摘　要

随着人类和一些模式生物基因组计划的相继完成或全面实施，生物学研究的重点正从积累数据向分析解释这些数据过渡，生物信息学（也称计算分子生物学）便应运而生。它的研究内容十分丰富，例如，序列比较、计算机辅助基因识别、分子进化和比较基因组学、RNA 和蛋白质结构预测、遗传密码及其起源、序列重叠群装配、基于结构的药物设计等等，都是生物信息学中重要的研究领域．其中大多数领域的研究工作都有一个共同的需求，就是常常需要给出生物学数据的数学上的描述，因此，生物大分子的数学描述便成为生物信息学中一个非常基础又十分重要的课题。

本文的主要工作包括以下几个方面：

在第一章，针对原有图形表示的缺陷，我们从不同的角度在不同的层次上给出了生物序列的三种图表示。首先，直接从 DNA 原始序列出发，通过赋予四种碱基四个 3 维空间中的向量给出了 DNA 序列一种 3 维图形表示，同时在 DNA 序列的特征序列的基础上提出了两种 2 维图形表示：“双水平线”图和“梯状”图，这两种 2 维表示都既考虑了序列本身的线性结构，又考虑了四种碱基的化学结构．最后，面向所有生物序列的图形表示，从整体上提出了有向图的概念．有向图表示不仅弥补了现有图形表示的许多不足，还为生物序列的数值刻画提供了新的途径。

在第二章，提出了一个基于矩阵范数的新的序列不变量 —— ALE 指标，它与目前应用最为广泛的序列不变量 —— 最大特征值等效但它的计算非常容易，这使得基于不变量的比较方法在完全基因组比较及其相关研究领域中的应用具有了可行性。同时，我们还就某种特殊情况下最大特征值所反映的信息是否全面进行了探讨，并提出了伪迹的概念．此外，在有向图的基础上提出了生物序列的上三角矩阵表示，并对现有序列不变量在上三角矩阵情况下的兼容性作了讨论．为了更好地反映序列中元素，尤其是它们之间的序关系所包含的信息，本章最后一节从一般数字序列出发构造出一种特殊的链（全序集），在此基础上提出了 DNA 序列的正规化相对熵，并简要讨论了 DNA 序列基于正规化相对熵的 12 维向量表示在酿酒酵母基因组的蛋白质编码基因识别中潜在的应用．

在第三章，利用代数学中的同态思想对 DNA 序列进行粗粒化描述，提出了 DNA 序列的逻辑表示，并将这一概念推广到蛋白质序列．同时，给出了 (0,1) 序列的广义 LZ 复杂度，并将其和正规化相对熵分别应用到 DNA 及蛋白质序列的相似性分析．此外，根据 RNA 二级结构的特点给出了 RNA 二级结构的影子序列，并结合序列复杂性，对 9 种病毒的 RNA 二级结构进行了比较．

在最后一章，利用 DNA 序列的正规化相对熵和 Fisher 线性判别法对酿酒酵母基因

组序列进行基因识别. 我们将识别的准确度提高到了 96%，得到了一个酿酒酵母基因组中基因总数为 5873 的估计，与普遍接受的 5800-6000 相符.

关键词：生物信息学；生物大分子；DNA；RNA；蛋白质；图表示；数值刻画；逻辑序列；影子序列；序列复杂度；序列比较；基因识别

Abstract

With the completion/development of the genome projects of human and some model organism, the focus of biology shifts from accumulation of biological data to the analysis and interpretation of them, and thus bioinformatics, also named computational molecular biology, emerges as a new and developing interdiscipline. The research area of bioinformatics is very wide, which includes sequence comparison, gene recognition by computers, molecular evolution and comparative genomics, RNA and protein structure prediction, codon origin and evolution of the genetic code, assembly of contigs, structure-based drug design, and so on. Most of them have a common requirement — the biological data must be transferred into a certain mathematical description, this leads to that the mathematical description of the biological macromolecules becomes a basic but very important topic in bioinformatics.

The main contents of this thesis are listed as follows:

In Chapter 1, we propose three kinds of graphical representations for biological sequences from different points of view. Firstly, we introduce a 3-D graphical representation of DNA primary sequences by taking four special vectors in a 3-D space to represent the four nucleic acid bases A, G, C, and T, respectively. Secondly, based on the characteristic sequences of a DNA primary sequence, we introduce two 2-D graphical representations of DNA sequences: one is the "two horizontal lines" graph, and the other is the "ladder-like" graph, each of which considers the sequences' structure as well as the chemical structure of DNA sequences. Finally, we introduce a directed graphical representation of biological sequences, which not only overcomes the serious drawback of the existing graphical representations, but also provides us with a new way of characterizing bio-sequences numerically.

In Chapter 2, we propose a new sequence invariant named "ALE-index", which is based on norms of a matrix. The ALE-index can be regarded as an approximation of the leading eigenvalue, the currently most widely used invariant. Different from the leading eigenvalue, the ALE-index is very simple for calculation so that it can be directly used to handle long biological sequences. Therefrom, it becomes practicable to compare the whole genomes by the invariant-based sequence comparison method. Meanwhile, we find that the information reflected only by the leading eigenvalue might not be comprehensive in a special case. So we suggest, in this case, use the so-called "pseudo-trace" instead of the leading eigenvalue to characterize DNA sequences. Moreover, we describe a scheme that transforms the directed graph of a biological sequence into an upper triangular matrix, and investigate whether or not the existing sequence

invariants are compatible for the upper triangular matrix representation. Finally, to reflect the information on elements of a sequence and, especially, the order relation among them, we construct a *chain* (totally ordered set) from a sequence of numbers, and then introduce the normalized relative-entropy. A potential application of a 12-component vector based on the normalized relative-entropy associated with a DNA sequence to discriminating protein coding and non-coding sequences in the yeast genome is briefly discussed.

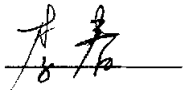
In Chapter 3, based on the ideas of homomorphism in algebra, we describe a DNA sequence in the way of coarse graining, and propose the logical representation (LR) for DNA primary sequences. Furthermore, we present a generalized LZ complexity for (0,1)-sequences. The examination of the similarity among DNA sequences of the full *beta*-globin genes of 11 species shows the utility of our approach. We also generalize the concept of the logical representation of DNA primary sequences to the protein primary sequences. Similarity and dissimilarity analysis based on the normalized relative-entropy of logical sequences of protein are given for eight protein sequences. Besides these, we introduce the shadow sequence for RNA secondary structure. By combining it with the symbolic sequence complexity, we compare RNA secondary structures of nine viruses.

In the last chapter, based on the normalized relative-entropy of DNA sequences, we use the Fisher discriminant method to find protein coding genes in the yeast genome. Cross-validation tests demonstrate that the accuracy of the algorithm is 96%. The total number of protein coding genes in the yeast *S. cerevisiae* genome is estimated to be less than or equal to 5873, significantly coincident with the widely accepted range 5800-6000.

Keywords: Bioinformatics; Biological macromolecule; DNA; RNA; Protein; Graphical representation; Numerical characterization; Logical sequence; Shadow sequence; Sequence complexity; Sequence comparison; Gene recognition.

独创性说明

作者郑重声明：本博士学位论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得大连理工大学或者其他单位的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

作者签名：  日期：2006.6.16

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：

2008 年 6 月 16 日

0 绪　论

20 世纪是科学技术迅速发展的世纪，物理和化学的发展使我们可以清楚地认识物质的组成，从分子、原子、电子等各个层次上深入地了解微观世界，天文技术、空间技术的发展使得我们可以了解地球以外的客观世界，以电子信息技术为龙头的工业技术的飞速发展，使得我们可以不断地改造世界，甚至为人类更加舒适地生活创造新的世界，而生命科学的发展，则使我们能从器官、组织、细胞、生物大分子等各个层次认识生命的物质基础。

0.1　生物信息学产生的背景

1953 年 4 月 25 日，詹姆斯·沃森与同在剑桥大学的合作伙伴弗朗西斯·克里克一起，在《自然》杂志上发表了一篇仅两页的论文，提出了 DNA 的结构和自我复制机制，揭开了分子生物学的新篇章。50 年后，人们迎来了又一个激动人心的时刻，那就是在 2003 年 4 月 14 日，美，英，日，法，德和中国科学家经过 13 年努力共同完成了人类基因组计划 (Human Genome Project, HGP)，比原计划提前两年，在人类揭示生命奥秘、认识自我的漫漫长路上又迈出重要一步。人类基因组计划是美国在 1990 年提出实施的一项伟大的科学计划，与阿波罗登月计划、曼哈顿原子弹计划同称为人类自然科学史上的三大计划，其目标是用大约 15 年时间，完成人类所有染色体中 3×10^9 个碱基对 (bp, base pair) 的序列测定。人类基因组计划的成果是一个人类遗传信息数据库，是一本指导人类进化的“说明书”。它不仅可以揭示人类生命活动的奥秘，而且人类几千种单基因遗传性疾病和严重危害人类健康的多基因易感性疾病的致病机理有望得到彻底阐明，为这些疾病的诊断、治疗和预防奠定基础。同时，人类基因组计划的实施还将带动医药业、农业、工业等相关行业的发展，产生极其巨大的经济效益和无法估量的社会效益。

随着 HGP 的顺利完成，和诸如大肠杆菌、啤酒酵母、线虫、果蝇、小鼠、鸡、拟南芥、水稻、玉米等等模式生物的基因组计划的相继完成或全面实施，DNA / 蛋白质序列数据正以惊人的速度增长，在此基础上派生和整理出来的数据库已达 500 余个。这一切构成了一个生物学数据的海洋。我们知道，生物学是一门实验科学，也是一门发现科学。通过实验发现新的现象、新的生物学规律，经过分析和归纳总结，提炼出新的生物学知识。在这个过程中，需要对实验数据进行处理和理论分析，在此基础上解释实验现象，认识实验现象发生的本质，探索固有的生物学规律，进而了解和掌握生命的物质基础和

生命的本质. 生物数据积累速度不断加快, 对生物数据的科学分析方法和实用分析工具提出了更新、更高的要求.

传统分子生物学实验往往是集中精力研究一个基因、一条代谢路径, 手工分析完全能够胜任. 然而, 一方面, 现在我们面对的是海量并且仍在不断迅速增加的生物学数据. 一次只分析一个生物分子的传统的生物学已经无法满足要求. 换句话说, 现在需要的是同时分析成千上万个生物分子, 是自动分析. 同时, 面对这么多生物分子数据, 不可能用实验的方法去详细研究每一条序列, 必须先进行信息处理和分析, 去粗取精, 去伪存真. 通过预处理, 发现有用的线索, 在此基础上进行有针对性、有明确目的的分子生物学实验. 另一方面, 从生物分子数据本身来看, 各种数据之间存在着密切的关系, 如 DNA 序列与蛋白质序列、基因突变与疾病等, 这些联系反映了生物学的规律. 但是, 这些关系可能是非常复杂的, 是我们未知的, 是简单的统计方法难以分析的. 对于这些复杂的关系, 必须运用现代信息学的方法去分析, 去研究.

综上所述, 生物学已经不再单单是原来的观察和实验的科学, 理论和计算对生物学的进步正在发挥越来越重要的作用, 这就催生了生物信息学 (又称计算分子生物学, 现在人们常常不加区分地使用这两个名称) 这门崭新的交叉科学. 这是一门运用数学、信息科学、计算机科学和系统科学的理论与方法研究生命现象、分析和处理呈指数增长的生物学原始数据并进行加工、分析和建立计算模型的一门学科. 普遍认为, 生物信息学是当今生命科学和自然科学中最关键、最重要的部分, 是 21 世纪自然科学的核心领域之一 [1]- [8] .

0.2 生物信息学的研究对象

生物体是一个复杂的系统, 生命过程是一个极端复杂的过程, 需要物质和能量的支持. 生物体也是一个信息系统, 该系统控制着生物的遗传、生长和发育. 所有的信息存贮在生物体内, 存贮在遗传物质中. 在生命科学研究方面, 人们已经逐渐认识到, 不仅需要用物理、化学和生物学方法研究生命的物质基础、能量转换、代谢过程等, 还需要用信息科学方法研究生命信息特别是遗传信息的组织、复制、传递、表达及其作用, 否则难以理解生命的工作机制, 难以揭示生命的奥秘. 从生物学的观点来看, 细胞是生命的基本单位, 而从信息科学的观点来看, 细胞则是存贮、复制和传递遗传信息的系统.

生物系统通过存贮、修改、解读遗传信息和执行遗传指令形成特定的生命活动, 生长发育, 产生生物进化. 从信息学的角度来看, 生物分子是生物信息的载体. 生物分子至少携带着三种信息, 即遗传信息、与功能相关的结构信息、进化信息. 俗话说 “种瓜得瓜, 种豆得豆”, 这是对生物遗传现象的生动描述. 地球上的所有生物, 上至 “万物之灵” 的人类, 下至细菌的 “寄生虫” —— 噬菌体, 都表现着遗传现象, 能够复制出新的一代, 这是生命延续和种族繁衍的保证. 生物的复制由基因所决定, 复制是生命的基本

特征，但不是生命的全部特征．计算机程序可以自动复制大量的拷贝，但是这些程序不是活动的生命，活动的生命是不断变化的。绝大多数生命体可以从周围的环境中摄取物质，获取能量，并将所摄取的物质转换为其自身的一部分。计算机程序虽然可以拷贝，但是这种拷贝往往是绝对真实的拷贝，毫厘不差．而生物体在繁殖和遗传的过程中并非一成不变，后代与亲代存在着差异．正因为有遗传差异的存在，才有生物的进化．

生物信息学主要研究两种信息载体，即核酸（ DNA 、 RNA 分子）和蛋白质分子。

0.2.1 核酸

核酸是遗传物质．核酸分为脱氧核糖核酸（ DNA ）和核糖核酸（ RNA ）。 DNA 主要存在于细胞核中，但细胞质里的线立体、叶绿体中也含有少量 DNA ， RNA 则主要分布在细胞质中。遗传的主要物质基础是 DNA ，但有时也是 RNA（如病毒的遗传物质）.

核酸是由称为核苷酸（ nucleotide ）的小分子生成的聚合物。核苷酸还可以进一步分解成核苷（ nucleoside ）和磷酸，核苷进一步水解生成碱基（ base ）和戊糖．所以，核酸的基本结构单位是核苷酸，其组成方式为碱基 - 戊糖 - 磷酸（见图 1 ）.

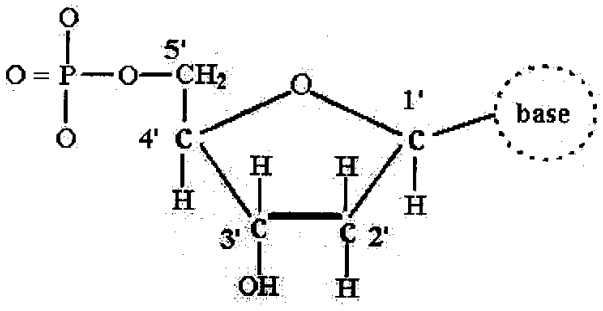


图 1: 核苷酸分子结构示意图

DNA 和 RNA 所含的戊糖不同：前者中的戊糖是脱氧核糖，而后者的则是核糖．DNA 和 RNA 在组成上的另一个区别体现在它们所含的碱基组成上． DNA 中的碱基有 4 种，分别是腺嘌呤（ adenine，简写作 A ）、鸟嘌呤（ guanine，简写作 G ）、胞嘧啶（ cytosine，简写作 C ）和胸腺嘧啶（ thymine，简写作 T ）． RNA 中没有胸腺嘧啶 T，取而代之的是尿嘧啶 U（ Uracil ）．五种碱基的分子结构示意图如图 2 所示．

可见，仅就 DNA 或者 RNA 分子而言，不同核苷酸之间的区别仅在于它们所含的碱基不同．因此， A 、 G 、 C 、 T（ U ）也常被用来直接表示相应的核苷酸．核苷酸相互连接形成长的多核苷酸链．由四种脱氧核苷酸连接而成的长链高分子多聚体为 DNA 分子的一级结构． DNA 分子中第一个核苷酸的 3′- 羟基与第二个核苷酸的 5′- 磷酸基脱水形

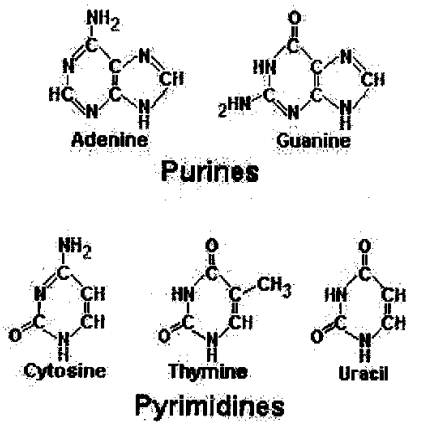


图 2: **五种碱基 A, G, C, T, U 的分子结构示意图**

成 3′,5′- 磷酸二酯键，第二个核苷酸的 3′- 羟基又与第三个核苷酸的磷酸基脱水形成 3′,5′- 磷酸二酯键，依此类推，形成线性多聚体。 DNA 分子中第一个核苷酸的 5′- 磷酸与最末一个核苷酸的 3′- 羟基都未参与形成 3′,5′- 磷酸二酯键，故分别称为 5′- 磷酸端 (或 5′- 端) 和 3′ 羟基端 (或 3′- 端)。

DNA 蕴涵的复制机制的关键特征是互补基对。这就是著名的 Watson-Crick 配对，即 A 与 T 配对， G 与 C 配对。这种配对是由于氢键作用，原理是 DNA 单链 (按从 5′ 到 3′ 的次序) 与相反方向写的互补链配对。例如，单链碱基序列 5′-ATGGTGCACC-3′ 和 3′-TACCACGTGG-5′ 配对：

```
5′ – A T G G T G C A C C – 3′
     | | | | | | | | | |
3′ – T A C C A C G T G G – 5′
```

0.2.2 蛋白质

蛋白质是生物体内占有特殊地位的生物大分子，它是生物体的基本构件，也是生命活动的重要物质基础，几乎一切生命现象都要通过蛋白质的结构与功能而体现出来。因此，在分子生物学中，深刻阐明蛋白质的结构与功能，是探索生命奥秘最基本的任务。

蛋白质是由氨基酸 (amino acid) 聚合而成的生物大分子。氨基酸是蛋白质的基本组成单位。自然界中的氨基酸种类很多，但参与蛋白质组成的常见氨基酸只有 20 种。这 20 种标准氨基酸的英文三字母和单字母表示见表 1。

氨基酸是带有氨基的有机酸，它的中心碳原子特称为 α 碳 (C_α)。 C_α 有四个键，分

表 1: 20 种标准氨基酸的三字母和单字母表示

氨基酸名称	英文缩写	简写	氨基酸名称	英文缩写	简写
甘氨酸	Gly	G	丝氨酸	Ser	S
丙氨酸	Ala	A	苏氨酸	Thr	T
缬氨酸	Val	V	天冬酰胺	Asn	N
异亮氨酸	Ile	I	谷酰胺	Gln	Q
亮氨酸	Leu	L	酪氨酸	Tyr	Y
苯丙氨酸	Phe	F	组氨酸	His	H
脯氨酸	Pro	P	天冬氨酸	Asp	D
甲硫氨酸	Met	M	谷氨酸	Glu	E
色氨酸	Trp	W	赖氨酸	Lys	K
半胱氨酸	Cys	C	精氨酸	Arg	R

别连着一个氨基 (NH_2), 一个羧基 (COOH), 一个氢原子和一个 R 基团 (如图 3) .

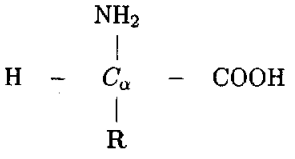


图 3: 氨基酸分子结构示意图

各种 α 氨基酸的区别在于侧链 R 基团不同， R 基团的特异性使不同氨基酸显示出不同的理化性质，进而决定了氨基酸在蛋白质分子的空间结构中可能的位置。

在蛋白质合成时，一个氨基酸的羧基和另一个氨基酸的氨基缩水形成肽键 (peptide bond) 。所以，蛋白质也是有方向的一维链，带氨基的一头称为 N 端或记为 N', 另一头带羧基称为 C 端，常用 C' 表示。

0.2.3 中心法则和遗传密码

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) 。信息从基因的核苷酸序列中被提取出来，用来指导蛋白质合成的过程对地球上的所有生物是相同的，分子生物学家称之为中心法则 (central dogma) 。

生物体的遗传信息以密码形式编码在 DNA 分子上，表现为特定的核苷酸排列顺序，并通过 DNA 的复制 (replication) 使遗传信息从亲代传向子代。在后代的生长发育过程中， DNA 分子中的遗传信息转录 (transcription) 到 RNA 分子中 (即 RNA 聚合酶以 DNA 为模板合成 RNA), 再由 RNA 翻译 (translation) 生成体内各种蛋白质，行使特

定的生物功能。翻译过程是在核糖体上进行的。这样，通过遗传信息从亲代传向子代，并在子代表达，使得子代获得了亲代的遗传性状。 RNA 也能通过复制过程合成出与其自身相同的分子。此外，生物界还存在由 RNA 指导下的 DNA 合成过程，即逆转录，这一过程发现于逆转录病毒中。通过基因转录和翻译得到的蛋白质分子可以反过来作用于 DNA，调控其它基因的表达。分子生物学的中心法则见图 4，它说明遗传信息由 DNA 分子到 RNA，再到蛋白质的传递过程。

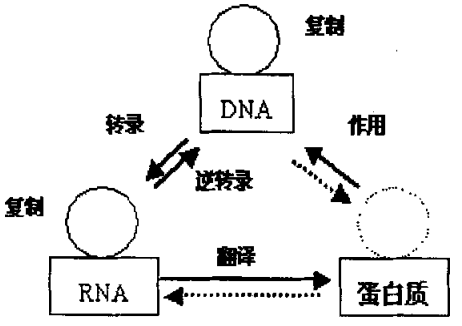


图 4: **分子生物学中心法则**

在翻译过程中，每三个碱基构成一个三联体，对应一个氨基酸或者一个终止密码子。我们称这种对应为遗传编码，可数学地表示为：

设 $\Omega = \{A, C, G, U(T)\}$ 是核苷酸集合，$\overline{\Omega} = \{(x_1x_2x_3) : x_i \in \Omega\}$, $\mathcal{R}$ 是氨基酸和终止密码子的集合，遗传编码就是一个映射 $\phi : \overline{\Omega} \longrightarrow \mathcal{R}$.

表 2 列出了这个对应，其中的符号 * 表示终止密码子，第二列中的四个 4×4 矩阵中的字母为 20 种氨基酸的单字母表示，第一、三列和第二列的第二行的字母为四种核苷酸碱基。

从表中可以看出，在 64 种密码子 (condon) 中有三个终止密码子 UAA, UAG 和 UGA，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 (这一现象称为密码的简并性 (degeneracy))：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。

表 2: 三联体通用遗传密码表

第一个核苷酸 (5′– 端)	第二个核苷酸 U	C	A	G	第三个核苷酸 (3′– 端)
U	F	S	Y	C	U
	F	S	Y	C	C
	L	S	*	*	A
	L	S	*	W	G
C	L	P	H	R	U
	L	P	H	R	C
	L	P	Q	R	A
	L	P	Q	R	G
A	I	T	N	S	U
	I	T	N	S	C
	I	T	K	R	A
	M	T	K	R	G
G	V	A	D	G	U
	V	A	D	G	C
	V	A	E	G	A
	V	A	E	G	G

0.3 生物信息学主要研究内容

生物信息学的研究内容非常丰富，例如序列比较、分子进化和比较基因组学、计算机辅助基因识别、 RNA 和蛋白质结构预测、遗传密码及其起源、序列重叠群装配、基于结构的药物设计等等，都是生物信息学中重要的研究领域．下面对其中与本文的工作密切相关的部分作简要介绍．

0.3.1 序列比较

生物信息学计算的核心是序列的比较，这包括同一序列内不同片段的比较，以及两个或多个序列的比较．序列比较的主要目的是通过序列之间的相似性阐明序列之间的同源关系、寻找序列的编码片断以及从已知序列预测新序列的结构和功能．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，即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，这需要有进化事实的验证．

法一：序列比对

Levenshtein 在 1966 年引入了“编辑距离”的概念，作为将一条字符串转变为另一条字符串所需要的最小操作数，这里的编辑操作是指插入一个字符、删除一个字符或者将一个字符替换成另一个字符 [9]．找出序列之间的差异（编辑距离）常常等价于找出序列之间的相似之处．

生物学家倾向于使用比对（ alignment ）这个词来指称 DNA/ 蛋白质序列的比较．在

早期的有关比对的文献中，人们试图找到整个字符串 V 和 W 之间的相似性，即全局比对. 这对于同一蛋白质家族成员之间的比较是有意义的，比如球蛋白，在从果蝇到人的各种生物体中都非常保守并且具有几乎相同的长度. 此外，在许多生物学的应用中，V 和 W 的子序列之间比对得分可能都大于整个序列之间的比对得分，这就是有名的局部比对问题. Smith-Waterman 算法是解决局部比对的一个好算法，缺点是速度较慢. 于是就有一些快速启发式算法应运而生. 多数快速启发式算法都采用了相同的过滤思想. 过滤是以观察为基础的，一个好的比对通常包括相同或非常相似的片段. 这样人们就可以先搜索这些短的子序列，并利用它们作为进一步分析的种子. 这种过滤思想最初是由 Dumas 和 Ninio 提出的，后来在 FASTA 和 BLAST 中得到了进一步发展，而 BLAST 正是当今生物信息学中的主流数据库搜索工具.

两条序列比对算法的要点是：

- **序列的扩张**：令 Ω 是一个有 k 种字母的字符集，$V = a_1a_2\ldots a_s$ 和 $W = b_1b_2\ldots b_t$ 是基于 Ω 的两条序列. 又令 $\Omega' = \Omega \cup \{-\}$ 为一扩展字符集（这里 "$-$" 是一个虚拟的元素，代表插入一个空格或者删除一个字母），然后视 V 和 W 为基于字符集 Ω' 的序列.
- **打分函数**：它是定义在 $k+1$ 个字母上的一种度量函数. 一般记作 $\rho(a,b)$.
- **Alignment 问题**：对于两个序列 V 和 W，寻求它们的扩张序列 V' 和 W'，使得它们的得分值为最大. 这里所用的是动态规划算法，即利用下面的递归关系做出一张序列比对得分表.

$$D_{i,j} = \max\begin{cases} D_{i-1,j} + \rho(a_i, -) \\ D_{i-1,j-1} + \rho(a_i, b_j) \\ D_{i,j-1} + \rho(-, b_j) \end{cases}$$

其中 a_i 和 b_j 分别表示序列 V 和 W 的第 i 和第 j 个基，$D_{i,j}$ 表示序列 V 的长为 i 的字首与序列 W 的长为 j 的字首比对后的得分.

序列两两比对的做法实际上是来自计算机算法中的字符串比较算法. 本质上是将两个序列的各个字符（代表核苷酸或者氨基酸残基）按照对应等同或者置换等关系进行对比排列，其结果是两个序列共有的排列顺序，这是序列相似程度的一种定性描述. 尽管人们在序列比对方面已经做了大量的工作，但有两个方面的问题一直在困扰着人们：一是没有什么合适的理论模型能很好地描述空位问题，因此打分矩阵中空位罚分缺乏理论依据而更多的带有主观色彩. 一般的处理方法是用两个罚分值，一个是对插入的第一个空位罚分，另一个是对空位的延伸罚分. 对于具体的比对问题，采用不同的罚分方法会有不同的效果. 二是比对算法的时间和空间复杂度一直没有达到令人满意的效果，特别是多重序列比对，目前尚缺乏快速而又十分有效的算法. 序列比对的这些不足，促使很多人试图寻找其它的方法来比较序列.

法二：基于不变量的方法

近年来，Randic 等人提出了一种基于序列不变量的序列比较方法，开辟了一条序列比较的新途径. 这种方法来源于计算化学中的化学指标计算，是一种间接的序列比较方法。最终，一条序列将由一个 k- 维向量来描述，这个向量常被称为序列的描述子 (descriptor)。我们可以按如下的步骤来实现序列到向量描述子的转换：

Step 1: 用比如图或曲线等数学对象来表示 DNA 序列；

Step 2: 从得到的数学对象构造矩阵；

Step 3: 从得到的矩阵提取不变量作为描述子向量的分量.

近 20 年来，一些针对 DNA 和蛋白质序列的图表示相继被提出 [10]- [36]. 这是一种生物分子数据可视化的方法，它不但使我们可以比较直观地考察生物序列，从相似的序列中发现它们的差别，还为生物序列的矩阵表示提供了一种途径：对于一个给定的图，人们总可以按照某些规则将它转化为矩阵. 目前文献中常见的由图转换而来的矩阵主要有 $ED, D/D, M/M, L/L$ 矩阵以及它们的“高阶”矩阵 [28]- [34], [37,38]. 矩阵一经给出，便可以从中提取不变量. 常用的不变量有平均矩阵元素，平均行和以及最大特征值等，其中最大特征值应用最为广泛 [28]- [34], [37]- [42].

一旦生物序列具有了向量的形式，两条序列之间的比较就被与这两条序列相对应的向量（描述子）之间的比较所代替。利用不变量来刻画和比较生物序列的优势在于不变量的刻画和比较相对简单，然而，这种方法应该说还只是处于起步阶段，其自身还存在着一系列有待解决的问题. 特别值得我们注意的是： (1) 从生物序列到它的数学表示的转换过程中某些信息的丢失； (2) 象最大特征值这样较有效的不变量其计算随着序列长度的增加会变得越来越难. 因此，正如 Randic 等 [42] 所指出的那样，如何构造合适的不变量来刻画序列又如何选择合适的不变量进行序列比较尚需做更深入的研究。

0.3.2 计算机辅助基因识别

基因是 DNA 序列上具有特定功能的一个片段，负责一种特定性状的表达. 它具有如下几个重要的特征：（1）基因是一种相对独立的遗传信息单位，这些信息单位可以通过各种方式在生物个体之间进行重新组合，并向后代传递；（2）基因是一段 DNA 分子，遗传信息贮存在 DNA 序列之中；（3）基因的信息内容通过相应的形式表现出来，即指导合成蛋白质或 RNA，进而产生生理功能，或影响其他基因的表达。

任何一条染色体上都带有许多基因，一条高等生物的染色体上可能带有成千上万个基因，一个细胞中的全部基因序列及其间隔序列统称为基因组 (genomes)。经过几十年的努力，科学家们已经得到了几十种生物体的完整基因组序列，获得了海量的数据。然而，得到完整的基因组序列仅仅是了解基因信息转录的第一步。生物信息学的目的是要对基因组数据进行分析和挖掘，从中得到具有普遍意义的规律和对生命现象的本质性解释。

所以，得到一个物种的完整基因组序列后还要做更进一步的数据处理，对一个完整的基因组序列一般要进行下面几个步骤的处理：

得到一个新的完整的基因组 DNA 序列 $\longrightarrow$ 翻译所有的 6 种可能的开放阅读框 (ORF) 并与已知的蛋白质数据库进行比较；如果可能的话，还要和同类型的物种的 EST 数据库或 cDNA 序列进行数据库的相似性搜索 $\longrightarrow$ 基因识别 $\longrightarrow$ 分析基因组中基因的调控序列。

基因识别的基本问题是给定基因组序列后，正确识别基因的范围和在基因组序列中的精确位置，这是当今最有挑战性也是最重要的课题之一．在基因组序列中寻找基因可以从两个方面入手，一是识别与基因相关的特殊信号，如启动子、起始密码子、终止密码子等等，通过信号大致确定基因所在的位置．二是预测基因的编码区域，或预测外显子所在的区域 [43]- [55] ．从本质上讲，基因识别就是要把基因组上编码蛋白质的区域和非编码蛋白质的区域区分开，这在理论方法上就是要找到在编码蛋白质的区域和非编码蛋白质的区域哪些数学、物理学特征是不一样的．统计获得的经验说明， 4 种碱基在三个编码位置的出现不是等概率的，而且密码子的使用频率也不是平均分布的，某些密码子会以较高的频率使用而另一些则较少出现．这样就使得编码区的序列呈现出可察觉的统计特异性，即所谓的“密码子偏好性”，利用这一特性对未知序列进行统计学分析可以发现编码区的粗略位置，相关文献见 [55]- [67] ．天津大学张春霆等人则通过他们提出的 Z- 曲线提取 DNA 序列的数字特征，比较成功地研究了真核和原核生物基因组中若干重要问题，包括人类基因组外显子识别，酵母基因组基因识别，细菌与古细菌基因组的 ab initio 基因识别， SARS-CoV 基因组基因识别等 [51,52], [68]- [74] ．此外， Nandy 的 2 维图形表示显示出内含子和外显子区域具有不同的特点，内含子呈细丝状结构，外显子构成密集的点丛，这对真核生物基因序列更为显著， Ghosh 等人正是基于这种图形技术针对人类 3 号染色体进行了基因识别 [13,15,16,19,22,23] ．

总的说来，原核生物计算机辅助基因识别相对容易些，结果好一些．这主要是因为原核生物的基因组比较小， DNA 量低．原核生物 DNA 分子的绝大部分是用来编码蛋白质的，只有非常小的一部分不转录，而且原核基因是连续基因，其编码区是一个完整的 DNA 片段．而真核生物基因组的规模远大于原核生物基因组，组织复杂，信息含量高．在整个 DNA 序列中，蛋白质编码区域仅占一小部分．大多数真核基因都是由蛋白质编码序列和非蛋白质编码序列两部分组成的，其中的编码序列称为外显子 (exon)，非编码序列称为内含子 (intron)．在一个结构基因中，编码某一蛋白质不同区域的各个外显子并不是连续地排列在一起的，而是常常被长度不同的内含子所隔离，形成镶嵌排列的断裂方式，所以，真核基因有时被称为分裂基因．在基因转录时，内含子和外显子一同被转录下来．然后， RNA 中的内含子被切掉，外显子随之连在一起成为成熟的 mRNA ，作为指导蛋白质合成的模板．除此之外，在 DNA 上还存在着假基因以及大量的重复序列．因此，真核生物基因序列的正确识别是个相当困难的问题．

0.3.3 分子进化和比较基因组学

地球上的一切生命形式，不管是现存的还是已经灭绝了的，都有一个共同的起源，它们的祖先可以追溯到大约在 40 亿年以前生存的一种或几种生物．因此所有动物、植物、细菌通过祖藉而相互关联．亲缘关系近的生物是由一个较近代的共同祖先传下来的，亲缘关系远的生物则由较远古的共同祖先传下来。

过去，人们研究物种之间的进化关系时主要利用形态学、解剖学、生理学以及古生物学等传统手段．随着分子生物学中的各种技术的发展以及生物分子数据的积累，系统发生分析进入了分子层次．在现代分子进化研究中，根据现有生物基因或物种多样性来构建生物的进化史是一个非常重要的问题．一个可靠的系统发生的推断，将揭示出有关生物进化过程的顺序，有助于我们了解生物进化的历史和进化机制．根据核酸和蛋白质的序列信息，可以推断物种之间的系统发生关系．其基本原理是：从一条序列转变为另一条序列所需要的变换越多，那么这两条序列的相关性就越小，从共同祖先分歧的时间就越早，进化距离就越大；相反，两个序列越相似，那么它们之间的进化距离就可能越小．

早期的工作主要是利用不同物种中同一种基因序列的异同来研究生物的进化，构建进化树．近年来由于较多模式生物基因组测序任务的完成，为从整个基因组的角度来研究分子进化提供了条件．但同时也对我们的工作提出了更高的要求．面对完全基因组，多重序列比对因其自身的不足（至少目前）已经无能为力 [75]．于是，人们开始探索新的方法． Nandy 和张春霆分别将他们提出的 DNA 序列 2 、 3 维曲线表示应用到分子进化和基因组比较研究中 [13,16,19,22,74,76]． Otu and Sayood [75], Li [77] 则提出了基于序列复杂性的数据压缩方法．需要指出的是， Randic 等提出的基于不变量的序列比较方法在此并没有失去它的用武之地，但前提是我们必须找到有效但又不会随着序列长度的增加而难于计算的新的不变量．

0.3.4 RNA 和蛋白质的结构研究

0.3.4.1 RNA 二级结构

RNA（即 mRNA 、 rRNA 、 tRNA 和 SnRNA ）主要行使两大功能：一是某些病毒的遗传物质，二是参与蛋白质的合成．这些与细胞分化、代谢、记忆的存储等有重要关系。

与 DNA 序列类似， RNA 可以看作是字符集 $\Omega = \{A, G, C, U\}$ 上的字符串． RNA 虽然是一种单链分子，但它却经常通过自身的回折使链中碱基配对从而形成多端的双股螺旋区，即为 RNA 的二级结构．在标准二级结构中所允许的基对有三种： G-C ， A-U ， G-U ．其中 G-C 基对间靠三个氢键连接， A-U 之间是两个氢键， G-U 之间则是一个氢键．正是这些基对保证了 RNA 分子结构的稳定性．也正是自由基（即没有与别的碱基配对形成氢键的碱基）与基对的同时存在，使得 RNA 分子蕴涵了相当丰富的信息．近年

来，人们对获得 RNA 的结构信息表现出浓厚的兴趣．这既包括 RNA 二级结构的预测，又包括基于二级结构的相似性分析，有关文献见 [78]- [85] ．

0.3.4.2 蛋白质的结构

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序列．二级结构涉及按线性顺序来说相互接近的氨基酸残基之间的空间关系．这些空间关系中有的是很有规则的，产生了周期性的结构，α 螺旋、β 折叠是典型的二级结构实例．蛋白质的三级结构是指多肽链借助各种相互作用力盘绕成具有特定肽链走向的紧密球状构象．维持蛋白质三级结构的作用力主要是氢键、疏水相互作用、离子键（即盐键）、范德华力以及共价二硫键．某些分子量较大的球状蛋白质在空间上可明显分出两个或多个相对独立的区域，这些区域称为结构域 (structure domain)，结构域的缔合形成具有一定空间结构的蛋白质．包含一条以上多肽链的蛋白质在结构上表现出一个新的层次，即四级结构．四级结构涉及这些多肽链结合在一起的方式．

生物信息学的一个基本观点是：分子的结构决定分子的性质和分子的功能。因此，生物大分子蛋白质的空间结构决定蛋白质的生物学功能．但是，蛋白质的空间结构又是由什么决定的呢？大量的实验结果证明：蛋白质的结构由蛋白质序列所决定．虽然影响蛋白质空间结构的另一个因素是蛋白质分子所处的溶液环境，但是决定蛋白质结构的信息则是被编码于氨基酸序列之中．因此，研究蛋白质的低级结构，特别是一级结构是生物信息学中一个非常基础又十分重要的内容．为了尽可能多地挖掘出蛋白质序列及二级结构中所包含的有用的信息， Randic 等 [28,34] 给出了蛋白质序列的几种图形表示， Feng and Zhang [25] 则将 DNA 序列的 Z-curve 推广到蛋白质，提出了蛋白质序列的 Zp- 曲线，同时， Zhang and Zhang [86] 通过将蛋白质二级结构类抽象为三个字符 α, β, c 进而给出了蛋白质二级结构的 S- 曲线表示．此外， Zhang, Feng, Bu, Chou 等还从其它多个角度对蛋白质的一级结构以及二级结构类预测作了比较深入的研究，相关文献见 [25], [86]- [95] ．

0.4 本文的主要工作

通过上面的介绍，我们可以看出生物信息学中许多领域的研究工作都有一个共同的特点，就是首先要给出生物学数据的数学描述，然后在此基础上分析和解释这些数据的生物学意义，并探索其固有的生物学规律．

生物大分子的图表示是生物学数据可视化的一条重要途径，是定性地分析生物学数据的一种强有力的工具．但现有图表示自身还存在着一些缺陷，致使在从生物学数据到

图形表示的转换过程中总是伴随着某些信息的丢失。为了避免这一点，我们给出 DNA 序列一种 3-D 图形表示和两种 2-D 图形表示："双水平线"图和"梯状"图，这些图形表示能从不同角度反映隐藏在 DNA 序列中的生物学上的特征。同时，我们还提出了有向图的概念，这是对所有生物序列都适用的一类图表示。这种表示不但弥补了现有图形表示的许多不足，而且还为生物序列的数值刻画提供了新的途径。

序列的数值刻画是对海量生物学数据进行定量分析的一种常见方法，它在本质上就是构造生物序列的特征 / 模式向量。然而，现有 DNA 序列数值刻画方法中许多都只是从碱基组成上着手，而序列之所以称为序列的另一个重要因素：元素之间的序关系，却在很大程度上被忽略了。为了更好地反映序列中元素，特别是它们之间的序关系所包含的信息，本文提出了 DNA 序列的正规化相对熵的概念，并在此基础上对酿酒酵母基因组序列进行基因识别，我们将识别的准确度提高到 96%，得到了一个酿酒酵母基因组中基因总数为 5873 的估计，与普遍接受的 5800-6000 相符。

序列比较是生物信息学计算的核心，传统的方法被生物学家专门称为序列比对 (alignment)，尽管人们在这方面已经做了大量的工作，但序列比对，尤其是多重序列比对，一直没能摆脱空位罚分缺乏理论依据和算法复杂度居高不下这两个方面的困扰，这促使人们开始寻求其他的方法来比较序列。最近，Randic 等人提出了一种基于序列不变量的序列比较方法，开辟了一条序列比较的新途径。目前常用的序列不变量都是基于相应矩阵的，其中最大特征值应用最为广泛。然而，特征值的计算随着序列长度的增加会变得越来越难。我们结合代数图论相关知识提出了一个基于矩阵的 m_1 范数和 F 范数的新的序列不变量 — ALE 指标，并对其特有的一些性质进行了讨论。最重要的是，ALE 指标与最大特征值等效但它的计算非常容易，这使得基于不变量的比较方法在完全基因组比较及其相关研究领域中的应用具有了可行性。我们还就某种特殊情况下最大特征值所反映的信息是否全面进行了探讨，并提出了伪迹的概念。此外，我们在有向图的基础上提出了生物序列的上三角矩阵表示，并对包括我们提出的 ALE 指标在内的现有序列不变量在上三角矩阵情况下的兼容性作了讨论。

对于复杂事物，人们往往无法同时顾及它的所有细节，于是，人们便有意无意地忽略某些细节以期自己所关心的特征更为突出，这实际上就是代数学中的同态思想和物理学中的粗粒化思想。基于这种思想，我们通过定义四个字上同态映射，给出了 DNA 序列的逻辑表示，并结合氨基酸的 6 种重要的物理化学特性将逻辑序列的概念推广到蛋白质序列中。同时，我们给出了 (0,1) 序列的广义 LZ 复杂度，并将其和正规化相对熵分别应用到 DNA 及蛋白质序列的相似性分析。此外，根据 RNA 二级结构中自由基和配对基同时存在这一特点，我们提出了 RNA 二级结构的影子序列的概念，并在此基础上利用符号序列标准 LZ 复杂度对 9 种病毒的 RNA 二级结构进行了比较。

1 生物大分子的图表示

1.1 引言

人类的视觉在人类的科学发现中发挥过突出的作用，而科学观察仪器（如天文望远镜、显微镜）则大大促进了人类的科学发现过程，这些仪器放大和扩展了人眼的功能。先进的可视化技术可以与科学观察仪器相比拟，使人们能够观察和分析无形的数据。人的创造性不仅取决于逻辑思维，而且取决于形象思维。数据的可视化，可以激发人的形象思维，使人们从表面上看来是杂乱无章的海量数据中找出隐藏的规律，为科学发现提供线索和依据.

可视化技术早已应用在分子生物学中，如用分子图形学技术显示分子的结构，分子结构数据本质上是一系列原子的空间坐标，如果使用者直接查看这些坐标数据，他将得不到任何帮助，但是如果通过计算机图形学方法处理这些坐标数据，以球体代表原子，以直线或者短棒代表化学键，在各个坐标点显示对应原子，那么使用者就能够观察、分析和理解分子结构。利用分子结构可视化技术，我们可以看到 DNA 分子的双螺旋空间缠绕，可以看到蛋白质的折叠，包括其中的 α- 螺旋和 β- 折叠。然而，基于生物大分子序列的图形表示及其应用只是在近些年，在人类和一些模式生物基因组计划的实施而产生了海量的生物学数据的情况下，才受到了广泛的关注并得以发展.

1.1.1 2-D 图表示

用平面图形来表示 DNA 序列，最常见的方法是基于笛卡尔坐标系的，这涉及到如何用两个坐标轴来表示 4 种核苷酸碱基和记数序列中碱基个数的问题。

1992 年，Peng 等人 [96] 提出了一种 DNA 序列的“随机步 (random-walk)”模型，他们用 x- 轴记数碱基个数，同时将 DNA 序列看作是由嘧啶与嘌呤构成的序列，并用 y- 轴的正负两个方向区分它们：若是嘌呤碱基（即 A 或 G）则向负走一步，若是嘧啶碱基（即 C 或 T）则向正走一步。他们以此为工具分析 DNA 序列，发现包含内含子的序列中碱基存在长程关联。但遗憾的是这个发现不能作为区分外显子和内含子的一般方法 [22], [97]- [99] 。

事实上，直接用两个坐标轴的 4 个方向分别代表 4 种核苷酸碱基应该说是一种更简单的方法。这种思路最早是由 Gates [11] 在 1986 年提出的：将 x 轴的正方向设为 C, 负方向设为 G, y 轴的正负方向分别被设为 T 和 A 。后来， Nandy [13,14] 将 x 轴的正方向

赋予 G, 负方向赋予 A, y 轴的正负方向分别赋予 C 和 T . 而 Leong and Morgenthlor [12] 则给出了 DNA 序列的 AC/GT- 图. 现以 Nandy 的方法为例说明如何画出一条单链 DNA 序列的平面图形: 从左到右每次观察 DNA 序列的一个碱基. 从坐标原点开始, 如果碱基是 A , 就沿 x 轴的负方向移动一个单位后描一个点; 如果是 G , 则从当前位置开始向右移动; 是 C 则向上, 是 T 则向下. 图 1.1 是序列片段 ATGGTGCACC 的 Nandy 2-D 图.

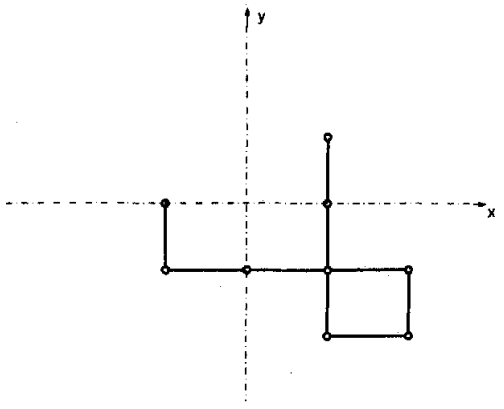


图 1.1: The 2-D Nandy's graphical representation of the sequence ATGGTGCACC

显然, 在 Nandy 2-D 图中沿 x- 轴方向的位移恰等于 A 和 G 的数量差别, 而在 y- 轴方向的位移则与 C 和 T 的数量差是一致的. 同时, Nandy 的图形具有很强的对称性. Gates 以及 Leong and Mogenthaler 的图形与此类似. 从群论的角度看, 这三种图实际上分别对应于对称群 S_4 的 3 个 4 阶正规子群. 它们从不同的角度描述了 DNA 序列, 并在序列分析、分子进化、内含子 / 外显子区分以及基因识别等领域得到了很好的应用 [11]- [23] .

然而, 我们应该意识到这种图形表示所伴随着的信息丢失. 因为, 反过来根据图 1.1 , 我们却无法判断出它到底和哪一条序列对应, 是 ATGGTGCACC , 还是 ATGGGTACCC? 其原因很明显, 就是因为这个 2-D 曲线中存在自交现象. 同时, 我们也不难发现, 序列片段 AG, AGA, AGAG, AGAGA, AGAGAG, AGAGAGA , $\cdots$, 从几何的角度看具有相同的图形表示. 上述情形都导致一个图形与多条序列对应, 我们称这种现象为图形的简并 / 退化 (degeneracy) . 从图论的角度看, 简并现象与图中圈 (circuit) 的存在是一致的. 在 2-D 图形表示中, 一个最小圈的长度是 2 .

显然, 若能使图中可能出现的圈的最小长度尽可能的大, 那么图的简并现象就将被尽可能的减少. 为此, Guo [26] 和 Liu [100] 建立了这样一种几何表示: 建立一个二维笛卡尔坐标系 xoy , 对 4 种碱基分别赋予下面四个方向, $(-1, \frac{1}{d}) \to A$, $(\frac{1}{d}, -1) \to T$, $(1, \frac{1}{d}) \to G$, $(\frac{1}{d}, 1) \to C$, 这里 d 取正整数. 这个图形也存在简并的情况, 但是在这个图中圈的最小长度与 d 有关, 即当 d 是偶数时, 圈的最小长度为 $4d$, 当 d 是奇数时, 圈的最小长度为 $2d$.

为了同样的目的， Wu 等人 [35] 提出了 DB- 曲线 (Dual-Base Curve)。一条 DB- 曲线本质上只强调 4 种碱基中的两种。例如， AC DB- 曲线的作图规则为：将向量 (1, 1) 赋予碱基 A， (-1, 1) 赋予 C， (0, 1) 方向赋予 T 和 G。 AC DB- 曲线强调的是 A 和 C 之间的关系。由于 $\binom{4}{2} = 6$，故按着这样的规则，完全描述一条 DNA 序列，需要 6 条 DB- 曲线的组合，另外 5 条是 AG, AT, TC, CG 和 TG DB- 曲线。

与如上在坐标系中赋予 4 种碱基 4 个方向的思路不同， Randic 等人 [30,31] 在 2003 年提出了一种新的图形表示。具体做法是：先画出相互间隔一个单位的 4 条水平线，并让 A, C, G, T 这 4 种碱基分别与这 4 条水平线对应，然后从左到右每次考察 DNA 序列的一个碱基，遇到的是哪种碱基，就在这种碱基所对应的水平线上描点，同时如果不是最后一个碱基的话还要右移一个单位。用直线连接所有相邻的点，最终得到锯齿状的 2-D 曲线。为方便起见，我们称 Randic 的这种图形表示为“四水平线”图。以序列片段 ATGGTGCACCTGACTCCTGA 为例，它的“四水平线”图见图 1.2。显然，“四水平线”图避免了简并现象的出现。不过，要完全描述一条 DNA 序列，这种方法可能需要 12 个“四水平线”图，这是由将 4 条水平线分配给 4 种碱基的不同方式所决定的。

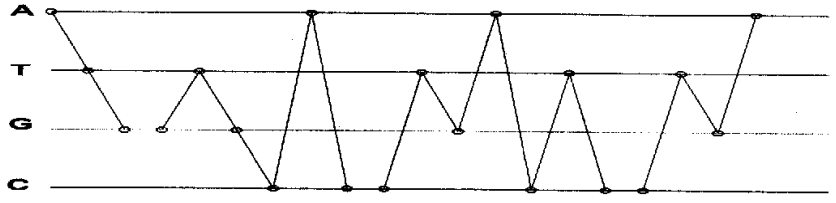


图 1.2: The graphical representation of the sequence ATGGTGCACCTGACTCCTGA

1.1.2 3-D 图表示

二十多年前， Hamori and Ruskin [10] 最先提出了 DNA 序列的一种三维图形表示，即 H- 曲线。其作图规则是：将东南、东北、西南、西北四个方向分别赋予 4 种碱基，同时用 z- 轴正方向记数碱基的个数。 Hamori and Ruskin 利用 H- 曲线对抗菌素 M13 进行研究，观察到所有基因的起始位点都由短的富含嘌呤的序列引导，这给出了一个直观上识别基因的敏感的信号。

Z- 曲线是张春霆等在 1994 年提出的 DNA 序列的另一种三维图形表示 [24]。在 Z- 曲线的基础上，他们将坐标系、投影、曲线、曲线微分等几何学概念与 DNA 序列建立起紧密的联系，并用这种思路研究了真核和原核生物基因组中若干重要问题，深受国际同行好评。此外，他们还将 Z- 曲线推广到蛋白质序列，构造了 Zp- 曲线 [25]。而另有学者给出了 RNA 的 Z- 曲线 [101]。

Z- 曲线的作图规则是：对于长为 N 的一条 DNA 序列，从第一个碱基开始依次考察此序列，每次只考察一个碱基。当考察到第 n 个碱基时 $(n = 1, 2, \ldots, N)$，计数四种碱基

A, C, G 和 T 出现的次数，分别记为 A_n, C_n, G_n 和 T_n，从而由下面的对应可以得到三维空间中点 p_n 的坐标

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n) \\ y_n = (A_n + C_n) - (G_n + T_n) \\ z_n = (A_n + T_n) - (C_n + G_n) \end{cases}$$

其中，$x_n, y_n, z_n \in [-N, N]$ 且 $n \in \{0, 1, \ldots, N\}$. 当 n 从 0 取到 N 时，依次得到 $P_0, P_1, P_2, \ldots, P_N$ 共 $N+1$ 个点. 将相邻的两点连接所得到的整条曲线被称为 Z- 曲线.

由上述点 P_n 的坐标公式，不难得到

$$\begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix} = \frac{n}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}$$

其中用到了 $A_n + C_n + G_n + T_n = n$ 这一事实. 这个对应被 Zhang 等人称为逆 Z- 变换. 基于此，Zhang 等指出在坐标与 DNA 序列可以相互唯一地重构的意义上 Z- 曲线是 DNA 序列的一个等价表示. 不过，在几何直观上，Z- 曲线却没能避免简并现象的出现. 例如，对于序列片段 AGCTAGCTAGCT，对应的点列 $P_0, P_1, P_2, P_3, P_4, \ldots$ 的坐标为

$$(0,0,0), (1,1,1), (2,0,0), (1,1,-1), (0,0,0), (1,1,1), (2,0,0), (1,1,-1), (0,0,0), \ldots$$

这导致了一个重叠圈的产生.

此外，在 2000 年，Randic 等 [29] 把 4 种碱基 A, C, G, T 设计到一个中心在坐标原点的四面体的四个顶点上，从而将 Gates, Nandy, Leong and Mogenthaler 的三种 2-D 图形统一到一个 3-D 空间曲线. 然而，正如 Guo 等 [26] 所言，这种图形表示的简并度仍然很高.

1.1.3 其他

除了 2 维平面和 3 维空间曲线表示外，文献中也可见 DNA 序列的“高维”表示。例如，Hamori and Ruskin [10] 曾提出了被称为 G- 曲线的 DNA 序列的 5 维空间表示。又如，Randic and Balaban [102] 将 4 维空间中 4 个坐标轴的正方向分别赋予 DNA 序列的 4 种碱基，从而得到了 DNA 序列的一种 4 维表示. 这些高维表示尽管已经在生物信息学的某些领域得到了应用，但正如它们的提出者所说，它们已经不再具备可视化这一重要优势。

针对上述图形表示的缺陷，我们提出了几种新的图形表示. 在这一章，我们侧重于介绍这些图的构造及其优点，具体应用将在下一章中给出.

1.2 DNA 序列的 3-D 图形表示

将生物序列转化为图形，关键在于建立碱基与点 / 向量之间的对应关系。这里，我们将 4 个 3 维空间中的向量分别赋予 DNA 序列的 4 种碱基：

$$(1,0,0) \mapsto A$$
$$(0,1,0) \mapsto C$$
$$(0,0,1) \mapsto G$$
$$(1,1,1) \mapsto T$$

对于任一给定的长为 n 的 DNA 序列 $S = S_1S_2 \ldots S_n$，从第一个碱基开始依次考察此序列，每次只考察一个碱基。当考察到第 i 个碱基时 $(i = 1,2,\ldots,n)$，一个 3 维空间中的点 $P_i(x_i, y_i, z_i)$ 可以按式 1.2.1 得到：

$$\begin{cases} x_i = \sum_{k=1}^{i} S_k^1 \\ y_i = \sum_{k=1}^{i} S_k^2 \\ z_i = \sum_{k=1}^{i} S_k^3 \end{cases} \tag{1.2.1}$$

其中 S_k^j $(j = 1,2,3)$ 表示 S_k 所对应的向量的第 j 个分量。当 i 从 1 取到 n 时，我们依次得到点 $P_1, P_2, \ldots, P_n$。用直线连接相邻的两点，我们便得到一条 3 维空间曲线。

表 1.1 给出了序列片段 ATGGTGCACC 的点的坐标。连接相邻点而得到的 3-D 曲线如图 1.3 所示。

表 1.1: 3-D Coordinates for the sequence ATGGTGCACC

Base	1	2	3	4	5	6	7	8	9	10
nucletic	A	T	G	G	T	G	C	A	C	C
x	1	2	2	2	3	3	3	4	4	4
y	0	1	1	1	2	2	3	3	4	5
z	0	1	2	3	4	5	5	5	5	5

如果将这个 3-D 曲线投影到 2 维平面，那么将得到 3 个不同的 2-D 图形表示，几乎每一个都可以看作是一条 DB- 曲线。例如，它在 xy- 平面的投影与 AC DB- 曲线相当。图 1.4 给出了大肠杆菌噬菌体 ϕX174 (GI: 9626372) 的前 20 个碱基构成的序列片段，GAGTTTTATCGCTTCCATGA，的 3 维图形到 xy- 平面的投影和同一序列片段的 AC DB- 曲线。我们清晰的看到，二者几乎具有相同的“痕迹”。事实上，从这两种表示方法对 4 种碱基赋予向量的规则处便可以料到这个结果是必然的。

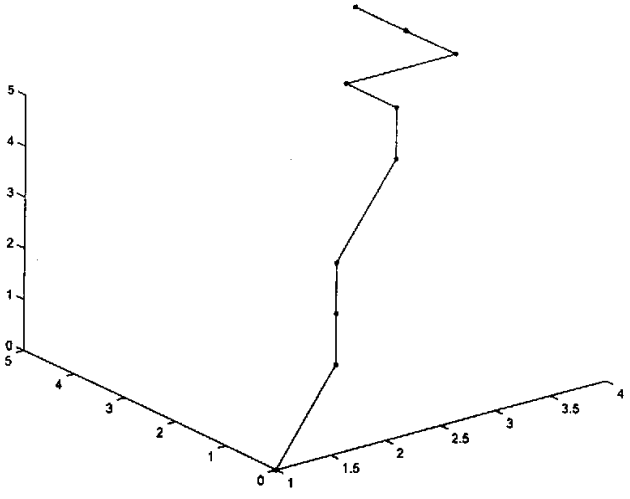


图 1.3: **The 3-D graphical representation of the sequence ATGGTGCACC**

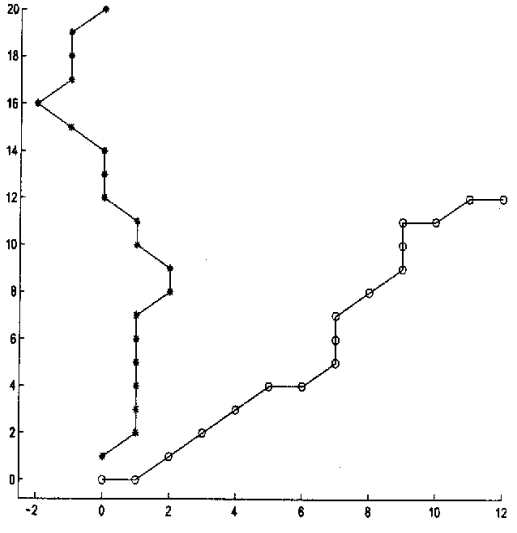


图 1.4: **AC DB-curve and the projection on xy-plane of our 3-D curve: (∗) AC DB-curve, (o) This work.**

我们的 3-D 曲线表示具有如下一些优点：

- 曲线避免了因与自身相交或重叠而导致的简并现象. 同时, 序列中重复片段在图上一目了然 (参图 1.5).
- H- 曲线总是在向四周膨胀的同时沿 z- 轴无限伸展直到和序列的长度相等为止. 而我

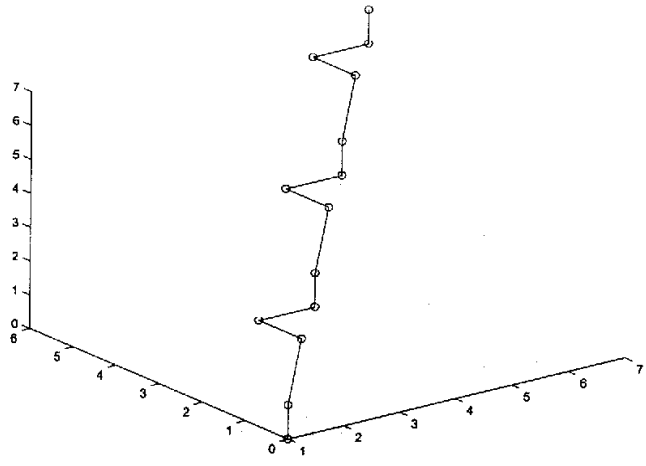


图 1.5: **The 3D graphical representation of sequence AGTCAGTCAGTCAG**

们的 3-D 曲线沿三个坐标轴正向的移动是等概率的，在这个意义上，它将比 H- 曲线占用更少的空间. 此外，两碱基的相对丰富用我们的 3-D 曲线有时将更容易观察到. 如 A 和 C 相对丰富的序列 ACACCCCACCAAACCAGGACAAACCACTAC，它的 3-D 图形如图 1.6 所示。

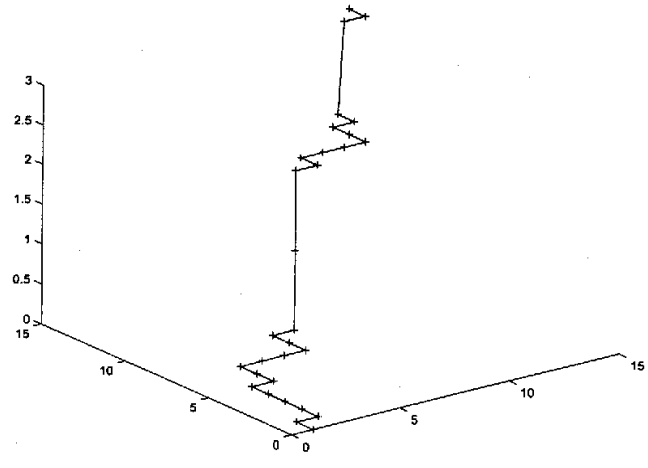


图 1.6: **The graphical representation of the sequence ACACCCCACCAAACCAGGACAAACCACTAC**

- 如上我们是通过赋予 4 种碱基 A, C, G, T 4 个向量 $(1,0,0),(0,1,0),(0,0,1),(1,1,1)$ 而得到 DNA 序列的一种 3-D 曲线。显然如果我们将 $(1,0,0),(0,1,0),(1,1,1),(0,0,1)$ 分别赋给 A, C, G, T，那么对于同一条 DNA 序列，我们将得到另一条 3-D 曲线。尽管将 4 个向量赋给 4 种碱基，看起来似乎有 $4!=24$ 种方式，但观察这 4 个向量的特点便不难发现，对于一条 DNA 序列，本质上只有 4 条不同的 3-D 曲线，这仅仅依赖于哪一个碱基被赋予了向量 $(1,1,1)$。而 DB- 曲线和“四水平线”图，前面我们已经指出，相应的数字分别是 6 和 12。

1.3 DNA 序列的 2-D 图形表示

在这一节，我们将通过 DNA 序列的特征序列间接地给出 DNA 序列的两种 2-D 图形表示。

1.3.1 特征序列

我们知道， DNA (RNA) 序列中的四个核苷酸碱基的环有两种，即单环嘧啶和双环嘌呤，记 R 为嘌呤，Y 为嘧啶，则有 $R=\{A,G\}$ 和 $Y=\{C,T\}$。同样的可以将这四个核苷酸碱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碱基分为弱氢键和强氢键两组：即 $W=\{A,T\}$ 和 $S=\{C,G\}$。对于上面的每一种分类，能做下面的操作使得每一个 DNA 序列对应一个 (0,1) 序列：若基属于 R, 则记它为 1，否则，若它属于 Y 则记为 0。在这样的操作下， DNA 序列就变为了 (0,1) 序列。对于同一个 DNA 序列，还可以根据另两种分类做相同的运算。用数学形式表示如下：

设 $S=S_1S_2\cdots$ 是一个 DNA 序列. 根据上面的分类, 定义三个字上的同态映射 $\phi_i(S)=\phi_i(S_1)\phi_i(S_2)\cdots$, $i=1,2,3$, 其中

$$\begin{aligned}
\phi_1(S_j)&=\begin{cases}1 & \text{if } S_j\in R\\ 0 & \text{if } S_j\in Y,\end{cases}\\
\phi_2(S_j)&=\begin{cases}1 & \text{if } S_j\in M\\ 0 & \text{if } S_j\in K\end{cases}\\
\phi_3(S_j)&=\begin{cases}1 & \text{if } S_j\in W\\ 0 & \text{if } S_j\in S.\end{cases}
\end{aligned}\tag{1.3.1}$$

这样从一个 DNA 序列就得到了三个 (0,1) 序列， He and Wang [103, 104] 将它们分别称为这个 DNA 序列的 (R,Y)-, (M,K)- 和 (W,S)- 特征序列，并且证明了这三个特征序列给出了这个 DNA 序列的所有信息。表 1.2 列出了人的 β-globin 基因的第一个外显子及其特征序列.

表 1.2: The three characteristic sequences of exon I of the human β-globin gene

Sequence	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
(M, K)-	1000001111001101100 1 0010110010011000110011100 0 000011100001 1100001001100000000001001110000110
(R, Y)-	1011010100011000001 1 1111111000100100100100001 0 111101111011 1010111011110011011011110000111011
(W, S)-	1100100100101010010 1 0010110101000011101000010 1 0000011001 0110010011011001100100101000001000010

1.3.2 基于特征序列的“双水平线”图

对任一给定的特征序列，我们按如下规则给出它的“双水平线”图：先画出相互间隔一个单位的两条水平线，并分别以 0 和 1 标记。然后从左到右每次考察特征序列的一个字母，遇到的是哪种字母，就在这种字母所对应的水平线上描点，同时如果不是最后一个字母的话还要右移一个单位。最后将所有相邻的点用直线连接，就得到一条特征序列的 2-D 曲线。表 1.2 中 DNA 序列的前 30 个碱基对应的三条特征序列的“双水平线”图如图 1.7 所示。

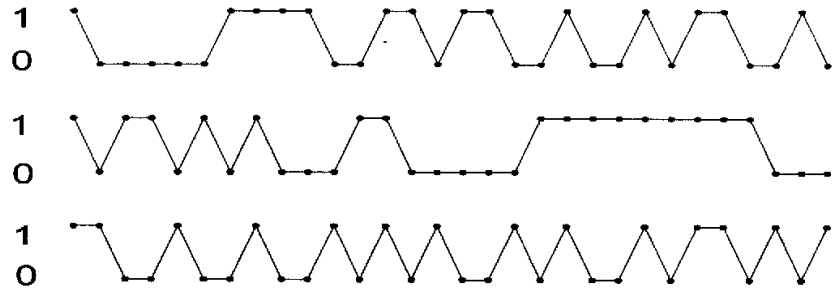


图 1.7: The 2-D graphical representations of the three characteristic sequences based on two horizontal lines.

显然，对两条平行的直线用 0 和 1 两个字母来标记在对称的意义上只有一种标记方法，因此，描述一条特征序列的“双水平线”图本质上只有一个。也就是说，特征序列和“双水平线”图是 1-1 的。从而，一条 DNA 序列可以唯一地被三个与特征序列相对应的“双水平线”图所表示，而且，从序列到图形的转化过程没有伴随信息的丢失。

不过仔细观察“双水平线”图，我们会发现它总是随着序列的增长而向右延伸，这从占用空间的角度来看是不太经济的。为了解决这个问题，我们可以在构图的规则上加以改进：对任一给定的长为 n 的特征序列 $b = b_1 b_2 \ldots b_n$，我们从左到右每次考察它的一个字母。当考察到第 i 个字母时 $(i = 1, 2, \ldots, n)$，一个平面上的点 $P_i(x_i, y_i)$ 可以按式 1.3.2 得到：

$$(x_i, y_i) = \begin{cases} (1_i, 1) & \text{if } b_i = 1 \\ (0_i, 0) & \text{if } b_i = 0, \end{cases} \tag{1.3.2}$$

其中 1_i 和 0_i $(i = 1, 2, \ldots, n)$ 分别表示 1 和 0 在子串 $b_1 b_2 \ldots b_i$ 中出现的次数. 用直线依次连接点 $P_1, P_2, \ldots P_n$, 我们便得到一条新的 2-D 曲线. 图 1.8 给出了表 1.2 中 DNA 序列的前 30 个碱基对应的三条特征序列的新的 2-D 曲线. 显然, 这种曲线是一种压缩的曲线, 因而更适合用来表示较长的序列.

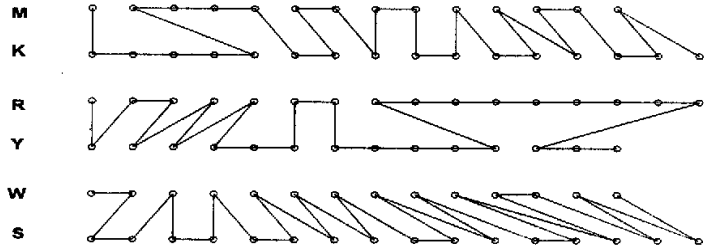


图 1.8: **The 2-D curves of the sequence ATGGTGCACCTGACTCCTGAGGAGAAGTCT.**

1.3.3 基于特征序列的“梯状”图

由于特征序列是一种二元序列, 序列中只有“ 0 ”和“ 1 ”两种字符, 所以除了按上述“水平线”方式构图外, 还很容易通过赋予“ 0 ”和“ 1 ”不同的方向给出平面图形表示, 如, 将 (1,0) 赋予“ 1 ”, 将 (0,1) 赋予“ 0 ”, 就可得到一条 xoy 平面曲线. 具体做法是: 从左到右观察特征序列, 从坐标原点开始, 如果遇到“ 1 ”, 就从当前位置沿 x 轴正方向移动一个单位, 如果是“ 0 ”则沿 y 轴正方向移动, 连接所有相邻的点, 得到一条 2-D 曲线, 根据其外形, 我们称之为“梯状”图 (参图 1.9).

p53 基因是一种重要的抑癌基因. 它在细胞周期调控、抑制细胞生长、诱导肿瘤细胞凋亡等方面有重要的作用. 研究发现, 在人类至少 50% 的肿瘤中都出现了 p53 基因缺失或突变, 而且突变热点主要位于第 4-8 个外显子 [105]- [109] . 表 1.3 列出了野生型 p53 (wt-p53) 的外显子 4-8 的编码序列.

我们在图 1.10 给出了野生型 p53 和 bladder, hepatocellular, brain, skin 四种癌细胞异常 p53 的外显子 4-8 的基于 (M,K)- 特征序列的“梯状”图, 相应数据取自 p53 突变数据库 IARC (the International Agency for Research on Cancer) . 从图 1.10 可以出, 脑瘤细胞 p53 基因的第 8 个外显子发生了明显的区域性突变, 而第 6 个外显子突变相对较弱; 皮肤癌中第 7 、 8 两个外显子突变较为明显, 其它变化不大; 膀胱癌的几乎与前两者相反,

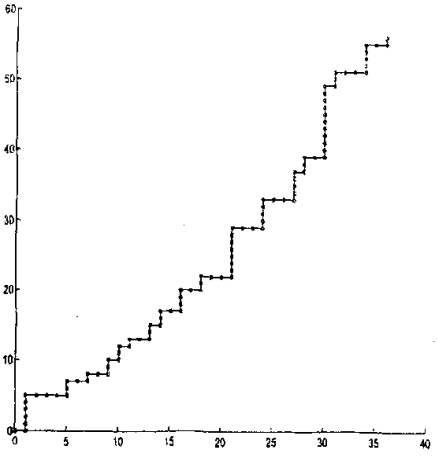


图 1.9: The 2-D ladder-like graphical representation of the (M,K)-characteristic sequence in Table 1.2

表 1.3: Exons 4-8 of wt-p53 gene

exon	coding sequence
4	tccccctttgccgtcccaagcaatggatgatttgatgctgtccccggacgatattgaacaatggttcactgaagacc caggtccagatgaagctcccagaatgccagaggctgctccccgcgtggcccctgcaccagcagctcctacaccggc ggcccctgcaccagccccctcctggcccctgtcatcttctgtcccttcccagaaaacctaccagggcagctacggt ttccgtctgggcttcttgcattctgggacagccaagtctgtgacttgcacg
5	tactcccctgccctcaacaagatgttttgccaactggccaagacctgccctgtgcagctgtgggttgattccacac cccgcccggcacccgcgtccgcgccatggccatctacaagcagtcacagcacatgacggaggttgtgaggcgctg cccccaccatgagcgctgctcagatagcgatg
6	gtctggcccctcctcagcatcttatccgagtggaaggaaatttgcgtgtggagtatttggatgacagaaacacttt tcgacatagtgtggtggtgccctatgagccgcctgag
7	gttggctctgactgtaccaccatccactacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggc ccatcctcaccatcatcacactggaagactccag
8	tggtaatctactgggacggaacagctttgaggtgcgtgtttgtgcctgtcctgggagagaccggcgcacagaggaa gagaatctccgcaagaaaggggagcctcaccacgagctgcccccagggagcactaagcgag

突变明显的是第 5 、 6 个外显子，而第 7 、 8 个外显子突变相对较弱；肝癌中突变明显的是第 4 个外显子，而第 7 个外显子中的变化则很小，而且，仅就第 7 个外显子而言，四种癌变细胞中也是肝癌的变化最小。这些信息无疑将有助于癌症的诊断、预后以及治疗，但它们若要从序列本身直接获得显然不是一件易事。

这一节我们提出了 DNA 序列的两类 2-D 图形表示，由于是基于特征序列的，所以二者都在避免了曲线的自交与重叠的同时，清晰地显示出不同类型碱基的分布情况和丰富程度，具有明显的生物学意义。更有意思的是，二者不但都源自二元的 (0,1)- 序列，而

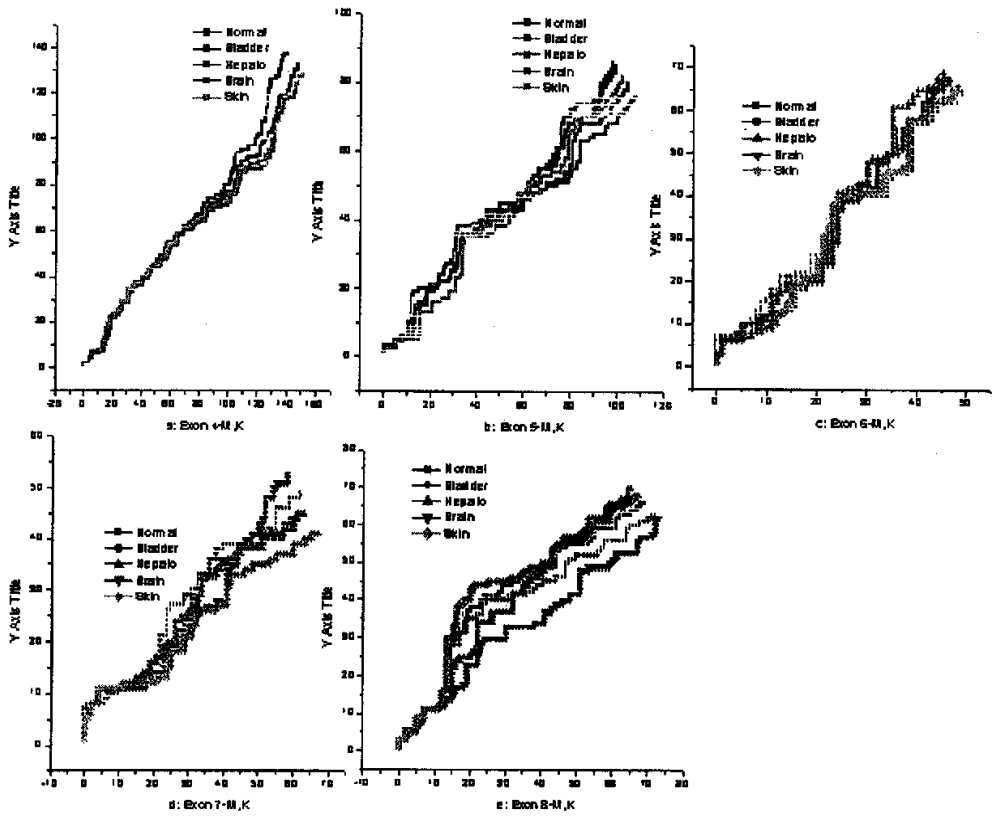


图 1.10: the 2-D ladder like graphical representations for exons 4-8 of the wt-p53 and mutant exons of the 4 cancers: bladder, hepatocellular, brain, skin cancers

且，在下一章将会看到，我们还可以在数值形式上找到它们共同的“归宿”，那就是在某种 (0,1)- 矩阵上的统一.

1.4 有向图表示

如前所述，现有的许多图形表示都存在着因圈的出现而导致的简并现象. 更一般地，这些图形所体现的都只是沿着 DNA 序列“travel”时的道路（path），而不是它的“历史” (history) [29]. 因此，在由 DNA 序列到它的图形表示的转化过程中，一些有用的信息就将被遗漏. 事实上，前面提到的图形表示中，不论其简并度高还是低，几乎都存在这个缺陷. 其原因在于，这些表示都是基于无向图的. 而另一方面，我们知道，生物大分子序列都是有方向的. 这一切都给了我们一个同样的提示，那就是利用有向图来描述生

物大分子序列！

定义：设 S 是一条 DNA 或 RNA 或蛋白质序列，其长为 n，$G(V,E)$ 是 S 的一个（无向）图，其中 V 是顶点集，E 是边集。通过分析前述图形表示的构造过程，我们发现这些图本质上都是一个“ walk ”：$G(V,E)=v_1e_1v_2e_2v_3e_3\ldots v_n$。其中顶点 v_i 对应于所考虑的生物序列的第 i 个基，$v_{i+1}\neq v_i$, 但 $v_j=v_i\ (j>i+1)$ 是允许的（正是这导致了图中圈的出现）。对任一边 $e_i=(v_i,v_{i+1})$，定义其方向为：$v_i\mapsto v_{i+1}$，即 v_i 定义为边（确切的说应该是弧）e_i 的起点，而 v_{i+1} 则被定义为终点。这样，我们便由图 $G(V,E)$ 得到一个有向图 $D(V,E)$。

以 DNA 序列片段 ATGGTGCACC 为例，我们以其 Nandy 的 2-D 图为“基图”，对应的有向图如图 1.11 所示。

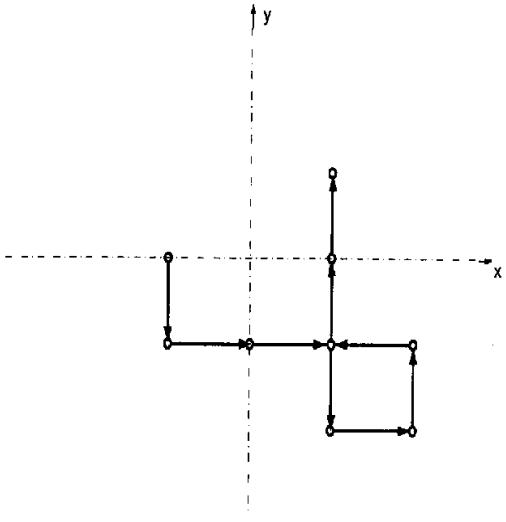


图 1.11: **The directed graph based on the 2-D Nandy's graphical representation of the sequence ATGGTGCACC**

显然，有向图表示的简并度比相应无向图的低得多，甚至在某些情况下将不再出现简并现象。而且，在有向图中，我们所看到的实际上正是沿着生物序列“ travel ”时的“历史”，从这个意义上讲，有向图表示更容易激发人们的形象思维，从而有利于人们迅速地抓住生物序列的特征。此外，在下一章将会看到，有向图表示还为生物序列的数值刻画提供了新的途径。

2 生物序列的数值刻画

2.1 引言

作为一种可视化技术，图形表示为我们研究生物大分子提供了一种定性的手段。与此相对应，数值刻画则提供了一种定量地研究生物学数据的方法。文献中数值刻画方法可以归结为如下几种形式：矩阵表示、序列不变量、子串计数等。

(1) 矩阵表示

矩阵在数学中已经是一个非常成熟的领域，如何利用矩阵分析生物序列是一个非常有潜力的课题。

Randic [39] 针对 DNA 序列提出了一种称为 S/S 的对称矩阵。假设 $S = S_1S_2\ldots S_n$ 是一条 DNA 序列，则它的 S/S 矩阵的 (i,j)- 元素定义为：

$$[S/S]_{ij} = \begin{cases} \frac{n_{ij}}{j-i} & \text{if } i < j \\ 0 & \text{if } i = j, \end{cases}$$

这里 n_{ij} 表示子串 $S_{i+1}\ldots S_j$ 中 S_j 所对应的碱基的个数。

显然，S/S 矩阵是从序列本身直接得到的。与此不同，还有一种矩阵是来自图形表示的，这可以说是图形表示在分析生物学数据方面的另一个贡献。这种基于图形的矩阵包括：ED、GD、PD、D/D 和 L/L 矩阵等 [28]- [34], [37,38]，这些矩阵也都是对称的，具体构造方法如下：

假设某生物大分子数据的图形是由 k-D 空间中 n 个点连接而成的曲线。则

- ED 矩阵的 (i,j)- 元素定义为曲线上两顶点 v_i 和 v_j 之间的 Euclidean 距离：

$$[ED]_{ij} = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \ldots + (x_{i_k} - x_{j_k})^2}$$

- GD 矩阵的 (i,j)- 元素定义为曲线上两顶点 v_i 和 v_j 之间的图论距离：

$$[GD]_{ij} = |j - i|$$

- PD 矩阵的 (i,j)- 元素定义为曲线上顶点 $v_i, v_{i+1}, \ldots, v_j$ 之间相邻两点的 Euclidean 距离之和：

$$[PD]_{ij} = \begin{cases} [ED]_{i,i+1} + [ED]_{i+1,i+2} + \ldots + [ED]_{j-1,j} & \text{if } i < j \\ 0 & \text{if } i = j \end{cases}$$

- D/D 矩阵的 (i,j)- 元素定义为 ED 和 GD 矩阵相应元素的商：

$$[D/D]_{ij} = \begin{cases} [ED]_{ij}/[GD]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

- L/L 矩阵的 (i,j)- 元素定义为 ED 和 PD 矩阵相应元素的商：

$$[L/L]_{ij} = \begin{cases} [ED]_{ij}/[PD]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

(2) 序列不变量

由上述矩阵的构造过程我们可以看到，这些矩阵的阶数是和生物序列的长度一致的。因而，当相比较的两个序列较长时，矩阵的直接利用并不方便。由矩阵论的知识我们知道，某些基于矩阵的不变量能很好地反映矩阵所包含的信息。

常用的不变量有 Wiener 数、平均矩阵元素、平均行（列）和、最大特征值等 [28]-[34], [37]- [42], [110]。Wiener 数、平均矩阵元素、平均行（列）和三者的差别仅在于它们的正规化子的不同，它们的计算相对简单，但它们并不能精确地反映矩阵所包含的信息。最大特征值是一个应用最为广泛而且已经被证明是很有效的不变量。但我们必须面对的问题是，特征值的计算会随着序列长度的增加而变得越来越难。因此，寻找一个既有效又易于计算的不变量将是一个很有意义的工作，这在人与一些模式生物基因组计划相继完成进而基于基因组比较的分子进化研究成为生物信息学中一个新的热点的今天，显得更为重要。

(3) 子串计数

根据碱基分布的不均一性，有人对 DNA 序列中的 4 种碱基出现的频率进行统计，并由此构造出 4 维向量来表示 DNA 序列。如果记 $\Omega = \{A, C, G, T\}$，并用 f_L $(L \in \Omega)$ 表示 L 在序列中出现的频率，即 $f_L = L_n/n$，其中 n 为序列长度，L_n 为序列中 L 出现的个数，则 4 维向量为 (f_A, f_C, f_G, f_T)。类似地，可以得到基于 20 种氨基酸出现频率的蛋白质序列的向量表示。

如果考察的不是单个碱基而是双碱基，显然可以得到 DNA 序列的一个 4^2 维向量。更一般地，如果我们考察 k- 子串出现的频率，则得到 DNA 序列的一个 4^k- 维向量。Li 等 [48], He and Wang [111] 在统计单碱基出现频率的基础上给出了 DNA 序列的 Shannon

熵，Randic 等人 [42] 给出了基于 3- 子串（三联体）的一种 DNA 序列的相似性分析方法，而 Karlin and Burge [112] 则提出用 2- 子串的让步比 (odds ratio) 对原核生物和真核生物的完全基因组进行分类比较. 沿着这个思路，He [113] 提出了 DNA 序列的筛比 (sieve ratio) 的概念，Hao 等 [114,115] 则提出了用剔除 k- 串的随机背景构建细菌、古细菌的进化树. 但仔细分析，我们发现用如上方法表示生物大分子序列是不充分的，因为它们都只是考虑序列中元素的组成，而序列之所以称为序列的另一个方面，即元素的先后顺序，却在很大程度上被忽略了，这必将导致生物大分子序列中某些重要信息的丢失.

在这一章，针对生物大分子数据现有数值刻画方法的这些不足，我们将提出矩阵的“伪迹”、“ ALE ” - 指标、以及序列的正规化相对熵等概念，并在有向图的基础上引出生物序列的上三角矩阵表示.

2.2 伪迹

分析上述矩阵的构造过程，不难发现 L/L 矩阵的元素总是属于闭区间 $[0,1]$ 的. 因此由 L/L 矩阵可以得到一个按元素收敛的矩阵序列 ${}^kL/{}^kL$ $(k=1,2,\ldots)$ ，其中 ${}^kL/{}^kL$ 表示矩阵 L/L 按 Hadamard 乘积自乘 k 次得到的矩阵. 显然，当 $k\to\infty$ 时，这个矩阵序列的极限是一个 (0,1)- 矩阵，我们记之为 ${}^bL/{}^bL$，它可以直接从 L/L 矩阵通过将所有小于 0 的元素用 0 代替而得到.

在第一章，我们曾提出了 DNA 序列的两类基于特征序列的 2-D 图形表示：“双水平线”图和“梯状”图，自然可以按照上述矩阵构造规则给出它们的 L/L 矩阵进而其极限矩阵 ${}^bL/{}^bL$. 对于同一特征序列，其基于“双水平线”图和“梯状”图的两个 L/L 矩阵显然是不同的，但不难发现这两个矩阵中值为 1 的元素的个数及其所在的行列位置完全一致，不同的只是那些非“ 1 ”的元素，即那些值处于半闭半开区间 $[0,1)$ 的元素. 因此，二者的极限矩阵 ${}^bL/{}^bL$ 必然是相等的，而且其矩阵元素“ 1 ”和“ 0 ”恰与特征序列中是否发生了二元字符之间的交替相一致，这一点犹如图与邻接矩阵之间的关系，这也充分说明了这个极限矩阵 ${}^bL/{}^bL$ 能够反映出序列的本质.

在一个矩阵的众多特征值中，最大特征值被认为足可以代替矩阵进而可以作为一种不变量来刻画 DNA 序列 [29]- [34], [37]- [42] . 但从表 2.1（取自 Randic [30] 表 2 ）可以看出，随着 k 的增大，相应 Hadamard 乘积矩阵的最大特征值与其它特征值之间的区别越来越不明显. 在这种情况下，仅仅由一个最大特征值所反映的信息恐怕是不充分的. 换句话说，我们还应该考虑其它特征值所起的作用. 另一方面，由群表示论可知，群的表示的特征标有着重要的作用. 而特征标将群的元素映射到某些矩阵的迹. 这表明矩阵的迹，即矩阵的主对角线元素之和或者所有特征值之和，能反映出矩阵所包含的主要信息. 但同时我也注意到 ${}^kL/{}^kL$ 矩阵的主对角线元素都是 0 ，从而它的迹等于 0 ，这就是说，迹在 ${}^bL/{}^bL$ 矩阵面前也失去了它应有的作用. 为了弥补上述缺陷，我们提出用最大与最小

特征值之和来刻画 DNA 序列，并称这个和为“伪迹”。

表 2.1: The eigenvalues λ_i $(i = 1, 2, \ldots, 10)$ of the ${}^kL/{}^kL$ $(k = 1, 2, 5, 10, 50)$ and ${}^bL/{}^bL$ matrices of the sequence ATGGTGCACC

Eigenvalue	L/L	${}^2L/{}^2L$	${}^5L/{}^5L$	${}^{10}L/{}^{10}L$	${}^{50}L/{}^{50}L$	${}^bL/{}^bL$
λ_1	6.7028	5.4435	3.8923	3.0925	2.4472	2.4284
λ_2	0.5885	1.2573	1.7199	1.9082	2.1035	2.0991
λ_3	-0.2551	0.2465	1.0552	1.3854	1.4027	1.4054
λ_4	-0.6792	-0.4591	-0.0831	0.1857	0.5353	0.5599
λ_5	-0.7990	-0.6491	-0.3857	-0.1888	0.0695	0.0785
λ_6	-0.9923	-0.9957	-0.9940	-0.9941	-1	-1
λ_7	-0.9984	-0.9973	-1.0018	-1.0019	-1	-1
λ_8	-1.0252	-1.0493	-1.1161	-1.1838	-1.0134	-1
λ_9	-1.1208	-1.2121	-1.3655	-1.4331	-1.6618	-1.6738
λ_{10}	-1.4213	-1.5846	-1.7212	-1.7701	-1.8831	-1.8975

为了验证我们的方法的有用性，我们取 11 个不同物种的 β-globin 基因的第一个外显子进行比较，它们被认为是非常保守的序列，即进化很慢的序列。表 2.2 列出了这 11 个序列。

表 2.2: The coding sequences of the exon 1 of *beta*-globin gene of 11 different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCT GGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCT GGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGT GGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGT GGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGT GGGGAAAGGTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTG GGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCA AGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAA GGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

既然基于“双水平线”图和基于“梯状”图的 ${}^bL/{}^bL$ 矩阵是一样的，所以我们不妨在“双水平线”图的基础上构造其 ${}^bL/{}^bL$ 矩阵，并计算相应的“伪迹”。以人的 β-globin 基因的第一个外显子的 (M,K)- 特征序列为例 (参表 1.2 和图 1.7)，它所对应的 ${}^bL/{}^bL$ 矩阵的最大和最小特征值分别为： 9.0226， -1.9983，进而其“伪迹”等于 7.0243。我们已经知道，每条 DNA 序列都可以唯一地由三个与其特征序列相对应的“双水平线”图所表示，而同时不难发现，任一 (0,1)- 序列的“双水平线”图与它所对应的 ${}^bL/{}^bL$ 矩阵之间是 1-1 的。因此，我们可以用一个 3 维向量来刻画 DNA 序列，这个向量的分量是与三条特征序列对应的 ${}^bL/{}^bL$ 矩阵的“伪迹”。表 2.3 列出了表 2.2 中的 11 个物种的 3 维向量表示。

表 2.3: The 3-component vectors associated with the 11 sequences in Table 2.2

Species	λ_{MK}	λ_{RY}	λ_{WS}
Human	7.0243	6.0317	2.1104
Goat	4.0512	7.0246	2.1106
Opossum	5.0384	8.0204	2.1085
Gallus	4.0509	6.0300	4.0510
Lemur	5.0390	5.0386	2.1085
Mouse	7.0242	6.0303	2.1103
Rabbit	7.0244	6.0305	2.1103
Rat	4.0522	6.0300	2.1140
Bovine	7.0246	7.0246	2.1671
Gorilla	7.0242	6.0317	2.1103
Chimpanzee	7.0239	6.0313	2.1100

从表 2.3 可以看出，相应于 (W,S)- 特征序列的分量整体偏小，这表明在如上每一条编码序列中强弱氢键交替出现得比较频繁。此外，在这个分量上我们还可以看到，gallus，11 个物种中唯一的一个非哺乳动物，与其他物种表现出明显的不同。这是否意味着强弱氢键会在哺乳动物与非哺乳动物的区分方面起到某种特殊的作用呢？

序列一旦具有了向量（描述子）的形式，我们便可以将序列之间的比较转化为向量之间的比较。通常认为，如果两个 k- 维向量具有相似的方向并且具有相近的模长，那么与之对应的 DNA 序列就是相似的。这就是说，我们可以从两个方面来考察序列之间的相似性： (1) 两个向量 $\vec{a}, \vec{b}$ 端点之间的欧氏距离 $d(\vec{a}, \vec{b})$， (2) 两个向量夹角的余弦 $\cos(\vec{a}, \vec{b})$。 $d(\vec{a}, \vec{b})$ 越小，或者 $\cos(\vec{a}, \vec{b})$ 越大，则相应的两条序列就越相似。因此，这里我们用二者的商 $D_c = \frac{d(\vec{a}, \vec{b})}{\cos(\vec{a}, \vec{b})}$ 来衡量序列之间的相似性。显然， D_c 值越小，那么相应的两条序列就越相似。

表 2.4 列出了表 2.2 中 11 条编码序列基于 3 维向量 D_c 值的相似性。由表 2.4 可以看出， opossum 和 gallus 与其它物种的相似性较小，因为它们所对应的元素的值都比较大。另一方面，在这个表中， human-gorilla, gorilla-chimpanzee, human-chimpanzee 对的值都比较小，说明它们之间的相似性很大。这与文献 [31,42,103,116] 是一致的。不过从表中也能看到， rabbit 与 chimpanzee， mouse 与 chimpanzee 以及 mouse 与 human 等也表现出

很大的相似性. 虽然在上述文献里也曾出现过这些物种有相似性的结论，但这与我们的直观是有一定的距离的，可能是由于构造序列不变量时某些人为因素造成了这种结果，而且，在对这 11 个物种进行比较时，我们用的仅仅是它们的一个基因的一段 (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的长，并且都含有很多的基因，物种的全部遗传信息不可能包含在某一个基因里，所以利用一小段序列比较，得到的只能这些物种的某些相似性而不是全部.

表 2.4: The similarity/dissimilarity matrix for the 11 coding sequences based on the quotient D_c of the 3-component vectors

Species	Goat	Oposs.	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimp.
Human	3.31309	2.93560	3.82206	2.23033	0.00141	0.00121	3.08168	0.00014	0.99730	0.00069
Goat		1.40404	2.25739	2.29211	3.31358	3.31371	0.99719	3.31299	3.07583	3.31283
Oposs.			3.08088	3.06093	2.93666	2.936655	2.22690	2.93552	2.27714	2.93561
Gallus				2.49408	3.82211	3.82230	1.98733	3.82203	3.88685	3.82197
Lemur					2.22963	2.22990	1.42335	2.23025	2.81644	2.22980
Mouse						0.00028	3.08166	0.00140	0.99871	0.00109
Rabbit							3.08186	0.00122	0.99851	0.00099
Rat								3.08157	3.19843	3.08127
Gorilla									0.99730	0.00058
Bovine										0.99772

2.3 ALE- 指标

上一节我们分析了某种特殊矩阵的最大特征值在有效性方面可能面临的问题并给出了对策. 实际上，以最大特征值为不变量还有一个不可回避的问题，那就是计算的复杂性. 人类和一些模式生物基因组计划的完成，为我们从整个基因组的角度进行研究提供了条件. 然而，每一个物种的基因组序列是非常的长. 从上述矩阵的构造方法可以看出，这些矩阵的阶数基本上都是和生物大分子序列的长度相当的. 因而，随着序列长度的增加，矩阵的阶数自然要增大，这势必导致特征值的计算变得越来越难.

在这一节，我们给出一个新的序列不变量： ALE- 指标，它的计算非常简便，而其有效性绝不亚于最大特征值.

2.3.1 ALE- 指标

仔细观察上面所提到的矩阵，我们发现，这些矩阵有一个共同的特点，就是它们都是“主对角线元素为 0 的、非负、实对称” 矩阵. 现设 $M = (a_{ij})_{n\times n}$ 是满足上述条件的矩阵，即对任意的 $i, j = 1, 2, \ldots, n$, 总有 $a_{ij} \geq 0, a_{ij} = a_{ji}$, 且 $a_{ii} = 0$ 。我们定义：

$$\chi = \chi(M) = \frac{1}{2}(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F) \tag{2.3.1}$$

其中 $\|M\|_{m1} \equiv \sum_{i,j=1}^{n} |a_{ij}|$, $\|M\|_F \equiv \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} = \sqrt{tr(M^T M)}$ (这里 trM 表示 M 的迹)。

如果设 λ_1 是矩阵 M 的最大特征值，则有下面的不等式：

$$\frac{1}{n}\|M\|_{m1} \leq \lambda_1 \leq \sqrt{\frac{n-1}{n}}\|M\|_F \tag{2.3.2}$$

其中，第一个不等式见 [117,118]，第二不等式见 [119]。而且，式 2.3.2 中的上下界都是可以取到的。例如，对于 $M = \begin{pmatrix} 0 & k \\ k & 0 \end{pmatrix}$ 其中 $k \geq 0$, 相应的上下界都存在并且相等。因此，$\chi(M)$ 实际上是矩阵 M 的最大特征值 λ_1 的一个近似 (an **A**pproximation of the **L**eading **E**igenvalue)，正是从这个意义上，我们称之为矩阵 M 的 ALE- 指标. 由于它本质上只涉及到矩阵的 m_1- 和 F- 范数，因此它的计算是非常容易的。特别是在实际计算中，尽管 ALE- 指标在形式上是对矩阵而言，但在程序设计中我们并不需要存储整个矩阵，而是在构造矩阵的每个元素的同时按着范数的定义随时自动累计即可算出该矩阵的 ALE- 指标。由此可见，相对特征值而言，算法复杂度大大降低了，也因此能够直接用来处理长的序列。

现在结合第一章中我们提出的 3-D 图形表示及 L/L 矩阵（为了叙述方便，在本小节下面部分我们将 L/L 矩阵简记为 Q 矩阵）来进一步说明 ALE- 指标与最大特征值的关系以及 ALE- 指标的性质。回顾一下，我们的 3-D 图形表示是通过赋予四种碱基 A, C, G, T 四个向量 $(1,0,0)$，$(0,1,0)$，$(0,0,1)$，$(1,1,1)$ 而得到的. 我们称对应于 $(1,1,1)$ 赋予碱基 N $(N \in \Omega = \{A,C,G,T\})$ 的方式而得到的图形为 N- 图。于是一条 DNA 序列 S 有四个图形： A- 图、C- 图、G- 图和 T- 图. 在这四个图形的基础上构造的 Q 矩阵分别记为 $Q_A(S)$, $Q_C(S)$, $Q_G(S)$ 和 $Q_T(S)$，如果无须区分，我们将用 $Q(S)$ 代指其中的某一个.

若记 $\Delta = \chi - \lambda_1$, $\delta = \Delta/\lambda_1$，我们发现，绝对误差 Δ 总是不小于 0 的，这就是说，ALE- 指标总是要大于或者等于最大特征值的. 此外，随着序列长度的增加，Δ 将会变大，但相对误差 δ 却在总体上呈现出变小的趋势。这表明，对于长的序列而言，ALE- 指标将能更好地反映最大特征值。通过表 2.5，我们可以管窥 ALE- 指标与最大特征值之间的这种关系。

2.3.2 性质

(1) 递增性

对于任一给定的 DNA 序列 $S = S_1 S_2 \ldots S_n$，通过追加一个碱基 N，可以得到一条“扩张”序列 $S^* = S_1 S_2 \ldots S_n N$。通过赋予 4 种碱基 4 个向量 $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,1,1)$，我们可以得到对应于新序列 S^* 的一个点列：

$$P_1(x_1,y_1,z_1), P_2(x_2,y_2,z_2), \ldots, P_n(x_n,y_n,z_n), P_N(x_N,y_N,z_N),$$

表 2.5: First exons of *beta*-globin genes of human, mouse, and gallus: ALE-index VS leading eigenvalue

	first 16 bases				first 39 bases				whole sequence			
	χ	λ_1	Δ	δ	χ	λ_1	Δ	δ	χ	λ_1	Δ	δ
	human	ATGGTGCACC TGACTCCTGA GGAGAAGTCT GCCGTTACTG CCCTGTGGGG CAAGGTGAAC GTGGATGAAG TTGGTGGTGA GGCCCTGGGC AG										
Q_A	11.5297	11.4812	0.0485	0.0042	28.4445	28.3648	0.0797	0.0028	66.4831	66.3696	0.1135	0.0017
Q_C	12.4599	12.4323	0.0276	0.0022	29.3916	29.3262	0.0654	0.0022	66.8936	66.7941	0.0995	0.0015
Q_G	12.0033	11.9703	0.0330	0.0028	29.9255	29.8740	0.0515	0.0017	73.8689	73.8361	0.0328	0.0004
Q_T	11.9270	11.8787	0.0483	0.0041	28.7920	28.7194	0.0726	0.0025	68.5937	68.5082	0.0855	0.0013
	mouse	ATGGTGCACC TGACTGATGC TGAGAAGTCT GCTGTCTCTT GCCTGTGGGC AAAGGTGAAC CCCGATGAAG TTGGTGGTGA GGCCCTGGGC AGG										
Q_A	11.4838	11.4349	0.0489	0.0043	27.9317	27.8567	0.0750	0.0027	66.4045	66.2872	0.1173	0.0018
Q_C	12.3398	12.3117	0.0281	0.0023	28.5527	28.4813	0.0714	0.0025	68.2165	68.1073	0.1092	0.0016
Q_G	12.1082	12.0729	0.0353	0.0029	29.8575	29.8118	0.0457	0.0015	73.1837	73.1364	0.0473	0.0007
Q_T	11.8919	11.8439	0.0480	0.0041	29.5363	29.4844	0.0519	0.0018	69.6615	69.5597	0.1018	0.0015
	gallus	ATGGTGCACT GGACTGCTGA GGAGAAGCAG CTCATCACCG GCCTCTGGGG CAAGGTCAAT GTGGCCGAAT GTGGGGCCGA AGCCCTGGCC AG										
Q_A	11.4416	11.3890	0.0525	0.0046	29.4872	29.4359	0.0513	0.0017	67.2666	67.1602	0.1064	0.0016
Q_C	11.9522	11.9170	0.0352	0.0029	29.0093	28.9500	0.0593	0.0020	69.5967	69.5109	0.0858	0.0012
Q_G	12.4406	12.4077	0.0329	0.0027	30.5050	30.4677	0.0373	0.0012	73.0014	72.9397	0.0617	0.0009
Q_T	11.8694	11.8222	0.0472	0.0040	27.7517	27.6655	0.0862	0.0031	64.9901	64.8698	0.1203	0.0019

进而其 Q 矩阵：

$$Q(S^*) = \begin{pmatrix} & & & b_1 \\ & & & b_2 \\ & & Q(S) & \vdots \\ b_1 & b_2 & \cdots & \end{pmatrix} \tag{2.3.3}$$

其中 $b_i = \frac{|P_N - P_i|}{|P_i - P_{i+1}| + |P_{i+1} - P_{i+2}| + \ldots + |P_{n-1} - P_n| + |P_n - P_N|}$ $(i = 1, 2, \ldots, n)$。由于 $0 < b_i \leq 1$，所以有

$$\chi(Q(S^*)) - \chi(Q(S)) > 0$$

这意味着 ALE- 指标将随着序列的“扩张”而增大（这从表 2.5 也能看到）。

(2) 连续度

设 $S = \ldots N_{i-1} N_i \ldots N_j N_{j+1} \ldots$。如果 $N_{i-1} \neq N_i = \ldots = N_j \neq N_{j+1} \ldots$，我们就称 $B = N_i \ldots N_j$ 是序列 S 的一个“块 (block)”，并称 $j - i$ 为这个块的连续度，记作 $dc(B)$。从而，序列 S 可以用 block 的形式表示出来：$S = B_1 B_2 \ldots B_k$。我们定义序列 S 的连续

度为：

$$dc(S) = \sum_{m=1}^{k} dc(B_m)$$

显然，如果序列 S 的长度为 n，则 $n - dc(S)$ 恰好是 S 的 block 的个数，为了方便，我们记之为 $NB(S)$，并记 S 中具有最大 dc 的 block 的个数为 $NB_L(S)$。

对于两条长为 n 的 DNA 序列 S_1, S_2，我们称 S_1 的连续度比 S_2 的高，如果下列两个条件中任意一个成立：

(i) $dc(S_1) > dc(S_2)$,

(ii) $dc(S_1) = dc(S_2)$，但是要么 $\max\{dc(B_{1i})\} > \max\{dc(B_{2j})\}$，要么 $\max\{dc(B_{1i})\} = \max\{dc(B_{2j})\}$，但是 $dc(S_1 \setminus rB) > dc(S_2 \setminus rB)$. 其中 B_{1i} $(i = 1, 2, \ldots, k_1)$ 和 B_{2j} $(j = 1, 2, \ldots, k_2)$ 分别表示序列 S_1 和 S_2 的"块"。$r = \min(NB_L(S_1), NB_L(S_2))$，$S_g \setminus rB$ $(g = 1, 2)$ 表示从 S_g 中去掉 r 个具有最大 dc 的块后剩下的部分。

不难发现，$dc(S)$ 越高，$\chi({}^bQ(S))$ 就越大，反之亦然。

(3) 极值

设 S 是一长为 n 的 DNA 序列，则

(i) $\chi(Q_A(S)) = \chi(Q_C(S)) = \chi(Q_G(S)) = \chi(Q_T(S)) = n - 1$ 的充分必要条件是 $dc(S) = n - 1$，即 S 只由一种核苷酸碱基组成，如，$S = AA\ldots A = Ploy(A)$。而且，对于所有长为 n 的 DNA 序列，$n - 1$ 是 χ 的最大值。

(ii) 如果 $dc(S) = 0$，即 $NB(S) = n$，则 ${}^bQ(S)$ 具有如下形式：

$${}^bQ(S) = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 0 \end{pmatrix}$$

易见，$\chi({}^bQ(S)) = (\frac{1}{n} + \sqrt{\frac{1}{2n}})(n-1)$，并且，这是 χ 的下界。

从而，对任一长为 n 的 DNA 序列，我们有：

$$(\frac{1}{n} + \sqrt{\frac{1}{2n}})(n-1) \leq \chi(Q(S)) \leq n - 1$$

从性质 (2) 和 (3) 可见，我们不但能够从一条序列的 ALE- 指标，χ，获得该序列在构形上的信息，而且还可以通过两条序列本身直接大致地比较它们的 χ 值的大小。

2.3.3 应用

在这一节的最后，我们结合 5 个物种的 *beta, gamma, epsilon*-globin, *neurogenin* 和 *neuroD* 基因（见表 2.6）给出 ALE- 指标初步的应用。为了避免在序列比较时两条序列长度的不同所造成的影响，可以考虑使用加权的 ALE- 指标。这里我们采用最常见的一种加权形式，即正规化： $\chi' = \chi/n$，其中 n 是序列的长度，也是相应矩阵的阶。如前所述，ALE- 指标，进而其正规化形式的计算是非常容易的，因而可以直接用来处理长的序列。以序列 M91037 (Homo sapiens G-*gamma* globin and A-*gamma* globin genes, NCBI) 为例，尽管它的长度为 11393 bp，它的四个对应于矩阵 Q_A, Q_C, Q_G, Q_T 的正规化的 ALE- 指标还是非常容易地被计算出来，它们分别是：

$$\chi'_A = 0.757477, \quad \chi'_C = 0.704654, \quad \chi'_G = 0.722373, \quad \chi'_G = 0.754462$$

表 2.6: Database source

species	sequences	database	ID/AC	location
human	*beta*-globin	EMBL	HSHBB	62187-63610
	epsilon-globin	EMBL	HSHBB	19289-20961
	gamma globin	NCBI	M91037	
	neuroD	NCBI	U50822	
	neurogenin 1	NCBI	NM_006161	
chimpanzee	*beta*-globin	EMBL	PTGLB1	4189-5532
mouse	*beta*-globin	EMBL	MMBGL1	275-1462
	neuroD	NCBI	NM_010894	
	neurogenin 1	NCBI	BC062148	
rat	*beta*-globin	EMBL	RNGLB	310-1505
	neuroD	NCBI	D82945	
	neurogenin	NCBI	U67777	
gallus	*beta*-globin	EMBL	GGGL02	465-1810
	epsilon-globin	EMBL	GGHBBRE	20349-21873
	neuroD	NCBI	AF060885	
	neurogenin 1	NCBI	AJ012660	

在表 2.7 我们列出了 human, chimpanzee, mouse, rat 和 gallus 这五个物种的 *beta*-globin 基因的 CDS 、 Introns 以及整个基因的相应的正规化的 ALE- 指标。此外，我们还分别在表 2.8 、 2.9 给出了 human 和 gallus 的 *epsilon*-globin 基因、以及 human, mouse, rat 和 gallus 的 *neurogenin* 和 *neuroD* 基因的相应的正规化的 ALE- 指标。从这些表中我们可以看到，除了 gallus 这唯一的一个非哺乳动物外，其它物种的 CDS 的 χ'_A 、 χ'_T 大体上都比 introns 的小，而 χ'_G 、 χ'_C 却大。

表 2.7: Normalized ALE-indices of matrices Q_A, Q_C, Q_G, and Q_T for *beta*-globin genes of five species: human, chimpanzee, mouse, rat and gallus

species		human	chimpanzee	mouse	rat	gallus
CDS	χ'_A	0.706851	0.705343	0.720117	0.721493	0.715427
	χ'_C	0.742856	0.736867	0.747247	0.737042	0.776458
	χ'_G	0.763747	0.768885	0.751660	0.748930	0.742461
	χ'_T	0.734740	0.740234	0.726074	0.736247	0.719555
Introns	χ'_A	0.773660	0.773118	0.744233	0.744911	0.746021
	χ'_C	0.706172	0.705246	0.712964	0.716074	0.716587
	χ'_G	0.699275	0.700753	0.720568	0.713307	0.792600
	χ'_T	0.815062	0.817109	0.809794	0.799505	0.712577
Whole sequences	χ'_A	0.753499	0.753763	0.732444	0.734419	0.733348
	χ'_C	0.710812	0.709336	0.720937	0.719882	0.730350
	χ'_G	0.712307	0.714382	0.725045	0.720675	0.776980
	χ'_T	0.793699	0.795884	0.782648	0.779037	0.713454

表 2.8: Normalized ALE-indices of matrices Q_A, Q_C, Q_G, and Q_T for *epsilon*-globin genes of human and gallus

species		CDS	introns	whole sequences
human	χ'_A	0.722906	0.782308	0.761616
	χ'_C	0.739140	0.677548	0.696289
	χ'_G	0.746910	0.731284	0.731807
	χ'_T	0.735290	0.774535	0.758863
gallus	χ'_A	0.718527	0.743595	0.734998
	χ'_C	0.768351	0.724440	0.733497
	χ'_G	0.749461	0.750527	0.749797
	χ'_T	0.714530	0.725275	0.721668

在图 2.1-2.4 我们给出了表 2.7-2.9 中相应 ALE- 指标的图形. 可以看到, 对于每个哺乳动物, CDS 的四个正规化的 ALE- 指标形成上凸曲线, 而 introns 和整个基因的都是下凹的. 对于 gallus , CDS 的四个正规化的 ALE- 指标形成上凸曲线, 但 introns 和整个 *beta* 以及 *epsilon*-globin 基因的却不再是下凹曲线。 gallus 是非哺乳动物而其它的都是哺乳动物大概是导致这一明显差异的一个原因. 此外, 图 2.1 , 2.2 , 2.4 还显示出, human

表 2.9: Normalized ALE-indices of matrices Q_A, Q_C, Q_G, and Q_T for CDS of *neurogenin* and *neuroD* genes

species		human	mouse	rat	gallus
neurogenin	χ'_A	0.708583	0.699440	0.697353	0.709295
	χ'_C	0.811860	0.801830	0.801563	0.829484
	χ'_G	0.783072	0.772424	0.777495	0.778177
	χ'_T	0.687691	0.702011	0.701598	0.704637
neuroD	χ'_A	0.736484	0.742439	0.743348	0.722028
	χ'_C	0.770810	0.774895	0.769644	0.809090
	χ'_G	0.745069	0.742812	0.741603	0.784720
	χ'_T	0.703437	0.702002	0.704582	0.680224

和 chimpanzee 相似， mouse 和 rat 相似，而 gallus 则同其它物种明显的不相似。这是与事实相符的。

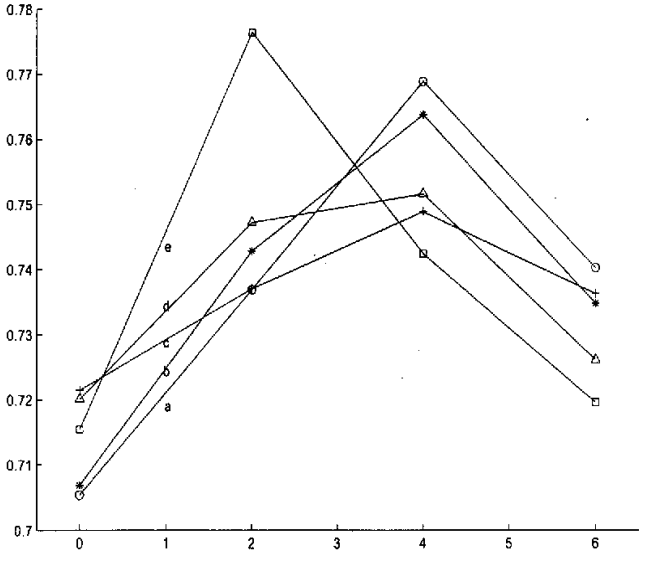


图 2.1: The plots of normalized ALE-indices for CDS of *beta*-globin genes of five species: (a) chimpanzee, (b) human, (c) rat, (d) mouse, (e) gallus.

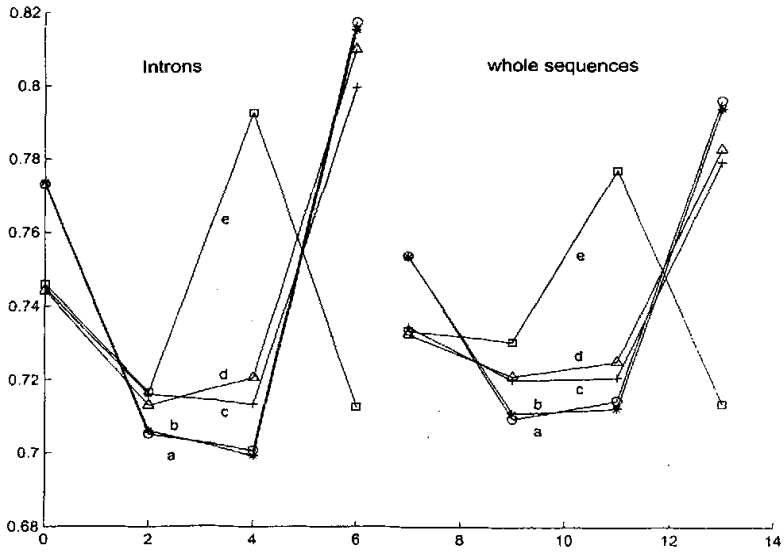


图 2.2: The plots of normalized ALE-indices for introns and the whole sequences of *beta*-globin genes of five species: (a) chimpanzee, (b) human, (c) rat, (d) mouse, (e) gallus

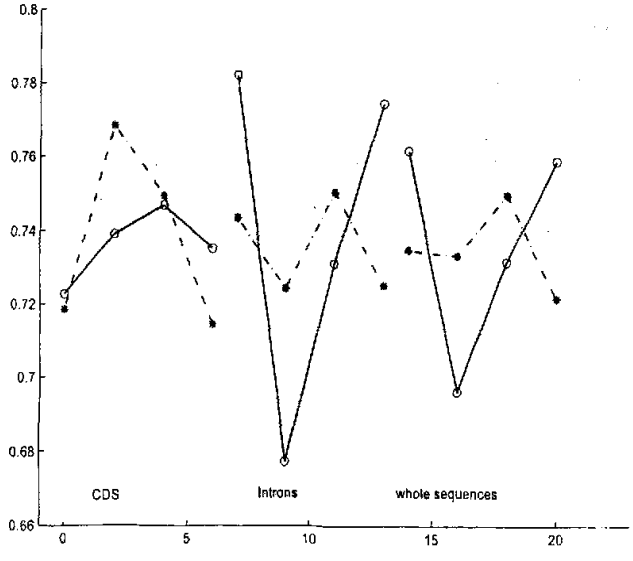


图 2.3: The plots of normalized ALE-indices for *epsilon*-globin genes of human (real line), and gallus (dashed line)

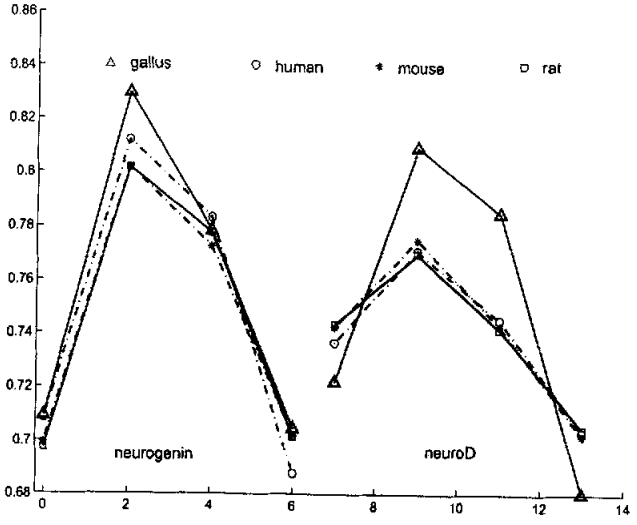


图 2.4: **The plots of normalized ALE-indices for CDS of *neurogenin* and *neuroD* genes of human, mouse, rat, and gallus**

2.4 上三角矩阵表示

在本章开头，我们已经介绍了几种基于图表示的矩阵。其中 GD 矩阵的 (i,j)- 元素定义为图 / 曲线上两顶点 v_i 和 v_j 之间的图论距离. 基于无向图的 GD 矩阵是个对称矩阵，然而，对于我们提出的有向图表示来说，虽然它在字面上仍是“图 / 曲线上两顶点 v_i 和 v_j 之间的图论距离”，但从它的数学表达式我们一眼就能看出这个矩阵已经发生了重要变化：

$$[GD]_{ij} = \begin{cases} j-i & \text{if } j \geq i \\ \infty & \text{otherwise} \end{cases} \tag{2.4.1}$$

我们再来考察 D/D 矩阵，它的 (i,j)- 元素定义为 ED 和 GD 矩阵相应元素的商：

$$[D/D]_{ij} = \begin{cases} [ED]_{ij}/[GD]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j, \end{cases} \tag{2.4.2}$$

由于 $[GD]_{ij} = \infty$ $(i > j)$，所以有向图的 D/D 矩阵也不再是对称矩阵，而成了上三角矩阵. 我们知道，一旦一个对称矩阵 M 被给出，人们经常以基于矩阵的不变量，如平均矩阵元素、平均行和、最大特征值、 Wiener 数、以及我们提出的 ALE- 指标等作为序列的

不变量. 为了方便，我们将上述不变量依次记为 $Ie(M)$ 、 $Ir(M)$ 、 $\lambda(M)$ 、 $Iw(M)$ 、 $\chi(M)$ ，并讨论它们对上三角矩阵表示来说是否具有兼容性.

2.4.1 序列不变量的相容性

设 $M=(a_{ij})_{n\times n}$ ，其中 $a_{ij}=a_{ji}\geq 0$ 且 $a_{ii}=0$ 对所有的 $i,j=1,2,\ldots n$ 都成立. 与其相对应的上三角矩阵记为 $\hat{M}=(b_{ij})_{n\times n}$ ，其中

$$b_{ij} = \begin{cases} a_{ij} & \text{if } j\geq i \\ 0 & \text{otherwise} \end{cases} \tag{2.4.3}$$

从而有：

(1) $Ie(\hat{M})=\frac{1}{2}Ie(M)$, $Ir(\hat{M})=\frac{1}{2}Ir(M)$, $Iw(\hat{M})=\frac{1}{2}Iw(M)$. 这表明上三角矩阵的平均矩阵元素、平均行和以及 Wiener 数仍可以用作序列不变量.

(2) 最大特征值. 易见， $\lambda(\hat{M})\equiv 0$ ，而 $\lambda(M)$ 通常是不等于 0 的. 因此，虽然 M 和 $\hat{M}$ 可以相互唯一确定，但是 $\lambda(M)$ 却不能由 $\lambda(\hat{M})$ 直接反映出来. 这意味着，在上三角矩阵情形，最大特征值不能再被用作不变量来刻画序列.

(3) ALE- 指标 χ . 观察 ALE- 指标 χ 的定义，我们发现，尽管 $\hat{M}$ 不再是对称矩阵，但我们仍然可以计算它的“形式上的” ALE- 指标. 并且，不难发现，

$$\|M\|_{m1}=2\|\hat{M}\|_{m1},$$

$$\|M\|_F=\sqrt{2}\|\hat{M}\|_F$$

如果我们用 α 表示 $(\frac{1}{2n}\|\hat{M}\|_{m1},\sqrt{\frac{n-1}{4n}}\|\hat{M}\|_F)$, 则下面的等式总成立：

$$\chi(M)=\frac{1}{2}(\frac{1}{n}\|M\|_{m1}+\sqrt{\frac{n-1}{n}}\|M\|_F)=(2,\sqrt{2})\cdot\alpha$$

$$\chi(\hat{M})=\frac{1}{2}(\frac{1}{n}\|\hat{M}\|_{m1}+\sqrt{\frac{n-1}{n}}\|\hat{M}\|_F)=(1,1)\cdot\alpha$$

其中“ $\cdot$ ”表示两个向量之间的内积. 这意味着一个上三角矩阵的“形式上的” ALE- 指标与相应的对称矩阵的 ALE- 指标能够相互唯一确定，从而上三角矩阵的 ALE- 指标可以用作序列不变量.

2.4.2 有向图及上三角矩阵的应用

现在，让我们回顾一下以前的基于不变量的序列比较方法：

给出序列的（**无向**）**图** 表示 → 构造 **对称矩阵** → 以 **最大特征值** 为不变量 → 序列的描述子.

至此，我们实际上已经给出了一条与此平行但在各个环节都有不同程度的改进的路线：

给出序列的 **有向图** 表示 → 构造 **上三角矩阵** → 以 **ALE- 指标** 为不变量 → 序列的描述子。

在这一小节，我们将沿此对人和其它七个物种（见表 2.10 ）之间的相似性进行分析。我们选用 Nandy 的 2-D 图为 "基图"，并按照第 1 章第 4 节有向图的定义给出相应的有向图，进而按照上面的方法得到它的上三角矩阵，并计算其 ALE- 指标。我们知道，Nandy 的 2-D 图形表示实际是通过将向量 $(-1,0)$, $(1,0)$, $(0,1)$, 和 $(0,-1)$ 分别赋予 4 种碱基 A, G, C, 和 T 而得到的。将这样的 4 个向量赋予 4 种碱基，本质上只有三种不同的方式，因而完全描述一条 DNA 序列，需要三个图形，它们恰是：Nandy 的 (AG-CT) 图，Gates 的 (AT-GC) 图 [11]，和 Leong and Mogenthaler 的 (AC-GT) 图 [12]。从而，一条 DNA 序列可以由一个 3 维向量来刻画，这个向量的分量就是与这三个图相对应的上三角矩阵的正规化的 ALE- 指标。在表 2.11 我们给出了表 2.10 中的 8 条序列的 3 维向量表示。

表 2.10: The coding sequences of the exon 1 of *beta*-globin gene of 8 different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
(92 bases)	GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCT
(92 bases)	GGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCT
(92 bases)	GGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGT
(92 bases)	GGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGT
(92 bases)	GGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTG
(90 bases)	GGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCA
(86 bases)	AGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA
(105 bases)	GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

表 2.11: The 3-component vectors associated with the 8 sequences in Table 2.10

Digraphs	Human	Opossum	Gallus	Lemur	Rat	Rabbit	Goat	Chimpanzee
AG-CT	0.180420	0.139922	0.174217	0.171890	0.164949	0.185368	0.186073	0.184112
AT-GC	0.193267	0.157891	0.155884	0.206649	0.176081	0.210271	0.199680	0.194083
AC-GT	0.176060	0.137588	0.184739	0.175107	0.156429	0.190593	0.200011	0.170708

这里，我们用两个向量的 Euclidean 距离来衡量它们之间的相似性。显然，两个向量的 Euclidean 距离越小，那么相应的两条 DNA 序列就越相似。表 2.12 列出了人与其它七个物种之间的距离。同时，我们在表 2.12 中还列出了文献中利用不同的 DNA 序列描述子得到的结果。这些描述子包括：基于三联体出现频率的 64 维向量 [42], 由 12 个 (4×4)

矩阵的最大特征值构成的 12 维向量 [42], 由 12 个对应于“四水平线图”的 L/L 矩阵的正规化最大特征值构成的 12 维向量 [31]. 从表 2.12 可以看出, 这些不同方法所得到的相似性基本上是一致的. 需要指出的是, 在最近的一个基于 15 维向量描述子的结果中, 出现了 human 与 chimpanzee 之间的相似程度明显低于 human 与包括 rat, rabbit 特别是 gallus 在内的一些物种之间的相似程度, 这从进化的意义上讲是不太可能的现象. 而观察我们的结果, 我们看到在八个物种中, human 与 chimpanzee 之间的相似性是最大的, 因为它们之间的距离最小, 我们有理由相信这绝不是偶然的.

表 2.12: The similarity/dissimilarity of the coding sequences of the exon 1 of *beta*-globin gene of 7 species with that of human based on (A) the 3-component vectors, (B) the 64-component vectors, (C) the 12-component vectors, (D) the 12-component vectors, and (E) the 15-component vectors, which represent the coding sequences

Species	Opossum	Gallus	Lemur	Rat	Rabbit	Goat	Chimpanzee
A^a	0.0661	0.0389	0.0159	0.0303	0.0229	0.0254	0.0066
B^b	11.402	10.630	10.100	8.246	6.708	8.944	–
C^c	4.4910	5.0150	2.9700	4.8570	3.1710	4.9960	–
D^d	0.1480	0.1090	0.0870	0.0430	0.0420	0.0610	0.017
E^e	0.009273	0.004180	0.009288	0.004277	0.004669	0.007723	0.004679

[a] This work
[b] From [Ref [42], Table 9]
[c] From [Ref [42], Table 12]
[d] From [Ref [31], Table 3]
[e] From [Ref [120], Table 8]

2.5 正规化相对熵

我们认为, 对于序列而言, 元素之间的序关系同元素本身一样包含着重要的信息. 事实上, 反映序列中元素之间的序关系的重要性的例子在日常生活中是屡见不鲜的. 例如, 在支票或者帐单上, 数字序列 00500000 意味着 5000 元, 而 00000050 却代表 50 分. 一般而言, 在一个数字序列中第一个非 0 数字越靠左, 那么这个数字序列对应的值就越大. 这就是说, 数值的大小对元素的序关系是非常敏感的. 类似的现象在生物大分子序列中同样可见, 例如, 人的 β-globin 基因的第一个外显子序列为 ATGGTGCAC...GAT.... 如果我们将处于第 64 和 65 位置上双核苷酸碱基 GA 与 66 位上的碱基 T 对调, 则序列变为: ATGGTGCAC...TGA..., 这将导致仅前 63 个碱基被翻译成一条短的蛋白质 (确切的说应该是肽链), 这是因为第 22 个三联体, TGA , 是一个终止密码子. 两条序列中所含的 4 种碱基的个数丝毫不差, 但“产品”却截然不同. 是什么造成它们会有如此巨

大的差异呢？唯一的答案还是字符之间的序关系。所有这一切都意味着，序列中元素之间的序关系，如果不比元素本身更重要的话，那就和元素本身同样重要。然而，文献中许多基于子串计数的方法都只是考虑序列中元素的组成，而元素的先后顺序却在很大程度上被忽略了，这必将导致生物大分子序列中某些重要信息的丢失。在这一节，我们将结合信息论的知识给出序列的正规化相对熵来反映序列中元素，特别是它们之间的序关系所包含的信息。

2.5.1 定义

我们知道，一个有限的非负实数序列 $x_1, x_2, \ldots, x_k$ 可以看作是一个标号的多重集 $S = \{x_1, x_2, \ldots, x_k\}$。为了反映元素 x_i 的序，我们考察由 S 的元素的部分和组成的集合 $D(S)$:

$$x_1,\ x_1 + x_2,\ \ldots,\ x_1 + x_2 + \ldots + x_k.$$

显然，$D(S)$ 的元素满足 $x_1 \le x_1 + x_2 \le \ldots \le x_1 + x_2 + \ldots + x_k$. 这就是说，$(D(S), \le)$ 是一个链 (chain)，即全序集。为了方便，我们将 $D(S)$ 记为

$$D(S) = \{X_1, X_2, \ldots, X_k\},$$

其中 $X_i = \sum_{j=1}^{i} x_j \quad (i = 1, 2, \ldots, k)$。易见，标号的多重集 S 和链 $D(S)$ 能够相互唯一确定。

作为离散集，$D(S)$ 可以看作是一个点分划，进而可以构造一个离散概率分布 $P = \{p_1, p_2, \ldots, p_k\}$, 其中 $\sum_{i=1}^{k} p_i = 1$，$p_i = \frac{X_i}{X_1 + X_2 + \ldots + X_k}$。根据信息论的知识，这个离散概率分布的 Shannon 熵可以按式 2.5.1 计算 [121]- [123]：

$$H(S) = H(D(S), P) = -\sum_{i=1}^{k} p_i \log_2 p_i. \tag{2.5.1}$$

这样定义的熵具有下述四个性质：

(1) 对于任意具有 k 个分量的离散概率分布，其熵都属于闭区间 $[0, \log_2 k]$；

(2) 若 $S = \{x_1, x_2, \ldots, x_k\}$，则对所有 $m > 0$ 有 $H(mS) = H(S)$，这里 $mS = \{mx_1, mx_2, \ldots, mx_k\}$；

(3) 设 $\varepsilon_1 = \{1, 0, \ldots, 0, 0\}$, $\varepsilon_k = \{0, 0, \ldots, 0, 1\}$。并设 $S = \{x_1, x_2, \ldots, x_k\}$。则 $H(S) = 0$ 的充分必要条件是存在一个 $m > 0$ 使得 $S = m\varepsilon_k$；又，$H(S) = \log_2 k$ 的充分必要条件是存在一个 $m > 0$ 使得 $S = m\varepsilon_1$；

(4) 设 $\varepsilon_1 = \{1, 0, \ldots, 0, 0\}$, $\varepsilon_2 = \{0, 1, \ldots, 0, 0\}$, $\ldots$, $\varepsilon_k = \{0, 0, \ldots, 0, 1\}$。则 $H(\varepsilon_1) > H(\varepsilon_2) > \ldots > H(\varepsilon_k)$。请注意，$\varepsilon_i$ 对应于数字序列中的第 i 个位置，因此，这一性质意味着按式 2.5.1 定义的熵可以反映出序列中位置的重要程度。

此外，相对熵，即熵的相对比可以按式 2.5.2 计算。

$$\eta(S)=\frac{H(S)}{H_{max}}, \tag{2.5.2}$$

其中 $H_{max}=\log_2 k$ 。

如果我们将一个 k 元标号多重集看作一个 k- 维空间中的向量（事实上，人们经常这样做），那么由性质 (2) 和公式 2.5.2 可知，方向相同的 k 元多重集（向量）具有相等的相对熵。另一方面，从几何学的角度，要想精确地描述出一个 k- 维空间中的向量 $v=(x_1,x_2,\ldots,x_k)$，需要两个参数：一个是方向，另一个是向量的模长 $\|v\|=\sqrt{\sum_i x_i^2}$。注意到这一点，并且考虑到向量的维数或者说集合的大小可能产生的影响，我们用式 2.5.3 来刻画一个有限的非负实数序列 $S=\{x_1,x_2,\ldots,x_k\}$

$$h_L=h_L(S)=\frac{H(S)\Big/H_{max}}{\|S\|\Big/\sqrt{k}} \tag{2.5.3}$$

其中 $\|S\|=\sqrt{\sum_i x_i^2}$。为了方便，我们称 $h_L(S)$ (或 h_L) 为序列 S 的正规化的相对熵.

2.5.2 应用

对一个 DNA 序列，我们按照序列中每个基所在的位置，将它分为三个子序列如下：首先按从左到右的方向取这个 DNA 序列的在 $3p+1(p=0,1,2,\cdots)$ 位置上的所有碱基形成一个特别的片段，我们称这个序列为这个 DNA 序列的 1- 子序列. 用类似的方法，我们定义 2- 子序列和 3- 子序列分别为这个 DNA 序列的在位置 $3p+i$ ($p=0,1,2,\cdots$； $i=2$ 或 3) 的所有碱基构成的子序列. 假设 i- 子序列 ($i=1,2,3$) 具有 k_i 个基，则可以赋予它 4 个 k_i 元标号多重集：

$$S_{iY}=\{f_{iY1},f_{iY2},\ldots,f_{iYk_i}\}\quad(Y=A,C,G,T), \tag{2.5.4}$$

其中 f_{iYj} $(j=1,2,\ldots,k_i)$ 表示基 Y 在 i- 子序列的前 j 个基中出现的频率。进而我们按式 2.5.3 计算它们的正规化相对熵。易见，这样的 4 个多重集能够唯一地表示相应的子序列，从而，一条 DNA 序列可以由如下 12- 维向量来刻画：

$$(h_L(S_{1A}),h_L(S_{1C}),h_L(S_{1G}),h_L(S_{1T}),h_L(S_{2A}),h_L(S_{2C}),h_L(S_{2G}),h_L(S_{2T}),h_L(S_{3A}),h_L(S_{3C}),h_L(S_{3G}),h_L(S_{3T}))$$

MIPS(the Munich Information Center for Protein Sequences) 数据分类库酿酒酵母基因组的第一类 ORF，即已知的蛋白质，共有 3392 条序列（线立体的除外）. 我们从中随机

选择 100, 200, 500, 1000, 2000 和 3000 条序列并计算它们的 12- 维向量，在表 2.13 左侧，我们分别列出了它们的平均向量。

此外，我们还从酿酒酵母基因组的 16 条染色体中按下面的方法截取了非编码序列：(1) 找出 MIPS 数据分类库中被注释为 ORF 的序列的起始位置；(2) 计算在相邻两个 ORF 之间的 DNA 序列的长度，并舍去那些长度小于 300bp 的 DNA 序列；(3) 从剩下的所有长度不小于 300bp 的 DNA 序列中，从它们的第一个碱基开始搜索密码子 "ATG"；然后从密码子 "ATG" 开始往下游的方向一个一个密码子的搜索，在第 101 个密码子以后搜索结束密码子 (TAA,TGA,TAG)，遇到一个结束密码子时搜索结束，从而得到一条基因间的 DNA 序列。这样的序列经常含有几个结束密码子因而不可能是 ORF，进而常被视为非编码序列 [51,52], [124]。连续地在往下游的方向搜索更多的基因间的 DNA 序列，直到不能再找到为止。

对于上面的过程，可能找到的基因间的 DNA 序列非常多。我们随机选择了相应数量的基因间的 DNA 序列，并将它们的平均向量列在表 2.13 右侧。

表 2.13: The average vectors of the samples of protein coding and non-coding sequences

Coding sequences			Non-coding sequences		
100 samples	200 samples	500 samples	100 samples	200 samples	500 samples
2.8258657	2.8238464	2.8290745	2.8908449	2.9293518	2.9171448
6.2348228	6.3999550	6.3167015	5.6045277	5.4888218	5.6350777
3.4668537	3.4521938	3.4916172	5.3039043	5.7005470	5.7704155
4.7187012	4.6894428	4.6570442	3.2195938	3.0748743	3.0732355
2.8679161	2.8512489	2.8612064	3.0050877	3.0475996	3.1110214
4.3858952	4.2684013	4.3323254	5.8133408	5.6908134	5.7138748
7.0147690	7.2860519	7.2360497	5.5501366	5.5311268	5.6881065
3.4246105	3.4725067	3.4308743	2.9509781	2.9826641	2.8873745
3.4595089	3.4693546	3.4606354	3.0400379	3.0447432	3.0896740
5.2680897	5.1698189	5.1823372	5.6689056	5.6443970	5.7276951
4.7152298	4.8524124	4.8538475	4.8508583	4.7729156	4.8191879
3.1366604	3.0472683	3.0697285	3.1485361	3.1894255	3.1488590
1000 samples	2000 samples	3000 samples	1000 samples	2000 samples	3000 samples
2.8512194	2.8542795	2.8526176	2.9074736	2.9098501	2.9119383
6.2753380	6.2258923	6.1966835	5.6363426	5.6218521	5.6087345
3.4868256	3.4774303	3.4800640	5.7025833	5.6901657	5.6535740
4.5959625	4.6469532	4.6504536	3.0790478	3.0861153	3.0879122
2.8632723	2.8576616	2.8672752	3.0766981	3.0753771	3.0699488
4.3055317	4.3101707	4.2996674	5.6618824	5.6854729	5.6791756
7.0958061	7.1534592	7.1515091	5.5952923	5.6053730	5.6072609
3.4473171	3.4507023	3.4498695	2.9177189	2.9160429	2.9227104
3.4972797	3.4885124	3.4884438	3.0891022	3.0873640	3.0814230
5.2561626	5.2652763	5.2503307	5.6861366	5.6992138	5.6772896
4.8595955	4.8613106	4.8759853	4.7368885	4.7458519	4.7582882
3.0955257	3.0793430	3.0726254	3.1634178	3.1609743	3.1604469

从表 2.13 可以看到，编码序列的向量彼此相似，非编码序列的也是如此，但在编码序列和非编码序列之间却存在着明显的差别。这意味着这个 12- 维向量已经抓住了所考察的 DNA 序列的特征，并很有可能会在酿酒酵母基因组的蛋白质编码基因识别中派上用场。进一步，我们还就每一“编码位置”计算正规化相对熵所占的比例（见表 2.14 ）。由表 2.14 可以看出，非编码序列的三个“编码位置”大体上表现出相同的规律，而编码序列的却彼此相异。这意味着编码氨基酸和“ STOP ”的 64 个三联体是编码序列的字，但对“沉默”的非编码序列而言，它们很可能不是字，或者，如果是字的话它们不会是全部字。

表 2.14: The ratio of normalized relative-entropies, h_L's, at each of the "codon positions".

Codon position	Coding sequences (%)	Non-coding sequences (%)
1	17 : 36 : 20 : 27	17 : 32 : 33 : 18
2	16 : 24 : 41 : 19	18 : 33 : 32 : 17
3	21 : 31 : 29 : 19	18 : 34 : 29 : 19

3 逻辑序列与序列复杂度

核酸序列和蛋白质序列分别是基于 4 和 20 种字符的字符串，这种表示形式具有最高的“解像力”，即序列的每个细节都排列得非常清楚，但序列的长度一旦超过了一定的限度，这样的表示将很难给人留下整体的印象．如果我们借用代数学中的同态思想将其“粗粒化”，省略掉序列的某些细节，突出特征，然后将其表示成适当的数学对象，这将有助于研究序列的结构和隐藏在其中的规律．

在这一章，我们从不同的角度对 4 种核苷酸碱基 A, G, C, T 和 20 种氨基酸残基进行二分类，进而将原始序列简约为二元 (0,1) 序列．同时，我们给出了 (0,1) 序列的广义 LZ 复杂度，并将其和正规化相对熵分别应用到 DNA 及蛋白质序列的相似性分析．此外，我们还给出了 RNA 二级结构的“影子序列”，从而将 2 维结构转化为 1 维线性表示．我们在此基础上，结合序列复杂性，对 9 种病毒的 RNA 二级结构进行了比较．

3.1 DNA 序列的逻辑表示

四种核苷酸碱基 A, G, C, T，可以从逻辑学的角度分成两类： A 与非 A (即或 G 或 C 或 T)．从而，我们能按下面的操作将每一个 DNA 序列转换为一个 (0,1) 序列：若碱基是 A, 则记它为 1，否则，若它属于非 A 则记为 0．在这样的操作下， DNA 序列就变为了 (0,1) 序列，对应于这种操作得到的二元序列，我们称之为 DNA 序列的 A- 序列．在这种 (0,1) 表示中， DNA 序列自身的某些结构上的信息可能会丢失，但这种变换可以使和腺嘌呤相关的信息表现得更为明显．类似地，可以定义另外三条 (0,1) 序列： G- 序列，C- 序列和 T- 序列．上述过程用数学形式表示如下：

设 $X = x_1x_2\cdots$ 是一个 DNA 序列．定义四个字上的同态映射 $\phi_i(X) = \phi_i(x_1)\phi_i(x_2)\cdots$,

设 $X = x_1x_2\cdots$ 是一个 DNA 序列．定义四个字上的同态映射 $\phi_i(X) = \phi_i(x_1)\phi_i(x_2)\cdots$, $i = 1, 2, 3, 4$, 其中

$$\phi_1(x_j) = \begin{cases} 1 & \text{if } x_j = A \\ 0 & \text{if } x_j \neq A, \end{cases} \tag{3.1.1}$$

$$\phi_2(x_j) = \begin{cases} 1 & \text{if } x_j = C \\ 0 & \text{if } x_j \neq C, \end{cases} \tag{3.1.2}$$

$$\phi_3(x_j)=\begin{cases}1 & \text{if } x_j=G\\ 0 & \text{if } x_j\neq G,\end{cases} \tag{3.1.3}$$

$$\phi_4(x_j)=\begin{cases}1 & \text{if } x_j=T\\ 0 & \text{if } x_j\neq T,\end{cases} \tag{3.1.4}$$

我们称这四条 (0,1) 序列为相应 DNA 序列的逻辑表示 (logical representation)，并简记为 LR．其中的每一条都称为逻辑序列，分别用 LR^A, LR^C, LR^G, LR^T 来表示。例如，人的 β-globin 基因的第一个外显子的前 60 个碱基构成的序列片段的逻辑表示如表 3.1 所示。

表 3.1: The logical representation (LR) of the DNA sequence

Sequence (S)	ATGGTGCACC TGACTCCTGA GGAGAAGTCT GCCGTTACTG CCCTGTGGGG CAAGGTGAAC
$LR^A(S)$	10000001000001000000100101100000000001000000000000000110000110
$LR^C(S)$	00000010110001011000000000001001100001001110000000010000000001
$LR^G(S)$	00110100000100000010110100100010010000010000101111000011010000
$LR^T(S)$	01001000001000100100000000010100001100100001010000000000010000

由逻辑表示的定义及表 3.1，我们有

(1) 逻辑表示给出了相应 DNA 序列的所有信息，这是因为一个 DNA 序列被它的四个逻辑序列中任意三个所唯一决定。

(2) 逻辑表示清楚地显示出四种碱基的分布，因此，人们可以从逻辑表示直接得到核苷酸在 DNA 序列及其片段中的相对丰富程度以及位置上的信息。

(3) 逻辑表示可以用来数字辨别点突变，特别是辨别一个替换到底是转换 ($A\leftrightarrow G, C\leftrightarrow T$) 还是颠换 ($T\leftrightarrow A, T\leftrightarrow G, C\leftrightarrow A, C\leftrightarrow G$)．为此，我们将一个单个核苷酸碱基 Y 的逻辑表示写成一个 4- 维向量的形式：

$$y=(LR^A(Y), LR^G(Y), LR^C(Y), LR^T(Y))\doteq(y_a, y_g, y_c, y_t)$$

对于任意两个核苷酸碱基 Y 和 Z，如果 $y\cdot z=1$，其中 "$\cdot$" 表示两个向量之间的内积，则它们是匹配基；而 $y\cdot z=0$ 则意味着是一个替换．要进一步辨别到底是哪种类型的替换，我们只须考察 $y_a+z_a+y_g+z_g\doteq yz_{ag}$，显然，如果 $yz_{ag}=1$，那么这个替换就一定是个颠换；否则，就是转换．

3.1.1 逻辑表示同其它表示的比较

(1) 逻辑表示与特征序列

对任一给定的 DNA 序列 S，如果用 "$CS_{(\cdot)}$" 表示相应的特征序列 (详见 [103,104])，则由逻辑表示的定义，我们可以立即得到：

$$\begin{aligned} CS_{(M,K)} &= LR^A(S) \vee LR^C(S) \\ CS_{(R,Y)} &= LR^A(S) \vee LR^G(S) \\ CS_{(W,T)} &= LR^A(S) \vee LR^T(S) \end{aligned} \tag{3.1.5}$$

其中，“$\vee$”表示逻辑和运算。

(2) 逻辑表示与 Randic's 2-D 图

在第 1 章，我们曾介绍过 Randic 等人提出的 DNA 序列的“四水平线” 2-D 图形表示（详见 [30] ）. 以 ATGGTGCACCTGACTCCTGA，人的 β-globin 基因的第一个外显子的前 20 个碱基构成的序列片段为例，其 Randic's 2-D 图形如图 3.1 所示.

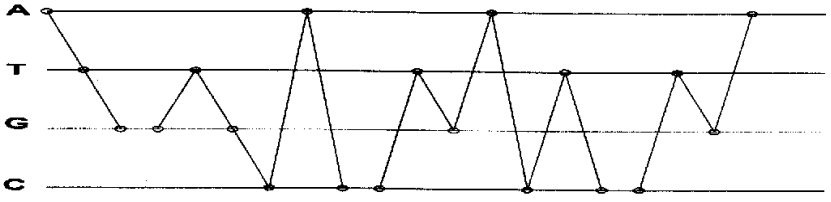


图 3.1: The graphical representation of the sequence ATGGTGCACCTGACTCCTGA

现在我们在表 3.2 按 A-T-G-C 的顺序给出这个序列片段的逻辑表示. 如果忽略“0”而仅将“1”一个接一个地连接起来，我们将立即得到这个序列的同样的 2-D 图形表示.

表 3.2: The LR of the sequence ATGGTGCACCTGACTCCTGA

$LR^A(S)$	1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1
$LR^T(S)$	0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0
$LR^G(S)$	0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0
$LR^C(S)$	0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 1 1 0 0 0

(3) 逻辑表示与 Randic's 4-D 表示

在文献 [102] 中， Randic and Balaban 按下面的方式将 4-D 空间中的四个坐标轴正方向分别赋予 4 种核苷酸碱基，从而给出了 DNA 序列的一种 4-D 表示：

$$\begin{aligned} A &\ (1,0,0,0) \\ T &\ (0,1,0,0) \\ G &\ (0,0,1,0) \\ C &\ (0,0,0,1) \end{aligned}$$

以人的 β-globin 基因的第一个外显子的前 15 个碱基构成的序列片段为例，它的 4-D 坐标如表 3.3 所示。

表 3.3: 4-D coordinates for the first 15 bases of the first exon of human β-globin gene

NO.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
base	A	T	G	G	T	G	C	A	C	C	T	G	A	C	T
e_A	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3
e_T	0	1	1	1	2	2	2	2	2	2	3	3	3	3	4
e_G	0	0	1	2	2	3	3	3	3	3	3	4	4	4	4
e_C	0	0	0	0	0	0	1	1	2	3	3	3	3	4	4

现在，我们用 $LR_i^Y(S)$ 表示逻辑序列 $LR^Y(S)$ $(Y = A,T,G,C)$ 中的第 i 个 "基"，并定义

$$\begin{cases} x_{1i} = \sum\limits_{k=1}^{i} LR_k^A(S) \\ x_{2i} = \sum\limits_{k=1}^{i} LR_k^T(S) \\ x_{3i} = \sum\limits_{k=1}^{i} LR_k^G(S) \\ x_{4i} = \sum\limits_{k=1}^{i} LR_k^C(S) \end{cases}$$

则我们可以由逻辑序列立即得到相同的坐标（参表 3.2 ）.

(4) 逻辑表示与 Z- 曲线

在第 1 章，我们曾介绍了张春霆等人提出的 DNA 序列的一种 3-D 图形表示— Z- 曲线（详见 [24] ）. 其三维空间中点 p_n 的坐标为：

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n) \\ y_n = (A_n + C_n) - (G_n + T_n) \\ z_n = (A_n + T_n) - (C_n + G_n) \end{cases}$$

其中，$x_n, y_n, z_n \in [-N, N]$ 且 $n \in \{0, 1, \ldots, N\}$ 。

由逻辑表示的定义及表 3.1 ，易见

$$Y_n = \sum_{k=1}^{n} LR_k^Y(S), \quad (Y = A,T,G,C)$$

因此，我们可以从逻辑表示直接得到上述点的坐标进而得到 Z- 曲线。

3.1.2 逻辑序列的 S/S 矩阵及其压缩矩阵

在第 2 章，我们曾提到了 DNA 序列的 S/S 矩阵（详见 [39] ）. 在这一小节，我们将给出逻辑序列的 S/S 矩阵. 假设 $b_1b_2b_3\ldots b_n$ 是一条 (0,1) 序列. 我们定义它的 $S/S=(s_{ij})_{n\times n}$ 矩阵为：

$$\begin{cases} s_{ji}=s_{ij}=\frac{n_{ij}}{j-i} & (i<j) \\ s_{ii}=0 \end{cases} \tag{3.1.6}$$

这里 n_{ij} 表示子串 $b_{i+1}\ldots b_j$ 中 b_j 所对应的“逻辑基”的个数.

由于一条 DNA 序列的逻辑表示包含 4 条逻辑序列，所以按上述定义，从一条 DNA 序列可以得到 4 个 S/S 矩阵. 例如，人的 β-globin 基因的第一个外显子的前 20 个碱基构成的序列片段： $A_1T_2G_3G_4T_5G_6C_7A_8C_9C_{10}T_{11}G_{12}A_{13}C_{14}T_{15}C_{16}C_{17}T_{18}G_{19}A_{20}$. 我们在表 3.4 给出了它的对应于 G- 序列 (见表 3.2 $LR^G(S)$) 的 S/S 矩阵.

表 3.4: The upper triangles of the matrix S/S corresponding to $LR^G(S)$

S/S	0_1	0_2	1_3	1_4	0_5	1_6	0_7	0_8	0_9	0_{10}	0_{11}	1_{12}	0_{13}	0_{14}	0_{15}	0_{16}	0_{17}	0_{18}	1_{19}	0_{20}
0_1	0	1/1	1/2	2/3	2/4	3/5	3/6	4/7	5/8	6/9	7/10	4/11	8/12	9/13	10/14	11/15	12/16	13/17	5/18	14/19
0_2		0	1/1	2/2	1/3	3/4	2/5	3/6	4/7	5/8	6/9	4/10	7/11	8/12	9/13	10/14	11/15	12/16	5/17	13/18
1_3			0	1/1	1/2	2/3	2/4	3/5	4/6	5/7	6/8	3/9	7/10	8/11	9/12	10/13	11/14	12/15	4/16	13/17
1_4				0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	2/8	7/9	8/10	9/11	10/12	11/13	12/14	3/15	13/16
0_5					0	1/1	1/2	2/3	3/4	4/5	5/6	2/7	6/8	7/9	8/10	9/11	10/12	11/13	3/14	12/15
1_6						0	1/1	2/2	3/3	4/4	5/5	1/6	6/7	7/8	8/9	9/10	10/11	11/12	2/13	12/14
0_7							0	1/1	2/2	3/3	4/4	1/5	5/6	6/7	7/8	8/9	9/10	10/11	2/12	11/13
0_8								0	1/1	2/2	3/3	1/4	4/5	5/6	6/7	7/8	8/9	9/10	2/11	10/12
0_9									0	1/1	2/2	1/3	3/4	4/5	5/6	6/7	7/8	8/9	2/10	9/11
0_{10}										0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	2/9	8/10
0_{11}											0	1/1	1/2	2/3	3/4	4/5	5/6	6/7	2/8	7/9
1_{12}												0	1/1	2/2	3/3	4/4	5/5	6/6	1/7	7/8
0_{13}													0	1/1	2/2	3/3	4/4	5/5	1/6	6/7
0_{14}														0	1/1	2/2	3/3	4/4	1/5	5/6
0_{15}															0	1/1	2/2	3/3	1/4	4/5
0_{16}																0	1/1	2/2	1/3	3/4
0_{17}																	0	1/1	1/2	2/3
0_{18}																		0	1/1	1/2
1_{19}																			0	1/1
0_{20}																				0

观察表 3.4，我们看到 S/S 矩阵的次对角线元素总是等于 1 的，而其它元素总是小于或等于 1 . 事实上，这可以从 S/S 矩阵的定义直接推导出来. 这样的矩阵有一个优点，那就是由它可以构造出一个按元素收敛的矩阵序列： ${}^kS/{}^kS\ \ (k=1,2,\ldots)$ ，它的 (i,j)-矩阵元素为 s_{ij}^k . 但同时我们也看到， S/S 及其“生成”矩阵的阶数都是和相应序列的长度相等的. 这样，长的序列就将产生非常“大”的矩阵，靠直接观察矩阵本身来获得有用的信息显得不太现实. 为了处理这个问题，我们可以从下面两个方面着手.

(1) 不变量. 在第 2 章，我们已经指出，一旦得到了一个与序列相对应的实对称矩阵，那么一些基于矩阵的不变量，如平均矩阵元素、平均行 / 列和、最大特征值、 ALE-指标等，就可以用作不变量来描述这条序列. 例如，表 3.4 所对应的矩阵的 ALE- 指标可以很容易地计算出来： $\chi = 14.4126$ ，正规化后为 $\chi' = 0.72063$.

(2) 压缩矩阵. 与针对 DNA 序列而构造的 4×4 矩阵类似，这里，我们考虑逻辑序列的 S/S 矩阵的 2×2 的压缩矩阵.

假设所考虑的逻辑序列由 m 个 " 1 " 和 n 个 " 0 " 组成. 于是，通过对 S/S 矩阵进行列或行互换使所有与 " 1 " 对应元素聚集在一起，同时使所有与 " 0 " 对应元素聚集在一起，我们就可以得到一个分块矩阵：

	0...0	1...1
0...0	$00_{n\times n}$	$01_{n\times m}$
1...1	$10_{m\times n}$	$11_{m\times m}$

将每一块都用一个适当的数来代替，于是一个 2×2 的压缩矩阵就得到了. 许多矩阵不变量都可以选作相应子阵的 "代表"，当然，一般而言， $m \neq n$ ，因此，通常不用特征值. 我们在表 3.5 给出了表 3.4 中 S/S 矩阵的一个分块矩阵.

进一步，如果用平均矩阵元素来代替每个块 (子阵)，那么我们就得到了一个压缩矩阵：

$$\begin{pmatrix} 0.75469 & 0.69154 \\ 0.69154 & 0.29307 \end{pmatrix}_{2\times 2}$$

3.2 蛋白质序列的逻辑表示

生物信息学的一个普遍接受的观点是：蛋白质序列中包含了蛋白质的结构和功能方面的所有信息. 虽然有时我们会遇到有些来自不同来源的蛋白质，它们具有不同的功能但拥有相似的序列，然而在一般情况下，人们总认为如果序列间有越高的相似性它们的结构就越相似. 但是，与核酸序列类似，蛋白质序列是基于字符集 Ω = {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，直接从序列本身提取出有效的信息是很难的. 为此，人们便从各种不同的角度来研究蛋白质序列. 例如， Randic [28,34], Feng and Zhang [25] 给出了蛋白质序列的几种图形表示. 而 Feng [92] 则提出了蛋白质序列的 "表征向量 (attribute vector) "，具体讲，就是用一个 20 维向量来表示蛋白质序列，这个向量的分量是 20 种氨基酸在序列中出现的频率的平方根. 由于这样的向量的模长是 1 ，所以所有蛋白质都在 20-D 空间中单位球面上有一个代表点，而且同一家族或者序列

表 3.5: The partitioned matrix corresponding to the S/S matrix in table 3.4

O_1	O_2	O_3	O_7	O_4	O_5	O_{10}	O_{11}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	O_{18}	O_{19}	I_2	I_4	I_6	I_{12}	I_{17}
0	1/1	2/4	3/6	4/7	5/8	6/9	7/10	8/12	9/13	10/14	11/15	12/16	13/17	14/19	1/2	2/3	3/5	4/11	5/18
1/1	0	1/3	2/5	3/6	4/7	5/8	6/9	7/11	8/12	9/13	10/14	11/15	12/16	13/18	1/1	2/2	3/4	4/10	5/17
2/4	1/3	0	1/2	2/3	3/4	4/5	5/6	6/8	7/9	8/10	9/11	10/12	11/13	12/15	1/2	1/1	1/1	2/7	3/14
3/6	2/5	1/2	0	1/1	2/2	3/3	4/4	5/6	6/7	7/8	8/9	9/10	10/11	11/13	2/4	2/3	1/1	1/5	2/12
4/7	3/6	2/3	1/1	0	1/1	2/2	3/3	4/6	5/6	6/7	7/8	8/9	9/10	10/12	3/5	3/4	2/2	1/4	2/11
5/8	4/7	3/4	2/2	1/1	0	1/1	2/2	3/4	4/5	5/6	6/7	7/8	8/9	9/11	4/6	4/5	3/5	1/6	2/10
6/9	5/8	4/5	3/3	2/2	1/1	0	1/1	2/3	3/4	4/5	5/6	6/7	7/8	8/10	5/7	5/6	4/4	1/2	2/9
7/10	6/9	5/6	4/4	3/3	2/2	1/1	0	1/2	2/3	3/4	4/5	5/6	6/7	7/9	6/8	6/7	5/5	1/1	2/8
8/12	7/11	6/8	5/6	4/5	3/4	2/3	1/2	0	1/1	2/2	3/3	4/4	5/5	6/7	7/10	7/9	6/7	1/1	1/6
9/13	8/12	7/9	6/7	5/6	4/5	3/4	2/3	1/1	0	1/1	2/2	3/3	4/4	5/5	8/11	8/10	7/8	2/3	1/5
10/14	9/13	8/10	7/8	6/7	5/6	4/5	3/4	2/2	1/1	0	1/1	2/2	3/3	4/5	9/12	9/11	8/9	3/5	1/4
11/15	10/14	9/11	8/9	7/8	6/7	5/6	4/5	3/3	2/2	1/1	0	1/1	2/2	3/4	10/13	10/12	9/10	4/4	1/3
12/16	11/15	10/12	9/10	8/9	7/8	6/7	5/6	4/4	3/3	2/2	1/1	0	1/1	2/3	12/14	11/13	10/11	5/5	1/2
13/17	12/16	11/13	10/11	9/10	8/9	7/8	6/7	5/6	4/4	3/3	2/2	1/1	0	1/2	12/15	12/14	11/12	6/6	1/1
14/19	13/18	12/15	11/13	10/12	9/11	8/10	7/9	6/7	5/6	4/5	3/4	2/3	1/2	0	13/17	13/16	12/14	7/9	1/1
1/2	1/1	1/2	2/4	3/5	4/6	5/7	6/8	7/10	8/11	9/12	10/13	11/14	12/15	13/17	0	1/1	2/3	3/9	4/16
2/3	1/2	1/1	2/3	3/4	4/5	5/6	6/7	7/9	8/10	9/11	10/12	11/13	12/14	13/16	1/1	0	1/2	2/8	3/15
3/5	2/4	1/1	1/1	2/2	3/3	4/4	5/5	6/7	7/8	8/9	9/10	10/11	11/12	12/14	2/3	1/2	0	1/6	2/13
4/11	4/10	2/7	1/5	1/4	1/3	1/2	1/1	1/3	2/2	2/3	4/4	5/5	6/6	7/8	3/9	2/8	1/6	0	1/7
5/18	5/17	3/14	2/12	2/11	2/10	2/9	2/8	1/6	1/5	1/4	1/3	1/2	1/1	1/1	4/16	3/15	2/13	1/7	0

高度一致的蛋白质的代表点将在这个球面上聚集在一起。另有一些研究人员构造了蛋白质序列的几种 $(20+m)$ 维向量表示，其中前 20 个分量用以反映氨基酸的组成，而后 m（m 是个待定值）个分量用于反映序列中元素的序关系 [91,93–95]。此外，在蛋白质折叠研究中广泛使用的 HP 格点模型，则是对蛋白质序列的一种“粗粒化”表示，这种简化模型反映了蛋白质系统的一些基本特性并表征了蛋白质的复杂性 [125]- [128]。

受上述工作的启发，我们将 DNA 序列的逻辑表示推广到蛋白质序列中，并结合我们提出的正规化相对熵给出了蛋白质序列的 12 维向量表示。

3.2.1 蛋白质序列的逻辑表示

氨基酸是蛋白质的基本组成单位，其自身的特性必然会对蛋白质产生重要的影响。在表 3.6，我们列出了 20 种氨基酸的 6 种重要性质，分别是：分子量 (Mw)，等电点 (pI

value), 疏水 / 极性 (HP), 相对距离 (RD), 不可替代性 (nS), 以及遗传密码简并度 (DD). 从表可以知道, 20 种氨基酸的平均分子量是 136.89, 我们以此为界将氨基酸分成两类:

Mw: $Class_1^1$={GASPVTCLIND}; $Class_1^2$={EFHKMQRWY}.

类似地, 我们得到

pI values: $Class_2^1$={GASVCLINDQEMFYW}, $Class_2^2$={PTKHR};

HP: $Class_3^1$={STCNDQKEHRY}; $Class_3^2$={AFGILMPVW};

RD: $Class_4^1$={APVTLIQKEMHRY}, $Class_4^2$={CDFGNSW};

nS: $Class_5^1$={CQMHFYW}, $Class_5^2$={ADEGIKLNPRSTV};

DD: $Class_6^1$={LSRAPTVGI}; $Class_6^2$={CDEFHKMNQWY}.

表 3.6: Some important properties of 20 amino acids

Amino acid	Mw [a]	pI-value [a]	HP [b]	RD [c]	nS [c]	DD [d]
G	75.07	5.97	Hydrophobic	2078	0.56	high
A	89.09	6.02	H	1889	0.52	h
S	105.09	5.68	Polar	2000	0.64	h
P	115.13	6.30	H	1720	0.61	h
V	117.15	5.97	H	1680	0.54	h
T	119.12	6.53	P	1469	0.56	h
C	121.12	5.02	P	3355	1.12	low
L	131.17	5.98	H	1822	0.58	h
I	131.17	6.02	H	1765	0.65	h
N	132.10	5.42	P	1943	0.79	low
D	133.10	2.97	P	2209	0.77	l
Q	146.15	5.65	P	1538	0.86	l
K	146.19	9.74	P	1797	0.81	l
E	147.13	3.22	P	1812	0.76	l
M	149.21	5.75	H	1689	1.25	l
H	155.16	7.59	P	1507	0.94	l
F	165.10	5.48	H	1916	0.86	l
R	174.20	10.76	P	1697	0.6	high
Y	181.19	5.66	P	1787	0.98	l
W	204.22	5.89	H	2317	1.82	l
average	136.89	6.08	-	1899.5	0.811	-

a: see [129]; b: see [130]; c: see [131]; d: see [132]

设 $X = x_1x_2\cdots$ 是一个蛋白质序列. 定义 6 个字上的同态映射 $\phi_i(X) = \phi_i(x_1)\phi_i(x_2)\cdots$, $i = 1, 2, \ldots, 6$, 其中

$$\phi_i(x_j) = \begin{cases} 0, & \text{if } x_j \in Class_i^1 \\ 1, & \text{otherwise} \end{cases} \tag{3.2.1}$$

于是, 一条蛋白质序列可以转换为 6 个 (0,1) 序列, 我们称之为蛋白质序列的逻辑序列. 以 human neurocan 的前 80 个氨基酸残基构成的序列片段为例, 它的 6 条逻辑序列见表 3.7. 从逻辑序列我们能够直接得到蛋白质序列中氨基酸相关性质的一些有用的信息. 例如, 观察第一条逻辑序列, 我们看到 0 的个数明显比 1 的个数多, 这意味着这 80 个残基

表 3.7: The first 80 residues of human neurocan and its six (0,1)-sequences

MGAPFVWALGLLMLQMLLFVAGEQGTQDITDASERGLHMQ KLGSGSVQAALAELVALPCLFTLQPRPSAARDAPRIKWTK
10001010000010110010001100100000011001111000000100001000000010010100001000101101
00010000000000000000000001000100001001001000000000000000010001001110001001101011
11111111111111011111110010001001000110100110101011110111111011010101011001101010
01001010010000000010010010010010100100000011110000000000001010000001000100000100
01110101111101001101111011011111111110001111111011111111110101101111111111111011
10001010000010110010001100110010010001111000000100001000001010010000000100001101

中多数氨基酸的分子量都比较小. 而第三条逻辑序列则表明其中大多数氨基酸都是疏水的.

3.2.2 应用

这一小节, 我们将以表 3.8 中八种神经基因为例, 在逻辑序列的基础上比较它们之间的相似性.

表 3.8: The 8 proteins (taken from NCBI)

Sequence	ACCESSION
Human neurocan	AAC80576
Chimpanzee neurocan	XM_524162
Mouse neurocan	S52781
Rat neurocan-precursor	S28764
Rattus neurocan	AAC15766
Mus brevican	NP_031555
Brevican Rattus	NP_037048
Versican V0 isoform	AAC40166

设 S 是一条长为 k 的蛋白质序列, 则对于 S 的第 i 个逻辑序列, 我们赋予它两个 k-元标号多重集:

$$S_{ib} = \{f_{ib1}, f_{ib2}, \ldots, f_{ibk}\} \quad (b = 0, 1),$$

其中 f_{ibj} $(j = 1, 2, \ldots, k)$ 表示 "逻辑基" b 在第 i 个 (0,1) 序列的长为 j 的串首中出现的频率. 进而, 我们可以按第 2 章中正规化相对熵的定义计算出这两个 k- 元标号多重集的正规化相对熵. 因为 S 共有 6 个逻辑序列, 所以, S 最终可以表示为一个 12 维向量:

$$(h_L(S_{10}), h_L(S_{11}), h_L(S_{20}), \ldots, h_L(S_{60}), h_L(S_{61}))^t$$

其中 t 表示向量的转置.

例如, 表 3.7 中的具有 80 个氨基酸残基的序列的 12 维向量为:

$$(1.43646,\ 2.74372,\ 1.08096,\ 7.57235,\ 3.32583,\ 1.22778,\ 1.32902,\ 3.33458,\ 3.18932,\ 1.32322,\ 1.47143,\ 2.62922)^t$$

我们知道，一旦生物序列具有了向量的形式，那么序列之间的比较就可以转化为向量之间的比较．一般认为，两个向量之间欧氏距离越小，或者它们夹角的余弦越大，两个序列就越相似．我们分别在表 3.9 和 3.10 给出了表 3.8 中八种神经基因的基于欧氏距离和夹角余弦的相似性矩阵．

表 3.9: The similarity/dissimilarity matrix for the 8 protein sequences based on Euclidean-distance of 12-D vectors

	AAC80576 human	XM_524162 chimpanzee	S52781 Mouse	S28764 Rat	AAC15766 Rattus	NP_031555 B-mus	NP_037048 B-rattus	AAC40166 Versican
human	0	0.143070	0.597195	0.507541	0.584235	1.044902	1.055947	1.762949
chimpanzee		0	0.693458	0.590378	0.679171	1.095801	1.119811	1.860002
Mouse			0	0.175484	0.446265	0.713226	0.742529	1.508062
Rat				0	0.394056	0.787227	0.826247	1.543034
Rattus					0	1.150685	1.175360	1.615851
B-mus						0	0.233506	1.532224
B-rattus							0	1.581986
Versican								0

表 3.10: The similarity/dissimilarity matrix for the 8 protein sequences based on cosine of the angle among 12-D vectors

	AAC80576 human	XM_524162 chimpanzee	S52781 Mouse	S28764 Rat	AAC15766 Rattus	NP_031555 B-mus	NP_037048 B-rattus	AAC40166 Versican
human	1	0.999901	0.998732	0.998981	0.998110	0.997248	0.997798	0.986047
chimpanzee		1	0.998272	0.998652	0.997481	0.997109	0.997551	0.984512
Mouse			1	0.999874	0.999365	0.998244	0.998392	0.988991
Rat				1	0.999377	0.998039	0.998153	0.988663
Rattus					1	0.995570	0.995899	0.988299
B-mus						1	0.999768	0.988576
B-rattus							1	0.988002
Versican								1

观察表 3.9 ，我们看到矩阵中最小的元素对应于 human-chimpanzee neurocan 对，同时 mouse(S52781), rat(S28764), rattus(AAC15766) 之间的相似性也较大． Mus brevican (NP_031555) 和 Rattus brevican (NP_037048) 之间表现出很大的相似性，而它们与其它神经基因之间的相似性则较小．此外，与 Versican V0 isoform (AAC40166) 所对应的元素都明显偏大，这表明它与其它神经基因之间的相似性很差．由表 3.10 可以得到同样的结论．

为了验证我们的结果，我们用 Clustal X1.8 对八个蛋白质序列进行了多重序列比对，并将比对的结果用 NJplot 以树的方式给出（见图 3.2 ）．不难看出，我们的结果与该树所反映出来的相似性是一致的，这表明我们的方法是可靠的．而相比之下，我们的方法更简单。

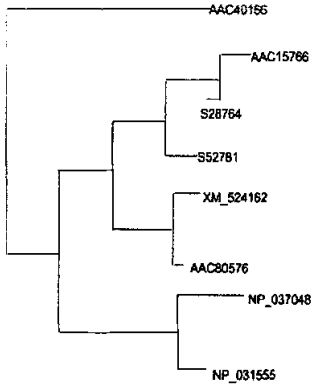


图 3.2: The relationship tree of 8 proteins

3.3 LZ 复杂度及其应用

20 世纪 60 年代，柯尔莫哥洛夫 (A.N. Kolmogorov) 定义一个 (0,1) 序列的复杂度为能够产生这一序列的最短程序的 bit 数. 这种复杂性可以称之为算法复杂性 (Algorithm Complexity)，当给定某一算法，对于不同的序列，将产生不同的程序长度，用它来衡量序列的复杂程度如何，但是这个定义却没有通用的算法，因而 Kolmogorov 复杂度很难由计算机实现 [77,133,134]。1976 年，Lempel and Ziv 将这一理论应用到有限序列，并给出了数学讨论及算法描述 [75], [135]- [139]。

3.3.1 有限序列的 LZ 复杂度

设 Ω 是一个有限字符集，Ω 上的所有有限长序列的集合记为 Ω^*，即 $\Omega^* = \{S : 0 \leq L(S) < \infty\}$, 其中 $L(S)$ 表示序列 S 的长度，并记 Ω 上所有长为 n 的序列的集合为 Ω^n, 即 $\Omega^n = \{S \in \Omega^* : L(S) = n\}$。对于序列 $S = s_1 s_2 \dots s_n \in \Omega^n$，它的起于第 i 个字符止于第 j 个字符的子串记为 $S[i,j]$, 即 $S[i,j] = s_i \dots s_j$。

设 $W \in \Omega^m, V \in \Omega^n$，则通过将 V 连接在 W 后面可以得到一条新的序列 $S = WV = w_1 w_2 \dots w_m v_1 v_2 \dots v_n \in \Omega^{m+n}$，其中 $W = S[1,m], V = S[m+1, m+n]$。序列 S 称为序列 W 的一个 “扩张”，而 W 则称为 S 的一个前缀. 现定义算子 π：

$$S\pi^i = S[1, L(S) - i], \quad i = 0, 1, \dots$$

特别地，$S\pi^0 = S$, 而当 $i \geq L(S)$ 时，$S\pi^i = \Phi$（空串，即不含任何字符的字符串）。

设 $S = s_1 s_2 \dots s_n \in \Omega^n$ 是一个非空序列，则可以按如下算法从空串 Φ “生成” S：

(1) 从空串 Φ 出发开始添加 s_1。如果 $n > 1$，还要在 s_1 后面加上一个记号 “ · ”;

(2) 设已生成前缀 $W = s_1 s_2 \dots s_r$, $0 < r < n$。观察 $V = s_{r+1}$ 是否可以用复制的方法从 $WV\pi$ 中得到：如果 V 不能由 $WV\pi$ 的某个子串复制得到，就连接 W 和 V 从而得到新的前缀 WV，并在它的后面加上一个记号 "$\cdot$"；如果 $V = s_{r+1}$ 可以由 $WV\pi$ 的某个子串复制得到，则继续观察 $V = s_{r+1}s_{r+2}$ 是否可以用复制的方法从 $WV\pi$ 中得到，如果能办到，则继续观察 $V = s_{r+1}s_{r+2}s_{r+3}$，并提出同样的问题 $\dots$。这样下去有两种可能，一种可能是 $V = s_{r+1} \dots s_n$，则分析结束，此时新的"前缀" $WV = S$；另一种可能是出现 $k < n$，使得 $V = s_{r+1} \dots s_k$ 不再能从 $WV\pi$ 的任何一个子串复制得到，这时就得到新的前缀 WV，并在它的后面加上一个记号 "$\cdot$"。

(3) 重复步骤 (2) 直到生成 S：

$$S = S[1 : i_1] \cdot S[i_1 + 1 : i_2] \cdot \dots \cdot S[i_{k-1} + 1 : i_k] \cdot \dots \cdot S[i_{m-1} + 1 : n] \tag{3.3.1}$$

上述过程本质上就是在进行两种操作：S 的已生成部分的子串的"最长复制"和额外"添加"一个字符.

例如，序列 $S = 0001101001000101$ 可以按下面步骤生成：

(i) 从空串 Φ 出发开始添加 0：$\Phi + 0 \mapsto 0\cdot$;

(ii) 最长复制 + 额外添加一个字符 1：$0\cdot + 001 \mapsto 0 \cdot 001\cdot$

(iii) 最长复制 + 额外添加一个字符 0: $0 \cdot 001 \cdot + 10 \mapsto 0 \cdot 001 \cdot 10\cdot$

(iv) 最长复制 + 额外添加一个字符 0: $0 \cdot 001 \cdot 10 \cdot + 100 \mapsto 0 \cdot 001 \cdot 10 \cdot 100\cdot$

(v) 最长复制 + 额外添加一个字符 0: $0 \cdot 001 \cdot 10 \cdot 100 \cdot + 1000 \mapsto 0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000\cdot$

(vi) 最长复制：$0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000 \cdot + 101 \mapsto 0 \cdot 001 \cdot 10 \cdot 100 \cdot 1000 \cdot 101$

Lempel and Ziv 称通过"复制"与"添加"两种操作从空串生成的 S 为 S 的历史，而对应于上述通过"最大复制"与"添加"所得到的式 3.3.1 形式的历史称为 S 的"exhaustive"历史，其中被记号 "$\cdot$" 隔开的成分 (component) 的个数恰是生成序列 S 所需要的最少步骤数，这就是序列 S 的 LZ 复杂度，记作 $c(S)$。例如，序列 $S = 0001101001000101$ 的 LZ 复杂度为 $c(S) = 6$。

3.3.2 基于 LZ 复杂度的 RNA 二级结构相似性分析

按照 LZ 复杂度的定义，对于任意两个序列 W 和 V，由 V 生成 W 的所必须的步骤数为 $c(VW) - c(V)$，而且 $c(VW) - c(V) \le c(W)$ 总是成立的，这意味着由 V 扩张到 VW 所必须的步骤数总是不会超过由空串 Φ 生成 W 所必须的步骤数。此外，序列 W 和 V 越相似，那么 $c(VW) - c(V)$ 和 $c(WV) - c(W)$ 就都越小 [135]。因此，我们可以用式 3.3.2 来衡量两条序列之间的相似性：

$$d(W,V) = f(W,V) - \min\{f(W,W), f(V,V)\} \tag{3.3.2}$$

其中，$f(W,V)=\frac{c(WV)-c(W)+c(VW)-c(V)}{c(WV)+c(VW)}$. 为了方便，我们称 $d(W,V)$ 为序列 W 与 V 的相对距离.

下面我们将在此基础上给出文 [140] 报告的 9 种病毒的 RNA 二级结构相似性. 这 9 种病毒的 RNA 二级结构见图 3.3 .

我们知道，单链 RNA 通过自身的回折使链中碱基配对从而形成多端的双股螺旋区即为 RNA 的二级结构. RNA 二级结构中的碱基分为两类，没有与别的碱基匹配形成氢键的称为自由基 (free base)，否则就称为配对基. 我们分别用 B,D,H,V 来表示处于配对中的碱基 A, C, G, U，于是我们可以从 RNA 二级结构得到一条基于字符集 $\Omega=\{A,C,G,U,B,D,H,V\}$ 的特殊序列 *，我们称之为相应二级结构 **R** 的"影子"序列 (shadow sequence)，并记作 S_R . 表 3.11 给出了图 3.3 中的 9 个 RNA 二级结构的影子序列.

表 3.11: The shadow sequences of the RNA secondary structures of nine viruses in Fig.3.3 (from 5′ to 3′)

	Shadow sequences (S_R)
S_{AIMV-3}	AUGCVDBVHDBAAACVHDBVHBAUGCDDDUAAHHHAUGC
S_{APMV-3}	AUGCDDBDBBDGUGAAHVVHVHHAUGCDDDGUUAHHHAAGC
S_{AVII}	AUGCDVBBUBDVDVDVDVCAGHHBHBHBHVVVBHAUGCDVDDAAAHHBHAUGC
S_{CILRV}	AUGCDVBVBVVVVDVDUCCUHBHBBBBVBVBHAUGCDVDDAAAHHBHAUGC
S_{CVV-3}	AUGCDDBAADVDVDVDVCAUHHBHBHBHAAVHHAUGCDVDDGAAHHBHAUGC
S_{EMV-3}	CDVBBVUDVDVDVDVCACHHBHBHBHBVVBHAUGCDVDCAAGHBHAUGC
S_{LRMV-3}	GUUCDVBVVDVDVDVDUCAGHBHBHGBHBBVBHAUGCDVDDAAAHHBHUCGC
S_{PDV-3}	AUGCDDVDBDDGUAAHHVHBHHAUGCDDDVUAABHHHAUGC
S_{TSV-3}	GUGCDBHVBHVBVBUAAVBVBDVBDVHAUGCDVDDVUUAUBHHBHAUGC

易见，影子序列 S_{AIMV-3} 的 " exhaustive " 历史为：$H_E(S_{AIMV-3})=A\cdot U\cdot G\cdot C\cdot V\cdot D\cdot B\cdot VH\cdot DBA\cdot AAC\cdot VHDBV\cdot HB\cdot AUGCD\cdot DDU\cdot AAH\cdot HHA\cdot UGC.$ 从而其 LZ 复杂度为 $c(S_{AIMV-3})=17$. 此外，

$$\begin{aligned} c(S_{CVV-3}) &= 19 \\ c(S_{AIMV-3}S_{AIMV-3}) &= 18 \\ c(S_{AIMV-3}S_{CVV-3}) &= 30 \\ c(S_{CVV-3}S_{AIMV-3}) &= 29 \\ c(S_{CVV-3}S_{CVV-3}) &= 20 \end{aligned}$$

根据式 3.3.2，我们得到序列 S_{AIMV-3} 和 S_{CVV-3} 之间的相对距离为：

$$d(S_{AIMV-3},S_{CVV-3})=0.3398.$$

*这里的 B,D,H,V 只是为了方便而引入的记号，与 IUPAC 编码中的含义不同.

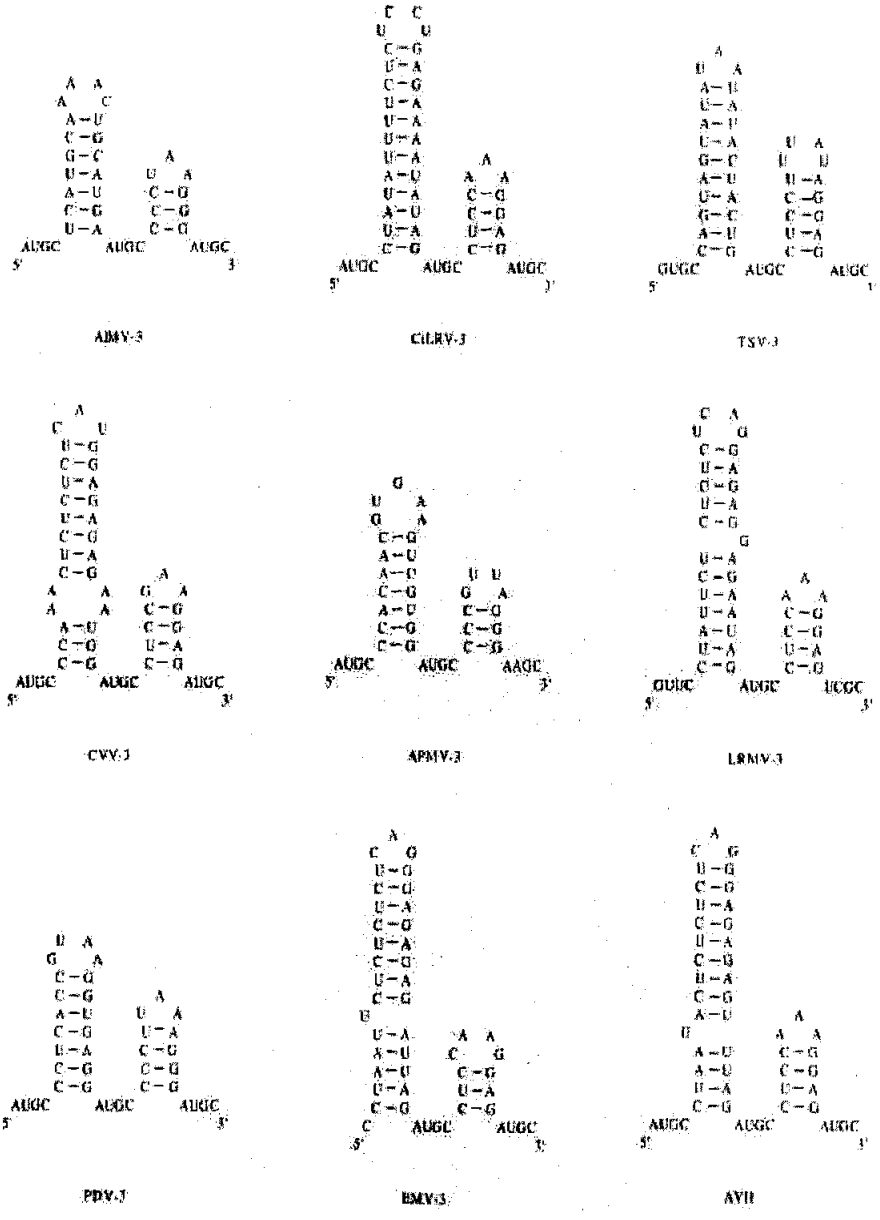


图 3.3: Secondary structure at the 3′-terminus of RNA 3 of alfalfa mosaic virus (AlMV-3 [141]), citrus leaf rugose virus (CiLRV-3 [142]), tobacco streak virus (TSV-3 [143,144]), citrus variegation virus (CVV-3 [142]), apple mosaic virus (APMV-3 [145]), prune dwarf ilarvirus (PDV-3 [146]), lilac ring mottle virus (LRMV-3 [147]), elm mottle virus (EMV-3 [148]), and asparagus virus II (AVII (EMBL/GenBank/DDBJ databases; accession no. X86352)

按照同样的方法，我们计算出表 3.11 中的 9 条序列之间的相对距离（见表 3.12 ），从中我们看到 (AVII, EMV-3), (CVV-3, AVII), (LRMV-3, AVII), (LRMV-3, CILRV-3), 以及 (AVII, CILRV-3) 对的元素都比较小，这表明它们之间的相似性较大。另一方面，与 AIMV-3 和 APMV-3 对应的元素都比较大，这表示它们与其它二级结构之间的相似性较低。这与文献 [83]- [85] 的结果是一致的。

表 3.12: The similarity/dissimilarity matrix for the secondary structures at the 3′ terminus belonging to nine viruses of Fig.3.3

Species	AIMV-3	APMV-3	AVII	CILRV-3	CVV-3	EMV-3	LRMV-3	PDV-3	TSV-3
AIMV-3	0	0.3224	0.3471	0.3608	0.3398	0.3694	0.3613	0.2992	0.3514
APMV-3	0.3224	0	0.3315	0.3636	0.3229	0.3627	0.3641	0.2928	0.3452
AVII	0.3471	0.3315	0	0.2630	0.2496	0.2372	0.2552	0.3315	0.2950
CILRV-3	0.3608	0.3636	0.2630	0	0.3094	0.3094	0.2487	0.3452	0.3084
CVV-3	0.3398	0.3229	0.2496	0.3094	0	0.3059	0.3021	0.3229	0.3196
EMV-3	0.3694	0.3627	0.2372	0.3094	0.3059	0	0.2917	0.3627	0.3196
LRMV-3	0.3613	0.3641	0.2552	0.2487	0.3021	0.2917	0	0.3554	0.3298
PDV-3	0.2992	0.2928	0.3315	0.3452	0.3229	0.3627	0.3554	0	0.3255
TSV-3	0.3514	0.3452	0.2950	0.3084	0.3196	0.3196	0.3298	0.3255	0

3.3.3 广义 LZ 复杂度及其应用

按照 Lempel and Ziv 的方法，在一个序列的 " exhaustive " 历史的产生过程中所允许的操作可以概括为子串的 "最长复制" 并额外 "添加" 一个字符，这里的复制确切地说是正向复制 (Direct copying) 。此外，一些学者在 DNA 序列分析研究中还提出了对称 (Symmetric) 、反向互补 (Inverted complementary) 和正向互补 (direct Complementary) 等操作，并称通过这四种方式 "复制" 而得到的复杂度为 " DSIC " 复杂度 [149]- [151] 。例如，序列 $S = ATGCATCGTACATC$ 的 DSIC 复杂度为 6 ，相应历史为：

$$H(S) = A \cdot T \cdot G \cdot CAT \cdot CGTA \cdot CATC.$$

其生成过程如图 3.4 所示。其中 " + " 表示添加一个字符，实箭头从左到右表示正向复制，否则表示对称（反向复制），虚箭头从左到右表示正向互补，否则表示反向互补。虚直线表示相应操作的对象。

从数学上讲，复制就是恒等置换，而互补则是 $\Omega = \{A, C, G, T\}$ 上这样一个置换 p ：

$$p(A) = T, p(T) = A, p(G) = C, p(C) = G.$$

这只是四元字符集上的 $4! = 24$ 个置换中的两个。而对于二元字符集，定义在它上面的置换显然只有恒等置换 $p_1(0) = 0, p_1(1) = 1$ 和对换 $p_2(0) = 1, p_2(1) = 0$ 。下面，我们将在此基础上给出 (0,1) 序列的广义 LZ 复杂度。我们同样考虑正反两个方向，同时，类似于 Lempel and Ziv 方法，在生成序列的历史时要求每一个复制之后必须额外 "添加" 一个字

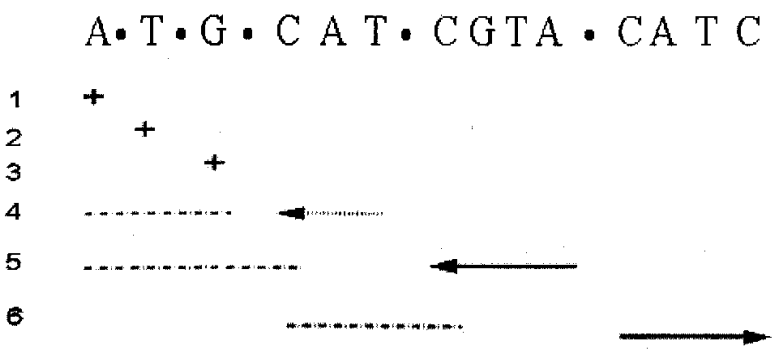


图 3.4: The example of the DSIC complexity decomposition

符，除非是历史的最后一个成分. 需要指出的是，“ DSIC ”方法复制之后并没有考虑“添加”操作. 额外“添加”一个字符看似简单，实际上却在序列的生成中有着重要的作用，因为它保证了序列的历史中每一个成分的唯一性.

对于任意一条 (0,1) 序列 S，由正向 p_1 置换和“添加”两种操作，可以得到它的标准 LZ 复杂度；而由反向 p_1 置换，正向 p_2 置换以及反向 p_2 置换同“添加”操作一起，可以得到另外三个广义 LZ 复杂度. 这样，一条 (0,1) 序列 S 有 4 个复杂度与之对应，我们将这 4 个复杂度放在一起从而构成一个复杂性向量，记作 $(c_{dp_1}(S), c_{ip_1}(S), c_{dp_2}(S), c_{ip_2}(S))$. 例如，序列 $S = 0001101001000101$ 的复杂性向量为 $(6,5,6,7)$.

我们知道，一条 DNA 序列的逻辑表示是 4 条 (0,1) 序列，从而，可以用一个 16 维向量来刻画一条 DNA 序列，这个向量的分量就是 4 条 (0,1) 序列所对应的广义复杂度. 我们在表 3.13 列出了 11 个物种的 *beta*-globin 基因，并在表 3.14 给出了它们的复杂性向量.

表 3.13: The full *beta*-globin genes of 11 species

Species	Database	ID	Location	Length (bp)
Human	EMBL	HSHBB	62187-63610	1424
Chimpanzee	EMBL	PTGLB1	4189 - 5532	1344
Gorilla	EMBL	GGBGLOBIN	4538 - 5881	1344
Lemur	EMBL	LMHBB	154 -1595	1442
Rat	EMBL	RNGLB	310 - 1505	1196
Mouse	EMBL	MMBGL1	275 - 1462	1188
Goat	EMBL	CHHBBAA	279 - 1749	1471
Bovine	EMBL	BTGL02	278 - 1741	1464
Rabbit	EMBL	OCBGLO	277 - 1419	1143
Opossum	EMBL	DVHBBB	467 - 2488	2022
Gallus	EMBL	GGGL02	465 - 1810	1346

我们通过对表 3.13 中的 11 个物种之间的相似性检查来说明 DNA 序列的复杂性向量

表 3.14: The 16-D complexity vectors associated with 11 full *beta*-globin genes in Table 3.13

Species	human	chim.	gorilla	lemur	goat	bovine	mouse	rat	rabbit	oposs.	gallus
$c_{dp_1}(LR^A)$	113	105	106	111	114	118	89	96	89	158	100
$c_{ip_1}(LR^A)$	109	104	105	105	117	116	91	92	83	152	98
$c_{dp_2}(LR^A)$	224	216	215	230	234	230	206	209	203	278	241
$c_{ip_2}(LR^A)$	224	217	214	228	232	229	203	206	201	281	241
$c_{dp_1}(LR^C)$	100	94	93	99	104	104	93	92	87	140	108
$c_{ip_1}(LR^C)$	103	97	97	95	103	104	91	88	84	135	101
$c_{dp_2}(LR^C)$	246	235	236	245	260	229	202	201	207	328	205
$c_{ip_2}(LR^C)$	245	234	235	246	264	232	204	200	209	325	205
$c_{dp_1}(LR^G)$	108	100	102	113	120	116	97	94	92	136	123
$c_{ip_1}(LR^G)$	99	93	95	110	113	112	91	90	92	136	116
$c_{dp_2}(LR^G)$	228	218	219	223	229	229	197	194	171	328	170
$c_{ip_2}(LR^G)$	227	216	221	224	228	231	195	197	167	329	169
$c_{dp_1}(LR^T)$	127	123	123	128	127	123	109	113	108	172	100
$c_{ip_1}(LR^T)$	127	122	121	130	124	123	108	107	107	172	96
$c_{dp_2}(LR^T)$	186	185	185	182	208	191	165	162	154	255	268
$c_{ip_2}(LR^T)$	185	182	182	182	210	192	165	160	152	251	266

表示的有用性. 为了避免在相互比较时序列长度的不同所造成的影响, 我们用正规化的复杂性向量, 即复杂性向量除以相应序列的长度, 代替原来的复杂性向量. 潜在的假设是: 如果两个向量方向相近且模长相似, 那么这两个向量所对应的 DNA 序列就是相似的. 而这样的向量之间的相似性可以由向量终点之间的欧氏距离来衡量, 从而, 欧氏距离越小, 那么相应的两条 DNA 序列就越相似. 我们在表 3.15 给出了表 3.13 中 11 个物种的基于正规化复杂性向量的相似性矩阵. 观察表 3.15 , 我们看到 gallus 与其它物种的相似性最小, 因为它所对应的元素都明显偏大, 甚至可以说和其它元素根本就不在同一个量级, 这与 gallus 是非哺乳动物而其它物种都是哺乳动物这一点是完全吻合的. 同时, 表中对应于 opossum 的元素也相对较大, 出现这样的结果并不奇怪, 因为在这 10 个哺乳动物中, opossum 是和其它物种亲缘关系最远的一个. 另一方面, 我们看到 chimpanzee-gorilla, chimpanzee-human, human-gorilla, mouse-rat 对的值都非常小, 而且 human, chimpanzee, gorilla 之间的距离比它们同其它物种之间的距离都明显的小, mouse 和 rat 也是如此. 而一段时间以来, 一些文献报告的结果中始终存在着从进化意义上不能令人满意的现象, 如 [120] 中出现了 human 与 chimpanzee 之间的相似程度明显低于 human 与 gallus 之间的相似程度, 而在文 [31,103,120,139] 中都存在着 mouse-rat 的相似性小于 mouse-human 的相似性, 这在我们的 "伪迹" 法结果中也比较明显. 而在本节的相似性矩阵中, 这几个不理想的结果都得到了修正.

3.4 小结

在这一章, 我们提出了 DAN 序列的逻辑表示的概念, 并将其推广到蛋白质序列. 这实际上是源于代数学中的同态思想和物理学中的粗粒化思想. 以这样一种简约的方式描述序列, 毫无疑问会牺牲序列的某些细节, 但这却可以使序列的某些生物学上的特征更

表 3.15: The similarity/dissimilarity matrix for the 11 full *beta*-globin genes

Species	human	chim	gorilla	lemur	goat	bovine	mouse	rat	rabbit	oposs	gallus
human	0	0.0117	0.0123	0.0146	0.0230	0.0244	0.0292	0.0276	0.0387	0.0344	0.1209
chim.		0	0.0053	0.0221	0.0195	0.0304	0.0229	0.0235	0.0352	0.0426	0.1152
gorill			0	0.0231	0.0194	0.0309	0.0238	0.0248	0.0378	0.0418	0.1170
lemur				0	0.0274	0.0232	0.0336	0.0304	0.0364	0.0371	0.1214
goat					0	0.0338	0.0297	0.0333	0.0349	0.0489	0.1049
bovine						0	0.0365	0.0328	0.0501	0.0336	0.1137
mouse							0	0.0118	0.0316	0.0587	0.1089
rat								0	0.0319	0.0563	0.1124
rabbit									0	0.0708	0.1104
oposs.										0	0.1361
gallus											0

为突出. 人类基因组计划的成果, 是一本多达 30 亿个字母的 "天书", 其中既没有段落, 也没有标点, 但却包含着人类生老病死及遗传进化的全部信息。破译这部世界上最巨量信息的 "天书" 是二十一世纪最重要的任务之一, 但 "天书" 的复杂性使我们无法同时考虑它的所有细节, 因此, 利用同态的思想对其进行粗粒化描述以使我们所关心的特征更为突出, 然后将其表示成适当的数学对象进而研究序列的结构并探索其固有的生物学规律, 对理解这本 "天书" 将是十分有意义的。

除逻辑序列之外, 在这一章我们还给出了 RNA 二级结构的影子序列 (shadow sequence), 我们有理由相信, 可以按照某种方式进一步给出 RNA 二级结构的逻辑表示。在逻辑序列和影子序列的基础上, 我们结合序列的正规化相对熵、符号序列标准 LZ 复杂性以及 (0,1) 序列的广义 LZ 复杂性进行了 DNA 、 RNA 及蛋白质序列的相似性分析, 结果令人比较满意。这种方法可能会在生物信息学其他研究领域中有更多的应用, 在今后的工作中, 我们将继续探讨 (0,1) 序列的特点, 并结合二态离散特征在分子进化、 DNA 序列拼接与组装以及蛋白质折叠等方面进行较为深入的研究。

顺便指出, (0,1) 序列还有一个在计算方面的优势: 由于序列仅有两个状态, 计算机处理时可以只用一个比特的比较, 这在带有位操作的计算机语言 (如 C 语言) 中, 将大大提高运算速度. 这其实也正是人们在进行大规模序列分析时所期望的。

4 蛋白质编码基因识别

4.1 引言

生命是大自然最伟大的创造物，经过亿万年的进化，生命的形式从简单的有机物发展到现在高度复杂但有序的生物系统．蛋白质是构造生命机器的基本元件，大量结构不同、功能各异的蛋白质在遗传信息的控制之下，被不断地合成出来，并有机地组成复杂的生物体．遗传信息存贮在基因组中，具体说就是存贮在 4 种字符组成的核酸序列中．

随着分子生物学中心法则的确立，人们逐渐认识到遗传信息的载体主要是 DNA（在少数情况下 RNA 也充当遗传信息载体），控制生物体性状的基因则是一系列 DNA 片段．一方面， DNA 通过自我复制，在生物体的繁衍过程中传递遗传信息．另一方面，基因通过转录和翻译，使遗传信息在生物个体中得以表达，并使后代表现出与亲代相似的生物性状．在基因表达过程中，基因上的遗传信息首先通过转录从 DNA 传到 RNA，然后再通过翻译从 RNA 传递到蛋白质．基因控制着蛋白质的合成，基因的 DNA 序列到蛋白质序列存在着一种明确的对应关系，而这种对应关系就是遗传密码．

遗传密码的发现开创了在分子水平上的生命信息科学，启动了人类探索遗传语言奥秘的进程．许多科学家认为基因组 DNA 序列并非是一种简单的生物分子序列，而可能是一种语言，该语言描述遗传信息，控制生物体的性状，规定生物个体的生老病死．为了深刻揭示这种遗传语言的奥秘，科学家们开始测序人类及其它模式生物基因组，希望解读和破译遗传信息，使人类在分子水平上全面地认识自我．由于生物技术的高速发展，人类基因组计划已经提前至 2003 年全部完成，我们得到的是一本关于人类遗传信息的长达数百万页的“天书”．之所以称它为天书，不单是因为它所包含的信息量巨大，更重要的是目前人类对它了解甚少，还无法读懂它．天书中只有 4 种字符 (A,G,C,T)，既没有段落，也没有标点符号，是一个长度为 3×10^9 的一维序列．

人类基因组是科学家研究的第一个脊椎动物染色体基因组，人类基因组已成为其它脊椎动物中的代表．它比线虫和果蝇基因组大 30 倍左右，比酵母的大 250 倍左右．尽管它的长度比较大，但它的基因数目似乎只有果蝇和线虫基因组基因数目的两倍或三倍．人类基因组大约有 2 万 — 2.5 万个基因 [152]，这些基因分布在染色体中的 DNA 序列上，或者说就隐藏在“天书”中．到目前为止只有一部分基因已明确定位．那么如何在“天书”中找到其它的基因呢？一种方法是通过分子生物学实验确定基因的位置，另一种方

法就是通过信息分析寻找基因。科学家已经发现基因的蛋白质编码区域与非编码区域在序列的统计特征上有明显的差异，因此，从理论上讲，可以应用组合数学、概率统计以及信息论等相关知识来构造模式向量，并利用模式识别方法区分特性，从而识别基因。但正如我们在绪论中所指出，对具有较多内含子的真核生物基因组序列的正确识别其实是个相当困难的问题。而相比之下，一些模式生物基因组结构相对比较简单，单位 DNA 片段上基因的密度高，易于进行基因识别。而且，从进化的角度上讲，生物的许多基因有很大的同源性，对模式生物基因的分析有助于阐明人类基因的结构与功能。

酿酒酵母 (*Saccharomyces cerevisiae*) 是一个非常重要的模式生物. 在来自欧洲、北美和日本 100 多个实验室的大约 600 名科学家的共同努力下, 酿酒酵母基因组测序工作于 1996 年完成 [153]- [155] . 它是最早被完全测序的一个真核细胞 (eukaryote) 基因组. 酿酒酵母是一个单细胞生物体, 在这个基因组中共有 16 条染色体 (chromosome), 总长 12.068Mbp. 历史上, 人们提出了许多在基因组序列中识别蛋白质编码基因的方法 (详见 [23], [43]- [61], [68]- [74], [156]- [159]), 仅就酿酒酵母基因组而言, 早在上个世纪 80 年代, CBI(the Codon Bias Index) [60] 和 CAI(the Codon Adaptation Index) [61] 两种指标就被先后提出并广泛用于描述其编码序列。但正如 [51,52,160] 所指出，它们并不能充分反映编码序列的编码性质。因此， Zhang 等 [51,52] 提出了两个新的编码指标： delta (△) 和 YZ-score 。他们基于这两个指标对酿酒酵母基因组的序列进行基因识别，识别率分别为 93% 和 95% 。而整个酿酒酵母基因组中蛋白质编码基因的数目一直存在着争议，大多数研究人员认为应该在 5800-6000 这个范围 [153]- [155] , 但有些人则认为应该少于 4800 [161] . 与这两个都不同, Zhang 等报告的结果是 5600 左右 [51,52] 。

在这一章, 我们将基于 DNA 序列的正规化相对熵对酿酒酵母基因组序列进行基因识别. 我们所用的数据取自 Munich 的蛋白质序列信息中心 (MIPS , the Munich Information Center for Protein Sequences) , 网址是 http://pedant.gsf.de/ , 释放的时间是 2001 年 10 月 10 日. 在 MIPS 数据库中, 酿酒酵母基因组总共有 6449 个 ORF , 它们被分成了 6 类, 分别是第 1 类: 已知的蛋白质 (known proteins) 、第 2 类: 强相似于已知的蛋白质 (strong similarity to known proteins) 、第 3 类: 相似于或弱相似于已知的蛋白质 (similarity or weak similarity to known proteins) 、第 4 类: 相似于未知的蛋白质 (similarity to unknown proteins) 、第 5 类: 不相似类 (no similarity) 、第 6 类: 有问题的 ORF(questionable ORFs) . 在这六类中分别包含了 3410(18), 229, 820(2), 1003, 516, and 471(8) 个序列, 括号里面的数字表示的是线粒体基因 ORF 的个数. 考虑到一方面这里的线粒体 ORF 个数太少因而缺少统计的意义, 另一方面线粒体基因的密码子与通用密码子稍有不同, 所以我们在执行基因识别算法时忽略掉这些序列. 这样六类 ORF 序列的个数应为 3392, 229, 818, 1003, 516 和 463 .

在酿酒酵母基因组中, 有 4% 的蛋白质编码基因含有内含子 [153] , 我们在执行识别算法时没有去掉它们, 我们的识别率为 96% , 估计的蛋白质编码基因总数为 5873 , 与普遍接受的 5800-6000 相符. 因此, 我们的方法有望被用以识别结构更为复杂些的基因. 当

然，这种方法不能直接被应用到具有较多内含子的高等真核生物基因组。

4.2 DNA 序列基于正规化相对熵的数值刻画

在第 2 章，我们曾按照序列中碱基所在的位置将一个 DNA 序列分为三个子序列：首先按从左到右的方向取这个 DNA 序列的在 $3p+1(p=0,1,2,\cdots)$ 位置上的所有碱基形成一个新的序列，我们称这个序列为这个 DNA 序列的 1- 子序列。类似地，由这个 DNA 序列的 $3p+i$ $(p=0,1,2,\cdots;\ i=2$ 或 3) 位置上的所有碱基构成的子序列，分别称为这个 DNA 序列的 2- 子序列和 3- 子序列。假设 i- 子序列 $(i=1,2,3)$ 具有 k_i 个基，则我们可以赋予它 4 个 k_i- 元标号多重集：

$$S_{iY}=\{f_{iY1},f_{iY2},\cdots,f_{iYk_i}\}\quad (Y=A,C,G,T),$$

其中 f_{iYj} $(j=1,2,\ldots,k_i)$ 表示基 Y 在 i- 子序列的前 j 个基中出现的频率。进而，按式 2.5.3，我们得到 12 个正规化相对熵：

$$h_L(S_{iY})\quad (i=1,2,3;\ Y=A,C,G,T)$$

此外，$\rho=\bar{a}^2+\bar{c}^2+\bar{g}^2+\bar{t}^2$，对 DNA 序列分析来说是一个很有用的统计量 [51,67]，其中 $\bar{a},\bar{c},\bar{g},\bar{t}$ 分别表示 DNA 序列中碱基 A,C,G,T 出现的平均频率。基于此，我们用如下的 12- 维向量来刻画 DNA 序列：

$$x=\rho(h_L(S_{1A}),h_L(S_{1C}),\ldots,h_L(S_{3C}),h_L(S_{3G}),h_L(S_{3T}))^\tau$$

其中 τ 表示向量的转置。

4.3 Fisher 线性判别法

Fisher 线性判别式在我们这种情形实际上是一个由权向量 w 所描述的 12- 维空间中的超平面。我们首先需要两个样本集：一个正样本集 (由真正的蛋白质编码基因序列组成的样本集) 和一个负样本集 (由非编码序列组成的样本集)。用 G_1 表示正样本集，G_2 表示负样本集，我们用这两个样本集作为算法中的训练集。现记 $n_g=|G_g|$ $(g=1,2)$，并用 $x_k^g=(x_{k1}^g,x_{k2}^g,\ldots,x_{k,12}^g)^\tau$ 表示样本集 G_g 中第 k 个样本的 12- 维向量，则两个样本集的几何均值向量为：

$$m_g=\frac{1}{n_g}\sum_{k=1}^{n_g}x_k^g,\quad g=1,2$$

用 S_w 表示两个样本集的离散度矩阵之和，则有

$$S_w=\sum_{g=1}^{2}\sum_{k=1}^{n_g}(x_k^g-m_g)(x_k^g-m_g)^\tau$$

进而， Fisher 权向量 w 可以由下式给出：

$$w = S_w^{-1}(m_1 - m_2)$$

其中 S_w^{-1} 表示矩阵 S_w 的逆矩阵. 从而, 对于任意一个 12- 维向量 $x_k^g = (x_{k1}^g, x_{k2}^g, \ldots, x_{k,12}^g)^\tau$, $k = 1, 2, \ldots, n_g$, 它的投影点是

$$y_k^g = w^\tau x_k^g .$$

请注意, 权向量 w 实际是指出了从 12- 维空间到 1- 维空间的最好的一个投影方向, 因此, 如果用一个常数乘以 w 其结果所表示的方向是不变的. 从这个意义上讲， w 的取值是不唯一的. 不失一般性, 这里取满足条件 $\|w\| = 1$ 的 w . 基于训练集中的数据, 合适的阈值 y_0 就可以被确定出来并用于编码和非编码序列的判别. 这里我们采用下面的式子来确定 y_0 :

$$y_0 = \frac{1}{2}(\frac{n_1\tilde{m}_1 + n_2\tilde{m}_2}{n_1 + n_2} + \frac{1}{2}(\tilde{m}_1 + \tilde{m}_2))$$

其中 $\tilde{m}_g = \frac{1}{n_g}\sum_{k=1}^{n_g} y_k^g$, $g = 1, 2$. 有了权向量 w 和阈值 y_0 之后, 对于测试集中的每一个 ORF , 就可以按照下面的决策规则非常容易地判别它是否是编码序列: 如果 $f(x) > 0$, 那么这个 ORF 就是一个编码序列; 否则, 我们就认为这个序列是一个非编码的 DNA 序列, 这里 $f(x) = w^\tau x - y_0$.

4.4 算法的评估

功能序列分析通常是基于两个集合的辨别: 功能序列集和不执行这部分功能的集合 (非功能集) . 从方法论的角度, 我们应该分初始的数据集为两个独立的训练集和测试集, 这里的测试集仅仅被用于算法的评估. 关于正确性的评估问题, 由于一般的数据库中通常有很多的多余信息而且有些序列会被重复许多次, 所以在测试集的选取时一定要注意避免出现训练集和测试集之间都有相似的序列或片断使得它们之间有很强的相互作用.

在早期的基因识别程序中, 训练集和测试集中序列的选取标准没有一个很好的定义方法, 而且常常训练集和测试集没有选在同一个数据库中, 所以很难说明这些程序的正确性.

为了避免这些问题, 一种可能的方法就是采用标准的执行 (performance) 矩阵和数据库. Fickett 和 Tung [156, 162] 提出了一种标准的检查程序来评估编码的测度; Burset 和 Guigo [163] 提出了基因识别的穷尽测试和几个正确性评估的标准测度. 另外, Kulp 等人 [164] 提出了用训练集和测试集来比较基因识别算法.

4.4.1 敏感度、特异性和准确度的定义

敏感度 (sensitivity) 和特异性 (specificity) 这两个测度经常被用来描述一个算法或一个识别函数的准确度 (Accuracy) . 所用记号及其含义如下: 用 TP 表示在正样本集中预

测正确的样本的个数，即在编码 ORF 中被正确地预测为编码 ORF 的个数；FN 表示在正样本集中预测错误的样本的个数，即编码 ORF 中被预测为非编码 DNA 序列的个数。于是敏感度 (S_n) 的定义为

$$S_n = \frac{TP}{TP + FN}. \tag{4.4.1}$$

这就是说，敏感度 S_n 表示的是编码基因中被正确预测的部分占编码基因总量的百分率。

类似地，用 TN 表示在负样本集中被正确预测的个数，即非编码 DNA 序列被预测为非编码 DNA 序列的数目；用 FP 表示在负样本集中被错误预测的个数，即非编码 DNA 被预测成编码 ORF 的数目。进而特异性 (S_p) 的定义为

$$S_p = \frac{TN}{TN + FP}. \tag{4.4.2}$$

可见，特异性 S_p 表示的是非编码 DNA 序列中被正确预测的部分占非编码 DNA 总数的百分率。

正如 Guigo [46] 所指出的那样，一个好的程序应该不但要将是基因的预测为基因，而且还要在没有基因的地方预测不出基因来。Claverie [165] 也曾指出，序列分析方法总是要求敏感度和特异性都要保持在一个可以接受的水平之上。因此，预测的准确度 (AC) 可以定义为敏感度和特异性的平均值：

$$AC = \frac{1}{2}(S_n + S_p). \tag{4.4.3}$$

需要指出的是，如上定义的特异性 S_p 在理论上可能存在问题，因为，在一个基因组中非编码序列远比编码序列多，从而 $TN \gg FP$，这将导致 S_p 总是趋近于 1。为了解决这个问题，有人给出了特异性的另一个替代定义 [159,163]：

$$\hat{S}_p = \frac{TP}{TP + FP}. \tag{4.4.4}$$

不过，我们这里的测试集所包含的非编码序列数都没有超过 1100 条，因此使用定义 4.4.2 是合适的。

4.4.2 算法的评估

交互验证 (cross-validation) 是评估一个算法的有效性的一种常见方法。在这一章里，我们取 MIPS 数据分类库酿酒酵母基因组的第一类 ORF，即已知的蛋白质，共有 3392 个序列作为正样本集。然后，从酿酒酵母基因组的 16 个染色体中，我们随机选择了 7200 条长度不小于 300bp 的基因间的 DNA 序列。

这些序列选取的方法如下： (1) 找出 MIPS 数据分类库中被注释为 ORF 的序列的起始位置；(2) 计算在相邻两个 ORF 之间的 DNA 序列的长度，并舍去那些长度小于 300bp 的 DNA 序列；(3) 从剩下的所有长度不小于 300bp 的 DNA 序列中，从它们的第一个碱基开始搜索密码子 "ATG"；然后从密码子 "ATG" 开始往下游的方向一个一个密码子的搜索，在第 101 个密码子以后搜索结束密码子 (TAA,TGA,TAG)，遇到一个结束密码子时搜索结束，从而得到一条基因间的 DNA 序列. 这样的序列经常含有几个结束密码子因而不可能是 ORF，进而常被视为非编码序列 [51,52,124]。连续地在往下游的方向搜索更多的基因间的 DNA 序列，直到不能再找到为止。(4) 对于六种可能的序列（一个序列的正向 / 反向的三种阅读方式）重复上面的步骤.

对于上面的过程，可能找到的基因间的 DNA 序列非常多. 我们随机截取了 7200 个基因间的 DNA 序列作为非编码序列的样本.

我们将正样本集中的 3392 条序列随机地分成两个部分：一部分包含 2392 个序列，另一部分包含 1000 个序列，分别作为训练和测试过程的正样本集. 同时，从上述 7200 条非编码序列中随机选择 2392 条作为训练过程的负样本集，然后再从剩下的非编码序列中随机选择 1000 条作为测试过程的负样本集.

利用训练集中的序列，我们就可以求出 Fisher 权向量 w 和阈值 y_0，然后，再由测试集中的序列就可以计算出算法的准确度 (AC)，并以此来评价算法的优劣.

我们重复上述过程两次，并将测试结果列在表 4.1 的左侧. 此外，我们还进行了另外三次交互验证，这三次交互验证中训练集和测试集中的正样本集的样本数依然分别是 2392 和 1000, 但负样本集的样本数则分别是 2500, 2500, 2600 和 1000, 1000, 1100。我们将相应的结果列在了表 4.1 的右侧. 从表中可以看出，每一次交互验证的准确度都在 96% 以上.

表 4.1: The accuracy of the gene-finding algorithm for five different test sets

Test set	1	2	3	4	5	Average
S_n (%)	96.3	96.5	96.2	96.2	96.2	96.28
S_p (%)	96.4	96.3	96.1	96.3	96.2	96.26
AC(%)	96.35	96.40	96.15	96.25	96.20	96.27

4.5 识别酿酒酵母基因组第 2-6 类中的基因

上面我们进行了 5 次交互验证，每一次都在训练过程中得到相应的权向量 w 和阈值 y_0，我们将这些值列在表 4.2，并取它们的平均值分别作为最终的权向量和阈值（见表 4.2 最后一列）.

表 4.2: The Fisher vector w and threshold y_0 for five different training sets

Training set	1	2	3	4	5	Final value
w	-0.496483	-0.496095	-0.471393	-0.499913	-0.496364	-0.492863
	-0.009156	-0.008589	-0.013183	-0.010457	-0.010212	-0.010336
	-0.456127	-0.495064	-0.453578	-0.457073	-0.443482	-0.461827
	0.398475	0.403335	0.443601	0.398042	0.448223	0.419027
	-0.304537	-0.226403	-0.207726	-0.236464	-0.195482	-0.234510
	-0.267572	-0.271753	-0.249058	-0.268509	-0.243988	-0.260606
	0.153286	0.163533	0.163544	0.146736	0.163753	0.158432
	0.060704	0.067681	0.080169	0.048575	0.075859	0.066708
	0.139042	0.147729	0.137534	0.145724	0.155000	0.145246
	-0.088173	-0.076957	-0.078387	-0.078259	-0.088988	-0.082289
	-0.078678	-0.070429	-0.081238	-0.094788	-0.084976	-0.082157
	-0.402206	-0.39612	-0.452865	-0.440882	-0.431811	-0.425479
y_0	-1.25682	-1.17777	-1.09254	-1.26773	-1.0693	-1.172832

为了验证这样得到的最终的权向量和阈值的可靠性，我们利用它们对前面提到的 5 个测试集重新测试，并将测试的结果列在表 4.3 。从表中可以看出，每一次测试的结果都和表 4.1 中的相似，而且准确度也总是大于 96% 的。这表明，可以利用这样得到的最终权向量 w 和阈值 y_0 对酿酒酵母基因组第 2-6 类中的 ORF 进行基因识别。对一个待测序列，根据前面提到的决策规则，如果 $f(x) = w^{\tau}x - y_0 > 0$ ，那么这个 ORF 就被认为是一个编码序列；否则，就认为它是非编码的。我们在表 4.4-4.6 分别列出了酿酒酵母基因组第 2-4 类、第 5 类和第 6 类中被预测为非编码的 ORF 。

表 4.3: The accuracy of the gene-finding algorithm for five different test sets using the final Fisher vector and threshold

Test set	1	2	3	4	5	Average
S_n (%)	96.3	96.3	96.2	95.9	96.1	96.16
S_p (%)	96.6	96.5	96.0	96.3	96.1	96.30
AC(%)	96.45	96.40	96.10	96.10	96.10	96.23

进一步，利用这些表我们重新估计酿酒酵母基因组中基因的个数。例如，在酿酒酵母基因组第 2-4 类中，总共有 2050 个 ORF ，其中有 1841 和 209 个 ORF 被分别预测为编码和非编码序列。假设算法的敏感度和特异性都是 96% ，则可得到如下线性方程组：

$$\begin{cases} TP/(TP+FN) = 0.96 \\ TN/(TN+FP) = 0.96 \\ TN+FN = 209 \\ TP+FP = 1841 \end{cases}$$

解这个方程组得 $TP \approx 1835$， $FP \approx 6$， $TN \approx 133$， $FN \approx 76$, 所以其中编码序列个数应该是 $TP+FN = 1835+76 = 1911$ 。对于酿酒酵母基因组的第 5 类和第 6 类，我们用同样的

表 4.4: The 209 ORFs of the $2^{nd}-4^{th}$ classes (" Similarity " to known/unknown proteins) in the MIPS database, which are recognized as non-coding

ybr210w	yfl027c	ymr245w	yil175w	ybl029c-a	ydl123w	yhr162w	ykr087c
yel004w	yfl040w	ymr317w	yll065w	ydr275w	ybr004c	yjl003w	ybr191w-a
yfl052w	yfr057w	ynl109w	yhl045w	yfr012w	yjl097w	yir040c	yjr108w
ygl002w	ygl104c	ynl176c	ycl065w	ydr286c	yar060c	ylr149c-a	yjr115w
ymr040w	ygl128c	ynl230c	yhl044w	ylr463c	ynl213c	ybr099c	ydr105c
ynr036c	ygl160w	ynr059w	ykl219w	ydl159w-a	ypl159c	ybr103c-a	ybr228w
ypl183w-a	ygr101w	yol079w	yhl041w	ydr316w	yml007c-a	yer140w	ypr074w-a
ypr196w	ygr284c	yol155c	yhl042w	yml047c	ymr326c	yjr013w	ydr126w
yal066w	yhr130c	yol163w	yhl034w-a	ydr352w	ydl114w-a	yer147c-a	ynr046w
ybl089w	yhr181w	yor053w	yel067c	ydr367w	ylr408c	ylr156w	yjr161c
ybr220c	yil025c	yor350c	ynl326c	ydr525w-a	ynl194c	ydr013w	ynr061c
ybr293w	yil042c	ypl123c	yel053w-a	ydr437w	ycr097w-a	ygl010w	yjr162c
ycr001w	yil169c	ypl236c	ypl264c	yar029w	yil029c	ydr018c	ynr075w
ydl018c	yjl170c	ypr094w	ydl222c	ydr459c	ycr102w-a	ylr159w	ynr077c
ydl206w	yjl193w	ypr140w	yal034c	yar031w	yer074w-a	ylr161w	ybr302c
ydl228c	ykr030w	ydl247w-a	ypl246c	ykl133c	yer079c-a	ynl067w-a	ypr151c
ydr115w	yll005c	ykl225w	yel033w	ydr504c	yol003c	ynl056w	ylr361c
ydr205w	yll037w	ykl223w	yll018c-a	yjr014w	yjl052c-a	yjr044c	yor289w
ydr307w	ylr010c	ybl109w	yfl015c	yil090w	ykl051w	ygr016w	ylr368w
ydr319c	ylr050c	ynl338w	ycl005w	ydr544c	ypl103c	ykr051w	ydr210w
ydr366c	ylr064w	ygl260w	ycl002c	ycr038w-a	ydl054c	yhr212c	
ydr413c	ylr184w	ynl337w	ybl059w	yhr067w	ykl018c-a	ygr033c	
ydr486c	ylr283w	yfl062w	yal018c	ykl106c-a	ydl027c	yhr217c	
yel045c	ylr311c	yml132w	ydl185c-a	yhr069c-a	ymr071c	ykr065c	
yer041w	ylr365w	ybl108w	ykl165c-a	ylr047c	yir030w-a	yor137c	
yer085c	ylr426w	yol159c-a	ybl049w	yol048c	yor044w	ypr016w-a	
yer097w	ymr068w	ynl336w	ynl260c	yol047c	ygl041c	ygr053c	

表 4.5: The 128 ORFs of the 5^{th} class (No similarity) in the MIPS database, which are recognized as non-coding

ypr170w-a	yol053w	ymr151w	ylr112w	yjl136w-a	yhl006c	yel059w	ydr010c
ypr064w	yol038c-a	ymr148w	ylr104w	yjl118w	yhl005c	yel014c	ydl196w
ypl200w	yol029c	ymr141c	yll059c	yjl077c	ygr291c	yel010w	ycr025c
ypl041c	yol026c	ymr134w	yll030c	yjl064w	ygr290w	ydr525w	ycr022c
yor364w	ynl324w	ymr103c	ykr049c	yjl028w	ygr168c	ydr524w-a	ycr006c
yor268c	ynl211c	ymr057c	ykr032w	yjl027c	ygr102c	ydr396w	ybr292c
yor255w	ynl179c	ymr003w	ykl206c	yir020c	ygr026w	ydr350c	ybr144c
yor248w	ynl174w	yml107c	ykl158w	yir014w	ygl188c	ydr344c	ybr056w-a
yor097c	ynl146w	yml090w	ykl102c	yil152w	ygl138c	ydr278c	ybr027c
yor068c	ynl143c	yml084w	ykl097c	yil071c	ygl006w-a	ydr274c	ybl071c
yor060c	ynl017c	ylr404w	ykl061w	yil012w	yfr042w	ydr179w-a	ybl048w
yor029w	ymr320w	ylr402w	ykl044w	yhr192w	yfl021c-a	ydr102c	yar070c
yor024w	ymr252c	ylr400w	yjr157w	yhr173c	yer172c-a	ydr065w	yar053w
yol160w	ymr191w	ylr366w	yjr120w	yhr139c-a	yer135c	ydr042c	yar047c
yol159c	ymr187c	ylr145w	yjr023c	yhr095w	yer091c-a	ydr029w	yar030c
yol118c	ymr157c	ylr122c	yjl215c	yhl037c	yer066c-a	ydr015c	yal064w

方法求得相应数值并列在表 4.7 中. 由表 4.7 可知, 在第 2-4 类的 2050 个 ORF 中, 仅有 6.78%(139 条) 可能是非编码的, 而在第 5 类和第 6 类中, 非编码的 ORF 所占百分比分别是 22.67% 和 63.07% . 整个基因组中基因的总数应该是 5873 , 它实际上是第一类中基因的个数 (3392) 与第 2-6 类中由算法识别出的基因个数 (1911+399+171 = 2481) 的总和, 这个值与普遍接受的 5800-6000 这个范围是吻合的.

表 4.6: The 287 ORFs of the 6^{th} class (Questionable) in the MIPS database, which are recognized as non-coding

ydr521w	ydr509w	ydr467c	ydr526c	ydr442w	ydr431w	ydr426c	ydr417c
ydr401w	ydr355c	ydr360w	ylr465c	ylr458w	yor379c	ygr269w	ygr265w
ygr259c	ylr434c	yor345c	ygr228w	ydr241w	yor333c	yor331c	ygr219w
ydr230w	yor325w	ymr316c-a	yor309c	ypr177c	yor300w	ygr190c	ylr379w
ylr374c	ydr209c	ygr182c	ydr199w	ydr203w	ymr290w-a	ygr176w	yor282w
yor277c	ydr187c	ypr150w	yor263c	ypr142c	ylr338w	ypr136c	ygr151c
ypr126c	yor235w	ylr322w	ydr157w	ydr154c	ylr317w	ygr139w	ygr137w
yor225w	ymr244c-a	ydr149c	ybr266c	ypr099c	ydr136c	ydr133c	ygr115c
ygr114c	ylr282c	ylr279w	ylr280c	ygr107w	ypr077c	ybr232c	ydr114c
ylr269c	ydr112w	ybr226c	ybr224w	ylr261c	ynr025c	ypr053c	ypr050c
yor170w	yor169c	ymr193c-a	ypr039w	ypr038w	ylr252w	ynr005c	ydr094w
ybr206w	ygr073c	ygr069w	ygr064w	ynl013c	ymr172c-a	yor146w	ygr051c
yor139c	ygr039w	ymr158w-b	ymr153c-a	ydr053w	ydr048c	yor121c	yer181c
ylr202c	ylr198c	ymr135w-a	ykr047w	ygr018c	yor102w	ygr011w	yer165c-a
ykr033c	yjr038c	ypl025c	ylr171w	ylr169w	ypl034w	yhr193c-a	ypl035c
yor082c	ybr116c	ypl044c	yjr020w	yjr018w	yer148w-a	ybr113w	yer145c-a
ydr008c	ybr109w-a	ynl089c	ygl024w	yer138w-a	yer137w-a	ymr086c-a	ybr165w-a
ydl009c	yor055w	ybr090c	ydl016c	ynl105w	ylr140w	yjl009w	ygl042c
ymr075c-a	ynl114c	yor041c	yjl015c	ydl026w	yir023c-a	ynl120c	yjl022w
ylr123c	yhr145c	yjl032w	yir017w-a	ykl030w	ymr052c-a	ygl072c	ykl036c
ygl074c	ydl050c	ypl102c	yhr125w	ygl088w	ylr101c	ydl062w	ykl053w
ybr051w	ypl114w	ydl068w	yer087c-a	yer084w-a	yer084w	ynl171c	yer076w-a
yil020c-a	ygl102c	yol013w-b	yil029w-a	ypl136w	yil030w-a	ynl184c	ykl076c
yer067c-a	ylr076c	ykl083w	ynl198c	yil047c-a	ycr087w	yfr056c	ylr062c
ynl205c	yol037c	yfr052c-a	yml009c-a	yil060w	yer046w-a	yml012c-a	yhr071c-a
yil066w-a	yil068w-a	yil071w-a	yer038w-a	ykl111c	ycr064c	yfr036w-a	ygl149w
yhr063w-a	ynl226w	ynl228w	ykl118w	ygl152c	ycr049c	ynl235c	yml034c-a
yhr049c-a	ycr041w	ypl185w	ykl131w	yer014c-a	ydl151c	ydl152w	yml047w-a
yil100c-a	ydl158c	ykl147c	yhr028w-a	ydl163w	ykl153w	ypl205c	yml057c-a
yjl135w	ydl172c	ygl193c	ynl266w	yjl150w	yel009c-a	ykl169c	ydl187c
ynl276c	yel018c-a	ybl053w	yal019w-a	ygl204c	yfl012w-a	yfl013w-a	yll020c
ybl062w	yhl002c-a	ypl238c	yhl006w-a	yal026c-a	yjl175w	yml089c	ybl070c
yjl182c	yal031w-a	ygl217c	ygl218w	yml094c-a	ybl077w	yfl032w	ypl251w
yol134c	yml101c-a	yjl195c	yhl019w-a	ykl202w	ydl221w	yal042c-a	ygl235w
ycl041c	ypl261c	yol150c	ybl094c	ybl096c	yhl030w-a	ynl319w	yal056c-a
yml116w-a	yil163c	yal059c-a	yil171w-a	yjl220w	ybl107w-a	yel075w-a	

表 4.7: The numbers of predicted coding and non-coding ORFs of the $2^{nd} - 6^{th}$ classes

分类	2-4	5	6
ORF 总数	2050	516	463
TP	1835	383	164
FN	76	16	7
TN	133	112	280
FP	6	5	12
编码基因总数	1911	399	171
非编码基因总数	139	117	292
非编码序列所占百分比	139/2050=6.78%	22.67%	63.07%

5 参考文献

[1] 孙啸，陆祖宏，谢建明，生物信息学基础，北京：清华大学出版社， 2005.

[2] 张阳德，生物信息学，北京：科学出版社， 2004.

[3] 赵国屏，生物信息学，北京：科学出版社， 2002.

[4] 裘娟萍，钱海丰，生命科学概论，北京：科学出版社， 2004.

[5] 贺林，解码生命 —— 人类基因组计划和后基因组计划，北京：科学出版社， 2000.

[6] C. 丹尼斯， R. 加拉格尔编，林侠，李彦，张秀清，张孟，包静月等译，人类基因组 —— 我们的DNA, 北京：科学出版社， 2003.

[7] 卢大儒，基因治疗，北京：化学工业出版社， 2003.

[8] 陈竺，强伯勤，方福德，基因组科学与人类疾病，北京：科学出版社， 2001.

[9] P.A. 帕夫纳著，王翼飞译，计算分子生物学 —— 算法逼近，北京：化学工业出版社， 2004.

[10] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem., 258 (1983), 1318.

[11] M.A. Gates, A simple way to look at DNA, J. Theor. Biol., 119 (1986), 319-328.

[12] P.M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, Comput. Applic. Biosci., 12 (1995), 503-511.

[13] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, Curr. Sci., 66 (1994), 309.

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

[15] A. Nandy, P. Nandy, Graphical analysis of DNA sequences structure: II. Relative abundance of nucleotides in DNAs, gene evolution and duplication, Curr, Sci., 68 (1995), 75-85.

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

[17] A. Nandy, S.C. Basak. Simple numerical descriptor for quantifying effect of toxic subastances on DNA sequences, J. Chem. Inf. Comput. Sci., 40 (2000), 915-919.

[18] A. Nandy, P. Nandy, S.C. Basak, Quantitative descriptor for SNP related gene sequences, Internet Electron. J. Mol. Des., 1 (2002), 367-373.

[19] A. Nandy, Investigation on evolutionary changes in base distributions in gene sequences, Internet Electron. J. Mol. Des., 1 (2002), 545-558.

[20] A. Nandy. P. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, Chem. Phys. Lett., 368 (2003), 102-107.

[21] C. Raychaudhury, A. Nandy, Index scheme and similarity measures for macromolecular sequences, J. Chem. Inf. Comput. Sci., 39 (1999), 243-247.

[22] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences-A review, J. Biosci., 23 (1998), 55-71.

[23] S. Ghosh, A. Roy, S. Adhya, A. Nandy, Identification of new genes in human chromosome 3 contig 7 by graphical representation technique, Curr. Sci., 84 (2003), 1534-1543.

[24] R. Zhang, C.T. Zhang, Z curves, an intuitive tool for visualizing and analyzing DNA sequences, J. Biomol. Struc. Dyn., 11 (1994), 767

[25] Z.P. Feng, C.T. Zhang, A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins, Int. J. Biochem. Cell Biol., 34 (2002) 298-307.

[26] X.F. Guo, M. Randic, S.C. Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy, Chem. Phys. Lett., 350 (2001), 106.

[27] X.F. Guo, A. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, Chem. Phys. Lett., 369 (2003), 361-366.

[28] M. Randic, G. Krilov, Characterization of 3-D sequences of proteins, Chem. Phys. Lett., 272 (1997), 115-119.

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

[30] M. Randic, M. Vracko, N. Lers, D. Plavsic, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett., 368 (2003), 1-6.

[31] M. Randic, M. Vracko, N. Lers, D. Plavsic, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett., 371 (2003), 202-207.

[32] M. Randic, M. Vracko, J. Zupan, M. Novic, Compact 2-D graphical representation of DNA, Chem. Phys. Lett., 373 (2003), 558.

[33] M. Randic, Graphical representations of DNA as 2-D map, Chem. Phys. Lett., 386 (2003), 468.

[34] M. Randic, J. Zupan, A.T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, Chem. Phys. Lett., 397 (2004) 247-252

[35] Y.H. Wu, A.W. Liew, H. Yan, M. Yang, DB-Curve: a novel 2D method of DNA sequence visualization and representation, Chem. Phys. Lett., 367 (2003), 170.

[36] J.A. Berger, S.K. Mitra, M. Carli, A. Neri, Visualization and analysis of DNA sequences using DNA walks, J. Franklin Inst., 341 (2004), 37-53.

[37] M. Randic, M. Vracko, On the similarity of DNA primary sequences, J. Chem. Inf. Comput. Sci., 40 (2000), 599-606.

[38] Z. Bajzer, M. Randic, D. Plasic, S.C. Basak, Novel map descriptors for characterization of toxic effects in proteomics maps, J. Mol. Graph. Model., 22 (2003), 1.

[39] M. Randic, On characterization of DNA primary sequences by a condensed matrix, Chem. Phys. Lett., 317 (2000), 29-34.

[40] M. Randic, Condensed Representation of DNA Primary Sequences, J. Chem. Inf. Comput. Sci., 40 (2000), 50-56.

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

[42] M. Randic, X.F. Guo, S.C. Basak, On the charaterization of DNA primary sequences by triplet of

nucleic acid bases, J. Chem. Inf. Comput. Sci., 41 (2001), 619-626.

[43] C. Mathe, M.F. Sagot, T. Schiex, P. Rouze, Survey and Summary: Current methods of gene prediction, their strengths and weaknesses, Nucleic Acids Res., 30 (2002), 4103-4117.

[44] J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, Nucleic Acids Res., 29 (2001), 2607-2618.

[45] R.J. Carter, I. Dubchak, S.R. Holbrook, A computational approach to identify genes for functional RNAs in genomic sequences, Nucleic Acids Res., 29 (2001), 3928-3938.

[46] R. Guigo, Computational gene identification: an open problem, Comput. Chem., 21 (1997), 215-222.

[47] T.A. Thanaraj, Positional characterization of false positives from computational prediction of human splice sites, Nucleic Acids Res., 28 (2000), 744-754.

[48] W. Li, P. Bernaola-Galvan, F. Haghighi, I. Grosse, Applications of recursive segmentation to the analysis of DNA sequences, Comput. Chem., 26 (2002), 491-510.

[49] Y.D. Cai and P. Bork, Homology-Based Gene Prediction Using Neural Nets, Analytical Biochem., 265 (1998), 269-274.

[50] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol., 268 (1997), 78-94.

[51] C.T. Zhang, J. Wang, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve, Nucleic Acids Res., 28 (2000), 2804-2814.

[52] C.T. Zhang, J. Wang, R. Zhang, Using a Euclid distance discriminant method to find protein coding genes in the yeast genome, Comput. Chem., 26 (2002), 195-206.

[53] Q. Liu, Y.S. Zhu, B.H. Wang, Y.X. Li, A HMM-based method to predict the transmembrane regions of β-barrel membrane proteins, Computational Biol. and Chem., 27 (2003), 69-76.

[54] S.L. Salzberg, M. Pertea, A.L. Delcher, M.J. Gardner, H. Tettelin, Interpolated Markov models for eukaryotic gene finding, Genomics, 59 (1999), 24-31.

[55] R. Staden, Computer methods to locate signals in nucleic acid sequences, Nucleic Acids Res., 12 (1984), 505-519.

[56] R. Staden, A.D. McLachlan, Codon preference and its use in identifying protein coding regions in long DNA sequences, Nucleic Acids Res., 10, (1982) 141-156.

[57] R. Staden, Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes, Nucleic Acids Res., 12 (1984), 551-567.

[58] A.D. McLachlan, R. Staden, D.R. Boswell, A method for measuring the non-random bias of codon usage table, Nucleic Acids Res., 12 (1984), 9567-9575.

[59] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res., 10 (1982), 5303-5318.

[60] J.L. Bennetzen, B.D. Hall, Codon selection in yeast, J. Biol. Chem., 257 (1982), 3026-3031.

[61] P.M. Sharp, W.H. Li, The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential application, Nucleic Acids Res., 15 (1987), 1281-1295.

[62] N. Sueoka, Y. Kawanishi, DNA G+C content of the third codon position and codon usage biases

of human genes, Gene, 261 (2000), 53-62.

[63] S.M. Leisner, D.A. Neher, Third Position Codon Composition Suggests Two Classes of Genes Within the Cauliflower Mosaic Virus Genome, J. theor. Biol., 217 (2002), 195-201.

[64] B.R. Morton, U. Sorhannus, M. Fox, Codon adaptation and synonymous substitution rate in diatom plastid genes, Molecular Phylogenetics and Evolution, 24 (2002), 1-9.

[65] A. Fuglsang, The relationship between palindrome avoidance and intragenic codon usage variations: a Monte Carlo study, Biochem. Biophys. Res. Commun., 316 (2004), 755-762.

[66] 顾万君，马建民，周童，孙啸，陆祖宏. 不同结构的蛋白编码基因的密码子偏性研究，生物物理学报， 18 (2002), 81-86.

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

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

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

[70] M. Yan, Z.S. Lin, C.T. Zhang, A New Fourier Transform Approach for Protein Coding Measure Based on the Format of the Z Curve, Bioinformatics, 14 (1998), 685-690.

[71] F.B. Guo, H.Y. Ou, C.T. Zhang, ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes, Nucleic Acids Res., 31 (2003), 1780-1789.

[72] L.L. Chen, C.T. Zhang, Gene recognition from questionable ORFs in bacterial and archaeal genomes, J. Biomol. Struct. Dyn., 21 (2003), 99-110.

[73] 张春霆，用几何学方法分析 DNA 序列，中国科学基金， 3 (1999), 152-153.

[74] 张春霆，人与其他生物基因组若干重要问题的生物信息学研究，《自然科学进展》， 14 (2004), 1367-1374.

[75] H.H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, Bioinformatics, 19 (2003), 2122-2130.

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

[77] T. Jiang, Y. Xu, M.Q. Zhang, Current topics in computational molecular biology, 计算分子生物学前沿课题，清华大学出版社， The MIT Press, 2002: 157-171.

[78] V. Bafna, S. Muthukrisnan, R. Ravi, comparing similarity between RNA strings, Comput. Sci., 937 (1995), 1-14.

[79] F. corper, B. Michot, RNAlign program: alignment of RNA sequences using both primary and secondary structres, Comput. Appl. Biosci., 10 (1995), 389-399.

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

[81] B. Shapiro, An algorithm for comparing multiple RNA secondary structures, Comput. Appl. Biosci., 4 (1988), 387-393.

[82] B. Shapiro, K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, Comput. Appl. Biosci., 6 (1990), 309-318.

[83] B. Liao, T.M. Wang, A 3D graphical representation of RNA secondary structures, J. Biomol. Struct. Dynam., 21 (2004), 827-832.

[84] B. Liao, K.Q. Ding, T.M. Wang, On A Six-Dimensional Representation of RNA Secondary Structures, J. Biomol. Struc. Dynam., 22 (2005), 455-463.

[85] Y.H. Yao, X.Y. Nan, T.M. Wang, A Class of 2D Graphical Representations of RNA Secondary Structures and the Analysis of Similarity Based on Them, J. Comput. Chem., 26 (2005), 1339-1346.

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

[87] 张春霆，蛋白质结构分类与结构类预测研究，中国科学基金， 5 (2000), 298-299.

[88] C.T. Zhang, K.C. Chou, G.M. Maggiora, Predicting protein structural classes from amino acid composition: application of fuzzy clustering, Protein Eng., 8 (1995)5, 425-435.

[89] C.T. Zhang, Z.S. Lin, Z.D. Zhang, M. Yan, Prediction of the helix/strand content of globular proteins based on their primary sequences, Protein Engineering, 11 (1998), 971-979.

[90] C.T. Zhang, R. Zhang, Skewed distribution of protein secondary structure contents over the conformational triangle, Protein engineering, 12 (1999), 807-809.

[91] W.S. Bu, Z.P. Feng, Z.D. Zhang, C.T. Zhang, Prediction of protein (domain) structural classes based on amino-acid index, Eur. J. Biochem., 266 (1999), 1043-1049.

[92] Z.P. Feng, Prediction of the Subcellular Location of Prokaryotic Proteins Based on a New Representation of the Amino Acid Composition, Biopolymers, 58 (2001), 491-499.

[93] Z.P. Feng, C.T. Zhang, Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids, Int. J. Biol. Macromol., 28 (2001) 255-261.

[94] Z.P. Feng, An overview on predicting the subcellular location of a protein, *In Silico* Biology, 2 (2002), 0027.

[95] K.C. Chou, Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition, Proteins, 43 (2001), 246-255.

[96] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long range correlations in nucleotide sequences, Nature, 356 (1992), 168-170.

[97] S. Nee, Uncorrelated DNA walks, Nature, 357 (1992), 450.

[98] V.V. Prabhu, J.M. Claverie, Correlations in intronless DNA, Nature, 359 (1992), 782.

[99] C.A. Chatzidimitriou-Dreismann, D. Larhammar, Long range correlations in DNA, Nature, 361 (1993), 212-213.

[100] Y.C. Liu, X.F. Guo, J. Xu, L.Q. Pan, S.Y. Wang, Some note on 2-D graphical representation of DNA sequence, J. Chem. Inf. Comput. Sci., 42 (2002), 529-533.

[101] 韩乐，莫忠息， RNA-Z 曲线及其在病毒基因识别中的应用，生物数学学报， 19 (2004), 245-250.

[102] M. Randic, A.T. Balaban, On A Four-Dimensional Representation of DNA Primary Sequences, J. Chem. Inf. Comput. Sci., 43 (2003), 532.

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

[104] P.A. He, J. Wang, Numerical Characterization of DNA Primary Sequence, Internet Electron. J. Mol. Des., 1, (2002) 668.

[105] 张艳, 何凤田, p53基因在肿瘤基因治疗中的研究进展, 世界华人消化杂志, 11 (2003), 1593-1596.

[106] 张云, 刘泽军, 新的抑癌基因 ASPP 对 p53 作用的研究进展, 生命科学, 16 (2004), 79-80.

[107] 顾健人, 曹雪涛, 基因治疗, 北京: 科学出版社, 2001.

[108] S. yamamoto, A. Romanenko, M. Wei, C. Masuda, W. Zaparin, W. Vinnichenko, A. Vozianov, C.C. Lee, K. Morimura, H. Wanibuchi, M. Tada, S. Fukushima, Specific p53 Mutations in Urinary Bladder Epithelium after the Chernobyl Accident, Cancer Research, 59 (1999), 3606-3609.

[109] C.A. Garcia, A. Ahmadian, B. Gharizadeh, J. Lundeberg, M. Ronaghi, P. Nyren, Mutation detection by pyrosequencing: sequencing of exons 5-8 of the p53 tumor suppressor gene, Gene, 253 (2000), 249-257.

[110] X.H. Li, Z.G. Li, M.L. Hu, A novel set of Wiener indices, J. Mol. Graph. Model., 22 (2003), 161-172.

[111] P.A. He, J. Wang, Some properties for DNA curve, J. Phys. A: Math. Gen., 37 (2004) 7135-7142.

[112] S. Karlin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics, 11 (1995), 283.

[113] P.A. He, The Sieve Ratio for Characterization and Similarity Analysis of DNA Sequences, Comb. Chem. High T. Scr., 8 (2005), 449-453.

[114] B.L. Hao, J. Qi, B. Wang, Prokaryotic Phylogeny Based on Complete Genomes without sequence alignment, Modern Physics Letters B, 17(2) (2003), 1-4.

[115] J. Qi, B. Wang, B.L. Hao, Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach, J. Mol. Evol., 58 (2004), 1-11.

[116] Y.C. Liu, The Numerical Characterization and Similarity Analysis of DNA Primary Sequences, Internet Ele. J. Mol. Des., 1 (2002), 675-684.

[117] D. London, Inequalities in quadratic forms, Duke Math. J., 33 (1966), 511-522.

[118] R. Shrock, S.H. Tsai, Upper and lower bounds for ground state entropy of antiferromagnetic Potts models, Phys. Rev. E., 55 (1997), 6791-6794.

[119] N. Biggs, Algebraic Graph Theory, 1^{st} ed., Cambridge, England: Cambridge University Press, 1974.

[120] B. Liao, T.M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett., 388 (2004), 195-200.

[121] S. Guiasu, Information Theory with Application, McGraw-Hill, New York, 1997.

[122] N.F.G. Martin, J.W. England, Mathematical theory of entropy, Addison-Wesley Pub. Co., Reading, MA, 1981.

[123] D. Bonchev, M. Randic, Shannon's entropy of proteomic 2D-gel maps, Chem. Phys. Lett., 372 (2003), 548.

[124] W.H. Li, D. Graur, Fundamentals of molecular evolution, Sinauer Associates, Sunderland, MA, 1991.

[125] K.F. Lau, K.A. Dill, A lattice statistical mechanics model of the conformation and sequence space of proteins, Macromolecules, 22(10) (1989), 3986-3997.

[126] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, H.C. Lee, Mean-Field HP Model, Designability and Alpha-Helices in Protein Structures, Phys. Rev. Lett, 84 (2000), 386-389.

[127] 李冬冬, 王正志, 杜耀华, 晏春, 蚂蚁群落优化算法在蛋白质折叠二维亲 - 疏水格点模型中的

应用，生物物理学报， 20(5) (2004), 371-374.

[128] 王仲君，蛋白质折叠过程中自回避搜索的算法研究，交通与计算机， 23(4) (2005), 43-46 .

[129] 沈同，王镜岩，生物化学，北京：高等教育出版社， 1990.

[130] D. Voet, J.G. Voet, C.W. Pratt., *Fundamentals of biochemistry*, New York: Wiley, 2000.

[131] 马飞，武耀廷，许晓风，遗传密码子和氨基酸若干物理化学特性的相关性研究，安徽农业大学学报， 30(4) (2003), 439-445.

[132] 陈志华，陈惟昌，邱红霞，王自强，氨基酸的分子结构与遗传密码简并及二维集合分类，生物物理学报， 17 (2001), 187-193.

[133] Thomas M. Cover, Joy A. Thomas 著，阮吉寿，张华译，信息论基础，北京：机械工业出版社，2005.

[134] 张红煊，朱贻盛，王自明，李颖洁，基于复杂度和复杂率的心动过速和心室纤颤检测，中国生物医学工程学报， 20 (2001), 423-429.

[135] A. Lempel, J. Ziv, On the Complexity of Finite Sequences, IEEE T. Inform. Theory, 22 (1976), 75-81.

[136] J. Ziv, A. Lempel, A universal algorithm for sequential data compression, IEEE T. Inform. Theory, 23 (1977), 337-343.

[137] J. Ziv, A. Lempel, Compression of indiviual sequences via variable-rate coding, IEEE T. Inform. Theory, 24 (1978), 530-536.

[138] 张泯泯，张宏，童勤业，基于混沌图像的防伪技术，电子技术应用， 9 (2003), 6.

[139] N. Liu, T.M. Wang, A relative similarity measure for the similarity analysis of DNA sequences, Chem. Phys. Lett., 408 (2005), 307-311.

[140] C. Reusken, J.F. Bol, Structural elements of the 3′-terminal coat protein binding site in alfalfa mosaic virus RNAs, Nucleic Acids Res., 24(14) (1996), 2660-2665.

[141] E.C. Koper-Zwarthoff, F.T. Brederode, P. Walstra, J.F. Bol, Nucleotide sequence of the 3′-noncoding region of alfalfa mosaic virus RNA 4 and its homology with the genomic RNAs, Nucleic Acids Res., 7 (1979), 1887-1900.

[142] S.W. Scott, X. Ge, The complete nucleotide sequence of RNA 3 of citrus leaf rugose and citrus variegation ilarviruses, J. Gen. Virol., 76 (1995), 957-963.

[143] E.C. Koper-Zwarthoff, F.T. Brederode, P. Walstra, J.F. Bol, Nucleotide sequence of the putative recognition site for coat protein in the RNAs of alfalfa mosaic virus and tobacco streak virus, Nucleic Acids Res., 8 (1980), 3307-3318.

[144] B.J.C. Cornelissen, H. Janssen, D. Zuidema, J.F. Bol, Complete nucleotide sequence of tobacco streak virus RNA 3, Nucl. Acids Res., 12 (1984), 2427-2437.

[145] R.H. Alrefai, P.J. Shicl, L.L. Domier, C.J. D'Arcy, P.H. Berger, S.S. Korban, The nucleotide sequence of apple mosaic virus coat protein gene has no similarity with other Bromoviridae coat protein genes, J. Gen. Virol., 75 (1994), 2847-2850.

[146] S.W. Scott, X. Ge, The complete nucleotide sequence of the RNA 3 of lilac ring mottle ilarvirus, J. Gen. Virol., 76 (1995), 1801-1806.

[147] E.J. Bachman, S.W. Scott, G. Xin, V.B. Vance, The complete nucleotide sequence of prune dwarf ilarvirus RNA3: implications for coat protein activation of genome replication in ilarviruses, Vi-

rology, 201 (1994), 127-131.

[148] F. Houser-Scott, M.L. Baer, K.F. Liem, J.M. Cai, L. Gehrke, Nucleotide sequence and structural determinants of specific binding of coat protein or coat protein peptides to the 3′ untranslated region of alfalfa mosaic virus RNA 4, J. Virol., 68 (1994), 2194-2205.

[149] V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin, A.S. Frolov, Investigating extended regulatory regions of genomic DNA sequences, Bioinformatics, 15 (1999), 644-653.

[150] V.D. Gusev, L.A. Nemytikova, N.A. Chuzhanova, On the complexity measures of genetic sequences, Bioinformatics, 15 (1999), 994-999.

[151] Y.L. Orlov, V.N. Potapov, Complexity: an internet resource for analysis of DNA sequence complexity, Nucleic Acids Res., 32 (2004), W628-W633.

[152] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature, 431 (2004), 931-945.

[153] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Life with 6000 genes, Science, 274 (1996), 546.

[154] The Yeast Genome Directory, Nature (Suppl.), 387 (1997), 5.

[155] H.W. Mewes, K. Albermann, M. Bahr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S.G. Oliver, F. Pfeiffer, A. Zollner, Overview of the yeast genome, Nature (Suppl.), 387 (1997), 7-8.

[156] J.W. Fickett, Finding genes by computer: the state of the art, Trends Genet, 12 (1996), 316-320.

[157] R. Staden, Graphic methods to determine the function of nucleic acid sequences, Nucleic Acids Res., 12 (1984), 521-538.

[158] C. Li, P.A. He, J. Wang, Artificial Neural Network Method for Predicting Protein Coding Genes in the Yeast Genome, Internet Electron. J. Mol. Des., 2 (2003), 527-538.

[159] M.Q. Zhang, Identification of protein coding regions in the human genome by quadratic discriminant analysis, Proc. Natl. Acad. Sci. USA, 94 (1997), 565-568.

[160] M.A. Basrai, P. Hieter, J.D. Boeke, Small Open Reading Frames: Beautiful Needles in the Haystack, Genome Res., 7 (1997), 768-771.

[161] P. Mackiewicz, M. Kowalczuk, A. Gierlik, M.R. Dudek, S. Cebrat, Origin and properties of non-coding ORFs in the yeast genome, Nucleic Acids Res., 27 (1999), 3503-3509.

[162] J.W. Fickett, C.S. Tung, Assesment of protein coding measures, Nucl. Acids Res., 20 (1992), 6441-6450.

[163] M. Burset, R. Guigo, Evaluation of gene structure prediction programs, Genomics, 34 (1996), 353-367.

[164] D. Kulp, D. Haussler, M. Rees, F. Eeckman, A generalized hidden Markov model for the recognition of human genes in DNA, Proc. Int. Con. Intell. Syst. Mol. Biol., 4 (1996), 134-142.

[165] J. Claverie, Computational methods for the identification of genes in vertebrate genomic sequences, Human Molecular Genetics, 6 (1997), 1735-1744.

读博期间发表、完成论文及获奖情况

论文

1 Chun Li, Jun Wang. Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences, Comb. Chem. High T. Scr., 6 (2003), 795-799. （SCI 收录，第 1 章第 3 节，第 2 章第 2 节）

2 Chun Li, Jun Wang. On a 3-D representation of DNA primary sequences, Comb. Chem. High T. Scr., 7 (2004) 23-27.（SCI 收录，第 1 章第 2 节）

3 Chun Li, Jun Wang. New invariant of DNA sequences, J. Chem. Inf. Model., 45 (2005) 115-120.（SCI 收录，EI 收录，第 2 章第 3 节）

4 Chun Li, Jun Wang. Relative entropy of DNA and its application, Physica A: Statistical Mechanics and its Applications, 347 (2005), 465-471.（SCI 收录，EI 收录，第 2 章第 5 节）

5 Chun Li, Nadia Helal, Jun Wang. Recognition of protein coding genes in the yeast genome based on the relative-entropy of DNA, Comb. Chem. High T. Scr., 9 (2006), 49-54.（SCI 收录，第 4 章）

6 Chun Li, Nannan Tang, Jun Wang. Directed graphs of DNA sequences and their numerical characterization, Journal of Theoretical Biology, 2006.（in press, SCI 刊源，第 1 章第 4 节，第 2 章第 4 节）

7 Chun Li, Jun Wang. A naturally logical representation for DNA primary sequences, Comb. Chem. High T. Scr., 2006.（in press, SCI 刊源，第 3 章第 1 节）

8 Nadia Helal, Mahmoud Dorrah, Chun Li. Ladder-like graphical representation of p53 alterations in some human cancers, Isotope and Radiation Research, 2006.（accepted, 第 1 章第 3 节）

9 Chun Li, Yi Zhang, Jun Wang. Numerical characterization of protein primary sequence,（已投稿，第 3 章第 2 节）

10 Chun Li, Ai-hua Wang, Lili Xing. On the similarity of RNA secondary structure,（已投稿，第 3 章第 3 节）

11 Chun Li, Jun Wang. Generalized LZ complexity of (0,1)-sequences and its application,（已投稿，第 3 章第 3 节）

获奖情况

第一届“纪念向坊隆”村井隆研究生专项奖学金 (2005)

创新点摘要

1. 由于序列比对 (alignment) 一直受空位罚分缺乏理论依据和算法复杂度居高不下的困扰， Randic 等人提出了一种基于序列不变量的序列比较方法．目前常用的序列不变量都是基于相应矩阵的，其中最大特征值应用最为广泛．然而，以最大特征值为不变量有一个不可回避的问题，那就是它的计算会随着序列长度的增加而变得越来越难．本文结合代数图论相关知识提出了一个基于矩阵范数的新的序列不变量 —— ALE 指标，它与最大特征值等效但它的计算非常容易，这使得基于不变量的比较方法在完全基因组比较及其相关研究领域中的应用具有了可行性 (论文 3)．

2. 提出了生物序列的有向图表示．有向图表示不仅弥补了原有生物序列图形表示的许多缺陷，而且还为生物大分子的数值刻画提供了新的途径．本文就在提出有向图的同时，给出了生物序列的上三角矩阵表示，并讨论了现有序列不变量在上三角矩阵情况下的兼容性 (论文 6)．

3. 生物序列的数值刻画在对海量生物学数据进行数学分析方面有着重要的作用．然而，现有 DNA 序列数值刻画方法中许多都只是从碱基组成上着手，而序列之所以称为序列的另一个重要因素：元素之间的序关系，却在很大程度上被忽略了。为了更好地反映序列中元素，尤其是它们之间的序关系所包含的信息，本文从一般数字序列出发构造出一种特殊的链 (全序集)，进而提出了 DNA 序列的正规化相对熵的概念 (论文 4)，并在此基础上对酿酒酵母基因组序列进行基因识别，得到了一个酿酒酵母基因组中基因总数为 5873 的估计，与普遍接受的 5800-6000 相符 (论文 5)．

致 谢

本文是在导师王军教授悉心指导下完成的。在五年的研究生学习期间，王老师在学习和科研方面给了我大力的支持，并在生活上给了我很大的帮助。在此毕业之际，谨对王老师多年以来所给予的栽培和厚爱表示深深的谢意。

感谢王天明教授、郑斯宁教授、邱瑞锋教授、冯红副教授、王毅教授、贺明峰教授、侯中华教授、于洪全副教授、代万基老师和蔡晶老师对我的关心与帮助.

感谢印度 Jadavpur University 的 Ashesh Nandy 教授在我的科研工作方面所给予的鼓励。

感谢浙江理工大学贺平安副研究员、新加坡 National University of Singapore 杨家亮博士、埃及 National Center for Nuclear Safety and Radiation Control, Atomic Energy Authority 的 Nadia Helal 博士，同他们的交流与合作使我受益匪浅。

感谢张之正博士、郑德印博士、吕可波博士、廖波博士以及博士生吉日木图、王欣、张华军、吴军、庄举娟、张彩环、孙怡东、张屹、姚玉华、袁春新、刘娜、李玉双、王晓霞、魏传安、张俊和硕士生唐南南、张伟、康金慧、王琛颖、王晓元等同门几年来对我的支持与帮助.

感谢渤海大学有关领导和同事对我的支持.

感谢我的家人对我的鼓励，特别感谢我的妻子，她在我上学期间，一人承担了家庭的重担使我能够在校安心学习，我的每一点进步都离不开她的理解和支持.

附　注

1. 如果对 DNA 序列的四种碱基实施 Zhang [166] 中的 $\boldsymbol{R}_c$ 操作，并对相应 Z- 曲线坐标（见本文第 18 页）做 "变换" ψ：

$$\hat{x}_n = (x_n + n)/2,\ \hat{y}_n = (y_n + n)/2,\ \hat{z}_n = (z_n + n)/2$$

则可以得到在本质上和本文提出的 3D 图形一样的坐标（见本文第 19 页式 1.2.1 ）.

从数学上讲，DNA 序列的任何两种曲线表示，只要它们都能通过坐标与 DNA 序列相互唯一地重构，那么显然这两种曲线表示之间就可以通过某种变换联系起来. 但二者是不是等价形式，取决于人们观察问题的角度：从代数的角度讲，如果坐标变换是可逆的，那二者一定可以互相转换，或者说是代数等价的. 但从几何角度讲，一般的代数变换并不能保持图形不变，只有正交变换才能做到这一点. 而我们讨论的 DNA 序列的图形表示一定得是具有相同的图形才认为是等价的. 而上述 "变换" ψ 中 n 是个变量，从而 ψ 不是线性变换，更进一步，ψ 不是正交变换. 所以，本文的 3D 图形并非 Z- 曲线的等价形式. 事实上，这一点从几何直观上也可以看出来. 如本文第 18 页所述，Z- 曲线没能避免曲线自身的重叠与自交，即 Z- 曲线是有圈的. 而本文的 3D 图形中这种现象是不会出现的（见本文第 20 页）. 这是本文的 3D 图形与 Z- 曲线之间的一个重要区别.

2. Voss [167] 通过定义 "二元操作子" 函数 U_k 而将 K 个字符集上的符号序列分解成 K 个能标识相应符号在序列中位置的二元序列，并对 DNA 序列进行了讨论，进而在 DNA 序列的长程关联等方面得出一些结论. 这种序列在 Buldyrev 等 [168] 中被称为碱基位置序列（ base position sequences, bp-sequences ）. 本文的逻辑表示和 Voss 的碱基位置序列提出的角度和后继的工作都不一样. 当然，从本质上讲，二者的构造规则相同.

致谢：

感谢匿名评阅人指出了本文的 3D 图形与 Z- 曲线之间，以及逻辑表示与 Voss 的碱基位置序列之间的一些联系，并对今后的工作提出了很有价值的建议.

参考文献：

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

[167] R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett., 68 (1992), 3805-3808.

[168] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, F. Sciortino, H.E. Stanley, Long-range fractal correlations in DNA, Phys. Rev. Lett., 71 (1993), 1776.