\} \approx \frac{19}{19+21} = \frac{19}{40};$$
$$P_{-1,1}^{RY} = P\{Y_{t+1} = 1 | Y_t = -1\} \approx \frac{21}{19+21} = \frac{21}{40};$$
$$P_{1,-1}^{RY} = P\{Y_{t+1} = -1 | Y_t = 1\} \approx \frac{21}{30+21} = \frac{21}{51};$$
$$P_{1,1}^{RY} = P\{Y_{t+1} = 1 | Y_t = 1\} \approx \frac{30}{30+21} = \frac{30}{51}.$$

对应的各个状态的信息熵为:

$$H_1(p) = -(\frac{19}{40} \times \log\frac{19}{40} + \frac{21}{40} \times \log\frac{21}{40} + \frac{21}{51} \times \log\frac{21}{51} + \frac{30}{51} \times \log\frac{30}{51}) = 1.3694$$

这样我们用同样的方法得到 9 个物种对应的 RY− 随机序列、MK− 随机序列以及 WS− 随机序列的一步转移概率及信息熵见表 3.1 、3.2 .

其次，用随机数字序列 X_n^u 的均值、方差、互协方差等数字特征可以表征 DNA 序列. 我们利用随机序列的均方差:

$$D(X_n^u) = (\frac{1}{L}\sum_{n=0}^{L}(X_n^u - \mu_{X_n^u})^2)^{\frac{1}{2}} \tag{3.5}$$

其中，$\mu_{X_n^u}$ 为序列 X_n^u 的均值. 计算了表 2.1 中 9 个不同物种的 β−Globin 基因的第一个外显子的碱基的 RY− 、MK− 以及 WS− 随机序列 X_n^u 的均方差值 $D(X_n^u)$ 见表 3.3 .

表 3.1: 表 2.1 中 9 个物种对应的随机序列 Y_m^u 的一步转移概率

Species	$P^{RY}_{-1,-1}$	$P^{RY}_{-1,1}$	$P^{RY}_{1,-1}$	$P^{RY}_{1,1}$	$P^{MK}_{-1,-1}$	$P^{MK}_{-1,1}$	$P^{MK}_{1,-1}$	$P^{MK}_{1,1}$	$P^{WS}_{-1,-1}$	$P^{WS}_{-1,1}$	$P^{WS}_{1,-1}$	$P^{WS}_{1,1}$
Human	$\frac{19}{40}$	$\frac{21}{40}$	$\frac{21}{51}$	$\frac{30}{51}$	$\frac{38}{56}$	$\frac{18}{56}$	$\frac{17}{35}$	$\frac{18}{35}$	$\frac{27}{54}$	$\frac{27}{54}$	$\frac{26}{37}$	$\frac{11}{37}$
Goat	$\frac{16}{34}$	$\frac{18}{34}$	$\frac{18}{51}$	$\frac{33}{51}$	$\frac{32}{52}$	$\frac{20}{52}$	$\frac{19}{33}$	$\frac{14}{33}$	$\frac{27}{52}$	$\frac{25}{52}$	$\frac{24}{33}$	$\frac{9}{33}$
Opossum	$\frac{21}{42}$	$\frac{21}{42}$	$\frac{21}{49}$	$\frac{28}{49}$	$\frac{31}{51}$	$\frac{20}{51}$	$\frac{19}{30}$	$\frac{21}{30}$	$\frac{18}{49}$	$\frac{31}{49}$	$\frac{30}{42}$	$\frac{12}{42}$
Gallus	$\frac{18}{39}$	$\frac{21}{39}$	$\frac{21}{52}$	$\frac{30}{52}$	$\frac{28}{49}$	$\frac{21}{49}$	$\frac{20}{42}$	$\frac{22}{42}$	$\frac{33}{58}$	$\frac{25}{58}$	$\frac{24}{33}$	$\frac{9}{33}$
Lemur	$\frac{18}{38}$	$\frac{20}{38}$	$\frac{20}{53}$	$\frac{33}{53}$	$\frac{36}{58}$	$\frac{22}{58}$	$\frac{21}{33}$	$\frac{12}{33}$	$\frac{21}{60}$	$\frac{29}{60}$	$\frac{28}{41}$	$\frac{13}{41}$
Mouse	$\frac{21}{42}$	$\frac{21}{42}$	$\frac{21}{51}$	$\frac{30}{51}$	$\frac{38}{57}$	$\frac{19}{57}$	$\frac{18}{36}$	$\frac{18}{36}$	$\frac{25}{54}$	$\frac{29}{54}$	$\frac{28}{39}$	$\frac{11}{39}$
Rabit	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{21}{53}$	$\frac{32}{53}$	$\frac{40}{57}$	$\frac{17}{57}$	$\frac{17}{32}$	$\frac{15}{32}$	$\frac{25}{53}$	$\frac{28}{53}$	$\frac{27}{36}$	$\frac{9}{36}$
Rat	$\frac{17}{39}$	$\frac{22}{39}$	$\frac{22}{52}$	$\frac{30}{52}$	$\frac{34}{54}$	$\frac{20}{54}$	$\frac{19}{37}$	$\frac{18}{37}$	$\frac{26}{51}$	$\frac{25}{51}$	$\frac{24}{40}$	$\frac{16}{40}$
Grilla	$\frac{18}{39}$	$\frac{21}{39}$	$\frac{21}{53}$	$\frac{32}{53}$	$\frac{39}{57}$	$\frac{18}{57}$	$\frac{17}{35}$	$\frac{18}{36}$	$\frac{28}{56}$	$\frac{28}{56}$	$\frac{27}{36}$	$\frac{9}{36}$

表 3.2: 表 2.1 中 9 个物种对应的随机序列 Y_m^u 的一步转移概率的信息熵

	Human	Goat	Oposs	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
$H_1(p)$	1.3694	1.3407	1.3761	1.3737	1.3545	1.3706	1.3506	1.3662	1.3616
$H_2(p)$	1.3207	1.3479	1.2087	1.3749	1.3192	1.3297	1.3006	1.3519	1.3164
$H_3(p)$	1.3017	1.2784	1.2558	1.2696	1.3049	1.2853	1.2539	1.3660	1.2555

表 3.3: 表 2.1 中 9 个物种对应的随机序列 X_n^u 的均方差

	Human	Goat	Oposs	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
$D(X_n^{RY})$	4.2927	6.1207	2.9590	4.3764	5.7913	3.2557	6.5832	4.7743	5.1645
$D(X_n^{MK})$	6.6257	5.7392	3.2293	2.5565	6.9540	5.6154	6.9870	5.6254	6.7463
$D(X_n^{WS})$	4.3737	5.3573	2.3434	6.9892	2.7898	3.7331	4.5815	2.8290	5.0518

3.4　相似性分析

用 9 个物种对应的随机序列的一步转移的概率值、信息熵以及均方差值作为 DNA 序列的不变量来比较它们的相似性。取表 3.1 、 3.2 、 3.3 中的概率值、信息熵和均方差值，构造一个 12 维向量和两个 3 维向量. 通过比较这些向量之间的相似性来研究这 9 个序列的相似性和非相似性，即两个向量越相似则它们对应的序列也越相似. 我们计算两个向量的欧式距离，如果两个向量的欧式距离较小就认为它们有较大的相似性，反之，相似性就小. 我们计算这些向量的欧式距离值并列在表 3.4 、 3.5 、 3.6 中，这样我们通过观察这三个表来分析它们的相似性和非相似性.

表 3.4: 表 2.1 中 9 个物种的基因序列的相似形表 (概率值 12 维向量的欧式距离)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.1820	0.3218	0.1855	0.2609	0.0718	0.1363	0.1759	0.0734
Goat		0	0.3730	0.1906	0.1796	0.1786	0.1844	0.2302	0.1771
Opossum			0	0.3790	0.3572	0.2891	0.3491	0.3686	0.3328
Gallus				0	0.3280	0.2123	0.2545	0.2250	0.1922
Lemur					0	0.2265	0.2383	0.2594	0.2762
Mouse						0	0.1449	0.2093	0.0962
Rabit							0	0.2475	0.1019
Rat								0	0.2355
Grilla									0

表 3.5: 表 2.1 中 9 个物种的基因序列的相似形表 (信息熵 3 维向量的欧式距离)

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.0459	0.1212	0.0631	0.0159	0.0187	0.0552	0.0715	0.0471
Goat		0	0.1454	0.0435	0.0414	0.0351	0.0542	0.0913	0.0442
Opossum			0	0.1668	0.1228	0.1247	0.0954	0.1810	0.1087
Gallus				0	0.0687	0.0479	0.0794	0.0994	0.0614
Lemur					0	0.0275	0.0544	0.0703	0.0500
Mouse						0	0.0473	0.0838	0.0339
Rabit							0	0.1243	0.0193
Rat								0	0.1162
Grilla									0

表 3.6：表 2.1 中 9 个物种的基因序列的相似形表（均方差 3 维向量的距离）

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	2.2572	4.1757	4.8380	2.2051	1.5832	2.3281	1.9023	1.1110
Goat		0	5.0378	3.9794	2.8594	3.2957	1.5404	2.8667	1.4219
Opossum			0	4.9036	4.7005	2.7772	5.6802	3.0451	4.9567
Gallus				0	6.2430	4.6060	5.5042	5.1850	4.6828
Lemur					0	3.0184	1.9592	1.6736	2.3564
Mouse						0	3.6977	1.7674	2.5810
Rabit							0	2.8631	1.5139
Rat								0	2.5198
Grilla									0

观察表 3.4 、3.5 、3.6 我们看出物种 human−gorilla， mouse−rat 对距离值最小，即其 DNA 序列有最大的相似性，与实际相符．其次发现物种 gallus 是 9 个物种中与其它物种最不相似的一个物种，还有物种 opossum 与其它 8 个物种的相似性很小，这些结果与其它的表征相似信息的方法 [28−31] 基本一致．另一方面， human−mouse， mouse−gorilla， human−lemur 对的值也比较小，说明它们之间有很大的相似性，与实际不太符合，也许是由于构造序列不变量时某些人为的因素造成了这种结果，也许是由于这些物种之间确实有某些内在的相似性．在其他的文献里也曾出现过这些物种有相似的结论，见文献 [28−31]．再有对这 9 个物种比较时，我们用的仅仅是它们的基因的一段 (β-Globin 基因的第一个外显子)，然而，每一个物种的基因组序列是非常长，并且都含有很多的基因，所以物种的全部遗传信息不可能包含在某一个基因里，因此利用这些比较，我们只能得到这些物种的某些相似性而不是全部信息．

3.5 小结

DNA 序列的描述、比较和类比分析仍然是科学家们研究的重要课题．在这个领域里有许多研究．由于 DNA 序列的数据库已经积累了大量数据资料，破译这些 DNA 序列密码的生物学意义，弄清 DNA 序列与生物进化，细胞功能，遗传机里的关系需要新颖的概念和方法去研究．在本章中，用随机游动来描述 DNA 序列，得到了 DNA 序列的 1-D 游动曲线表示且这种表示方法没有圈．本章的重点是通过游动曲线给出了 DNA 序列对应的随机数字序列 - 马尔可夫链，基于随机序列的概率分布、信息熵以及均方值得到了 DNA 序列的一些新的数学不变量，进而利用这些不变量来比较了 9 个不同物种的 DNA 序列的相似性．其中随机序列的均方值来表征 DNA 序列结果比较好．这种方法简单明了、速度快，使得易于直观进行 DNA 序列相似性比较．另一方面，由于这种图形表示是非退化的，所以减少了 DNA 序列所包含信息的丢失．基于 DNA 序列的图形表示，一些新的不变量可被引入为 DNA 序列的特征数值．在 DNA 序列的描述和类比分析中被广泛应用．

4 利用物种 DNA 序列的图形表示构建系统进化树的方法

本章提出了一种新的构建进化树方法：在 *DNA* 序列的三维图形表示的基础上，利用图的不变量给出了序列之间的距离度量，进而定义了物种进化距离，并利用基于距离法的 *NJ* 算法构建了生物系统进化树。选取 *30* 个物种线粒体 *DNA* 序列为材料，得到与传统的根据物种形态和其他方法构建的系统进化树基本一致的树形。

4.1 引言

构建系统进化树的研究是生物信息学中的一个热点，根据进化树不仅可以研究从单细胞有机体到多细胞有机体的生物进化过程，而且可以粗略估计现存的各类种属生物的进化时间 [76] 。

随着分子生物学的发展，人们发现生命的密码蕴涵在 DNA 链中，四种核苷酸的排序变化反映了进化信息。利用 DNA 序列进行发育分析就是在分子水平上推断并评价进化关系，并用分支图的形式表现出来，这种图就是系统进化树，简称进化树。进化树可分为有根树和无根树。近年来，较为流行的构建进化树方法有三大类：最大节约法 (maximum parsimony, MP) ，最大似然法 (maximum likelihood, ML) 和距离法。一种称为贝叶斯推断的统计学方法也开始使用。

最大似然方法考察一组序列的多重比对，优化出拥有一定拓扑结构和树枝长度的进化树，这个进化树能够以最大的概率与考察的多重比对结果符合。距离树考察一组序列的两两比对，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度。最大节约方法考察一组序列的多重比对结果，优化出的进化树能够利用最少的离散步骤去解释多重比对中的碱基差异。距离方阵方法简单的计算两个序列的差异数量。这个数量被看作进化距离，而其准确性依赖于进化模型的选择。然后运行一个聚类算法，从最相似 (也就是说，两者之间的距离最短) 的序列开始，通过距离值方阵计算出实际的进化树，或者通过将总的树枝长度最小化而优化出进化树。

Snel 等人 [66] 在 1999 年提出了基因容量作为一种距离度量。类似的方法还被其他人采用，但是当生物体的基因容量非常相似时，这样的方法就会失效。 Hasan H.Otu 和 Khalid Sayood 还提出了一些利用序列的 Lempel–Ziv 复杂性 [67] 作为一个新的的距离度量来构建进化树 [68] 。结果比较理想。但是计算量大，因为每个物种的基因序列都很长。距离法是最容易理解的、重要的构建系统树算法之一，常用的方法有： UPGMA 法 [138] 、Fitch-Margoliash 法 [75] 和 NJ(Neighbor-joining Method) 法 [69] ，其中 NJ 算法是效率最高的算法。这些方法可以生成有根树。在此基于距离法的 NJ 算法，我们尝试将张春霆等和

廖波，王天明提出的 DNA 序列三维图形表示 [49, 30] 引入分子进化的研究.

张春霆等提出的 DNA 序列的三维图形表示：根据 DNA 序列中的四个核苷酸的化学性质和化学结构可将其分为三类：(a) 嘌呤 R={A, G}，嘧啶 Y={C, T}；(b) 酮基 M={A, C}，氨基 K={G, T}；(c) 弱氢键 W={A, T}，强氢键 S={C, G}，将 DNA 序列的四个基看作一个正四面体的四个顶点，建立一个 $oxyz$ 坐标系. 具体的做法可直接用数学形式描述如下：定义 A(1,1,1), T(−1,−1,1), C(−1,1,−1), G(1,−1,−1)； Z 曲线上的点 (x_n, y_n, z_n)：

$$\begin{cases} x_n = (\mathrm{A}_n + \mathrm{G}_n) - (\mathrm{C}_n + \mathrm{T}_n), \\ y_n = (\mathrm{A}_n + \mathrm{C}_n) - (\mathrm{G}_n + \mathrm{T}_n), \\ z_n = (\mathrm{A}_n + \mathrm{T}_n) - (\mathrm{C}_n + \mathrm{G}_n). \end{cases} \tag{4.1}$$

其中，A_n 表示序列中前 n 个基因中碱基 A 出现的个数，其它同理.

廖波，王天明提出的 DNA 序列三维图形表示：他们将 A(adenine)，G(guanine)，T(thymine) 和 C(cytosine) 分别置于 -x 轴，+x 轴，-y 轴和 +y 轴，而特征曲线沿着 Z 轴伸展. 同样对应于 A, G, C, T 的以上三种分类方法，每个 DNA 序列有且只有 3 个特征曲线表示，用数学形式表示如下：设 $\mathrm{G} = g_1 g_2 \cdots g_n$ 为任意 DNA 序列，存在三个映射 $\phi_j, j = 1, 2, 3,$，$\phi_j(\mathrm{G}) = \phi_j(g_1)\phi_j(g_2)\cdots\phi_j(g_n)$，这里

$$\phi_1(g_i) = \begin{cases} (-1, 0, \mathrm{A}_i), & \text{如果 } g_i = \mathrm{A}, \\ (1, 0, \mathrm{G}_i), & \text{如果 } g_i = \mathrm{G}, \\ (0, -1, \mathrm{T}_i), & \text{如果 } g_i = \mathrm{T}, \\ (0, 1, \mathrm{C}_i), & \text{如果 } g_i = \mathrm{C}. \end{cases} \tag{4.2}$$

$$\phi_2(g_i) = \begin{cases} (-1, 0, \mathrm{A}_i), & \text{如果 } g_i = \mathrm{A}, \\ (1, 0, \mathrm{C}_i), & \text{如果 } g_i = \mathrm{C}, \\ (0, -1, \mathrm{T}_i), & \text{如果 } g_i = \mathrm{T}, \\ (0, 1, \mathrm{G}_i), & \text{如果 } g_i = \mathrm{G}. \end{cases} \tag{4.3}$$

$$\phi_3(g_i) = \begin{cases} (-1, 0, \mathrm{A}_i), & \text{如果 } g_i = \mathrm{A}, \\ (1, 0, \mathrm{T}_i), & \text{如果 } g_i = \mathrm{T}, \\ (0, -1, \mathrm{G}_i), & \text{如果 } g_i = \mathrm{G}, \\ (0, 1, \mathrm{C}_i), & \text{如果 } g_i = \mathrm{C}. \end{cases} \tag{4.4}$$

其中 A_i, G_i, T_i, C_i 分别代表 A, G, T, C 在序列中出现的累积的个数. 映射 ϕ_1, ϕ_2, ϕ_3 分别对应于形式 ATGC，ATCG，AGTC. 这样他们就把一个 DNA 序列转化为一个点集，

称之为特征点集，依次连接特征点集中各点所得曲线称之为特征曲线．由此 DNA 序列与几何图形之间建立了一一对应关系．以上这两种方法在 DNA 序列的分类、分析、相似性比较和基因识别等研究领域中得到了广泛的应用 [30, 49−56] ．

4.2　材料与方法

4.2.1　材料

目前，分子数据的大量涌现为系统发育分析提供了丰富的素材，但并非所有的数据都适合对特定问题的分析，在构建系统进化树时要求事先做出一个优先决定，哪些数据是合理的，哪些是不合理的．由于物种的线粒体 DNA 序列的差异只与变异有关，而线粒体 DNA 以每一百万年百分之二点二的速度变异，它是保守序列，因此我们选取文献 [68] 利用的 30 个物种的线粒体 DNA 序列作为研究对象．物种的名称及序列编号，见表 4.1 ．从网站 http://www.ncbi.clm.nih.gov 上免费下载．

表 4.1: 物种名称及序列代码对照表

Species	access No.	Species	access No.	Species	access No.
human	V00662	gray seal	X72004	squirrel	AJ238588
common chimpanzee	D38113	cat	U20753	fat dormouse	AJ001562
pigmy chimpanzee	D38116	fin whale	X61145	guinea pig	AJ222767
gorilla	D38114	blue whale	X72204	donkey	X97337
orangutan	D38115	cow	V00654	Indian rhinoceros	X97336
gibbon	X99526	rat	X14848	dog	U96639
baboon	Y18001	mouse	V00711	sheep	AF010406
horse	X79547	opossum	Z29573	pig	AJ002189
white rhinoceros	Y07726	wallaroo	Y10524	hippopotamus	AJ010957
harbor seal	X63726	platypus	X83427	rabbit	AJ001588

4.2.2　方法 1

我们用图形表示的数值刻划方法．根据 A.Nandy 提出的 DNA 序列的几何图形表示 [18-21] 给出了一种可比较的指标，称为曲线的散度均值：

$$\mu_x = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x(n_i),\quad \mu_y = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} y(n_i),\quad \mu_z = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} z(n_i)$$

这里 $x(n_i)$, $y(n_i)$ 和 $z(n_i)$ 分别表示几何图形表示中的第 i 个点的 x, y, z 坐标。

DNA 序列图形的半径：

$$G_R = (\mu_x^2 + \mu_y^2 + \mu_z^2)^{1/2}.$$

两个 DNA 序列 G_1, G_2 之间的距离：

$$d(G_1,\ G_2) = [(\mu_x - \mu_x')^2 + (\mu_y - \mu_y')^2 + (\mu_z - \mu_z')^2]^{1/2}.$$

以此来构造 DNA 序列之间的距离矩阵. 具体做法如下：如取 human 线粒体 DNA 序列的前 10 个字符 GATCACAGGT 和 gray seal 线粒体 DNA 序列的前 10 个字符 ACTAATGACT，根据上节式 (4.1) 求得对应的点集，示于表 4.2 . 表 4.2 对应的图形，示于图 4.1 、图 4.2 .

表 4.2：物种 human 和 gray seal 线粒体 DNA 序列的前 10 个字符对应的坐标

human	base	x_n	y_n	z_n	gray seal	base	x_n	y_n	z_n
1	G	1	-1	-1	1	A	1	1	1
2	A	2	0	0	2	C	0	2	0
3	T	1	-1	1	3	T	-1	1	1
4	C	0	0	0	4	A	0	2	2
5	A	1	1	1	5	A	1	3	3
6	C	0	2	0	6	T	0	2	4
7	A	1	3	1	7	G	1	1	3
8	G	2	2	0	8	A	2	2	4
9	G	3	1	-1	9	C	1	3	3
10	T	2	0	0	10	T	0	2	4

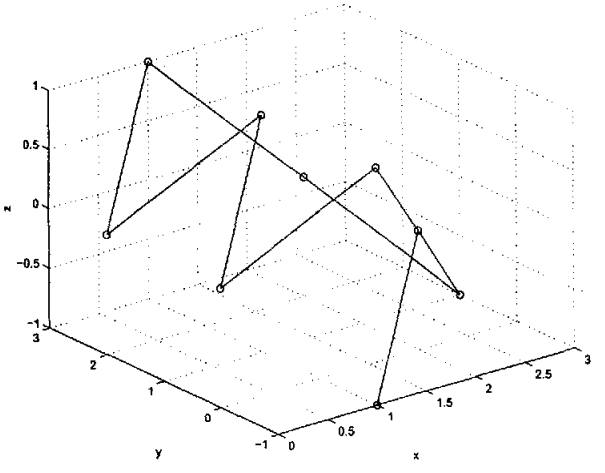


图 4.1 human 线粒体 DNA 序列的前 10 个字符对应的图

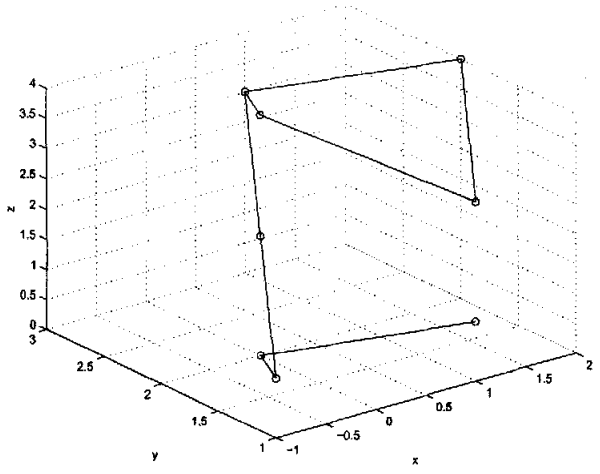


图 4.2　gray seal 线粒体 DNA 序列的前 10 个字符对应的图

则以上两个子序列对应曲线的散度均值向量分别为：$(\mu_x, \mu_y, \mu_z) = (1.3000, 0.7000, 0.1000)$，$(\mu'_x, \mu'_y, \mu'_z) = (0.5000, 1.9000, 2.5000)$，它们之间的距离：$d(\mathrm{G}_1, \mathrm{G}_2) = 2.8000$，然后以序列对应曲线的散度均值向量作为序列之间的距离度量来构造物种之间的进化距离矩阵。为了构造 30 个物种之间的距离矩阵，我们根据以上原理计算了 30 个物种线粒体 DNA 序列对应曲线的散度均值见表 4.3。

表 4.3：30 个物种的线粒体 DNA 序列对应曲线的散度均值

Species	μ_x	μ_y	μ_z	Species	μ_x	μ_y	μ_z
human	−0.7806	1.8988	0.9240	rat	−0.3740	1.6082	1.8117
c him	−0.7592	1.8566	1.1353	mouse	−0.3027	1.3440	2.1875
p chim	−0.7664	1.8482	1.0926	opossum	−0.3850	1.0487	2.6597
gorilla	−0.7509	1.7733	1.0522	wallaro	−0.5232	1.4880	1.7776
orangut	−0.7775	1.9928	0.7473	platypus	−0.6771	0.9423	2.1195
gibbon	−0.7136	1.9169	0.7665	squirrel	−0.6197	1.0657	2.1372
baboon	−0.6851	1.8570	1.0841	dormou	−0.6019	0.9297	2.2357
horse	−0.4755	1.7065	1.4824	g pig	−0.4123	1.1004	1.8858
w rhin	−0.3875	1.7996	1.5819	donkey	−0.4942	1.7093	1.4695
h seal	−0.2199	1.7180	1.4288	I rhin	−0.3813	1.7931	1.7388
g seal	−0.2242	1.7044	1.4273	dog	−0.4819	1.1956	1.7942
cat	−0.3459	1.4859	1.5998	sheep	−0.3381	1.5285	1.8338
f whale	−0.4297	1.5348	1.5652	pig	−0.1525	1.7564	1.7769
b whale	−0.4472	1.6035	1.5642	hippopo	−0.3247	1.8311	1.2850
cow	−0.3059	1.4259	1.7497	rabbit	−0.5958	1.2759	1.7255

接着构造 3 维向量 (向量元素为 30 个物种的线粒体 DNA 序列对应曲线的散度均值)，用其来表征 DNA 序列. 通过计算向量终点之间的欧式距离得到了 30 个物种的线粒体 DNA 序列之间的进化距离矩阵见表 4.4 。

表 4.4: 30 个物种的部分物种的线粒体 DNA 序列之间的距离矩阵

Species	human	c him	p chim	gorilla	orangut	gibbon	baboon	horse	w rhin	h seal
human	0	0.2165	0.4977	0.7285	0.8767	0.9520	0.9942	1.1984	1.3400	1.3935
c him		0	0.0441	0.2409	0.6406	0.8845	0.9448	1.0809	1.1924	1.2604
p chim			0	0.0865	0.4762	0.7681	0.8803	1.0663	1.2048	1.2781
gorilla				0	0.3766	0.6930	0.8399	1.0513	1.2102	1.2795
orangut					0	0.1011	0.4914	1.0977	1.4084	1.5031
gibbon						0	0.3245	0.9685	1.3242	1.4304
baboon							0	0.4746	0.9024	1.1211
horse								0	0.1622	0.4802
w rhin									0	0.2412
h seal										0

4.2.3 方法 2

我们同样根据 A.Nandy 提出的 DNA 曲线的散度均值来构造 DNA 序列之间的距离矩阵. 取 human 线粒体 DNA 序列的前 10 个字符 GATCACAGGT 和 gray seal 线粒体 DNA 序列的前 10 个字符 ACTAATGACT，在映射 ϕ_1 下对应的点集，示于表 4.5 .

表 4.5: 物种 human 和 gray seal 线粒体 DNA 序列的前 10 个字符对应的坐标

human	base	x	y	z	gray seal	base	x	y	z
1	G	1	0	1	1	A	-1	0	1
2	A	-1	0	1	2	C	0	1	1
3	T	0	-1	1	3	T	0	-1	1
4	C	0	1	1	4	A	-1	0	2
5	A	-1	0	2	5	A	-1	0	3
6	C	0	1	2	6	T	0	-1	2
7	A	-1	0	3	7	G	1	0	1
8	G	1	0	2	8	A	-1	0	4
9	G	1	0	3	9	C	0	1	2
10	T	0	-1	2	10	T	0	-1	3

图 4.3 —图 4.5 为 human 线粒体 DNA 序列的前 10 个字符 GATCACAGGT 和 gray seal 线粒体 DNA 序列的前 10 个字符 ACTAATGACT，在映射 ϕ_1, ϕ_2, ϕ_3 下的特征曲线。

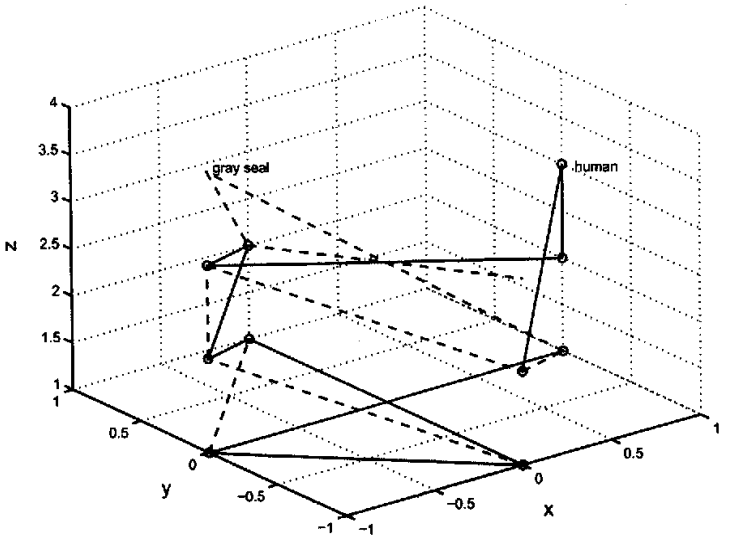


图 4.3　基于映射 $\phi_1$human 和 gray seal 线粒体 DNA 序列的前 10 个字符对应的图

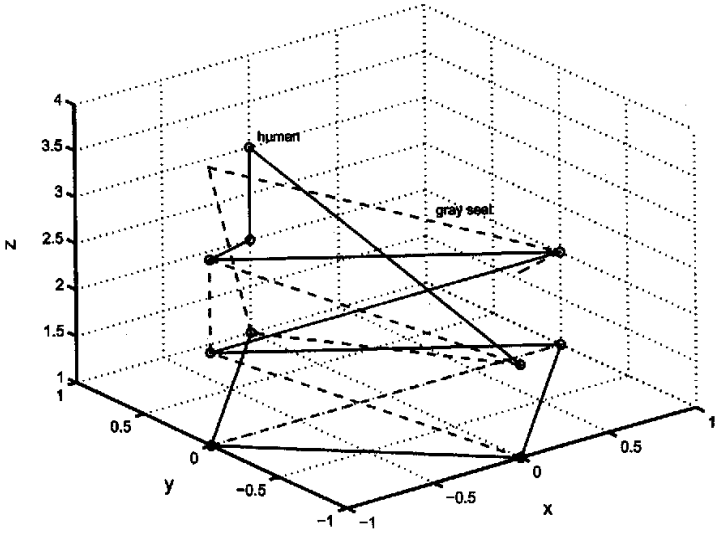


图 4.4　基于映射 $\phi_2$human 和 gray seal 线粒体 DNA 序列的前 10 个字符对应的图

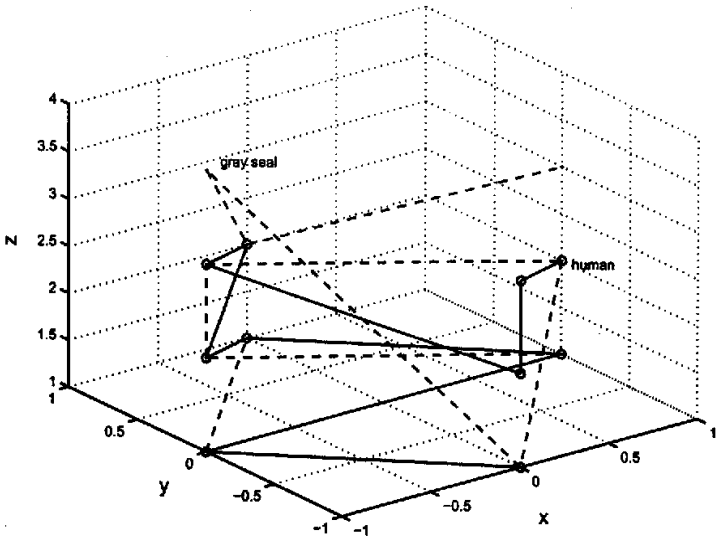


图 4.5　基于映射 $\phi_3$human 和 gray seal 线粒体 DNA 序列的前 10 个字符对应的图

我们利用曲线的散度均值构造以上两个子序列对应的九维向量：

$$(\mu_x^1,\mu_y^1,\mu_z^1,\mu_x^2,\mu_y^2,\mu_z^2,\mu_x^3,\mu_y^3,\mu_z^3)=(0,0,1.8,-0.1,0.1,1.8,-0.1,-0.1,1.8)$$

$$(\mu_x^{1\prime},\mu_y^{1\prime},\mu_z^{1\prime},\mu_x^{2\prime},\mu_y^{2\prime},\mu_z^{2\prime},\mu_x^{3\prime},\mu_y^{3\prime},\mu_z^{3\prime})=(-0.3,-0.1,2,-0.2,-0.2,2,-0.1,0.1,2)$$

则它们之间的距离定义为欧式距离： $d(\mathrm{G}_1,\mathrm{G}_2)=0.6000$ ，我们根据上节原理计算了 30 个物种线粒体 DNA 序列对应曲线的散度均值见表 4.7 ，和部分物种线粒体 DNA 序列之间的距离矩阵见表 4.6 。

表 4.6：30 个物种的部分物种的线粒体 DNA 序列之间的距离矩阵

Species	human	c him	p chim	gorilla	orangut	gibbon	baboon	horse	w rhin	h seal
human	0	0.0114	0.1069	0.3324	0.5771	0.7605	0.8722	0.9339	0.9696	0.9848
c him		0	0.0610	0.1236	0.3536	0.5965	0.7727	0.8792	0.9401	0.9696
p chim			0	0.0865	0.2484	0.4998	0.7071	0.8409	0.9203	0.9595
gorilla				0	0.0336	0.1849	0.4325	0.6602	0.8243	0.9111
orangut					0	0.0095	0.0984	0.3147	0.5707	0.7567
gibbon						0	0.0222	0.1527	0.4073	0.6403
baboon							0	0.0116	0.1421	0.3782
horse								0	0.0811	0.2853
w rhin									0	0.0627
h seal										0

表 4.7: 在映射 ϕ_1 ， ϕ_2 ， ϕ_3 下 30 个物种线粒体 DNA 序列对应曲线的散度均值

Species	mean in ϕ_1			mean in ϕ_2			mean in ϕ_3		
human	−0.0002	0.0001	2.2496	0.0000	−0.0001	2.2496	−0.0001	0.0002	2.2496
c chimp	−0.0002	0.0001	2.2562	0.0000	−0.0001	2.2562	−0.0001	0.0002	2.2562
p chimp	−0.0002	0.0001	2.2503	0.0000	−0.0001	2.2503	−0.0001	0.0002	2.2503
gorilla	−0.0002	0.0001	2.2151	0.0000	−0.0001	2.2151	−0.0001	0.0002	2.2151
orangut	−0.0002	0.0001	2.2345	0.0000	−0.0001	2.2345	−0.0001	0.0002	2.2345
gibbon	−0.0002	0.0001	2.2290	0.0000	−0.0001	2.2290	−0.0001	0.0002	2.2290
baboon	−0.0002	0.0001	2.2418	0.0000	−0.0001	2.2418	−0.0001	0.0002	2.2418
horse	−0.0002	0.0000	2.2485	0.0000	−0.0001	2.2485	−0.0001	0.0002	2.2485
white r	−0.0002	0.0000	2.2953	−0.0001	−0.0001	2.2953	−0.0001	0.0002	2.2953
g seal	−0.0002	0.0000	2.2591	−0.0001	−0.0001	2.2591	−0.0001	0.0001	2.2591
cat	−0.0002	0.0000	2.2551	−0.0001	−0.0001	2.2551	−0.0001	0.0001	2.2551
f whale	−0.0002	0.0000	2.2807	−0.0001	−0.0001	2.2807	−0.0001	0.0001	2.2807
b whale	−0.0002	0.0000	2.2175	−0.0001	−0.0001	2.2175	−0.0001	0.0001	2.2175
cow	−0.0002	0.0000	2.2250	−0.0001	−0.0001	2.2250	−0.0001	0.0001	2.2250
rat	−0.0002	0.0000	2.2142	−0.0001	−0.0001	2.2142	−0.0001	0.0001	2.2142
mouse	−0.0002	0.0000	2.2393	−0.0001	−0.0001	2.2393	−0.0001	0.0001	2.2393
opossu	−0.0002	0.0000	2.2536	−0.0001	−0.0002	2.2536	−0.0001	0.0001	2.2536
wallaroo	−0.0002	−0.0001	2.4193	−0.0001	−0.0002	2.4193	0.0000	0.0001	2.4193
platypus	−0.0002	0.0000	2.3001	−0.0001	−0.0001	2.3001	−0.0001	0.0001	2.3001
squirrel	−0.0002	−0.0001	2.3104	−0.0001	−0.0002	2.3104	0.0000	0.0001	2.3104
f dormo	−0.0002	−0.0001	2.2622	−0.0001	−0.0002	2.2622	0.0000	0.0001	2.2622
g pig	−0.0002	−0.0001	2.2842	−0.0001	−0.0002	2.2842	0.0000	0.0001	2.2842
donkey	−0.0002	0.0000	2.2449	−0.0001	−0.0001	2.2449	0.0000	0.0001	2.2449
I rhin	−0.0002	0.0000	2.2587	0.0000	−0.0001	2.2587	−0.0001	0.0002	2.2587
dog	−0.0002	0.0000	2.3016	−0.0001	−0.0001	2.3016	−0.0001	0.0001	2.3016
sheep	−0.0002	0.0000	2.2384	−0.0001	−0.0001	2.2384	0.0000	0.0001	2.2384
pig	−0.0002	0.0000	2.2624	−0.0001	−0.0001	2.2624	−0.0001	0.0001	2.2624
hippopo	−0.0002	0.0000	2.2812	−0.0001	−0.0001	2.2812	−0.0001	0.0001	2.2812
rabbit	−0.0002	0.0000	2.2115	0.0000	−0.0001	2.2115	−0.0001	0.0001	2.2115

4.3 进化树的构建

根据以上两种方法算得的 30 个物种的线粒体 DNA 序列之间的进化距离矩阵，利用 Neighbor-Joining 方法 [68] 构建了 30 个物种的进化树，如图 4.6 和图 4.7 所示。

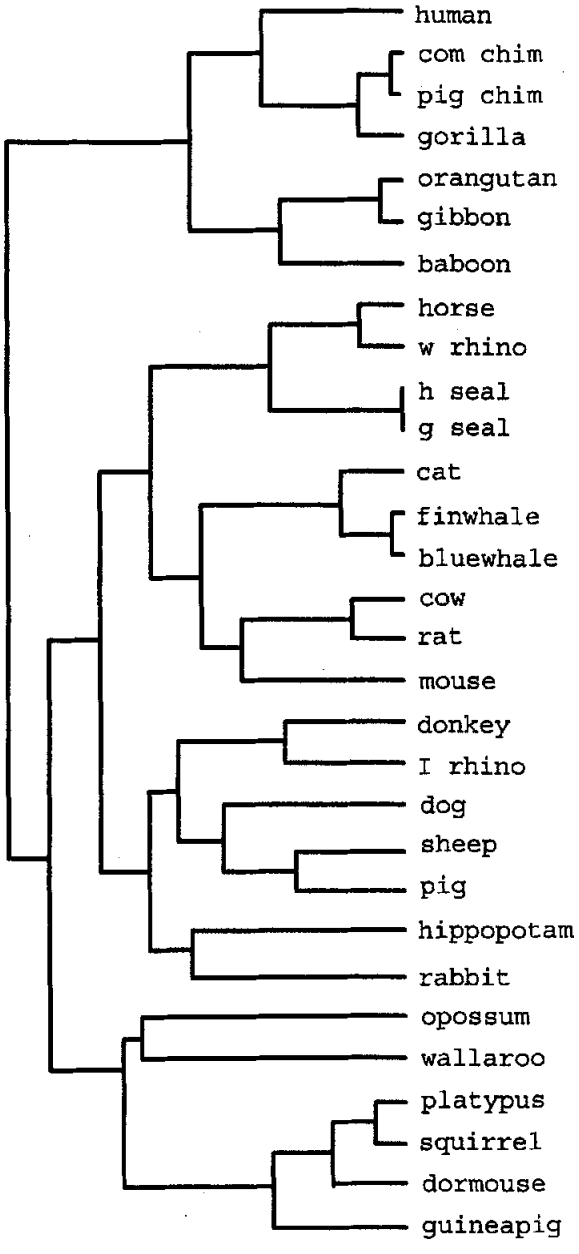


图 4.6 用方法 1 构建的 30 个物种线粒体 DNA 序列的系统树

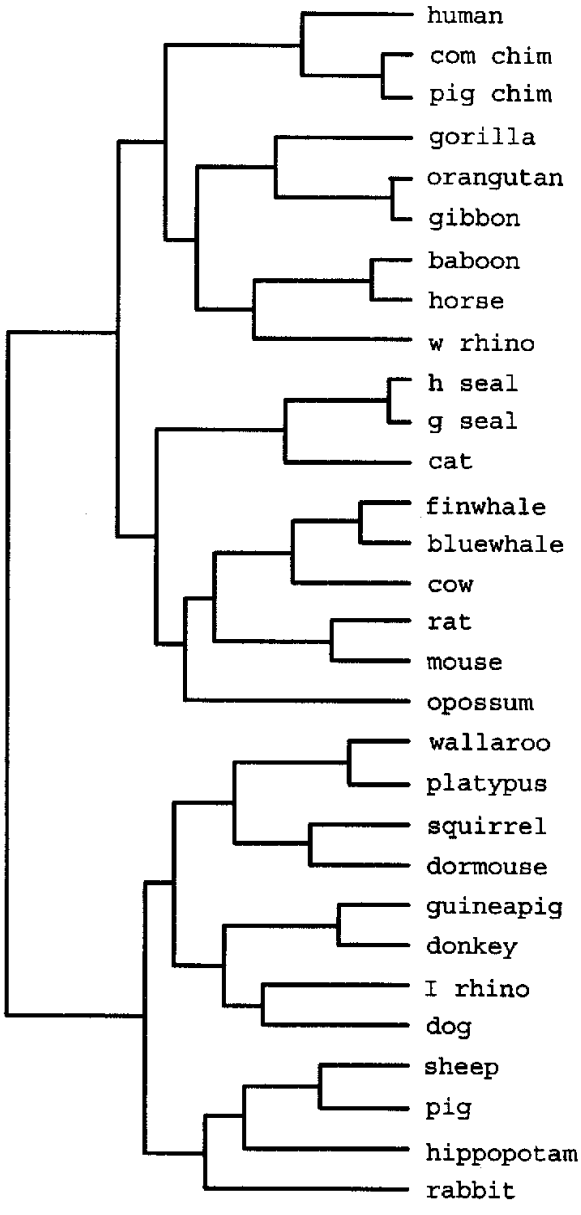


图 4.7　用方法 2 构建的 30 个物种线粒体 DNA 序列的系统树

从图 4.6 可以看出， common chimpanzee 和 pigmy chimpanzee 进化关系最为密切，其次是 gorilla，而后是 human， orangutan 和 gibbon 进化关系也比较近，接着是 baboon，而且以上七种生物归属一类. harbor seal 和 gray seal， fin whale 和 blue whale 进化关系也很近，另外， opossum 和 wallaroo，其次是 platypus， squirrel， fat dormouse， guinea pig 归为一类. donkey， Indian rhinoceros， sheep 和 pig， dog， horse 和 white rhinoceros，

hippopotamus 也基本上归为一类，但和其他生物的进化关系较远， rat 和 mouse 本来进化关系很近，但在此文章里 rat 和 cow 的进化关系最近，其次与 mouse 较近.

从图 4.7 可以看出， common chimpanzee 和 pigmy chimpanzee 进化关系最为密切，其次是 human，而后是 orangutan 和 gibbon， gorilla，进化关系也比较近，而且以上六种生物归属一类. rat 和 mouse， harbor seal 和 gray seal，其次是 cat， fin whale 和 blue whale 进化关系也很近，另外， platypus 和 wallaroo， squirrel 和 fat dormouse， guinea pig 和 donkey. sheep， pig， hippopotamus， rabbit 归为一类， baboon 没有与开始提到的 common chimpanzee 等六个物种归为一类， Indian rhinoceros 和 white rhinoceros，本该归为一类，但在此文章里 Indian rhinoceros 和 white rhinoceros，的进化关系较远.

以上两种结果中都出现了不符合实际的一些问题，造成这样的结果可能是由于把 DNA 序列转化成图形表示以及抽取图的不变量时失去了一些信息，这可能是我们取的参数的灵敏度不够好而造成的，这是有待改进之处. 综合看来，本章得到的进化关系与文献 [68] 里的结果基本一致，也与根据形态特征对物种分类构建的系统树基本保持了一致.

4.4 小结

从本章得到的结果来看，基于 DNA 序列的图形表示给出的距离度量能够作为推断物种之间进化关系的进化距离度量. 同时，此方法在具体处理过程中无需对序列进行比对分析，这与传统方法不同之处 [70–72]，并且，此方法可以推广到基因组层次上对序列进行比较，构建基因树. 当然，进化树的构建是一个统计学问题. 我们所构建出来的进化树只是对真实的进化关系的评估或者模拟. 如果采用的是一个适当的方法，那么所构建的进化树就会接近真实的 "进化树". 不同的算法有不同的适用目的. 本文提出的这个方法 [73, 74] 来做进化分析有简便，快速和可行等优点，且适用于所检验的序列的碱基数目较大 (大于几千个碱基) 的情况. 利用这一方法来揭示隐含在 DNA 序列中遗传信息还有很多的挑战，其中最重要的就是要如何能找到一些更适当的参数来突出 DNA 序列的主要特征进而对物种的基因组进行大规模理论分析并挖掘出生物序列中有用的信息.

5 RNA 二级结构的相似性分析

本章在复平面上用二维随机游动来描述了 RNA 二级结构序列，得到了对应的随机游动曲线和随机复数字序列. 在 6-D 空间中定义了使核苷酸集与点集之间一一对应的函数，进而利用这个函数在 6-D 空间中得到了 RNA 二级结构的 6-D 表示，然后基于 6-D 表示把它转化为矩阵表示和特征向量表示. 并利用 RNA 二级结构对应的随机数字序列的数字特征：模和相位以及矩阵不变量：矩阵的最大特征值、特征向量来分析了 AIMV-3 等 9 种病毒的 RNA 二级结构序列的相似性.

5.1 引言

今天的生命科学已不再是单纯的实验描述性科学. 在很多方面正在逐步由定性描述走向定量研究，不断出现一些新的研究手段. 近年来，在许多科学领域内对 DNA 碱基序列的研究中用定量分析的方法研究了符号序列，即将碱基与数字对应起来，几何图形与碱基对应起来，然后利用数字序列的特征数值以及图的不变量来描述 DNA 序列，且此方法在生物序列的研究中得到了广泛的应用 [1−35, 38−42, 46−75, 94−128] . 我们知道 DNA 是携带生物遗传信息的主要大分子，但 RNA 是大部分病毒的遗传物质，并且 RNA 还参与蛋白质的合成，与细胞分化，代谢，记忆的储存等有重要关系，因此研究 RNA 同等重要. 为了更好的了解 RNA 的功能就需要剖析 RNA 的结构.

用理论计算方法预测 RNA 的二级结构和比较 RNA 二级结构的相似性是当前结构研究中的一个十分活跃的领域. 且科学家们给出了很多算法预测 RNA 的二级结构和比较 RNA 二级结构的相似性. 而已有的比较相似性算法都是建立在字符串的比对上，他们的共同特点是给出插入，删除，替换的距离函数，通过计算结构之间的距离来比较相似性 [77−93] ，这种方法存在选取罚分函数的随意性，缺乏合适的理论模型而更多带有主观色彩，而选取罚分函数的好坏直接影响相似性分值，另外这些方法都忽略了组成基的化学性质和化学结构，且不适用于带假结以及较大的 RNA 二级结构相似性比较. 为了避免这些缺陷，使得很多人试图寻找其它方法来比较 RNA 二级结构序列，近来，廖波和王天明 [102−105] 根据 RNA 二级结构的组成特点给出了一种 3-D 和 6-D 表示法，并利用这些表示方法来比较 RNA 二级结构的相似性，这些方法计算简单，不受是否带假结的限制，得到了很好的结果.

我们在这章里，尝试着类似于用定量分析 DNA 序列的方法，根据廖波等给出的 RNA 二级结构的特征序列和张春霆的 Z 曲线表示法提出了用二维随机游动描述 RNA 二级结构序列，得到 RNA 二级结构序列的新的 2-D 和 6-D 表示方法，然后基于 2-D 和 6-D 表示把它转化为特征向量和矩阵表示，最后利用矩阵不变量及特征向量之间的距离来描述序

把它转化为特征向量和矩阵表示，最后利用矩阵不变量及特征向量之间的距离来描述序

列或结构的不变性来分析比较 RNA 二级结构的相似性，这些方法同样计算简单，不受是否带假结的限制，并且结果也很满意 [101]．本章我们以 9 种病毒的 RNA-3 末端的二级结构 (见图 5.1) 为例分别介绍了 2-D 和 6-D 表示法的应用．

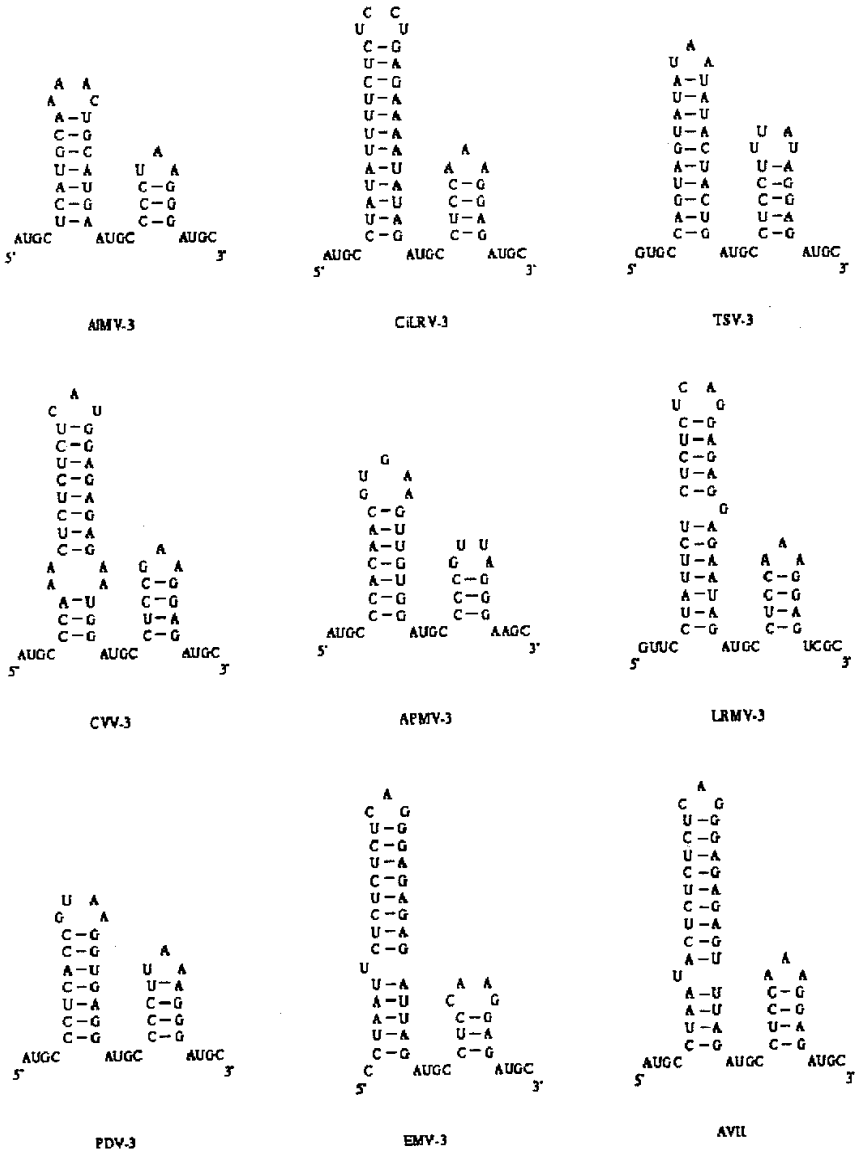


图 5.1　9 种病毒的 RNA-3 在 3' 末端的二级结构：alfalfa mosaic virus(AlMV-3 [73]), citrus leaf rugose virus(CiLRV-3 [74]), tobacco streak virus (TSV-3 [75,76]), citrus variegation virus(CVV-3 [74]), apple mosaic virus (APMV-3 [77]), prune dwarf ilarvirus(PDV-3 [78]), lilac ring mottle virus(LRMV-3 [79]), elm mottle virus(EMV-3 [80]) 和 asparagus virus II(AVII[81])，标号从 3' 端开始．

5.2　用二维随机游动描述 RNA 二级结构序列

根据廖波和王天明的 3-D 和 6-D 表示方法 [102–105]，我们先对二级结构进行处理，由于 RNA 二级结构是由自由基 A, C, G, U 和基对 A–U, G–C 组成的 (一般都视基对 G–U 为自由基)，为了区别自由基和基对，我们用 A', U', G', C' 分别代表基对 A–U 和基对 G–C 中的 A, C, G, U, 这样我们就将二级结构转化为基本序列，我们称之为二级结构的特征序列. 例如 AlMV-3 的子结构 (见图 5.2) 对应的特征序列为 $G'G'G'AAUC'C'C'$ (从 $3'$ 端到 $5'$ 端)

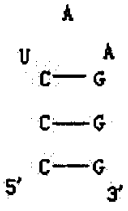


图 5.2　AlMV-3 的子结构

根据组成 RNA 的四种核苷酸基的化学结构和化学性质可将其分类，嘌呤 ((R={A, G})/ 嘧啶 (Y={C, U})；酮基 (M={A, C})/ 氨基 (K={G, U})；弱氢键 (W={A, U})/ 强氢键 (S={G, C}). 根据上面的分类和张春霆的 Z 曲线表示我们适当的将 RNA 二级结构 (见图 5.1) 映射到点的三维数列:

$$\phi^m = \{(\mathrm{Re}(\mathrm{P}_0^m), \mathrm{Im}(\mathrm{P}_0^m), 0), (\mathrm{Re}(\mathrm{P}_1^m), \mathrm{Im}(\mathrm{P}_1^m), 1), \cdots, (\mathrm{Re}(\mathrm{P}_\mathrm{L}^m), \mathrm{Im}(\mathrm{P}_\mathrm{L}^m), \mathrm{L})\}$$

其中，$m = 1, 2, 3$，L 为 RNA 二级结构对应序列 $G = g_1 g_2 \cdots$, 的长度，在复平面上第 n $(0 \le n \le \mathrm{L})$ 个碱基 g_n 对应的点 P_n^m 满足如下等式:

$$\begin{cases} \mathrm{P}_n^1 = (\mathrm{A}_n + \mathrm{G}_n) - (\mathrm{C}_n + \mathrm{U}_n) + [(\mathrm{A}'_n + \mathrm{G}'_n) - (\mathrm{C}'_n + \mathrm{U}'_n)]\mathrm{i}, \\ \mathrm{P}_n^2 = (\mathrm{A}_n + \mathrm{C}_n) - (\mathrm{G}_n + \mathrm{U}_n) + [(\mathrm{A}'_n + \mathrm{C}'_n) - (\mathrm{G}'_n + \mathrm{U}'_n)]\mathrm{i}, \\ \mathrm{P}_n^3 = (\mathrm{A}_n + \mathrm{U}_n) - (\mathrm{C}_n + \mathrm{G}_n) + [(\mathrm{A}'_n + \mathrm{U}'_n) - (\mathrm{C}'_n + \mathrm{G}'_n)]\mathrm{i}. \end{cases} \tag{5.1}$$

设 $r_n^m = \| \mathrm{P}_n^m \|$, $\varphi_n^m = arg\mathrm{P}_n^m$, $m = 1, 2, 3$. 例如,

$$r_n^1 = \sqrt{[(\mathrm{A}_n + \mathrm{G}_n) - (\mathrm{C}_n + \mathrm{U}_n)]^2 + [(\mathrm{A}'_n + \mathrm{G}'_n) - (\mathrm{C}'_n + \mathrm{U}'_n)]^2}$$

其中，$\mathrm{A}_n, \mathrm{G}_n, \mathrm{C}_n, \mathrm{U}_n, \mathrm{A}'_n, \mathrm{G}'_n, \mathrm{C}'_n$ 和 U'_n 分别为 1 到 n 这个子序列中碱基 A, G, C, U, A', G', C' 和 U'，所出现的次数. 我们假设 $\mathrm{A}_0 = \mathrm{G}_0 = \mathrm{C}_0 = \mathrm{U}_0 = \mathrm{A}'_0 = \mathrm{G}'_0 = \mathrm{C}'_0 = \mathrm{U}'_0 = 0$. 当 n 依次从 0 到 L 时，在复平面上我们便可以得到 $\mathrm{L}+1$ 个点，依次连接这些点就得到了从原点出发的 RNA 二级结构序列的一条随机游动曲线，同时也得到了与其对应的随机数字序列. 显然，对于任意一个 RNA 二级结构序列分别有称作 AG, AC, 和 AU 的三个随

机游动曲线和与其对应的三个随机数字序列．例如，病毒 AIMV-3 的二级结构序列对应的特征序列为 (从 5′ 端到 3′ 端)：

AUGCU′C′A′U′G′C′A′AAACU′G′C′A′U′G′A′AUGCC′C′C′UAAG′G′G′AUGC，它对应的三种 AG, AC, 和 AU 游动曲线，见图 5.3 一图 5.5 .

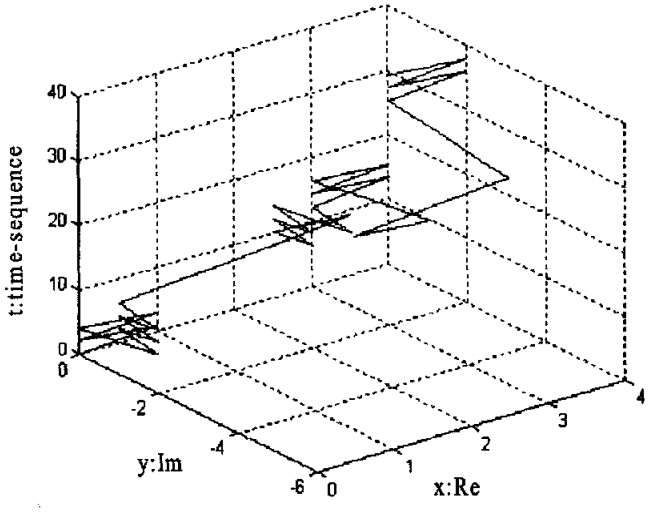


图 5.3 AIMV-3 的二级结构序列对应的 AG 随机游动曲线

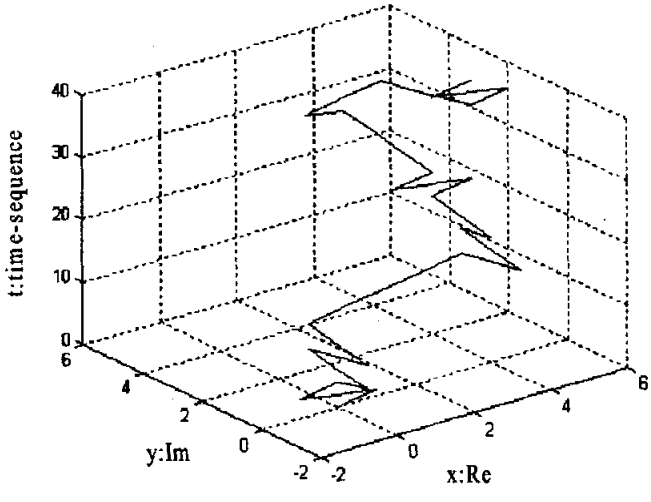


图 5.4 AIMV-3 的二级结构序列对应的 AC 随机游动曲线

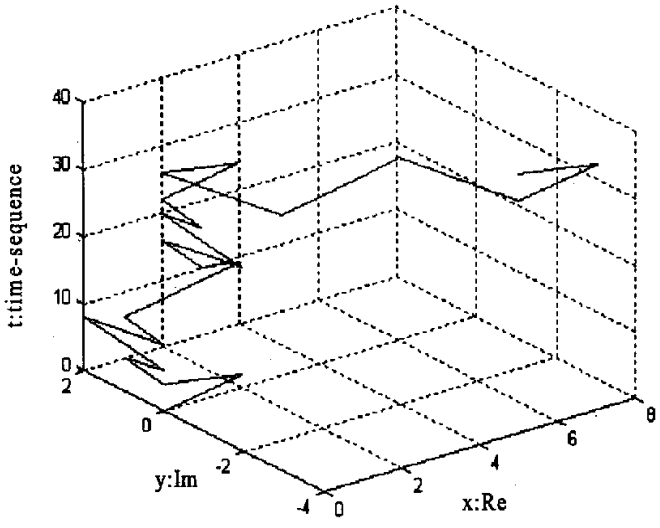


图 5.5　AIMV-3 的二级结构序列对应的 AU 随机游动曲线

图 5.3 —图 5.5 对应的三个随机数字序列分别如下：

$$P_n^1 = \{0,1,0,1,0,-i,-2i,-i,-2i,-i,1,-i,2-i,3-i,2-i,2-2i,2-i,2-2i,2-i,2-2i,$$
$$2-3i,2-2i,3-2i,2-2i,3-2i,2-2i,2-3i,2-4i,2-5i,1-5i,2-5i,3-5i,3-4i,$$
$$3-3i,3-2i,4-2i,3-2i,4-2i,3-2i\};$$

$$P_n^2 = \{0,1,0,-1,0,-i,0,i,0,-i,0,i,1+i,2+i,3+i,4+i,4,4-i,4,4+i,4,4+i,4+2i,5+2i,$$
$$4+2i,3+2i,4+2i,4+3i,4+4i,4+5i,3+5i,4+5i,5+5i,5+4i,5+3i,5+2i,6+2i,$$
$$5+2i,4+2i,5+2i\};$$

$$P_n^3 = \{0,1,2,1,0,i,0,i,2i,i,0,i,i+1,i+2,i+3,i+2,2+2i,2+i,2+2i,2+2i,i+2,2+2i,$$
$$3+2i,4+2i,3+2i,2+2i,2+i,2,2-i,3-i,4-i,5-i,5-2i,5-3i,5-4i,6-4i,$$
$$7-4i,6-4i,5-4i\},$$

因为在复平面内随机数字序列的模和相位比较突出反映其特征，所以我们在这里引进随机数字序列模和相位的均值的定义如下：

$$\overline{r^m} = \frac{1}{L}\sum_{n=1}^{L} r_n^m, \quad \overline{\varphi^m} = \frac{1}{L}\sum_{n=1}^{L} \varphi_n^m \tag{5.2}$$

假设 k 和 t 是两个物种，$\overline{r^m}(k), \overline{\varphi^m}(k)$ 和 $\overline{r^m}(t), \overline{\varphi^m}(t)$ 分别是其数值特征．利用等式 (5.2) 计算了 9 种病毒的 RNA 二级结构序列对应的三种随机数字序列的模和相位的均值见表 5.1 、表 5.2 。

表 5.1: 图 5.1 种 9 种病毒的二级结构序列对应的随机数字序列模的均值

	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
$\overline{r^1}$	2.8022	3.7803	2.5647	4.4746	2.7094	3.8878	2.7858	3.2509	2.5844
$\overline{r^2}$	3.4816	1.5282	2.5325	4.1438	3.2465	2.1424	3.0581	1.9414	2.1152
$\overline{r^3}$	2.9029	7.2121	5.8811	4.9118	3.1640	2.1839	6.5903	2.6325	2.5901

表 5.2: 图 5.1 种 9 种病毒的二级结构序列对应的随机数字序列相位的均值

	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
$\overline{\varphi^1}$	-0.8027	-1.0008	-0.3198	-0.6973	-0.5070	-1.0369	-0.7770	-1.0189	-0.7116
$\overline{\varphi^2}$	0.3935	-3.3735e-004	2.0834	0.3692	1.9756	1.6704	2.4204	2.4998e-004	0.4083
$\overline{\varphi^3}$	0.3833	1.2493	1.3837	-0.6780	-1.0223	1.2240	-0.7677	1.7647	0.9796

我们考虑三个模均值和三个相位均值来构造一个 6-D 向量：$(\overline{r^1}, \overline{r^2}, \overline{r^3}, \overline{\varphi^1}, \overline{\varphi^2}, \overline{\varphi^3})$. 假定如果在 6-D 空间中两个向量指向同一个方向，那么由向量表示的两个 RNA 二级结构序列是相似的. 这些向量的相似度可以通过计算向量终点的欧式距离来计算出. 两个向量之间的距离 d_{kt} 有如下等式给出：

$$d_{kt} = \sqrt{\Sigma_{m=1}^{3}[\overline{r^m}(k) - \overline{r^m}(t)]^2 + \Sigma_{m=1}^{3}[\overline{\varphi^m}(k) - \overline{\varphi^m}(t)]^2} \tag{5.3}$$

如果两个向量的欧式距离越小，它们对应的 RNA 二级结构序列就越相似，反之，相似性就小. 换句话说，进化相近的物种间的距离较小，异类进化的物种间的距离就较大. 我们计算了这些向量的欧式距离值并列在表 5.3 中.

表 5.3: 图 5.1 中 9 种病毒的序列的相似形表 (模和相位均值的 6 维向量的欧式距离矩阵)

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	4.9281	3.7306	2.8999	2.1675	2.4251	4.3830	2.1810	1.5416
CiLRV-3		0	3.0134	4.0691	5.4585	5.3353	3.7006	4.6572	4.8437
TSV-3			0	3.8100	3.7080	4.0351	2.4041	4.0434	3.7587
CVV-3				0	3.1163	4.1494	3.3272	4.2131	3.9784
APMV-3					0	2.9980	3.4809	3.7699	2.8514
LRMV-3						0	5.1055	1.9316	1.9034
PDV-3							0	5.4276	4.9028
EMV-3								0	1.1636
AVII									0

通过观察表 5.3，我们找到最相似的病毒对是 EMV-3 与 AVII，AIMV-3 与 AVII，LRMV-3 与 AVII，LRMV-3 与 EMV-3，AIMV-3 与 APMV-3，AIMV-3 与 EMV-3. 这结

果与实际结构比较吻合. 这些种类的分类说明他们的编码序列的数量足够大的, 可以根据表 5.3 所列的矩阵大致表现出来. 换句话说随着不同种类编码序列数量的持续增长, 可以通过距离矩阵进行聚类分析.

5.3　RNA 二级结构的 6-D 表示

同上节的表示方法一样, 用 A′, U′, G′, C′ 分别代表基对 A–U 和基对 G–C 中的 A, U, G, C, 这样我们就将二级结构转化为基本序列. 例如 LRMV-3 的子结构 (见图 5.6) 对应的特征序列为 CC′U′C′C′AAAG′G′A′G′U(从 5′ 端到 3′ 端) .

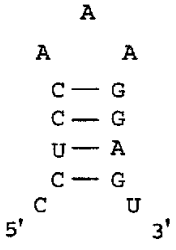


图 5.6　LRMV-3 的子结构

根据组成 RNA 的四种核苷酸基的化学结构和化学性质可将其分类, 嘌呤 (A, G)/ 嘧啶 (C, T) ; 酮基 (A, C)/ 氨基 (G, T) ; 弱氢键 (A, U)/ 强氢键 (C, G) , 以及张春霆等提出的 DNA 序列的三维图形表示 (Z 曲线)[49] , 我们定义如下的函数: 设 RNA 二级结构的特征序列长为 N , 用 $a_{n1}, g_{n1}, c_{n1}, u_{n1}, a_{n2}, g_{n2}, c_{n2}, u_{n2}$, 分别表示为 1 到 n 这个子序列中碱基 A, U, G, C 和 A′, U′, G′, C′ 所出现的次数,

$$\begin{cases} x_{ni} = (a_{ni} + g_{ni}) - (u_{ni} + c_{ni}) \\ y_{ni} = (a_{ni} + c_{ni}) - (u_{ni} + g_{ni}) \\ z_{ni} = (a_{ni} + u_{ni}) - (c_{ni} + g_{ni}) \end{cases} \qquad x_{ni}, y_{ni}, z_{ni} \in \mathrm{R}, \quad i = 1, 2. \tag{5.4}$$

我们令

$$v_{n1} = x_{n1}, \quad v_{n2} = y_{n1}, \quad v_{n3} = z_{n1}, \quad v_{n4} = x_{n2}, \quad v_{n5} = y_{n2}, \quad v_{n6} = z_{n2},$$
$$P_n = (v_{n1}, v_{n2}, v_{n3}, v_{n4}, v_{n5}, v_{n6}),$$

假设 $n = 0$ 时, $a_{ni} = g_{ni} = c_{ni} = u_{ni} = 0, i = 1, 2$, 当 n 从 0 到 N 时, 在 6-D 空间里便得到了 N+1 个点, 例如 LRMV-3 的子结构 (图 5.6) 的特征序列 CC′U′ ⋯ A′G′U 对应的点集: $\mathrm{P}_1 = (-1, 1, -1, 0, 0, 0)$, $\mathrm{P}_2 = (-1, 1, -1, -1, 1, -1)$, $\mathrm{P}_3 = (-1, 1, -1, -2, 0, 0)$, ⋯ , $\mathrm{P}_{11} = (2, 4, 2, -1, 1, -3)$, $\mathrm{P}_{12} = (2, 4, 2, 0, 0, -4)$, $\mathrm{P}_{13} = (1, 3, 3, 0, 0, -4)$. 这样我们将一个 RNA 二级结构的特征序列转化为 6-D 空间里的一个点集, 称之为特征点集, 依次连接特征点集中各点所得曲线我们称之为特征曲线, 但由于维数高我们不能画出直观图. 首先利用这些点集可以构造矩阵: Q 矩阵 [120](矩阵元素 $q_{i,j}$ 为顶点 i 和 j 之间的欧式距离和顶点 i 和

j 之间的边的几何距离之和的商），这样每一个 RNA 二级结构对应一个 Q 矩阵. 例如，表 5.4 列出了 LRMV-3 的子结构（见图 5.6 ）对应的特征序列为 $CC'U'C'C'AAAG'G'A'G'U$ 的 Q 矩阵.

表 5.4: LRMV-3 的子结构对应的特征序列为 $CC'U'C'C'AAAG'G'A'G'$ 的 Q 矩阵

base	C	C′	U′	C′	C′	A	A	A	G′	G′	A′	G′
C	0	1.0000	0.5774	0.6383	0.7071	0.6000	0.5774	0.5890	0.4895	0.4398	0.3559	0.3442
C′		0	1.0000	0.5774	0.6383	0.5401	0.5538	0.5932	0.4880	0.4449	0.3572	0.3559
U′			0	1.0000	1.0000	0.7454	0.7071	0.7211	0.5932	0.5408	0.4449	0.4398
C′				0	1.0000	0.7071	0.7454	0.7906	0.6429	0.5932	0.4880	0.4895
C′					0	1.0000	1.0000	1.0000	0.7906	0.7211	0.5932	0.5890
A						0	1.0000	1.0000	0.7454	0.7071	0.5538	0.5774
A							0	1.0000	0.7071	0.7454	0.5401	0.6000
A								0	1.0000	1.0000	0.6383	0.7071
G′									0	1.0000	0.5774	0.6383
G′										0	1.0000	0.5774
A′											0	1.0000
G′												0

其次利用这些点集的坐标，根据文献 [18] 在 6-D 坐标系上可以定义表示特征点集的 6 个分量 x, y, z, k, l, m 坐标值的均值即：

$$\mu_x = \sum x_i/\mathrm{N},\ \mu_y = \sum y_i/\mathrm{N}, \mu_z = \sum z_i/\mathrm{N},\ \mu_k = \sum k_i/\mathrm{N},\ \mu_l = \sum l_i/\mathrm{N},\ \mu_m = \sum m_i/\mathrm{N}$$

由此，我们称 6-D 向量：

$$\mathrm{P} = (\mu_x, \mu_y, \mu_z, \mu_k, \mu_l, \mu_m,)$$

为 RNA 二级结构的特征向量表示，其长度：

$$g_{\mathrm{R}} = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2 + \mu_k^2 + \mu_l^2 + \mu_m^2}$$

定义为图半径（特征曲线的半径）.

例如 LRMV-3 的子结构 (图 5.6) 对应的特征向量表示 $\mathrm{P} = (0.5385, 2.5385, 0.6923, -2.1538, 0.9231, -2.1538)$，图半径 $g_{\mathrm{R}} = 4.1645$.

因为矩阵不变量可以表征序列或结构不变性 [7−18]，所以我们在这里取 Q 矩阵的不变量作为 RNA 二级结构的不变量来比较它们之间的相似性，而矩阵不变量有很多种，如矩阵的最大特征值、矩阵的行列式值、矩阵的迹和矩阵的行或列和，等等. 在这里，我们利用 Q 矩阵的前 8 个正规化最大特征值来比较 RNA 二级结构的相似性. 以图 5.1 中 9 种病毒为例，计算了 9 种病毒的 RNA 二级结构序列对应的 Q 矩阵的前 8 个最大特征值以及相似性距离见表 5.5 、表 5.6 .

表 5.5: 9 种病毒的 RNA 序列对应的 Q 矩阵的前 8 个最大特征值

	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
λ_1	0.8222	0.6726	1.0108	0.7077	0.3231	0.7582	0.3653	0.6107	0.9967
λ_2	0.8982	0.9433	1.0427	0.9716	0.3386	0.8642	0.8044	0.9730	1.1843
λ_3	0.9322	1.0293	1.4003	1.3008	0.4996	1.1762	0.8911	1.3268	1.5081
λ_4	1.6074	1.2369	1.6863	1.5486	0.8497	1.6808	0.9807	1.5642	1.7850
λ_5	2.2338	1.5288	2.3590	1.9985	2.2661	1.9628	2.4634	2.2439	1.8332
λ_6	2.7188	4.8586	2.6626	4.8826	3.8611	5.3897	3.6109	4.9429	5.3314
λ_7	5.5183	7.2612	5.9263	6.2473	5.9379	7.8055	5.2977	6.2954	7.5175
λ_8	12.4814	18.0837	15.0726	17.4540	14.8231	15.6387	14.7009	14.2595	15.4574

表 5.6: 图 5.1 种 9 种病毒的序列的相似形表

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.0546	0.0311	0.0430	0.0573	0.0463	0.0494	0.0479	0.0486
CiLRV-3		0	0.0701	0.0313	0.0301	0.0563	0.0342	0.0680	0.0649
TSV-3			0	0.0504	0.0768	0.0595	0.0651	0.0508	0.0549
CVV-3				0	0.0442	0.0472	0.0343	0.0467	0.0503
APMV-3					0	0.0669	0.0233	0.0767	0.0782
LRMV-3						0	0.0681	0.0259	0.0160
PDV-3							0	0.0710	0.0754
EMV-3								0	0.0189
AVII									0

表 5.7: 9 种病毒的 RNA 二级结构序列对应的特征向量

access No.	eigenvector
AlMV-3	P=(1.7692, 2.8974, 2.4359, -1.0513, 0.5385, -0.0256)
CiLRV-3	P=(-1.9216, 0.5882, 0.8235, -2.6667, -0.2745, 7.0980)
TSV-3	P=(0.2857, -1.8367, 1.2653, -0.4694, 0.3878, 5.0816)
CVV-3	P=(2.6538, 3.4615, 3.6154, -2.8077, 0.6538, -3.3077)
APMV-3	P=(2.2927, -1.2195, 1.2683, -1.0000, 2.5610, -2.7561)
LRMV-3	P=(-0.9423, -1.9808, 0.1346, -3.4423, 0.1346, 1.8269)
PDV-3	P=(1.2195, -0.3902, 2.3415, -2.1220, 1.7317, 5.1463)
EMV-3	P=(-0.8367, 1.2857, -0.6327, -2.8163, -0.3673, 2.4490)
AVII	P=(0.4340, 0.3962, 1.0000, -1.9623, 0.4906, 1.9245)

接着利用上节特征向量的定义计算了 9 种病毒的 RNA 二级结构序列对应的特征向量见表 5.7 .

因为特征向量或图半径包含 RNA 序列的结构信息，因此要比较两个 RNA 二级结构序列就比较其对应的特征向量及图半径之间的距离即可，这里我们利用特征向量之间的距离来比较其相似性. 定义两个特征向量之间的距离为：

$$D_c = [(\mu_x - \mu_x^{'})^2 + (\mu_y - \mu_y^{'})^2 + (\mu_z - \mu_z^{'})^2 + (\mu_k - \mu_k^{'})^2 + (\mu_l - \mu_l^{'})^2 + (\mu_m - \mu_m^{'})^2]^{\frac{1}{2}} \quad (5.5)$$

如果两个 RNA 序列的 D_c 值越小，说明它们之间越相似，否则相似性就不大. 根据 D_c 的定义计算了 9 种病毒的 RNA 序列的相似性见表 5.8 .

表 5.8: 图 5.1 种 9 种病毒的序列的相似形表

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	8.6931	7.2407	4.0450	5.4895	6.7644	6.3175	5.3725	3.8387
CiLRV-3		0	4.5033	12.0890	11.3643	6.0487	12.9373	5.0426	5.7848
TSV-3			0	10.7294	8.4168	4.7226	10.6428	5.2555	4.1575
CVV-3				0	5.8958	9.0390	4.8506	8.3154	7.0201
APMV-3					0	6.7221	3.2636	7.6583	5.7670
LRMV-3						0	8.0618	3.5072	3.2588
PDV-3							0	8.8577	7.3890
EMV-3								0	2.6100
AVII									0

通过观察表 5.6 , 表 5.8 可得，EMV-3 与 AVII 相似，LRMV-3 与 AVII 相似，LRMV-3 与 EMV-3 ， APMV-3 与 PDV-3 相似，这结果与实际结构基本吻合. 在表 5.6 中 CiLRV-3 与 CVV-3 ， CiLRV-3 与 APMV-3 之间的距离比较小说明也比较相似，可是这与实际结构不太符合，造成这种结果的原因可能是由结构构造距离矩阵时可能会丢失一些信息，不过这种误差并不影响相似性的比较，我们基本能正确地判断它们之间的相似性. 比较这两个相似性表，表 5.8 的结果比较好，且更符合实际结构.

5.4 小结

本章我们给出了 RNA 二级结构的三种表示方法和两种度量 RNA 二级结构序列相似性的方法，与过去仅限于字符结构的比较方法相比，我们的比较方法既兼顾了 RNA 二级结构的组成字符串，也考虑了 RNA 的化学结构和化学性质. 方法简单并不受是否有假结的限制，能比较准确的判断 RNA 二级结构之间的相似性. 这种方法的优点： (1) 使数据具有了可视性，即直接识别 RNA 二级结构的相似与不同； (2) 能够识别较大的 RNA 二级结构的相似与不同； (3) 结构不变量容易计算且直接应用到 RNA 二级结构比较上. (4) 具有统计学意义. 总之，比起用动态规划算法来比较 RNA 二级结构速度快、简单易行，结果较好，将对评价同源性有很大的帮助. 对预测 RNA 的二级结构中有应用价值.

6 RNA 二级结构序列映射到“波谱线”和“Z型曲线”表示

本章给出了把 RNA 二级结构的特征序列映射为“波谱线”和“Z型曲线”表示的三个递归公式，利用这三个递归公式给出了 RNA 二级结构序列的 1-D、2-D 和 3-D 图形表示，进一步利用 1-D 图形表示给出了关于 RNA 二级结构序列频谱分析的方法。

6.1 引言

由于复杂多维数据的低维图形表示方法非常便于获取隐藏信息，且运用快捷，有直观趣味，所以在生物序列比较中广泛应用。特别是，基于 DNA 序列的图形表示法以及其相应的各种相关距离矩阵 (如， E 矩阵、 L/L 矩阵、 M/M 矩阵、 D/D 矩阵和 Q 矩阵等) 把 DNA 序列的完全图形表示映射到相应的数字表示正在逐步推广 [7, 8, 10, 12, 18-31, 41, 57]。在可供选择的生物序列的图形表示中，能引起特殊兴趣的是具有以下特征的图示：被限制在一个恰当定义的平面内 [95, 97–99, 109–113]；或者是已知空间的三维盒子 [111]，呈二维或三维 Z 型曲线形式 [109–111]；或者在某一轴线上的有限间隔上的以“波谱状”曲线形式出现 [94, 95]。任何连接起来的曲线都可以用图形表示出来，并且利用这种图形，可以计算许多与物理化学和生物特性依次相关联的，有价值的数学不变量，这一点是非常重要。应该补充的一点是，我们可以通过记录单独核酸基的相继出现，或者通过 DNA 序列的四维空间表示 [13]，得到一个 DNA 序列数值的特性描述，而不借助文献中已讨论过的，考虑碱基对出现频率的 DNA 序列图形表示方法。这些通过手工绘制或者是电脑合成的、简化了的，更具视觉吸引力的已获得的图形表示必须保留尽可能多的原始信息，最好是全部。无论是在简单的多项式中，还是更复杂的模拟神经的网络模型，该类图形表示的一个功能就是，它可以作为数学模型的输入。在这章中，我们把焦点集中在将 RNA 二级结构序列编译成具有可视性的简单的算法上。

核糖核酸 RNA(Ribonucleic Acid) 分子是一种生物聚合物，它既类似于蛋白质也类似于 DNA，但又不同于它们。RNA 是由四种碱基 A, U, G 和 C 构成的聚核苷酸链。其中，A, U, G 和 C 分别代表腺嘌呤，尿嘧啶，鸟嘌呤和胞嘧啶。这四种碱基相互之间通过形成基对产生能量使得 RNA 自身折叠而形成 RNA 特有的单链双螺旋结构，形似于 DNA 的双链结构，从而形成 RNA 空间三级结构。

RNA 分子有两类结构信息：其一是 RNA 一级结构，它是由碱基 A, U, G 和 C 组成的单链；其二是三级结构，它是碱基序列自身折叠形成的双螺旋结构。RNA 的结构特性对于了解它的功能是非常重要的。RNA 二级结构是三级结构的严格子集，它在一级结构和三级结构之间起着重要的作用，即起着由 RNA 一级结构推测三级结构的桥梁作用。因

此，RNA 二级结构有着非常重要的研究价值，而且它可以应用于诸如 tRNA 和蛋白质的相互作用、mRNA 的稳定化处理。RNA 二级结构的表示方法对于 RNA 二级结构研究有着极其重要的价值。

目前已有很多信号处理方法被用于核酸和蛋白质序列的分析，并成为生物信息学的重要内容。谱分析作为信号处理的常用方法，近年来也被用于 DNA 序列的分析，谱分析用于 DNA 序列分析有自身的优势，可以将原始数据中局部的、潜在的周期性信息变得清晰和可观察 [106]。

在 DNA 序列的分析中，一类主要的分析对象是由 A, C, G, T 四种字符组成的符号序列，本质上这是生物分子的一种符号表示，可以把它看作一种离散信号，并采用离散傅立叶变换做频谱分析 [107,108]。在本章我们同样可以把 RNA 二级结构序列看作一种离散信号，并采用离散傅立叶变换 (简写为 DFT) 取得序列的频谱图来表现 RNA 二级结构序列的频域特征。

6.2 算法

Zupan 和 Randic 提出的算法 [94] 是结合了唯一且通过等长度的线连接在一起的，一维、二维和三维表示的同步运算。此外，二维和三维图形表示是可逆的，即可以只通过两个或三个坐标，来复制整个初始单元序列。如果除了重视结束点之外，供应附加数字，甚至一维的图示也满足这种条件。他们提出的 2-D “ Z 型曲线” 表示是中心在原点的正方形的四个角上分别用四个坐标 (-1,-1), (1,-1), (1,1) 和 (-1,1) 来表示四种核苷酸 A, T, G 和 C。他们从正方形的中心位置 (0,0) 开始，按顺序向准备编译的第一个核苷酸移动，接着，继续从这个点依次向第二个核苷酸移动，如此类推。在这里展示以人类第一个外显子 β– 球蛋白的 DNA 序列中的前三个字符 ATG 为例说明此算法最开始的三个步骤，见图 6.1。

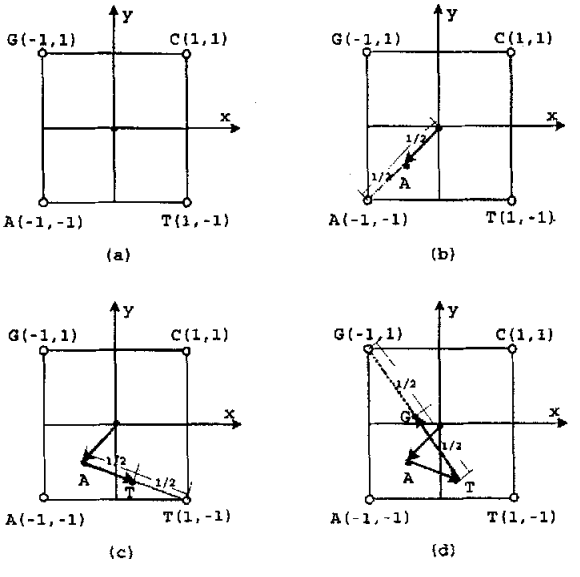


图 6.1　(a): 序列的初始点为坐标原点；(b): 序列 ATG 的第一个碱基 A 在原点到角 A 连线的中点上其坐标为 (-0.5,-0.5); (c): 第二个碱基 T 在碱基 A 和角 T 连线的中点上其坐标为 (+0.25,-0.75); (d): 第三个碱基 T 在碱基 T 和角 G 连线的中点线其坐标为 (-0.375,+0.125)

我们在这一章里结合 Zupan 和 Randic 提出的算法 [94]，提出了 RNA 二级结构特征序列 [101−105] 的一维、二维和三维图形表示．我们以第五章出示的 9 种病毒的 RNA 二级结构特征序列为例来说明此图形表示方法．见表 6.1 .

表 6.1: 9 个不同病毒 RNA 二级结构特征序列

Species	Coding sequence
AlMV-3	AUGCU′C′A′U′G′C′A′AAACU′G′C′A′U′G′A′AUGCC′C′C′UAAG′G′G′AUGC
CiLRV-3	AUGCC′U′A′U′A′U′U′U′C′U′C′UCCUG′A′G′A′A′A′A′U′A′U′A′G′AUGCC′U′C′C′AAAG′G′A′G′AUGC
TSV-3	GUGCC′A′G′U′A′G′U′A′U′A′UAAU′A′U′A′C′U′A′C′U′G′AUGCC′U′C′C′U′UUAUA′G′G′A′G′AUGC
CVV-3	AUGCC′C′A′AAC′U′C′U′C′U′C′U′CAUG′G′A′G′A′G′A′G′AAU′G′G′AUGCC′U′C′C′GAAG′G′A′G′AUGC
APMV-3	AUGCC′C′A′C′A′A′C′GUGAAG′U′U′G′U′G′G′AUGCC′C′C′GUUAG′G′G′AAGC
LRMV-3	GUUCC′U′A′U′U′C′U′C′U′C′U′C′UCAGG′A′G′A′G′GA′G′A′A′U′A′G′AUGCC′UC′C′AAAG′G′A′G′UCGC
PDV-3	AUGCC′C′U′C′A′C′C′GUAAG′G′U′G′A′G′G′AUGCC′C′C′U′UUAA′G′G′G′*AUG*C
EMV-3	CC′U′A′A′U′UC′U′C′U′C′U′C′U′CAGG′G′A′G′A′G′A′G′A′U′U′A′G′AUGCC′U′*C′CAAGG′A′G′AUGC*
AVII	AUGCC′U′A′A′UA′C′U′C′U′C′U′C′U′CAGG′G′A′G′A′G′A′G′U′U′U′A′G′AUGCCU′C′C′AAAG′G′A′G′AUGC

6.2.1　1-D 表示

我们先置 A, U, G 和 C 于 $+x$ 轴，A', U', G' 和 C' 于 -x 轴．即

$$\mathrm{A} \to x = +1,\ \ \mathrm{U} \to x = +2,\ \ \mathrm{G} \to x = +3,\ \ \mathrm{C} \to x = +4,$$
$$\mathrm{A}' \to x = -1,\ \ \mathrm{U}' \to x = -2,\ \ \mathrm{G}' \to x = -3,\ \ \mathrm{C}' \to x = -4.$$

当然，点的任何其他的分配和排列都是可以的．设 $\mathrm{S} = s_1 s_2 \cdots$ 为任意的 RNA 二级结构特征序列，如果，N 个单位长的 RNA 二级结构特征序列的第 i 个单位被记作 s_i，我们就可以得到获得 1-D 的“波谱线”图形表示 $\mathrm{R}(x_i)$ 的递推公式：

$$\mathrm{R}(x_{i+1}) = \frac{\mathrm{R}(x_i) + \mathrm{S}(x_{s_{i+1}})}{d} \tag{6.1}$$

其中，d 是非零实数．规定，$\mathrm{R}(x_0) = 0$.

例如，取 $d=2$，由递推公式 (6.1) 计算得出的病毒 EMV-3 RNA 二级结构特征序列的子序列 C′U′C′CAAGG′A′G′ 对应的坐标点集：$\mathrm{R}(x)=\{0,-2.0000,-2.0000,-3.0000,0.5000,0.7500,0.8750,1.9375,-0.5313,-0.7656,-1.8828\}$，以及图形表示见图 6.2 所示。

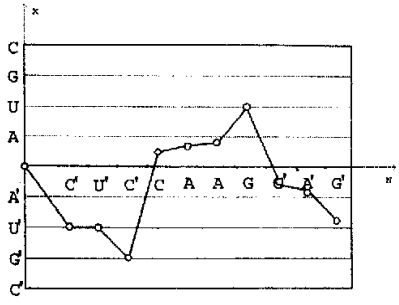


图 6.2　EMV-3 的 RNA 二级结构特征序列的子序列对应的 1-D 表示

显然，用这种方法确保了 1-D 表示的完全可逆性。当 $d\geq 2$ 时“波谱线”在 $x=\pm 4$ 之间游动。例如，无论 RNA 二级结构特征序列有多长，都可以通过图示的最后一点的坐标来重建 RNA 二级结构序列。利用递推公式 (6.1) 可以获得 RNA 二级结构序列的“波谱线”图形表示。图 6.3 给出了表 6.1 的 RNA 二级结构特征序列的 1-D “波谱线”图形表示。

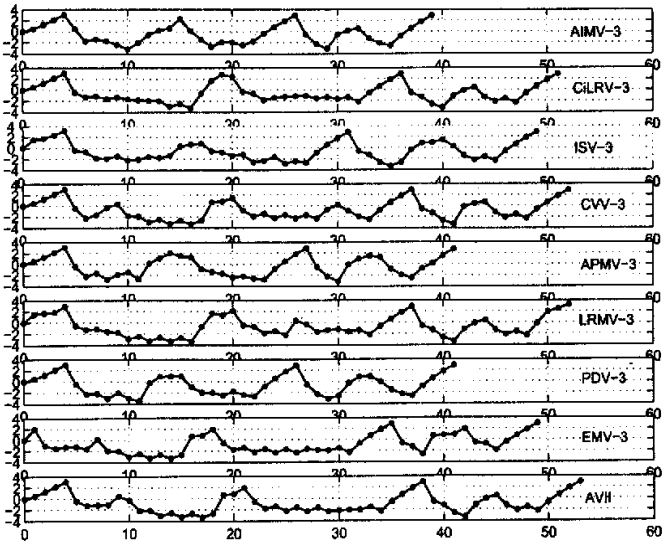


图 6.3　对应于表 6.1 的 9 个 RNA 二级结构特征序列的 1-D 表示

在图 6.3 中，各 RNA 二级结构特征序列的图形表示中没有圈形成。这些图形表示可以表示出 RNA 二级结构序列的连接顺序，序列信息也得到完整的表达。从图 6.3 可以看出，APMV-3 和 PDV-3 相似，EMV-3 和 AVII 相似，LRMV-3 和 AVII 相似。

6.2.2　2-D 表示

我们令

$$\mathrm{A}\to(-1,0),\quad \mathrm{U}\to(1,0),\quad \mathrm{G}\to(0,-1),\quad \mathrm{C}\to(0,1),$$
$$\mathrm{A}'\to(-1,-1),\quad \mathrm{U}'\to(1,1),\quad \mathrm{G}'\to(1,-1),\quad \mathrm{C}'\to(-1,1).$$

同样，点的任何其他的分配和排列都是可以的．我们可以得到获得 2-D 的“Z 型曲线”表示 $\mathrm{R}(x_i,y_i)$ 的递推公式：

$$\mathrm{R}(x_{i+1},y_{i+1})=\frac{\mathrm{R}(x_i,y_i)+\mathrm{S}(x_{s_{i+1}},y_{s_{i+1}})}{d} \tag{6.2}$$

其中，d 是非零实数．规定，$\mathrm{R}(x_0,y_0)=0$．

例如，取 $d=2$，由公式 (6.2) 计算得出的病毒 EMV-3 RNA 二级结构特征序列的子序列 C′U′C′CAAGG′A′G′ 对应的坐标点集：$\mathrm{R}(x,y)=\{(0,\ 0),\ (-0.5000,\ 0.5000),\ (0.2500,\ 0.7500),\ (-0.3750,\ 0.8750),\ (-0.1875,\ 0.9375),\ (-0.5938,0.4688),\ (-0.7969,\ 0.2344),\ (-0.3984,-0.3828),\ (0.3008,\ -0.6914),\ (-0.3496,\ -0.8457),\ (0.3252,\ -0.9229)\}$，以及 2-D“Z 型曲线”表示见图 6.4 所示．

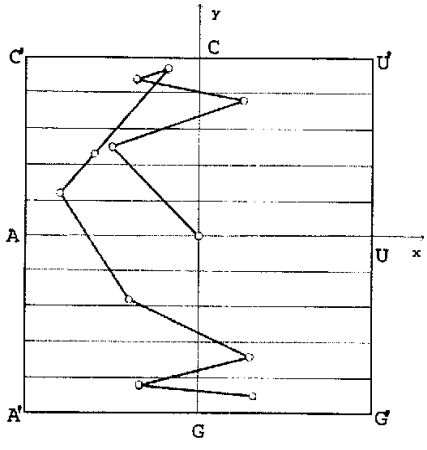


图 6.4　EMV-3 的 RNA 二级结构特征序列的子序列对应的 2-D 表示

6.2.3　3-D 表示

我们令

$$\mathrm{A}\to(-1,-1,1),\quad \mathrm{U}\to(1,-1,-1),\quad \mathrm{G}\to(1,1,1),\quad \mathrm{C}\to(-1,1,-1),$$
$$\mathrm{A}'\to(-1,-1-1),\quad \mathrm{U}'\to(1,-1,1),\quad \mathrm{G}'\to(1,1,-1),\quad \mathrm{C}'\to(-1,1,1).$$

同样，点的任何其他的分配和排列都是可以的．我们可以得到获得 3-D 的“Z 型曲线”表示 $\mathrm{R}(x_i,y_i,z_i)$ 的递推公式：

$$\mathrm{R}(x_{i+1},y_{i+1},z_{i+1})=\frac{\mathrm{R}(x_i,y_i,z_i)+\mathrm{S}(x_{s_{i+1}},y_{s_{i+1}},z_{s_{i+1}})}{d} \tag{6.3}$$

其中， d 是非零实数。规定， $\mathrm{R}(x_0, y_0, z_0) = 0$ 。

例如，取 $d = 2$，由公式 (6.3) 计算得出的病毒 EMV-3 RNA 二级结构特征序列的子序列 C′U′C′CAAGG′A′G′ 对应的坐标点集：$\mathrm{R}(x, y, z) = \{(0, 0, 0),\ (-0.5000, 0.5000, 0.5000),\ (0.2500, -0.2500, 0.7500),\ (-0.3750, 0.3750, 0.8750),\ (-0.6875, 0.6875, -0.0625),\ (-0.8438, -0.1563, 0.4688),\ (-0.9219, -0.5781, 0.7344),\ (0.0391, 0.2109, 0.8672),\ (0.5195, 0.6055, -0.0664),\ (-0.2402, -0.1973, -0.5332),\ (0.3799, 0.4014, -0.7666)\}$ ，以及 3-D “ Z 型曲线” 表示见图 6.5 所示。

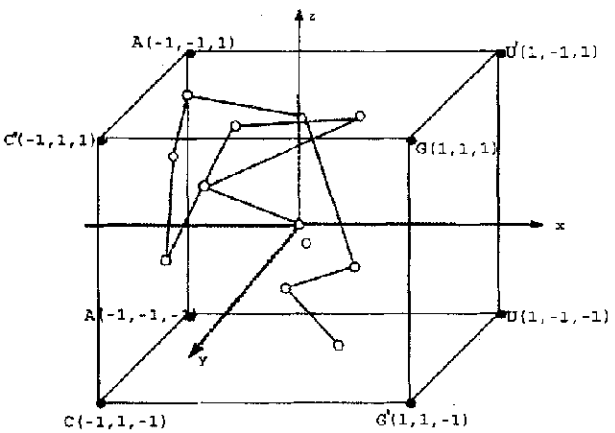


图 6.5 EMV-3 的 RNA 二级结构特征序列的子序列对应的 3-D 表示

我们注意到 2-D 和 3-D 两种表示方法不论序列有多长，他们对应的点列都在某一个正方形或立方体内。即此方法可以把任意长度序列的 “ Z 型曲线” 表示压缩在某一适当边长的正方形或立方体内。且能显示组成序列的核酸分布规律。例如，在 3-D 表示下的分布：序列中的所有核苷酸 A 都出现在一顶点为 $\mathrm{A}(-1,-1,1)$ 的 $\frac{1}{8}$ 小正方体内。由此我们大致能判断每个核苷酸的含量。显然，每一种分配下，表示方法是唯一且可逆的。

6.3 RNA 二级结构序列的频谱分析方法

我们根据 RNA 二级结构序列的 1-D 表示的递推公式 (6.1) 把 RNA 二级结构序列的特征序列映射成数值序列 $x(n)$，$n = 1, 2, \cdots, N$(N 是特征序列的长度)，$x(n)$ 表示特征序列在位置 n 的相应符号的映射值。

长度为 N 的序列 $x(n)$ 的离散傅立叶变换的基本形式 [140] 为：

$$\mathrm{DFT}[x(n)] = \mathrm{X}(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \qquad k = 0, 1, 2, \cdots, N-1 \tag{6.4}$$

$\mathrm{X}(k)$ 长度仍为 N ，提供了 $x(n)$ 在频率 k(周期 N/k) 处的频域信息。

我们以在表 6.1 中出示的 9 个不同病毒的 RNA 二级结构序列为分析对象。图 6.6 是对序列 $x(n)$ 做 DFT 得到的频谱图，横轴代表周期。

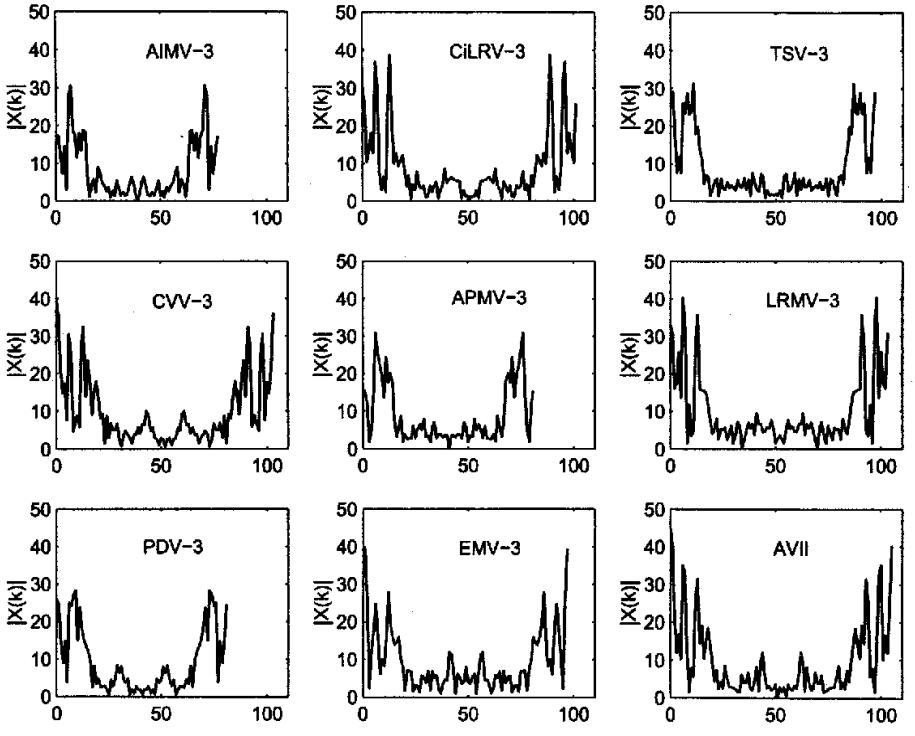


图 6.6　对应于表 6.1 的 9 个 RNA 二级结构特征序列的频谱图

由图 6.6 看出每个图在某些区域上都有明显的周期表现，例如，从 CVV-3 的频谱图可以看到在区域 10–18 和 24–29 有明显的周期变化。另外，序列的频谱分布的十分清晰，APMV-3 和 AIMV-3 对应的序列在频谱图中最高谱峰都在 3 附近且大小相近，还有 EMV-3 和 AVII 对应的序列在频谱图中最高谱峰都在 1 附近且大小相近，从频谱图也能判断，APMV-3 和 AIMV-3，EMV-3 和 AVII 相似，LRMV-3 与 EMV-3 和 AVII 也比较相似，CVV-3 和 AVII 也比较相似，这和文献 [101] 的结果以及时域表示 (图 6.3 的直观表示) 基本一致。

在关于核苷酸序列的频谱分析中，基于离散傅立叶变换的方法能够较好的反应序列的频谱特征。那么这种方法能否提供序列的其它频率特征是有待进一步分析。实际上这个方法依然考虑的是序列的局部特征，因此，如何从它们之间的相互作用，即系统的、综合的角度而不是局部的、静止的观点来分析序列也是一个更有效的方法。

6.4 小结

在本章的工作中，在直角坐标系下，通过一个简单的递推公式将 RNA 二级结构序列转换成 1-D， 2-D 或 3-D 图形表示。首先在直观上也能判断序列之间的相似与不同。其次，根据这些表示的可逆性，仅通过最后一点的坐标就可以解码为原始序列。即他们可以获得等量的如同原始序列一样多的信息。

递归公式 (6.1)， (6.2) 和 (6.3) 还有一个优点是它完全能够分别转化由任何单位字符构成的序列 (如，由 20 个氨基酸构成的蛋白质序列) 为 1-D， 2-D 或 3-D "波谱线" 或 " Z 型曲线" 表示。

我们提出的方法主要是利用数学的方法对 RNA 二级结构序列的结构特征和相似性进行分析。事实上， RNA 二级结构序列还包含了丰富的生物学信息，如果利用生物学信息对模型进行优化，将会得到更多更好的信息。

另外，对于一个确定的 RNA 二级结构特征序列的 "波谱线" 曲线表示可以进一步被视作任意 "光谱束" 或 "数字信号"，因此适用各种光谱操作程序或信号处理操作程序。以此来研究序列更多的特征和挖掘生物序列隐藏的生物学意义。

生物信息学是当今生命科学和自然科学的重大前沿领域之一，同时也将是 21 世纪自然科学的核心领域和最具活力的领域之一。其中， RNA 二级结构是生物信息学的一个重要分支，对于给定 RNA 分子的二级结构的分析与预测确实还留给我们很大的空间去探究，用于比较 RNA 二级结构的各种表示方法无疑对这个问题的研究有很重要的价值，二级结构各种表示的可视化是应该进一步研究的问题。

7 蛋白质序列的图形表示及其应用

本章在 DNA 三联体密码子表示的基础上，在半复平面上给出了蛋白质序列的非退化的 2-D 图形表示，同时利用复向量的主要特征—模和相位，给出了蛋白质序列的一种数值刻划. 还有在 3-D 空间里，把 20 种氨基酸分别分配给正 12 面体的 20 个顶点，根据正 12 面体的对称性得到了 20 种氨基酸的 3-D 表示，进而得到了蛋白质序列的 3-D 图形表示，利用图的不变量比较了 9 种动物的神经元基因序列的相似性以及构建了一组细胞色素 C 蛋白质的序列进化树.

7.1 引言

阐述蛋白质的序列 – 结构 – 功能之间的关系是后基因组时代的一项重要任务，大量的蛋白质序列数据库的数学分析是生物学家的一大挑战. 近年来，通过实施基因组测序计划，得到了许多蛋白质序列. 在解读这些蛋白质序列的过程中，人们发现各种功能的蛋白质之间往往有着千丝万缕的联系，一般认为，蛋白质在一级结构上具有同源性，则表示它们在分子进化历程中有共同的始祖分子，如果只是确认蛋白质序列相似，而未确定它们在进化上的联系，则可笼统称之为序列相似性. 因此，每当获得一个新的蛋白质序列，人们总是希望通过相似性比较证明它与某些已知蛋白质序列相似，并进而分析相似蛋白质在功能上的联系. 显然，如果新序列与我们已知的某种蛋白质具有同源性，将会大大节省我们重新测定新序列功能的时间和精力，而生物分子数据的庞大，就使其显得更为重要了. 因此，在蛋白质结构与功能的分析中，序列相似性一直是研究的重点，其最终的目的，是阐明相似性的生物学意义 [133, 140] .

生物序列的相似性比较绝非简单机械的比较，而必然是多种多样的，同时还需要运用许多数学和统计学方法进行辅助分析与评判. 序列相似性比较最常用的是序列对比法. 然而比对算法的最大问题是罚分参数的选择带有主观性. 基于这种情况，使得很多人试图寻找其他的方法来比较生物序列. 近年来，一些 DNA 序列的图形表示已在文献中详述 [1–77] ，不仅形成了 DNA 序列的数值特性，还有蛋白组图形表示的数值特性 [77–93, 101–109, 111–119] . 另外，也可以从给定的图形表示中获得的原始序列的完全重建，同样方便了 DNA 或蛋白质序列不变量的建立作为一种 DNA 或蛋白质研究的工具. 跟 DNA 序列相比，蛋白质序列则没有太多类似的图形表示，除了最近蛋白质的高度压缩图形表示 [99, 109, 110, 112–117] 以及 Randic[111] 描述的一个基于 8 行 8 列 (8×8) 的 64 个密码子的蛋白质的文字图形以及蛋白质序列的 Z 型曲线的建立.

在分析蛋白质序列相似性或构建蛋白质序列进化树或结构进化树等问题中，蛋白质序列的特征描述直接影响序列进化树或结构进化树的正确与否；同时比较不同描述方法

对相似性分析结果的影响，可以帮助我们理解序列与结构或序列与功能之间的关系。

在这一章里，我们基于蛋白质序列的图形表示，给出了蛋白质序列的一些新的特征描述，并把它作为比较和分析蛋白序列的描述工具，正如同比较和分析 DNA 序列一样.

7.2 基于核苷酸三联体密码子上蛋白质序列的 2-D 图形表示

7.2.1 DNA 序列的 2-D 表示

我们在复平面直角坐标系中的第一、第四象限里构造一个嘌呤 - 嘧啶图，如图 7.1 所示：即嘌呤 A 和 G 在第一象限，嘧啶 T 和 C 在第四象限.

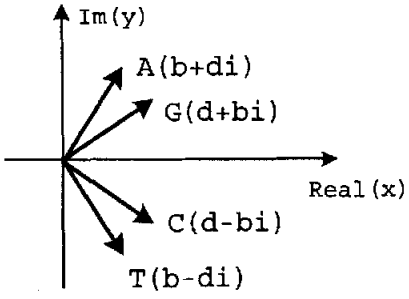


图 7.1 嘌呤 - 嘧啶图

用四个向量来表示如下：

$$(b+di) \longrightarrow \mathrm{A},\ \ (d+bi) \longrightarrow \mathrm{G},\ \ (b-di) \longrightarrow \mathrm{T},\ \ (d-bi) \longrightarrow \mathrm{C}.$$

或

$$re^{i\theta_1} \longrightarrow \mathrm{A},\ \ re^{i\theta_2} \longrightarrow \mathrm{G},\ \ re^{i\theta_3} \longrightarrow \mathrm{T},\ \ re^{i\theta_4} \longrightarrow \mathrm{C}.$$

其中，$r=\sqrt{b^2+d^2}$，$\theta_k=\arg\tan\frac{d}{b}$，$k=1,2,3,4$，其中 b, d 是非零正实数且 A 和 T 共轭，G 和 C 共轭，即 $\bar{\mathrm{A}}=\mathrm{T}$，$\bar{\mathrm{G}}=\mathrm{C}$，另外，A+T+C+G=2($b+d$)．这样我们可以把一个 DNA 序列转换成向量序列：$\vec{P_0},\vec{P_1},\vec{P_2},\cdots,\vec{P_n}$，或一个数字序列：

$$x(n)=x(0)+\sum_{j=1}^{n}y(j),\quad x(0)=0 \tag{7.1}$$

这里，$y(j)$ 满足如下条件：

$$y(j)=\begin{cases}(b+di) & \text{如果 } j=\mathrm{A},\\ (d+bi) & \text{如果 } j=\mathrm{G},\\ (b-di) & \text{如果 } j=\mathrm{T},\\ (d-bi) & \text{如果 } j=\mathrm{C}.\end{cases} \tag{7.2}$$

($j=0,1,2,\cdots,n$, 这里，n 是 DNA 序列的长度)

$$\vec{P_j} \longrightarrow (ba_j+dg_j+bt_j+dc_j)+(da_j+bg_j-dt_j-bc_j)\mathrm{i} \tag{7.3}$$

这里，a_j, g_j, t_j 和 c_j 分别为 1 到 j 这个子序列中碱基 A, G, T 和 C 出现的累积个数。我们定义 $a_0=g_0=t_0=c_0=0$，则我们有如下两个性质：

性质 1　对于给定的一个 DNA 序列就有唯一的 $x(n)$ 和它对应.

证明:　设在复平面里曲线 $x(n)$ 上的向量 $b_j+d_j\mathrm{i}$ 对应于 DNA 序列的第 j 个碱基，则有

$$(b_j+d_j\mathrm{i})=(ba_j+dg_j+bt_j+dc_j)+(da_j+bg_j-dt_j-bc_j)\mathrm{i} \tag{7.4}$$

或

$$\begin{cases} b_j=ba_j+dg_j+bt_j+dc_j \\ d_j=da_j+bg_j-dt_j-bc_j \end{cases} \tag{7.5}$$

如果 $b_j+d_j\mathrm{i}$ 同样可以表示成如下形式

$$\begin{cases} b_j=ba_j'+dg_j'+bt_j'+dc_j' \\ d_j=da_j'+bg_j'-dt_j'-bc_j' \end{cases} \tag{7.6}$$

由 (7.5) 和 (7.6) 我们有

$$\begin{cases} b(a_j-a_j')+d(g_j-g_j')+b(t_j-t_j')+d(c_j-c_j')=0 \\ d(a_j-a_j')+b(g_j-g_j')-d(t_j-t_j')-b(c_j-c_j')=0 \end{cases} \tag{7.7}$$

因为 b 和 d 是非零正实数，由 (7.7) 可得到

$$\begin{cases} a_j-a_j'+t_j-t_j'=0 \\ a_j-a_j'-t_j+t_j'=0 \\ g_j-g_j'+c_j-c_j'=0 \\ g_j-g_j'-c_j+c_j'=0 \end{cases} \tag{7.8}$$

进而得到 $a_j=a_j'$, $g_j=g_j'$, $t_j=t_j'$, $c_j=c_j'$.

在复平面上 2-D 曲线 $x(n)$ 上的任意给定点 $\vec{P}=b+d\mathrm{i}$，利用 $\vec{P}$ 的实部和虚部可以唯一确定从开始到点 $\vec{P}$ 的序列片断上 A, G, T 和 C 的出现个数的累积 a_p, g_p, t_p 和 c_p. 不断的应用曲线上的点，我们可以从 DNA 图恢复唯一的原始 DNA 序列.

性质 2　我们得到的 2-D 图形表示没有圈.

证明:　我们假设：(1) l 是形成一个圈的核苷酸个数；(2) 在一个圈中 A, G, T 和 C 出现的个数分别为 a', g', t' 和 c'. 因此，$a'+g'+t'+c'=l$. 因为，a'A, g'G, t'T 和 c'C 构成一个圈，所以有如下等式成立：$a'(b+d\mathrm{i})+g'(d+b\mathrm{i})+t'(b-d\mathrm{i})+c'(d-b\mathrm{i})=(0,0)$，即

$$\begin{cases} ba'+dg'+bt'+dc'=0 \\ da'+bg'-dt'-bc'=0 \end{cases} \tag{7.9}$$

显然，等式 (7.9) 成立当且仅当 $a^{'}=g^{'}=t^{'}=c^{'}=0$， $l=0$，这意味着这种图形表示中不存在圈。

例如，核苷酸序列 ATGGCATGCA 表示为向量集合：$x(10)=\{b+di, 2b, 2b+d+bi, 2b+2d+2bi, 2b+3d+bi, 3b+3d+(b+d)\mathrm{i}, 4b+3d+bi, 4b+4d+2bi, 4b+5d+bi, 5b+5d+(b+d)\mathrm{i}\}$。

设 $b=\frac{1}{2}, d=\frac{\sqrt{3}}{2}$，则 $x(10)=\{0.5+0.866\mathrm{i}, 1, 1.866+0.5\mathrm{i}, 2.732+\mathrm{i}, 3.598+0.5\mathrm{i}, 4.098+1.366i, 4.598+0.5\mathrm{i}, 5.464+\mathrm{i}, 6.33+0.5\mathrm{i}, 6.83+1.366\mathrm{i}\}$，那么序列 ATGGCATGCA 的 2-D 图形表示如图 7.2 所示。

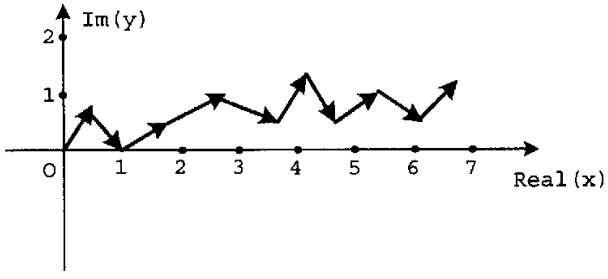


图 7.2　序列 ATGGCATGCA 的有向图

7.2.2　DNA 三联体的 2-D 表示

氨基酸是由核苷酸三联体决定，因此，4 个核苷酸可以产生 $4^3=64$ 个核苷酸三联体密码子，它们决定 20 个氨基酸及其综合链的起点和终点．此外，蛋白质是由氨基酸组成的．见表 7.1 。

在这里 A, G, T 和 C 分别代表腺嘌呤，鸟嘌呤，胸腺嘧啶和胞嘧啶，如果用尿嘧啶 (U) 来替换以上序列中的胸腺嘧啶 (T)，就得到与之对应的 RNA 序列。

由核苷酸的三联体生成氨基酸，令：

$$z(n)=h_0y(n)+h_1y(n-1)+h_2y(n-2) \tag{7.10}$$

这里，式中 $y(n), y(n-1), y(n-2)$ 表示 DNA 序列中三个相邻的核苷酸，$z(n)$ 表示对应的氨基酸．h_0, h_1, h_2 表示非零正实数．例如，设 $h_0=1, h_1=\frac{1}{2}, h_2=\frac{1}{4}, b=\frac{1}{2}, d=\frac{\sqrt{3}}{2}$，则生成氨基酸 Phe 的三联体 TTT 和 TTC 对应的向量为：

$$\mathrm{TTT}:(1+\frac{1}{2}+\frac{1}{4})(\frac{1}{2}-\frac{\sqrt{3}}{2}\mathrm{i})=\frac{7}{8}-\frac{7\sqrt{3}}{8}\mathrm{i},$$
$$\mathrm{TTC}:(1+\frac{1}{2})(\frac{1}{2}-\frac{\sqrt{3}}{2}\mathrm{i})+\frac{1}{4}(\frac{\sqrt{3}}{2}-\frac{1}{2}\mathrm{i})=\frac{6+\sqrt{3}}{8}-\frac{6\sqrt{3}+1}{8}\mathrm{i}.$$

根据以上方法，我们得到了 64 个核苷酸三联体密码子的向量表示和 20 个氨基酸的向量表示，如图 7.3 所示.

表 7.1: 64 个核苷酸三联体密码子

	T		C		A		G	
T	TTT	PHE	TCT	SER	TAT	TYR	TGT	CYS
	TTC		TCC		TAC		TGC	
	TTA	LEU	TCA		TAA	STOP	TGA	STOP
	TTG		TCG		TAG		TGG	TRP
C	CTT	LEU	CCT	PRO	CAT	HIS	CGT	ARG
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	GLN	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	ILE	ACT	THR	AAT	ASN	AGT	SER
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	LYS	AGA	ARG
	ATG	MET	ACG		AAG		AGG	
G	GTT	VAL	GCT	ALA	GAT	ASP	GGT	GLY
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	GLU	GGA	
	GTG		GCG		GAG		GGG	

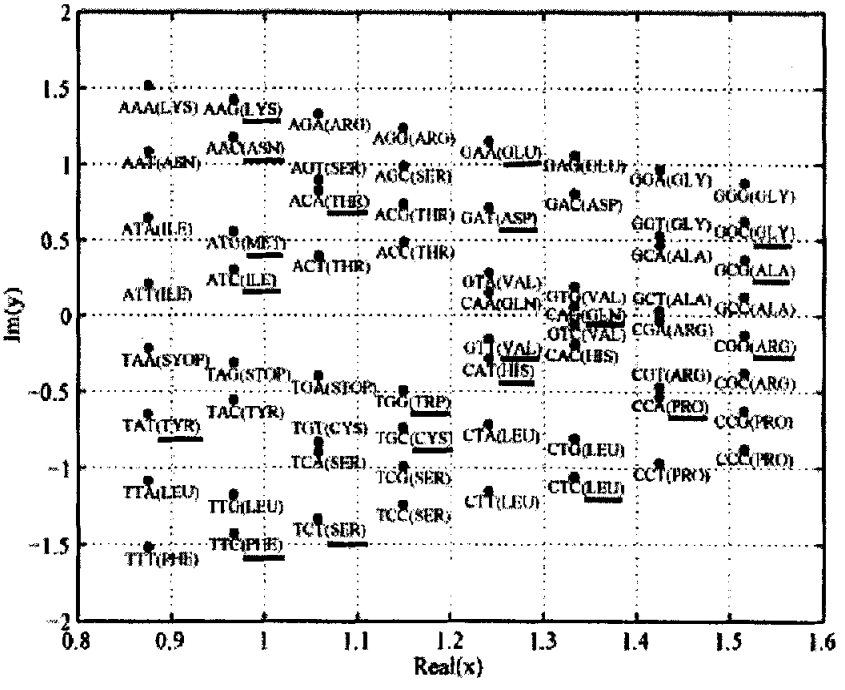


图 7.3　20 个氨基酸的对应向量表示

把四个向量 $(b+di)$, $(d+bi)$, $(b-di)$ 和 $(d-bi)$ 分别分配给四个核苷酸 A, G, T 和 C，以及利用公式 (7.10) 我们得到了如图 7.3 中所显示的一些简单结构： 64 个核酸三联体密码子在半复平面上的分布呈梯形状，且位于梯形四个角的三联体有相同的核苷酸，即它们是 AAA, GGG, TTT 和 CCC．在三联体中的第一、第二核苷酸相同的三联体成小梯形状，除了 CAA, CAG, GTT, GTC, 三联体的第一个核苷酸决定其所在的象限， 64 个核苷酸三联体密码子中 32 个三联体跟其它的 32 个三联体分别互为共轭．

根据公式 (7.10) 设

$$l(m)=z(0)+\sum_{k=1}^{m}z(k),\quad z(0)=0,\quad m=1,2,\cdots,n-2 \tag{7.11}$$

则同样把一个 DNA 三联体序列转化成向量序列： $\vec{V_0},\vec{V_1},\vec{V_2},\cdots\vec{V_m}$，

$$\begin{aligned}\vec{V_j}\longrightarrow&(h_0(ba_j+dg_j+bt_j+dc_j)+h_1(ba_{j-1}+dg_{j-1}+bt_{j-1}+dc_{j-1})\\&+h_2(ba_{j-2}+dg_{j-2}+bt_{j-2}+dc_{j-2}))+(h_0(da_j+bg_j-dt_j-bc_j)\\&+h_1(da_{j-1}+bg_{j-1}-dt_{j-1}-bc_{j-1})+h_2(da_{j-2}+bg_{j-2}-dt_{j-2}-bc_{j-2}))\mathrm{i}\end{aligned} \tag{7.11$'$}$$

其中， a_r, g_r, t_r 和 c_r $(r=j,\ j-1,\ j-2)$ 分别表示为 1 到 $j, j-1, j-2$ 这个子序列中碱基 A, G, T 和 C 出现的累积个数．我们定义 $a_0=g_0=t_0=c_0=0$，则我们有如下两个性质：

性质 3 对于给定的一个 DNA 三联体序列就有唯一的 $l(m)$ 和它对应．

性质 4 我们得到的一个 DNA 三联体序的 2-D 图形表示没有圈．

性质 3 与性质 4 的证明跟性质 1 与性质 2 的证明类似，所以在这里省略．

例如， DNA 序列 ATGGCATGCATC 对应的三联体向量：

$l(m)$={0.9665+0.5580i,2.1160+0.0670i, 3.6315+0.6920i, 5.0555+1.1585i, 6.2965+0.8750i, 7.2630+1.4330i, 8.4125+0.6920i, 9.8365+1.1585i, 11.0775+0.8750i, 12.0440+1.1830i}．对应的 2-D 图形表示，见图 7.4．

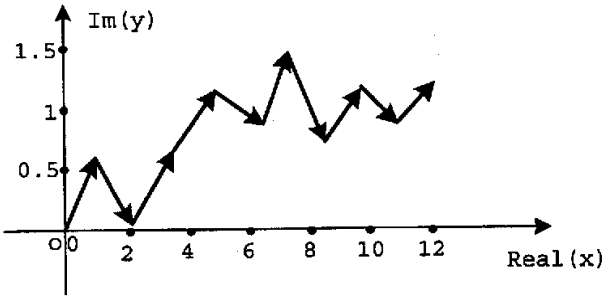


图 7.4 序列 ATGGCATGCATC 三联体对应的有向曲线

7.2.3 蛋白质序列的 2-D 图形表示

我们也可以考虑用蛋白质的一个图形表示法，来代替通常的用氨基酸序列表示蛋白质的方法．一旦所有由 (7.10) 在 (x,yi) 复平面上三联体的位置和 DNA/RNA 序列蛋白质

的表达已知，该蛋白质的图形表示也是很容易获得的．如果该 DNA 序列未知，我们可以使用虚拟基因编码和构造虚拟三联体，这样就可以使蛋白质的表示的获得变成可能 [109, 111]．在图 7.3 中，我们展示了 20 种建立在图示上的氨基酸．各类氨基酸不仅出现了一次，因为不仅用一个三联体把它们表示在标准基因编码中．这 20 种氨基酸都用带有下划线的文字将其显示在图 7.3 中，符合实际基因编码的 20 种氨基酸，图 7.3 可以看作是自然氨基酸的二维图形表示，并且可以将蛋白质序列直接转化成在复平面 (x, yi) 内用坐标表示的数字序列，随后可以用来构造序列不变量．

7.2.4　DNA 三联体的数值刻划

为了找到有向曲线的一些有灵敏度的不变量，我们将有向曲线的图形表示转换成其它的数学对象，例如：取组成有向曲线的向量的模：$r_n = \| \vec{P}_n \|$，和相位：$\varphi_n = arg\vec{P}_n$，因为在复平面内向量的模和相位比较突出反映其特征．我们可以利用模和相位的均值以及模和相位的标准协方差作为不变量来描述 DNA 三联体序列，如表 7.2 所示．

表 7.2: Homo sapiens X-linked nuclear protein (ATRX) gene 的部分 2-D 表示的向量、模和相位

	triplet	vector	module	phase
1	CCA	1.4240-0.5335i	1.5207	-0.3585
2	CAC	1.3325-0.1920i	1.3463	-0.1431
3	ACA	1.0580+0.8325i	1.3463	0.6667
4	CAC	1.3325-0.1920i	1.3463	-0.1431
5	ACC	1.1495+0.4910i	1.2500	0.4037
6	CCA	1.4240-0.5335i	1.5207	-0.3585
7	CAG	1.3325+0.0580i	1.3338	0.0435
8	AGT	1.0580+0.8995i	1.3887	0.7046
9	GTG	1.3325+0.1920i	1.3463	0.1431
10	TGT	1.0580-0.8325i	1.3463	-0.6667
11	GTC	1.3325-0.0580i	1.3338	-0.0435
12	TCC	1.1495-1.2410i	1.6916	-0.8237
13	CCT	1.4240-0.9665i	1.7210	-0.5963
14	CTG	1.3325-0.8080i	1.5583	-0.5451
15	TGG	1.1495-0.4910i	1.2500	-0.4037

我们基于蛋白质的 2-D 图形表示给出 Homo sapiens X-linked nuclear protein (ATRX) gene 的三联体 (氨基酸) 对应的坐标．从这些坐标可以计算出描述蛋白质的特征值．为了这个目的我们可以利用 L/L 矩阵和它相关的高阶矩阵 L^k/L^k 和 L^b/L^b [10]，这些矩阵的

最大特征值和对角线平均宽带或其它的矩阵不变量 [120] 更准确的描述 DNA 和蛋白质的特征．蛋白质的图形表示的唯一性和简单性以及相伴随的数字特征的描述为我们关于蛋白质的比较研究提供了一个新的领域．

7.3 蛋白质序列的 3-D 表示

7.3.1 蛋白质序列的 3-D 表示

在表 7.3 中按字母顺序列出了 20 个氨基酸的全称以及三个和一个字母的压缩形式．

表 7.3: 20 个氨基酸的全称和三个字母、一个字母的压缩形式

Full name	Three symbol	One symbol	Full name	Three symbol	One symbol
Alanine	ala	A	Leucine	leu	L
Arginine	arg	R	Lysine	lys	K
Asparagine	asn	N	Methionine	met	M
Aspartic acid	asp	D	Phenylalanine	phe	F
Cysteine	cys	C	Proline	pro	P
Histidine	his	H	Serine	ser	S
Glutamine	gln	Q	Threonine	thr	T
Glutamic acid	glu	E	Triptophan	trp	W
Glycine	gly	G	Tyrosine	tyr	Y
Isoleucine	ile	I	Valine	val	V

如图 7.5 所示，我们把正十二面体置于空间直角坐标系，设 $O(0,0)$, $R(O,\theta)(x,y) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$ 表示点 (x,y) 绕 O 旋转 θ 角度，

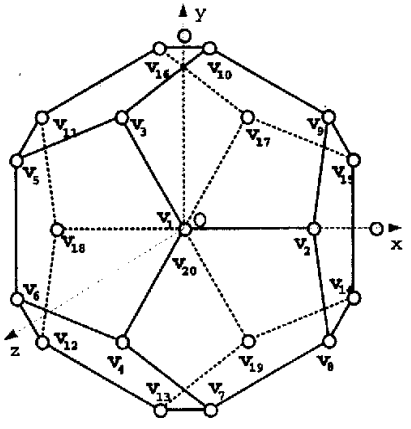


图 7.5　正十二面体

则正十二面体的 20 个顶点的坐标如下

$$v_1=(0,0,1); v_2=(\frac{2}{3},0,\frac{\sqrt{5}}{3});$$

$$v_3=(R(O,120°)(\frac{2}{3},0),\frac{\sqrt{5}}{3}); v_4=(R(O,240°)(\frac{2}{3},0),\frac{\sqrt{5}}{3});$$

$$v_5=(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{15}-\sqrt{3}}{6},\frac{1}{3}); v_6=(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{3}-\sqrt{15}}{6},\frac{1}{3});$$

$$v_7=(R(O,120°)(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{15}-\sqrt{3}}{6}),\frac{1}{3}); v_8=(R(O,120°)(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{3}-\sqrt{15}}{6}),\frac{1}{3});$$

$$v_9=(R(O,240°)(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{15}-\sqrt{3}}{6}),\frac{1}{3}); v_{10}=(R(O,240°)(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{3}-\sqrt{15}}{6}),\frac{1}{3});$$

$$v_{11}=-v_8; v_{12}=-v_9; v_{13}=-v_{10}; v_{14}=-v_5; v_{15}=-v_6;$$

$$v_{16}=-v_7; v_{17}=-v_4; v_{18}=-v_2; v_{19}=-v_3; v_{20}=-v_1.$$

我们在 3-D 空间上构造一个氨基酸图，如图 7.6 所示，根据文献 [50,111] 的方法把氨基酸分配给正十二面体的二十个顶点，使氨基酸与 3-D 空间上的点之间建立对应关系. 表 7.3 中的 20 个氨基酸 A, R, N, D, C, H, Q, E, G, I, L, K, M, F, P, S, T, W, Y, V 的单位向量表示如下：

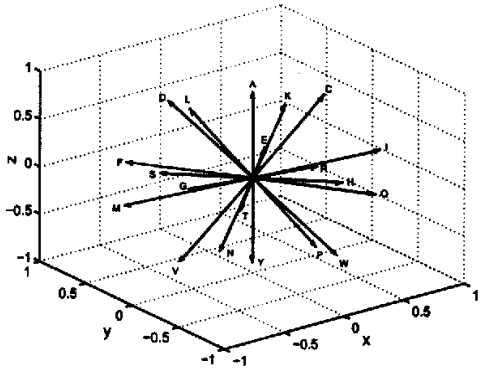


图 7.6　氨基酸图

A $\longrightarrow v_1=(0,0,1)$,
C $\longrightarrow v_2=(\frac{2}{3},0,\frac{\sqrt{5}}{3})$,
D $\longrightarrow v_3=(-\frac{1}{3},\frac{\sqrt{3}}{3},\frac{\sqrt{5}}{3})$,
E $\longrightarrow v_4=(-\frac{1}{3},\frac{-\sqrt{3}}{3},\frac{\sqrt{5}}{3})$,
F $\longrightarrow v_5=(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{15}-\sqrt{3}}{6},\frac{1}{3})$,
G $\longrightarrow v_6=(-\frac{3+\sqrt{5}}{6},\frac{\sqrt{3}-\sqrt{15}}{6},\frac{1}{3})$,
H $\longrightarrow v_7=(\frac{3-\sqrt{5}}{6},-\frac{\sqrt{3}+\sqrt{15}}{6},\frac{1}{3})$,
I $\longrightarrow v_8=(\frac{\sqrt{5}}{3},\frac{-\sqrt{3}}{3},\frac{1}{3})$,
K $\longrightarrow v_9=(\frac{\sqrt{5}}{3},\frac{\sqrt{3}}{3},\frac{1}{3})$,
L $\longrightarrow v_{10}=(\frac{3-\sqrt{5}}{6},\frac{\sqrt{3}+\sqrt{15}}{6},\frac{1}{3})$,
M $\longrightarrow v_{11}=(-\frac{\sqrt{5}}{3},\frac{\sqrt{3}}{3},-\frac{1}{3})$,
N $\longrightarrow v_{12}=(-\frac{\sqrt{5}}{3},-\frac{\sqrt{3}}{3},-\frac{1}{3})$,
P $\longrightarrow v_{13}=(\frac{\sqrt{5}-3}{6},-\frac{\sqrt{3}+\sqrt{15}}{6},-\frac{1}{3})$,
Q $\longrightarrow v_{14}=(\frac{3+\sqrt{5}}{6},\frac{\sqrt{3}-\sqrt{15}}{6},-\frac{1}{3})$,
R $\longrightarrow v_{15}=(\frac{3+\sqrt{5}}{6},\frac{\sqrt{15}-\sqrt{3}}{6},-\frac{1}{3})$,
S $\longrightarrow v_{16}=(\frac{\sqrt{5}-3}{6},\frac{\sqrt{3}+\sqrt{15}}{6},-\frac{1}{3})$,
T $\longrightarrow v_{17}=(\frac{1}{3},-\frac{-\sqrt{3}}{3},-\frac{\sqrt{5}}{3})$,
V $\longrightarrow v_{18}=(-\frac{2}{3},0,-\frac{\sqrt{5}}{3})$,
W $\longrightarrow v_{19}=(\frac{1}{3},-\frac{\sqrt{3}}{3},-\frac{\sqrt{5}}{3})$,
Y $\longrightarrow v_{20}=(0,0,-1)$.

当然， 20 个氨基酸共有 20! 中分配方案．以上只是其中的一个方案。其它的可以经过旋转或反射得到．在这里就不一一列出了．

一个蛋白序列可以看作是由 20 个字母组成的字符串。这 20 个字母就是表 7.3 中所列的 20 个氨基酸，即 $\mathcal{M}$={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}．这样我们可以把一个蛋白质序列映射到点列 $P_0, P_1, P_2, \cdots, P_N$ ，或一个数字序列：

$$S(N) = S(0) + \sum_{i=1}^{N} s(i), \quad S(0) = 0.$$

这里 s(i) 满足如下条件：

$$s(i) = \begin{cases} (0,0,1), & \text{如果 } i = \mathrm{A}, \\ (\frac{2}{3},0,\frac{\sqrt{5}}{3}), & \text{如果 } i = \mathrm{C}, \\ \quad\vdots & \quad\vdots \\ (\frac{1}{3},-\frac{\sqrt{3}}{3},-\frac{\sqrt{5}}{3}), & \text{如果 } i = \mathrm{W}, \\ (0,0,-1), & \text{如果 } i = \mathrm{Y}. \end{cases} \tag{7.12}$$

($i = 0, 1, 2, \cdots, N$, N 是蛋白质序列的长度)

$$P_i \longrightarrow S(i) = \sum_{k=1}^{i} s(k) \tag{7.13}$$

根据以上分配，点 P_i 的坐标 x_i, y_i, z_i ($i = 0, 1, 2, \cdots, N$ ，其中， N 是蛋白质序列的长度) 满足如下方程：

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} 0 & \frac{2}{3} & -\frac{1}{3} & \cdots & \frac{1}{3} & 0 \\ 0 & 0 & \frac{\sqrt{3}}{3} & \cdots & -\frac{\sqrt{3}}{3} & 0 \\ 1 & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{3} & \cdots & -\frac{\sqrt{5}}{3} & -1 \end{pmatrix} \begin{pmatrix} \mathrm{A}_i \\ \mathrm{C}_i \\ \mathrm{D}_i \\ \vdots \\ \mathrm{W}_i \\ \mathrm{Y}_i \end{pmatrix} \tag{7.14}$$

这里， A_i, C_i, D_i, $\cdots$, W_i, Y_i 分别表示为 1 到 i 这个子序列中氨基酸 A, C, D, $\cdots$, W, Y 出现的累积个数．我们定义 A_0=C_0=D_0=$\cdots$=W_0=$\mathrm{Y}_0 = 0$．当 i 依次从 1 变到 N，我们可以得到一系列的点 $P_0, P_1, P_2, \cdots, P_N$ ．依次连接这些点得到了蛋白质序列的 3-D 曲线——蛋白质曲线．这些点称为 3-D 曲线上的结点．一个蛋白质曲线由蛋白质序列的 3-D 曲线上每个结点的坐标决定．因此，蛋白质序列可以由一个 $3 \times N$ 的矩阵表示：

$$\left(\text{蛋白质序列} \right) = P = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_N \\ y_1 & y_2 & y_3 & \cdots & y_N \\ z_1 & z_2 & z_3 & \cdots & z_N \end{pmatrix} \tag{7.15}$$

x_i, y_i, z_i 也具有一定的生物学意义，分别代表 20 个氨基酸的分布．

例如， human neurocan gene (AAC80576) 蛋白质序列的前 20 个字符 MGAPFVWALGLLMLQMLLFV 对应的 3-D 曲线如图 7.6 所示.

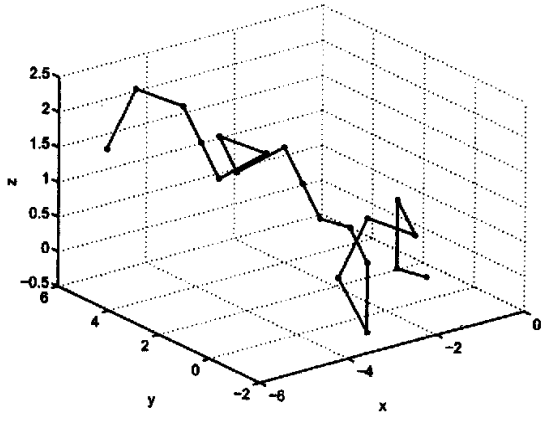


图 7.6　蛋白质序列 MGAPFVWALGLLMLQMLLFV 的 3-D 曲线

用 r_i 来表示从 O 原点到点 P_i 的向量， r_i 的分量 x_i, y_i 和 z_i 由方程 (7.12) 来计算. 设 $\Delta r_i = r_i - r_{i-1}$, 对于任意的 $i = 1, 2, \cdots, N$, 这里， N 是蛋白质序列的长度, 在蛋白质序列的第 i 个位置的每个氨基酸 (A, C, D, $\cdots$, W, Y) 其 Δr_i 的分量, 即 Δx_i, Δy_i 和 Δz_i 可以由方程 (7.12) 来计算. 例如, 第 i 个氨基酸是 W , 那么我们得到 $\Delta x_i = \frac{1}{3}$, $\Delta y_i = -\frac{\sqrt{3}}{3}$ 和 $\Delta z_i = -\frac{\sqrt{5}}{3}$. 这与第 $i-1$ 个氨基酸的构造状态无关. 这三个数 (如， $\Delta x_i = \frac{1}{3}$, $\Delta y_i = -\frac{\sqrt{3}}{3}$ 和 $\Delta z_i = -\frac{\sqrt{5}}{3}$) 称为 Δr_i 的方向数. 第 i 个位置上可能氨基酸的 Δr_i 的长度和方向数归纳在如下表 7.4 里.

表 7.4: 3-D 空间中 20 个氨基酸的 Δr_i 的长度和方向数

base	Δx_i	Δy_i	Δz_i	$\lvert\Delta r_i\rvert$	base	Δx_i	Δy_i	Δz_i	$\lvert\Delta r_i\rvert$
A	0	0	1	1	M	$-\frac{\sqrt{5}}{3}$	$\frac{\sqrt{3}}{3}$	$-\frac{1}{3}$	1
C	$\frac{2}{3}$	0	$\frac{\sqrt{5}}{3}$	1	N	$-\frac{\sqrt{5}}{3}$	$-\frac{\sqrt{3}}{3}$	$-\frac{1}{3}$	1
D	$-\frac{1}{3}$	$\frac{\sqrt{3}}{3}$	$\frac{\sqrt{5}}{3}$	1	P	$\frac{\sqrt{5}-3}{6}$	$-\frac{\sqrt{3}+\sqrt{15}}{6}$	$-\frac{1}{3}$	1
E	$-\frac{1}{3}$	$\frac{-\sqrt{3}}{3}$	$\frac{\sqrt{5}}{3}$	1	Q	$\frac{3+\sqrt{5}}{6}$	$\frac{\sqrt{3}-\sqrt{15}}{6}$	$-\frac{1}{3}$	1
F	$-\frac{3+\sqrt{5}}{6}$	$\frac{\sqrt{15}-\sqrt{3}}{6}$	$\frac{1}{3}$	1	R	$\frac{3+\sqrt{5}}{6}$	$\frac{\sqrt{15}-\sqrt{3}}{6}$	$-\frac{1}{3}$	1
G	$-\frac{3+\sqrt{5}}{6}$	$\frac{\sqrt{3}-\sqrt{15}}{6}$	$\frac{1}{3}$	1	S	$\frac{\sqrt{5}-3}{6}$	$\frac{\sqrt{3}+\sqrt{15}}{6}$	$-\frac{1}{3}$	1
H	$\frac{3-\sqrt{5}}{6}$	$-\frac{\sqrt{3}+\sqrt{15}}{6}$	$\frac{1}{3}$	1	T	$\frac{1}{3}$	$\frac{\sqrt{3}}{3}$	$-\frac{\sqrt{5}}{3}$	1
I	$\frac{\sqrt{5}}{3}$	$\frac{-\sqrt{3}}{3}$	$\frac{1}{3}$	1	V	$-\frac{2}{3}$	0	$-\frac{\sqrt{5}}{3}$	1
K	$\frac{\sqrt{5}}{3}$	$\frac{\sqrt{3}}{3}$	$\frac{1}{3}$	1	W	$\frac{1}{3}$	$-\frac{\sqrt{3}}{3}$	$-\frac{\sqrt{5}}{3}$	1
L	$\frac{3-\sqrt{5}}{6}$	$\frac{\sqrt{3}+\sqrt{15}}{6}$	$\frac{1}{3}$	1	Y	0	0	-1	1

类似于 DNA 序列的矩阵表示，我们同样根据图形表示把蛋白质序列转化为熵矩阵 Q[120] 表示，矩阵 Q 是一个对称矩阵，其定义如下：

$$[Q]_{ij} = [Q]_{ji} = \frac{e_{ij}}{\sum_{k=i}^{j-1} e_{k(k+1)}}, \quad i \neq j, \quad [Q]_{ii} = 0,$$

其元素 (i, j) 定义为 3-D 曲线上两个顶点 i 和 j 之间的欧式距离和顶点 i 和 j 之间的边的几何距离之和的商。e_{ij} 是 3-D 曲线上两个顶点 i 和 j 之间的欧式距离。特别利用矩阵不变量—矩阵最大特征值来描述蛋白质序列。

我们还可以把蛋白质序列通过它的图形表示转化成其它的数学对象—数字序列。然后取其数字序列的自相关系数和自协方差系数为分量构造特征向量 $\mathrm{R} = \{r_1, r_2, \cdots, r_n\}$ 和 $\mathrm{C} = \{c_1, c_2, \cdots, c_n\}$。并以此来描述蛋白质序列的特征。$r_n$ 和 c_n 的定义分别由如下 (7.16) 和 (7.17) 式给出：

$$r_n^j = \frac{1}{N-n} \sum_{i=1}^{N-n} h_i^j h_{i+n}^j, \quad n = 1, 2, \cdots, m, \; j = 1, 2, 3, \tag{7.16}$$

这里，N 是蛋白质序列的长度，m 是特征向量 R 的分量个数 $(m < N)$，$h_i^j (j = 1, 2, 3)$ 表示蛋白质序列上第 i 个氨基酸对应的 (x_i, y_i, z_i)。

$$c_n^j = \frac{1}{N-n} \sum_{i=1}^{N-n} (h_i^j - \bar{h^j})(h_{i+n}^j - \bar{h^j}), \quad n = 1, 2, \cdots, m, \; j = 1, 2, 3, \tag{7.17}$$

这里，$\bar{h^j}$ 表示序列 $h_1^j, h_2^j, \cdots h_N^j$ 的均值。

另外，还可以由组成蛋白质序列的氨基酸出现的频率和部分自相关系数相结合将蛋白质序列表示为如下的特征向量：

$$\mathrm{X} = \{p_1, p_2, \cdots, p_{20}, r_2^1, r_4^1, r_6^1, r_8^1, r_{10}^1, r_2^2, r_4^2, r_6^2, r_8^2, r_{10}^2, r_2^3, r_4^3, r_6^3, r_8^3, r_{10}^3\}, \tag{7.18}$$

这里，p_k $(k = 1, 2, \cdots, 20)$ 表示蛋白质序列上每个氨基酸出现的频率，r_i^j $(i = 2, 4, 6, 8, 10; j = 1, 2, 3)$ 由公式 (7.16) 给出。同样还可以由组成蛋白质序列的氨基酸出现的频率和部分自协方差系数相结合将蛋白质序列表示为如下的特征向量：

$$\mathrm{Y} = \{p_1, p_2, \cdots, p_{20}, c_2^1, c_4^1, c_6^1, c_8^1, c_{10}^1, c_2^2, c_4^2, c_6^2, c_8^2, c_{10}^2, c_2^3, c_4^3, c_6^3, c_8^3, c_{10}^3\}, \tag{7.19}$$

这里，c_i^j $(i = 2, 4, 6, 8, 10; j = 1, 2, 3)$ 由公式 (7.17) 给出。

因为 α 螺旋和 β 折叠蛋白质分别具有近似 4 和 2 的周期，所以我们在以上两个公式 (7.18) 和 7.19) 中都取了偶数个自相关系数 $r_i^j (i = 2, 4, 6, 8, 10; j = 1, 2, 3)$ 和自协方差系数 $c_i^j (i = 2, 4, 6, 8, 10; j = 1, 2, 3)$。

7.3.2 应用举例

首先，我们利用蛋白质序列的以上特征数值来分析蛋白质序列之间的相似性，以便得到某些未知蛋白质序列的信息。我们以 9 种神经基因 (Human neurocan gene (AAC80576), Rattus brevican (Rattus norvegicus, NP_037048), Gallus neurocan (AAD24546), Mouse neurocan (S52781), Mus brevican (Mus musculus, NP_031555), Rat neurocan(S28764), Rattus

neurocan (Rattus norvegicus， AAC15766)， Versican core protein precursor (Q9ERB4_1)，Versican-Rattus norvegicus (AAC40166) 为例，利用蛋白质序列对应熵矩阵不变量 (矩阵最大特征值) 和特征向量来比较相似性. 以上 9 种蛋白质序列数据从以下网站免费下载: http://www.ncbi.clm.nih.gov .

7.3.2.1　举例 1

我们利用熵矩阵 Q 的前 10 个最大特征值来构造 10 维向量，计算向量终点之间的欧式距离得到了 9 种神经基因的相似性表，见表 7.5 .

表 7.5: 9 种神经基因对应的 10 维向量终点之间的欧式距离 (矩阵正规化最大特征值)

Species	Human	B-Ratt	Gallus	Mouse	B-mus	Rat	Rattus	Versic	V-Ratt
Human	0	0.0213	0.0098	0.0148	0.0112	0.0077	0.0149	0.0487	0.0158
B-Ratt		0	0.0219	0.0250	0.0136	0.0211	0.0168	0.0354	0.0350
Gallus			0	0.0088	0.0143	0.0072	0.0221	0.0521	0.0222
Mouse				0	0.0159	0.0080	0.0268	0.0573	0.0243
B-mus					0	0.0104	0.0138	0.0447	0.0235
Rat						0	0.0193	0.0516	0.0197
Rattus							0	0.0360	0.0248
Versic								0	0.0601
V-Ratt									0

从表 7.5 中可以看出四种 human, mouse, gallus 和 rat 神经基因彼此较相似，B-Rattus, B-mus 和 Rattus 神经基因也较相似，同样我们发现 Versican 和 V-Rattus 与其它物种的神经基因相似性很差.

我们利用 (7.18) 和 (7.19) 把不同长度的蛋白质序列转化为向量 X^{35} 和 Y^{35}，通过计算向量终点间的欧式距离来比较其相似性. 9 种神经基因的相似性表，见表 7.6 和表 7.7 .

表 7.6: 9 种神经基因对应的 35 维特征向量终点之间的欧式距离 (氨基酸分布和自相关系数)

Species	Human	B-Ratt	Gallus	Mouse	B-mus	Rat	Rattus	Versic	V-Ratt
Human	0	0.5496	0.3678	0.4192	0.5175	0.4905	0.4575	0.6974	4.3116
B-Ratt		0	0.5105	0.4167	0.2855	0.4856	0.1859	0.2291	4.5153
Gallus			0	0.4507	0.3126	0.5550	0.5435	0.7306	4.4535
Mouse				0	0.3474	0.1078	0.4293	0.5677	4.1340
B-mus					0	0.4426	0.4109	0.5101	4.4462
Rat						0	0.4889	0.6046	4.0418
Rattus							0	0.2431	4.4834
Versic								0	4.5559
V-Ratt									0

表 7.7: 9 种神经基因对应的 35 维特征向量终点之间的欧式距离 (氨基酸分布和自协方差系数)

Species	Human	B-Ratt	Gallus	Mouse	B-mus	Rat	Rattus	Versic	V-Ratt
Human	0	0.2786	0.1981	0.1551	0.2615	0.1676	0.2173	0.3216	1.2926
B-Ratt		0	0.1435	0.1319	0.0680	0.1440	0.0764	0.0483	1.3338
Gallus			0	0.0737	0.0890	0.0871	0.1230	0.1880	1.2896
Mouse				0	0.1120	0.0415	0.0774	0.1759	1.2829
B-mus					0	0.1266	0.0987	0.1080	1.3249
Rat						0	0.0812	0.1822	1.2423
Rattus							0	0.1097	1.2866
Versic								0	1.3357
V-Ratt									0

观察表 (7.6) 和 (7.7)，我们找到 Mouse−Rat，B-Rattus−Versican，B-Rattus−B-mus，B-Rattus−Rattus 在 9 种神经基因中比较相似，同样我们发现 versican−Rattus norvegicus(AAC 40166) 与其他物种的神经基因相似性很差.

比较表 7.5– 表 7.7，我们可以找到对应的结果具有很大的相似性，但表 7.6 和表 7.7 的方法比较简洁快速.

7.3.2.2 举例 2

广泛存在于生物界的一种蛋白质细胞色素 C，比较它们的一级结构，可以帮助了解物种进化之间的关系. 物种间越接近，则细胞色素 C 的一级结构越相似；反之，则相差甚远.

表 7.8: 9 个细胞色素 C 蛋白质的代码及名称

Code	Proteins
3CYT	Albacore tuta heart ferricytochrome c(oxidized)
5CYT	Albacore tuta heart ferrocytochrome c(reduced)
1CCR	Rice embryo ferricytochrome c
2C2C	Rhodospirillum rubrum ferricytochrome c_2
3C2C	Rhodospirillum rubrum ferrocytochrome c_2
155C	Paracoccus denitrificans cytochrome c_{550}
351C	Pseudomonas aeruginosa ferricytochrome c_{551}
451C	Pseudomonas aeruginosa ferrocytochrome c_{551}
1CC5	Azotobacter vinelandii ferricytochrome c_5

在序列相似性的基础上重建大部分具有大量变异的现存蛋白质的历史是可能的. 蛋白质序列之间的关系可以用进化树来表示，一棵树有两个因素：拓扑结构和分支长度. 拓扑结构反映了序列之间的进化距离，分支长度与进化距离成正比. 一棵进化树构建的

正确与否，取决于描述各序列之间的差异性的距离是否合适，也就是说，描述序列的特征数值是否有很高的灵敏度.

我们基于蛋白质序列的 3-D 表示，并利用第 4 章的构建进化树的方法来构建在表 7.9 中出示的 9 个细胞色素 C 蛋白质的序列进化树，即 (1) 根据 3-D 图形表示把表 7.9 中出示的 9 个细胞色素 C 蛋白质序列转化为熵矩阵 Q 表示，再利用矩阵 Q 的最大特征值构造蛋白质序列之间的进化距离矩阵，最后利用 NJ 方法 [68] 构建进化树，见图 7.7；(2) 把蛋白质序列通过它的图形表示转化成一数字序列，然后提取其数字序列的部分自协方差系数和组成蛋白质序列的氨基酸出现的频率相结合的特征数值来构造蛋白质序列之间的进化距离矩阵，并利用 NJ 方法构建进化树，见图 7.8 .

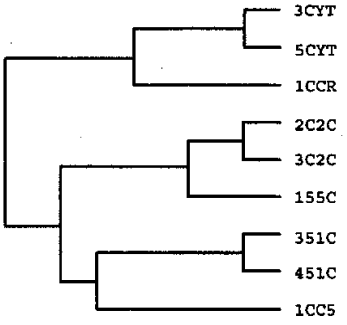


图 7.7　基于蛋白序列的 3-D 表示对应的矩阵正规化最大特征值下的进化树

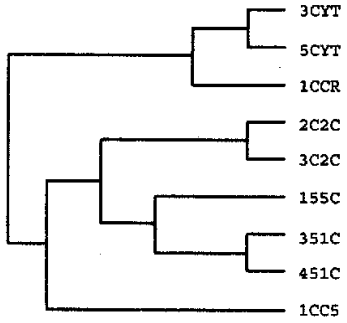


图 7.8　基于蛋白序列的 3-D 表示对应数字序列自协方差系数和氨基酸分布下的进化树

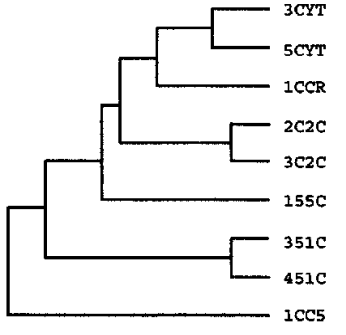


图 7.9　基于蛋白质序列比对下的进化树

观察图 7.7 和图 7.8 可知，都有蛋白质 3CYT 与 5CYT 、 2C2C 与 3C2C 、 351C 与 451C 之间的进化距离最近，与文献 [139] 的用序列比对打分矩阵构建的序列进化树 (图 7.9) 相一致. 图 7.7 的树有三个大分支 3CYT 、5CYT 和 1CCR，2C2C 、3C2C 和 155C，351C 、 451C 和 1CC5 . 这与图 7.9 的拓扑结构基本一致，而图 7.8 中 155C 与 351C 和 451C 在一个分支上， 1CC5 与 2C2C 和 3C2C 的进化距离比较近. 所以图 7.7 比较符合实际进化关系. 图 7.7 ，图 7.8 和图 7.9 的不同，可能是由于计算方法的不同所致，因为有

资料显示到目前为止，还没有一种数学方法，能够用来评估两棵树是否在本质上是不同的.

7.4 小结

我们在复平面上给出了 DNA 三联体或蛋白质序列的 2-D 表示，数据给我们更直观的视觉感受；有利于识别不同的 DNA 序列和蛋白质序列之间的相似性；容易识别 DNA 序列上蛋白质编码区域起始和终止位点，此表示法没有圈即是非退化的，很大程度上减少了信息的丢失. 另外，在复平面上用二维图形表示把 DNA 三联体和蛋白质序列转化成复数字序列，可看成离散的数字信号，这样更适用短时傅立叶变换 (DFT) 取得序列的频谱图来表现 DNA 序列三联体或蛋白质序列的频域特性以及周期特性等. 我们这样做的目的是将一个字符序列中非常多的特征转化为较少的几个量化的特征，这样可能得到更简单且有高灵敏度的比较指标.

在 3-D 空间中给出了蛋白质的特征数值表示 (定量表示)，进而首先比较了蛋白质序列的相似性分析，此方法简洁快速，得到的结果与实际比较吻合. 其次构建了蛋白质序列进化树. 我们提出的基于序列的图形表示构建序列进化树方法简单省时，既不需要蛋白质的晶体结构的坐标也不需要比对打分矩阵的参数优化选择，这是区别于其它结构进化树和序列进化树构建方法的本质特点. 因为蛋白质的序列进化树与结构进化树是一致的 [139]，因此构建序蛋白质列进化树对蛋白质结构预测有应用价值.

由上所述，序列相似性的比较仍是一个需要不断发展的领域，利用序列对比法进行多元比较，归纳各个蛋白质超家族的特征保守区并用于序列比较，都是序列比较的新发展. 随着序列资料的继续积累和扩充，序列相似性比较必将发挥更大的作用.

8 附录：分子生物学知识概论

本章介绍了分子生物学的基本概念，提供一些主要的信息，以便能从容应对本文中涉及的生物学背景。

分子生物学是在分子水平上研究生物的结构、组织和功能的科学．广义而言，分子生物学主要包括分子生物学技术、分子生物学技术的应用及这些技术研究所取得的理论成就等方面．狭义地讲，分子生物学的范畴偏重于核酸（或基因）的分子生物学，主要研究基因或 DNA 的复制、转录、翻译和调控等过程，同时也涉及与这些过程有关的蛋白质、酶的结构和功能的研究．

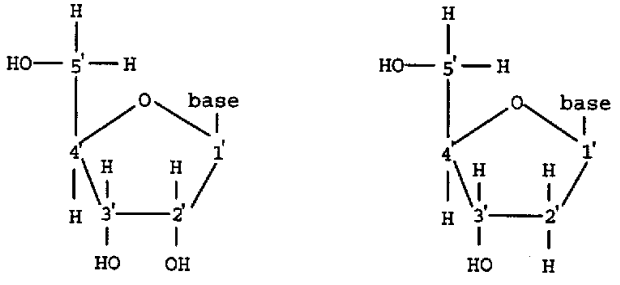
生命的基本单位是细胞，它是由细胞膜、细胞质和细胞核三者组成，遗传信息储存在细胞核中．细胞的分子有两类：大分子和小分子．大分子有三种类型： DNA, RNA 和蛋白质，它们是由某些小分子聚合在一起形成的．

8.1 核酸

生物体包含两类核酸：核糖核酸 (ribonucleic acid)，简写为 RNA; 脱氧核糖核酸 (deoxyribonucleic acid)，简写为 DNA．

8.1.1 DNA

DNA 是由称为核苷酸的小分子组成的链．实际上 DNA 是双链，先认识一下 DNA 的单链 (strand)，它是由重复的基本单元组成的骨架．这种基本单元由一个称做 $2'$- 脱氧核糖的糖分子和一个磷酸残基组成．糖分子含有 5 个碳原子，标记为 $1' \longrightarrow 5'$(图 8.1)．形成骨架的键是在一个单元的 $3'$ 碳原子和下一个单元的 $5'$ 碳原子之间．因此， DNA 分子具有方向性 (orientation)，一般从 $5'$ 开始到 $3'$ 结束．



核糖　　$2'$- 脱氧核糖

图 8.1　核酸中的糖， RNA 中是核糖， DNA 中是脱氧核糖

与骨架中与 $1'$ 碳原子相连的分子为碱基 (base)．在 DNA 分子中有 4 种碱基，分别是：腺嘌呤 (adenine， A)、鸟嘌呤 (guanine， G)、胞嘧啶 (cytosine， C)、和胸腺嘧啶 (thymine， T)．图 8.2 显示了每种碱基的分子结构，图 8.3 为单链 DNA 分子的示意图．碱基 A 和 G 是嘌呤，而碱基 C 和 T 属于嘧啶．DNA 分子的基本单元由糖、磷酸和碱基组成，该基本单元谓之核苷酸 (nucleotide)．

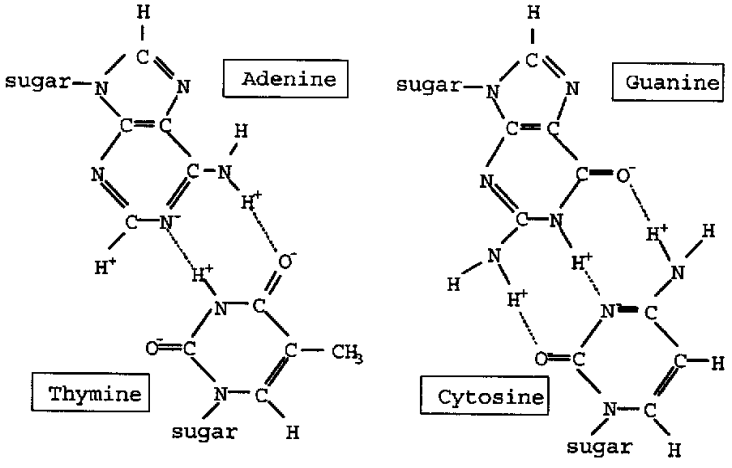


图 8.2 DNA 中氮化的碱基．两种 Watson−Crick 配对，注意腺嘌呤与胸腺嘧啶、鸟嘌呤与胞嘧啶之间所形成的键，图中用点线表示

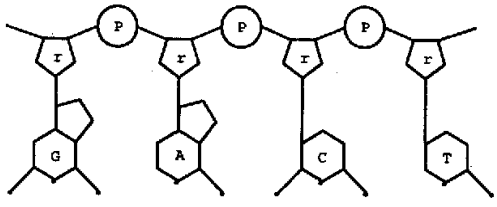


图 8.3 一个 DNA 链分子结构的图示

前已述及，DNA 分子是双链结构．两条链缠绕在一起形成双螺旋，此著名的双螺旋 (double helix) 结构是由 James Watson 和 Francis Crick 在 1953 年发现的．两条链结合的机制是一条链的碱基与另一条链的碱基配对，碱基 A 与碱基 T 配对，碱基 C 与碱基 T 配对，如图 8.2 和图 8.4 所示．

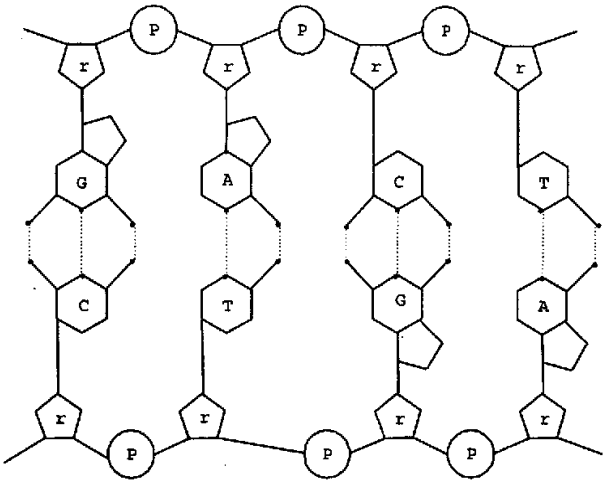


图 8.4　双链 DNA 链分子结构的图示

由图 8.4 我们能看出，在 A−T 配对时，有两个氢键连接，而在 G−C 配对时有三个氢键连接. 因此，我们把 A, T 称谓弱氢键碱基而把 G, C 称为强氢键碱基.

一般地，我们可以把 DNA 分子看成是字符集 Ω={A, C, G, T} 上的字符串，每一个字符代表一个碱基. 图 8.5 是 DNA 的 “串表示”，将一串字符置于另一串字符之上来表示双链 DNA .

图 8.5　双链 DNA 表达的字符串

8.1.2　RNA

RNA 分子与 DNA 分子非常相似，但有以下组成和结构上的不同：(1) RNA 中，核糖取代了 DNA 分子中的 2′- 脱氧核糖 (如图 8.1 所示)．(2) RNA 中，胸腺嘧啶 T 被尿嘧啶 U 取代，U 和 T 一样能够与 A 配对．(3) RNA 并不形成双螺旋.

RNA 分子同样可以看作是字符集 Ω*={A, C, G, U} 上的字符串. 同样具有方向性的，左端通常记为 5′，另一端记为 3′．在 RNA 分子中 A 与 U 配对，G 和 C 配对构成 RNA 二级结构 (如图 8.6 所示)，这对参与蛋白质的合成起着决定性的作用.

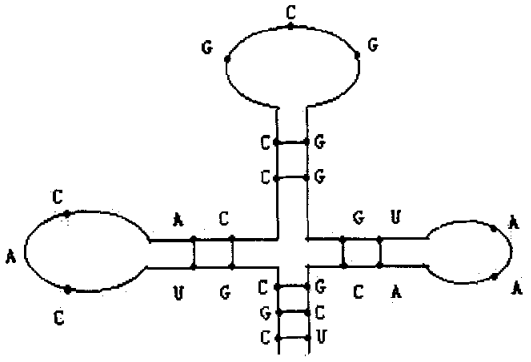


图 8.6 RNA 二级结构：三叶草

8.2 蛋白质

蛋白质是由一类称为氨基酸 (amino acid) 的简单分子组成. 图 8.7 是氨基酸的例子, 每一个氨基酸有有一个中心碳原子, 谓之 α 碳原子或 C_α . C_α 连接 1 个氢原子 (H) 、 1 个氨基 (-NH_2) 、 1 个羧基 (-COOH) 和 1 个侧链. 正是这个决定了氨基酸间的差异.

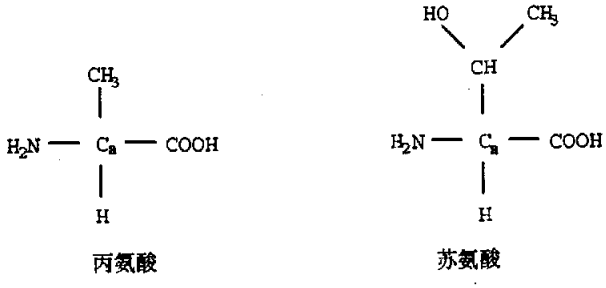


图 8.7 氨基酸的例子

我们在自然界中发现了 20 种不同的氨基酸, 表 8.1 列出了这些氨基酸, 这是蛋白质中最常见的 20 种, 另外还有几种非标准的氨基酸. 蛋白质也是聚合物, 是 20 种氨基酸字符集 Ω^{**} ={A, R, D, N, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V} 上的字符串. 蛋白质也有方向性.

核酸是遗传信息的携带者, 而蛋白质是信息转化为生物结构和功能的表达者. 蛋白质按外形和在生物组织中的位置和作用, 可分为三大类: 纤维蛋白 (fibrous protein) , 膜蛋白和球蛋白. 其中球蛋白的种类最多, 功能也最重要. 一般地, 球蛋白质的结构分为一级结构、二级结构和三级结构, 除此之外还有超二级结构和四级结构等. 它的一级结构

就是指这个蛋白质的氨基酸本原序列．二级结构是指蛋白质多肽主链在空间中的趋向，是一级结构通过折叠产生的．二级结构中主要由两类：α 螺旋和 β 折叠．蛋白质的三级结构是蛋白质的肽链中所有肽键和残基 (包括侧链) 键的相对位置．

表 8.1: 蛋白质中发现的 20 种常见氨基酸以及氨基酸的遗传密码

Genetic code	氨基酸	3 个字母	1 个字母
GCU,GCC,GCA,GCG	丙氨酸 (Alanine)	Ala	A
CGU,CGC,CGA,CGG	精氨酸 (Arginine)	Arg	R
GAU,GAC	天冬氨酸 (Aspartic acid)	Asp	D
AAU,AAC	天冬酰胺 (Asparagine)	Asn	N
UGU,UGC	半胱氨酸 (Cystein)	Cys	C
GAA,GAG	谷氨酸 (glutamic acid)	Glu	E
CAA,CAG	谷氨酰胺 (glutamine)	Gln	Q
GGU,GGC,GGA,GGG	甘氨酸 (glycine)	Gly	G
CAU,CAC	组氨酸 (histidine)	His	H
AUU,AUC,AUA	异亮氨酸 (isoleucine)	Ile	I
CUU,CUC,CUA,CUG , UUA,UUG	亮氨酸 (leucine)	Leu	L
AAA,AAG	赖氨酸 (lysine)	Lys	K
AUG	甲硫氨酸 (methionine)	Met	M
UUU,UUC	苯丙氨酸 (phenylalanine)	Phe	F
CCU,CCC,CCA,CCG	脯氨酸 (proline)	Pro	P
UCU,UCC,UCA,UCG	丝氨酸 (serine)	Ser	S
ACU,ACC,ACA,ACG	苏氨酸 (threonine)	Thr	T
UGG	色氨酸 (tryptophan)	Trp	W
UAU,UAC	酪氨酸 (tyrosine)	Tyr	Y
GUU,GUC,GUA,GUG	缬氨酸 (valine)	Val	V

8.3　分子遗传学机制

DNA 携带遗传材料，即生物功能所要求的信息 (某些病毒除外，它们的遗传材料是 RNA)，而且生物体通过 DNA 将遗传信息传给下一代．

8.3.1　基因和遗传密码

DNA 中仅有一部分连续的片断编码构建蛋白质信息．而每一种不同的蛋白质仅对应一段 DNA 序列，该段序列称为基因 (gene)．因为某些基因编码 RNA 分子，因此更正确的说基因是编码蛋白质或 RNA 的连续的 DNA 序列．

贮存在 DNA 上的遗传信息通过 mRNA 传递到蛋白质上，mRNA 与蛋白质之间的联系是通过遗传密码的破译来实现的．mRNA 上每 3 个核苷酸翻译成蛋白质多肽链上的一

个氨基酸，这 3 个核苷酸成为密码，也叫三联子密码。三联核苷酸和与氨基酸之间的对应关系称为遗传密码 (genetic code)，见表 8.1 .

从表 8.1 中可以看出，在 64 种三联体密码子 (condon) 中有三个终止密码子 UAA，UAG 和 UGA(可用 STOP 表示)，其余的 61 个密码子编码了 20 种氨基酸，因此很多氨基酸都有多种编码 (简并)：三种氨基酸有 6 重简并编码：亮氨酸 (L)、丝氨酸 (S) 和精氨酸 (R)；五种氨基酸有 4 重简并编码：缬氨酸 (V)、脯氨酸 (P)、丙氨酸 (A)、甘氨酸 (G) 和苏氨酸 (T)；有 3 重简并编码的是异亮氨酸 (I) 和终止密码子；有 9 种氨基酸有 2 重简并编码：苯丙氨酸 (F)、酪氨酸 (Y)、组氨酸 (H)、谷氨酰胺 (Q)、天冬酰胺 (N)、赖氨酸 (K)、天冬氨酸 (D)、谷氨酸 (E) 和半胱氨酸 (C)；只有甲硫氨酸 (M) 和色氨酸 (W) 是单重编码.

8.3.2 中心法则：转录、翻译、和蛋白质的合成

一个识别的基因或基因簇起始的机制是启动子 (promoter). 启动子是基因前面的一段 DNA 序列，指征位于其前面的基因. 密码子 AUG(编码甲硫氨酸) 则是基因开始的信号. 识别出基因的起始点后，基因到 RNA 的拷贝就开始了，合成的 RNA 为信使 RNA，简写为 mRNA，其序列与 DNA 中的一条链相同，但 U 代替了 T，该过程称为转录 (transcription). 遗传密码的翻译是由 tRNA 实现的，这个过程称为翻译 (translation)，它连接密码子和其所编码的氨基酸. 蛋白质是按一个氨基酸接一个氨基酸的方式合成起来的. 当出现终止密码子时，没有 tRNA 与之对应，合成便终止，mRNA 被释放，并被降解成核糖核苷酸，降解物可循环用于其他 RNA 合成. 图 8.8 总结了上述的过程，对细胞内遗传信息流动的观点通常用中心法则 (cantral dogma) 来说明.

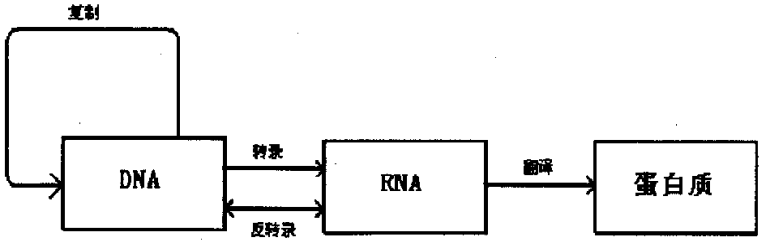


图 8.8 细胞内的遗传信息流动：分子生物学的中心法则

关于分子生物学更详细的知识，建议看 Lewin 在 1999 年写的书 [129] .

参考文献

[1] Hamori, E.; Ruskin, J., H curves, a novel method of representation of uncleotide series especially suited for long DNA sequence, J. Biol. Chem. 258(1983) 1318−1327.

[2] Hamori, E., Novel DNA sequence representations, Nature, 314(1985) 585−586.

[3] Hamori, E., Graphical representation of long DNA sequences by methods of H curves, current results and futureaspects, J. Bio. Techniques, 7(1989) 710−720.

[4] Gates, M. A., Simple DNA sequence representations, Nature, 316(1985) 219.

[5] Gates, M. A., A simple way to look at DNA, J. Thor. Biol. 119(1986) 319−328.

[6] Leong, M.; Morgenthalar, S., Random walk and gap plots of DNA sequences, Comput. Applic. Biosci. 21(1995) 503−511.

[7] Randic, M.; Vracko, M., On the similarty of DNA primary sequence, J. Chem. Inf. Comput. 40(2000) 599−606.

[8] Randic, M.; Vracko, M., A. Nandy, S. C. Basak, On 3-D Graphical representation of DNA primary sequence and their numerical characterization, J. Inf. Comput. 40(2000) 1235−1244.

[9] Randic, M., Condensed representation of DNA primary sequence, J. Chem. Inf. Comput. Sci. 40(2000) 50−56.

[10] Randic, M.; Vracko, M.; Nella, L.; Dejan, P., Novel 2-D Graphical representation of DNA sequence and their numerical characterization, Chemical Physics Letters, 368(2003) 1−6.

[11] Randic, M.; Vracko, M.; Zupan, J.; Novic, M., Compact 2-D graphical representation of DNA, Chemical Physics Letters, 373(2003) 558−562.

[12] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D., Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chemical Physics Letters, 371(2003) 202−207.

[13] Randic, M.; Balaban, A. T., On a four-dimensional representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. 43(2003) 532−539.

[14] Randic, M., On characterization of DNA primary sequences by a condensed matrix, Chemical Physic Letters, 317(2000) 29−34.

[15] Randic, M.; Basak, S. C., Characterization of DNA primary sequences based on the average distance between bases, J. Chem. Inf. Comput. Sci. 41(2001) 561−568.

[16] Balaban, A. T.; Plavsic, D.; Randic, M., DNA invarants based on nonoverlapping triplets of nucleotide bases, Chemical Physic Letters, 379(2003) 147−154.

[17] Nandy, A., Graphical representation of DNA sequence, Curr. Sci. 40(2000) 915−919.

[18] Nandy, A.; Nandy, P., On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, Chemical Physics Letters, 368(2003) 102−107.

[19] Nandy, A., A new graphical representation and analysis of DNA sequence structure: I.Methodology and Application to Globin Genes, Curr. Sci. 66(2004) 309−314.

[20] Nandy, A.; Nandy, P., Graphical analysis of DNA sequences structure: II. Relative abundance

of nucleotides in DNAs, gene evolution and duplication, Curr. Sci. 68(1995) 75–85.

[21] Nandy, A., Graphical analysis of DNA sequence structure: III. Indication of evolutionary distinctions and characteristics of introns and exons, Curr. Sci. 70(1996) 661–668.

[22] Nandy, A., Two-dimensional graphical representation of DNA sequences and intron–exon discrimination in intron–rich sequences, Comput. Appl. Biosci. 12(1996) 55–62.

[23] Nandy, A., Graphical representation of long DNA sequence, Curr. Sci. 66(1994) 821.

[24] Balaban, A. T., pure Appl. Chem. 55(1983) 199.

[25] Guo, X. F.; Randic, M.; Basak, S. C., A novel 2-D graphical representation of DNA sequence of low degeneracy, Chemical Physics Letters, 350(2001) 106–112.

[26] Guo, X. F.; Nandy, A., Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, Chemical Physics Letters, 369(2003) 361–366.

[27] Randic, M.; Guo, X. F.; Basak, S. C., On the characterization of DNA primary sequence by triplet of nucleic acid bases, J. Chem. Inf. Comput. Sci. 41(2001) 619–626.

[28] Liao, B.; Wang, T. M., New 2D Graphical representation of DNA sequences, Journal computational chemistry, 11(2005) 1364–1368.

[29] Liao, B.; Wang, T. M., Analysis of similarity of DNA sequences based on triplets, J. Chem. Inf. Comput. Sci. , 44(2004) 1666–1670.

[30] Liao, B.; Zhang, Y. S., Kequan Ding,Tianming Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, Journal of Molecular Structure: THEOCHEM, 717(2005) 199–203.

[31] Liao, B.; Tan, M. S.; Ding, K. Q., A 4D representation of DNA sequences and its application, Chemical Physic Letters, 402(2005) 380–383

[32] Stephn, S. T. Y.; Wang, J. S.; Niknejad, A. C.; Lu, X.; Jin, N.; Ho, Y. k., DNA sequence representation without degeneracy, Nucleic Acids Research, 31(2003) 3078–3080.

[33] Luo, L. F.; Li, W. J., A critical review on the correlation properties of coding DNA sequences, 内蒙古大学学报 (自然科学版), 27(1996) 622–626.

[34] 王守源，李晓琴，罗辽复，氨基酸分类与蛋白质二级结构相关性，内蒙古大学学报 (自然科学版), 33(2002) 423–427.

[35] Luo, L. F.; Lui, C., The comparative study of DNA walk in 3–, 2– and 1–dimensional base space, 内蒙古大学学报 (自然科学版), 27(1996) 781–789.

[36] 吴晓明，宋长新，王波，程敬之，隐马尔可夫模型用于蛋白质序列分析，生物医学工程杂志，19(2002) 455–458.

[37] Waterman, M. S., Introduction to Computational Biology: Maps, Sequences and Genomes. Chapman Hall, London, 1995.

[38] Bai, F. L.; Liu, Y. Z.; Wang, T. M., A representation of DNA primary sequences by random walk, Mathematical Biosciences, (Revised).

[39] He, P. A.; Wang, J., Characteristic sequences for DNA primary sequence, J. of Chem. Inf. Comput. Sci. 42(2002) 1080–1085.

[40] He, P. A.; Wang, J., Numerical zharacterization of DNA primary sequence, Internet Elec. J. Mol. Des. 12(2002) 668–674.

[41] 张庆友，许禄， DNA 编码序列的图形表示及相似度计算，高等学校化学学报， 23(2002) 1255–1258.

[42] 刘西奎，李艳，许进，DNA 序列二维图表示和有关分析，自然科学进展，14(2004) 1032–1038.

[43] 陈惟昌，陈志华，陈志义等，DNA 序列高维空间数字编码的运算法则，生物物理学报，17(2001)

542−549.

[44] 陈志华，陈惟昌，陈志义等， DNA 序列的“双符三阶”图形编码，生物物理学报， 19(2003) 167−170.

[45] 张新生，王梓坤，生命信息遗传中的若干数学问题，科学通报， 45(2000) 113−119.

[46] 张春霆，用几何学方法分析 DNA 序列，中国科学基金， 6(2000) 298−299.

[47] 张任，张春霆， Z 曲线，显示和分析 DNA 序列的直观工具，自然杂志， 17(2001) 34−38.

[48] Zhang, C. T.; Zhang, R., Analysis of distribution of bases in the coding sequences by a diagrammatic techniqut, Nucl. Acids Rev. 19(1991) 6313−6317.

[49] Zhang, R. and Zhang, C. T., Z−curve, an intuitive tool for visualizing and analyzing the DNA sequences, J. Biomol. str. Dyn. 11(1994) 767−782.

[50] Zhang, C. T., A symmetrical theory of DNA sequences and its applications, J. Theor Biol. 187(1997) 297−306.

[51] Zhang, C. T.; Chou, K. C., A graphic approach to analyzing codon usage in 1562 Escherihia coli protin coding sequences, J. Mol. Biol. 238(1994) 1−8.

[52] Zhang, C. T.; Zhang, R., Skewed distribution of protein secondary structure contents over the conformation triangle, Protein engineering, 12(1999) 807−809.

[53] Zhang, C. T.; Zhang, R., S curve, a graphic representation of protein secondary structure sequence and its applications, Biopolymers, 53(2000) 539−549.

[54] Zhang, C. T.; Zhang, R., A graphic approach to evaluate algorithms of secondary structure prediction, J. Biomol. Str. Dyn. 17(2000) 829−841.

[55] Zhang, C. T.; Lin, Z. S.; Yan, M.; Zhang, R., A novel approach to distinguish between intron-containing and genes based on the format of Z curve, J. Theo. Biol. 192(1998) 467−473.

[56] Zhang, C. T.; Zhang, R. and Ou, H. Y., The Z curve database: a graphic representation of genome sequences, Bioinformatics, 19(2003) 593−599.

[57] Yuan, C. X.; Liao, B.; Wang, T. M., New 3D graphical representation of DNA sequence and their numerical characterization, Chemical Physics Letters, 397(2003) 412−417.

[58] Peng, C. K.; Buldyrev, S. V.; Goldbergeretal, A. L., Nature, 356(1992) 168−170.

[59] Dodin, G.; Pierre, V.; Levoiretal, P., J. Theor. Biol. 206(2000) 323−326.

[60] Tsonis, A. A.; Kumar, P.; Elsneretal, J. B., Phys. Rev. E, 53(2)(1996) 1828−1834.

[61] Luo, L. F.; Tsai, L.; Zhou, Y. M., J. Theor. Biol. 130(1988) 351−361.

[62] Luo, L. F.; Tsai, L., Chem. Phys. Lett. 5(1988) 421−424.

[63] Arneodo, A.; d'Aubenton−Carafa, Y.; Bacry, E. etal, Physica D, 96(1996) 291−320.

[64] Arneodo, A.; d'Aubenton−Carafa, Y.; Bacry, E. etal, Eur. Phys. J. B, 1(1998) 259−263.

[65] Voss, R., Phys. Rev. Lett. 68(25)(1992) 3805−3808.

[66] Snel, B.; Bork, P.; Huynen, M. A., Genome phylogeny based on gene content, Nat. Genet. 21(1999) 108−110.

[67] Lempel, A.; Ziv, J., On the complexity of finite sequences, IEEET. Inform. Theory, 22(1976) 75−81.

[68] Hasan, H. O.; Khalid, S., A new sequence distance measure for phylogenetic tree construction, Bioinformatics,19(2003), 2122−2130.

[69] Saitou, N.; Nei, M., The neighbor−joining method:a new method for reconstructing phylogenetic tree, Mol. Biol. Evol. 4(1987) 406−425.

[70] Ming, L.; John, H. B., Paul, K., An information based sequence distance form unaligned whole genome protein squence, Bioinformatics, 18(2002) 100−108.

[71] Rowe, D. L.; Honeycutt, R. L., Phylogenetic relationships, ecological correlates, and molecular evolution within the cavioidea (Mammalia, Rodentia), Mol. Biol. Evol. 19(2002) 263−277.

[72] Benedetto, D.; Caglioti, E.; Lereto, V. etal., Language tree and zipping, Physical Review Letters, 88(2002) 1−5.

[73] Bai, F. L.; Wang, T. M., The construction of phylogenetic tree by graphic representation of DNA sequences, WSEAS Transactions on Information Scince and Applications, 2(2005).

[74] Bai, F. L.; Wang, T. M., Phylogenetic analysis by graphic representation of DNA sequences, Proceeding of the 5th WSEAS Int. Cont. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17−19, 2005 (463−467).

[75] 王浩，统计方法求系统树 [J]，云南大学学报 (自然科学版)， 24(2002) 199−201.

[76] 杨子恒，分子进化树的统计推断，遗传， 17(1995) 92−96.

[77] Reusken, C. B. E. M.; Bol, J. F., Structural elements of the 3′−terminal coat protein binding site in alfalfa mosaic virus RNAs, Nucleic Acids Research, 14(1996) 2660−2665.

[78] Bafna, V.; Muthukrisnan, S.; Ravi, R., Comparing similarity between RNA strings, Computer Science, 937(1995) 1−14.

[79] corpet, F.; Michot, B., RNAlign program: alignment of RNA sequences using both promary and secondary structures. Computer. Appl. Biosci. 10(1995) 389−399.

[80] Le, S. Y.; Nussinov, R., Mazel, J. V., Tree graphs of RNA secondary structures and their comparsion, Computer Biomed.Res. 22(1989) 461−473.

[81] Le, S. Y.; Onens, J.; Nussinov, R.; Chen, J. H.; shapiro, B.; Mazel, J. R., RNA secondary structures: comparsion and determination of frequenctly recurring sunstructures by consensus , Computer Biomed.Res. 5(1989) 205−210.

[82] Shapiro, B., An algorithm for comparing multiple RNA secondary structures, Computer. Appl. Biosci. 4(1998) 387−393.

[83] Shapiro, B.; Zhang, K., Comparing multiple RNA secondary structures using tree comparisons, Computer. Appl. Biosci. 6(1990) 309−318.

[84] Zhang, K., Computing similarity between RNA secondary structures , Pro. IEEE. Internat. Joint Symp On Intelligence and Sytems Rockviue, Maryland. May, (1998) 126−132.

[85] Koper−Zwarthoff, E. C.; Brederode, F. Th.; Walstra, P.; Bol, J. F., Nucleic Acids Research, 7(1979) 1887−1900.

[86] Scott, S. W. and Ge, X., J. Gen. Virol. 76(1995) 957−963

[87] Koper−Zwarthoff, E. C.; Brederode, F. Th.; Walstra, P.; Bol, J. F., Nucleic Acids Research, 8(1980) 3307−3318.

[88] Cornelissen, B. J. C.; Janssen, H.; Zuidema, D.; Bol, J. F., Nucleic Acids Research, 12(1984) 2427−2437.

[89] Alrefai, R. H.; Shicl, P. J.; Domier, L. L.; D'Arcy, C. J.; Berger, P. H.; Korban, S. S., J. Gen. Virol. 75(1994) 2847−2850.

[90] Scott, S. W. and Ge, X., J. Gen. Virol. 76(1995) 1801−1806.

[91] Bachman, E. J.; Scott, S. W.; Xin, G.; Bowman Vance, V.; Virology, 201(1994) 127−131.

[92] Houser−Scott, F.; Baer, M. L., Liem, K. F., Cai, J. M.; Gehrke, L., J. Virol. 68(1994) 2194−2205.

[93] EMBL/GenBank/DDBJ databases. Accession no. X86352.

[94] Zupan, J.; Randic, M., Algorithm for coding DNA sequences into "Spectrum−Like" and "Zigzag" representations, J. Chem. Inf. Model. 45(2005) 309−313.

[95] Randic, M.; Zupan, J., On graphical representations and graph theoretical characterization of

DNA and proteins, Advances in Quantum Chemistry(special issue on Chemical Graph Theory, D.J.Klein, guest editor), in press.

[96] Liu, Y.; Gou, X.; Xu, J.; Pan, L.; Wang, S., Some notes on 2−D graphical representations of DNA sequences, J. Chem. Inf. Comput. Sci. 42(2002) 529−533.

[97] Jeffrey, H. I., Chaos game representation of gene structure, Nucleic Acid Res. 18(1990) 2163−2170.

[98] Goldman, N., Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representation of DNA sequences, Nucleic Acid Res. 21(1993) 2487−2491.

[99] Basu, S.; Pan, A.; Dutta, C.; Das, J., Chaos game representation of proteins, J.Mol. Graphics Modelling, 15(1997) 279−289.

[100] Randic, M.; Nandy, A.; Basak, S. C.; Plavšić, D., On the numerical characterization of DNA sequences, J. Math. Chem. Submitted for publication.

[101] Bai, F. L.; Zhu, W.; Wang, T. M., Analysis of similarity RNA secondary structures, Chemical Physics Letters, 408(2005) 258−263.

[102] Liao, B.; Wang, T. M., A 3D Graphical representation of RNA secondary structure, J. Biomol. Struc. Dynamics, 21(2004) 827−832.

[103] Liao, B.; Ding, K. Q.; Wang, T. M., On a seven−dimensional representation of RNA secondary structures, Lecture Series on Compute and Computationa Science, 1(2004) 310−312.

[104] Liao, B.; Ding, K. Q.; Wang, T. M., On a six−dimensional representation of RNA secondary structures, J. Biomol. Struc. Dynamics, 22 (2005) 455−464

[105] Liao, B.; Ding, K. Q.; Wang, T. M., On a 6D graphical representation of RNA secondary structure, Journal Biomolecular structure Dynamics, 22(2005) 455−464.

[106] LiW, Are spectral analysis useful for DNA sequence analysis [P] DNA in Chromatin, At the Frontiers of Biology, Biophysics, and Genomics, Arcachon, France, 2002, March 23−293.

[107] Buldyrev, S. V.; Goldberger, A. L.; Havlin, S. etal., Long−range correlation properties of coding and noncoding DNA sequences: GenBank analysis [J]. Physical Review E, 51(1995) 5084−5091.

[108] Anastassiou, D., Frequency−domain analysis of biomolecular sequence, Bioinformatics, 16(2000) 1073−1081.

[109] Randic, M., 2-D Graphical representation of proteins based on virtual genetic dode, SAR QSAR Environ. Res. 15(3) (2004) 147−157.

[110] Randic, M.; Zupan, J., Highly compact 2D graphical representation of DNA sequences, SAR QSAR Environ. Res. 15(3) (2004) 191−205.

[111] Randic, M.; Zupan, J., Balaban, A. T., Unique graphical representation of protein sequences based on nucleotide triplet codons, Chemical Physics Letters, 397(2004) 247−252.

[112] Randic, M., Graphical representation of DNA as 2-D map, Chemical Physics Letters, 386(2004) 468−471.

[113] Randic, M., On graphical and numerical characterization of protemics maps, J. Chem. Inf. Comput. Sci. 41 (2001) 1330−1338.

[114] Randic, M.; Zupan, J. and Novic, M., On 3-D graphical representation of proteomics maps and their numerical charaterization, J. Chem. Inf. Comput. Sci. 41 (2001) 1339−1344.

[115] Randic, M., Quantitative charaterization of proteomics maps by matrix invariants, In: Conn, P.M., ed., Handbook of Proteomic Methods (Humana Press Inc., Totowa, NJ), (2003) 429−450.

[116] Randic, M.; Vracko, M. and Novic, M., On charaterization of dose variations of 2-D proteomics maps by matrix invariants, J. Proteome Res. 1 (2002) 217−226.

[117] Randic, M.; Zupan, J. and Novic, M.; Gute, B. and Basak, S. C., Novel matrix invariants for charaterization of changes of proteomics maps, SAR QSAR Environ. Res. 13 (2002) 689−703.

[118] Bai, F. L.; Wang, T. M., A 2-D graphical representation of protein sequences based on nucleotide triplet codons , Chemical Physic Letters ,(SCI) 413(2005) 458−462

[119] Liao, B.; Wang, T. M., Nunerical charaterization and similarity analysis of neurocan gene, Journal of Theoretical Biology, (Revised).

[120] Li, C. and Wang, J., New invariant of DNA sequence, J. Chem. Inf. Model. 45(2005) 115−120.

[121] Lan, M. L.; Carpendale, M. S. T., Supporting detail−in−context for the DNA Representation, H−Curves, 1998.

[122] Randic, M.; Vracko, M.; Novic, M., QSPR/QSAR Studies by molecular descriptors, in: M. V. Diudea(ED), Nova Science, Huntington, (2001) 145.

[123] Bajzer, Z.; Randic, M.; Plasic, D.; Basak, S. C., Novel map descriptors for characterization of toxic effects in proteomics maps, Journal of Molecular Graphics and Modelling, 22(2003) 1−9.

[124] Tan, Y. D., A new measure for DNA sequence information, J. Biomath.(Chinese), 15(1999) 45−54.

[125] Solovyev, V. V., Fractal graphical representatioin and analysis of DNA and protein sequences, BioSystems, 30(1993) 137−160.

[126] Speed, T. P., Biological sequence analysis, ICM2002, Vol III, 97−106.

[127] Switzer, C.; Moroney, S. E.; Benner, S. A., Enzymatic incorporation of a new base pair into DNA and RNA, J. Am. Chem. Soc. 111(1989) 8322−8323.

[128] Reijmers, T. H.; Wehrens, R.; Buydens, L. M. C., The influence of different structure representations on the clustering of an RNA nucleotides data set, J. Chem. Inf. Comput. Sci. 41(2001) 1388−1394.

[129] Lewin, B., Genes VII. Oxford University Press, 1999.

[130] Mount, D. W., Bioinformatic: sequence and genome analysis, 生物信息学：序列与基因组分析，科学出版社， 2002.

[131] 寿天德等，《现代生物学导论》中国科学技术出版社， (1998) 106−110

[132] 刘次全，信使 RNA 三维遗传信息的研究，中国科学基金， 4(1998), 32−34.

[133] 阎隆飞，孙之荣，蛋白质分子结构，清华大学出版社， 1999.

[134] 朱浩译，计算分子生物学导论，科学出版社， 2003

[135] Jiang, T.; Xu, Y.; Zhang, M. Q., Current topics computation molecular biology, 计算分子生物学前沿课题，清华大学出版社， 2002.

[136] 王翼飞等译，计算分子生物学－算法逼近，化学工业出版社， 2004.

[137] 沈世镒，生物序列突变与比对的结构分析，科学出版社， 2004.

[138] 郝柏林，张淑誉，生物信息学手册，上海科学技术出版社， 2000.

[139] 来鲁华等，蛋白质的结构预测与分子设计，北京大学出版社， 1993.

[140] 宗孔德，胡广书，数字信号处理 [M]., 北京清华大学出版社， 1998.

攻读博士学位期间发表学术论文情况

1. Analysis of Similarity between RNA secondary structures, Chemical Physic Letters, 408(2005) 258–263. (第五章第二节) (SCI)
2. A 2–D graphical representation of protein sequences based on nucleotide triplet codons, Chemical Physic Letters, 413(2005) 458–462. (第七章第二节) (SCI)
3. The construction of phylogenetic tree by graphic representation of DNA sequences, WSEAS Transactions on Information Scince and Applications, 2(2005). (第四章第二节) (EI)
4. Phylogenetic analysis by graphic representation of DNA sequences, Proceeding of the 5th WSEAS Int. Cont. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17–19, (2005) 463–467. (第四章第二节)
5. 近似求解 Cahn–Hillard 方程的拟谱方法，吉林大学学报 (理学版), 41(3)(2003) 262–268.
6. 拓扑指数在生物序列比较中的应用，生物数学学报，已接受，(第二章第二节)
7. A Representation of DNA primary sequences by random walk, Mathematical Biosciences, Revised. (第三章第二节) (SCI)
8. On graphical and numerical representation of protein sequences, J. Biomol. Struc. Dynamics, Revised. (第七章第三节) (SCI)
9. Algorithm for coding RNA secondary structure sequences into "Spectrum–like" and "Zigzag" representations, Journal of Mathematical Chemistry, submitted. (第六章第二节) (SCI)
10. 利用 DNA 序列的图形表示构建物种系统树的方法，生物数学学报，已投稿. (第四章第二节)

创新点摘要

1. 在生物序列的二维图形表示的基础上，利用 Balaban 指数和信息分布指数以及矩阵不变量—距离矩阵主对角线以外的次对角线之和的平均值给出了生物序列的比较方法。另外，利用 1-D 随机游动来描述 DNA 序列，把 DNA 序列转化成随机数字序列，利用随机数字序列的特征值，给出了 DNA 序列新的比较方法.
2. 在 DNA 序列和蛋白质序列的三维图形表示的基础上，利用图的不变量分别给出了物种线粒体 DNA 序列之间距离度量和蛋白质序列之间的距离度量，进而分别定义了物种之间和蛋白质序列之间的进化距离，并利用距离法构建了生物系统进化树和蛋白质序列进化树.
3. 分别给出了 RNA 二级结构的 1-D 、 2-D 、 3-D 和 6-D 表示法以及复平面上的 2-D 表示，并且利用这些图形表示的数据特征给出了比较 RNA 二级结构序列相似性的方法. 进一步利用 1-D 图形表示给出了关于 RNA 二级结构序列频谱分析的方法.
4. 在 DNA 三联体密码子表示的基础上，在半复平面上给出了蛋白质序列的非退化的 2-D 图形表示和蛋白质序列的一种数值刻划. 还有给出了蛋白质序列的 3-D 图形表示以及蛋白质序列的相似性分析方法.

致　谢

本学位论文是在导师王天明教授的悉心指导下完成的. 从选题, 到实际写作过程中, 都渗透着导师的汗水和心血. 王老师自始至终都给予了我极大的帮助. 王老师渊博的知识, 严谨求实的治学态度, 兢兢业业的工作精神, 宽容随和的待人风格使我受益匪浅, 终身难忘, 并激励我在今后的工作中不断超越自我. 在此作为学生的我向导师三年来的支持鼓励, 谆谆教导, 悉心栽培和生活上的关心表示深深的感谢. 由衷感谢师母三年来对我的关心与帮助.

作者十分感谢数学系各位老师的支持和帮助, 特别感谢王军教授, 唐焕文教授, 侯中华教授, 李凤泉教授, 张鸿庆教授、郑斯宁教授, 邱瑞锋教授, 王希诚教授, 卢玉峰教授, 王立伟教授, 冯红副教授, 杨彦春副教授, 蔡晶老师对我的关心和帮助.

感谢廖波博士后、姚玉华博士、王伟平博士、刘立伟博士、袁春欣博士、刘娜博士、郭颖博士、乌云高娃博士、白乙拉博士、吉日木图博士、刘迎照硕士、王文文硕士、代琦硕士、刘晓庆硕士、杨红硕士以及生物讨论班的各位师弟师妹在学业和生活方面的关心帮助和支持.

我要特别感谢丈夫在自己三年学习中给予的理解、支持和鼓励. 同时, 将女儿培养成为一名品学兼优的大学生.

感谢女儿对我的理解与支持.

最后要感谢那些在这里不能一一提到的同学和好友, 感谢所有帮助和关心过我的人们, 谢谢你们!

大连理工大学学位论文版权使用授权书

本学位论文作者及指导教师完全了解“大连理工大学硕士、博士学位论文版权使用规定”，同意大连理工大学保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅。本人授权大连理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，也可采用影印、缩印或扫描等复制手段保存和汇编学位论文。

作者签名：

导师签名：

2006 年 6 月 25 日