复杂非线性系统多尺度涨落的层次结构分析

任奎

2001 年 2 月 23 日

摘要

本文以湍流层次结构模型为基础，发展了对 Taylor-Couette 流动系统，时空广延动力系统（包括耦合映像格子和沙堆模型）以及生物 DNA 序列等几个典型的复杂系统的多尺度分析方法。

首先，我们以湍流为例详细介绍了标度律研究的基本概念和充分发展湍流的层次结构模型及其基本分析方法（ β - 检验和 γ - 检验方法），并简要地讨论了层次相似参数 β 和最高激发态的标度指数 γ 的物理意义。

通过对不同雷诺数下 Taylor-Couette 流的速度信号的详细分析，我们证明， Taylor-Couette 流的统计结构可以由层次结构模型很好地描述。层次相似参数 β 值与雷诺数无关，表明联系多尺度 (l)，多强度 (p) 涨落的机制是普适的。同时， β 值还反映了 Taylor-Couette 流比自由射流具有更强的间歇性。我们还发现，最高激发态的标度指数 γ 值在雷诺数 $R = 10^5$ 时出现跃迁，对应着实验观察到的有序的 Taylor 涡的破碎。这些统计分析令人鼓舞地给出了流体结构演化的图象。

接着，我们考察了两个典型的时空广延动力系统，耦合映像格子和自组织临界性的沙堆模型系统，发现其存在非线性的多标度行为，并用层次结构模型对其进行了很好地描述。层次相似参数 β 对两类系统的差别给出了定量的刻画。

我们进一步将层次结构分析的方法应用于对生物 DNA 序列的多尺度分析，发现 DNA 序列的碱基密度分布具有多尺度、强间歇的结构，并满足层次相似律，定量反映了基因序列在长期的进化过程中演化到一定的统计自组织状态。我们发现，基因组序列的平均 β 值比高等生物（如人类）的大，反映了原核生物在某种意义上具有更丰富的多样性（或随机性）。我们还对 DNA 序列词汇使用频率场进行了多尺度分析，证明了序列中不同长度的词汇的相关性中存在类似层次相似的结构。

最后，我们引进了一个刻画 DNA 编码序列的新方法 – 多变量熵距离法 (MED 方法) 。我们证明， MED 方法对 DNA 编码序列的发现具有非凡的效率（平均综合得分 >98% ）。我们提出， MED 方法可能是对复杂系统描述的普遍适用的一种方法。

关键词：层次结构模型，标度律，复杂系统，多尺度分析，湍流， DNA 序列，自组织临界性。

Abstract

This thesis is devoted to a study of complex systems using multiscaling analysis. We have focused on three complex systems, the Taylor-Couette flow, the coupled map lattice and SOC sandpile model, and the biological DNA sequence in the framework of the Hierarchical Structure model proposed recently by She and Lévêque.

In Chapter two, the basic concepts in the scaling analysis of turbulence are introduced. The SL Hierarchical Structure model and its methods of analysis (β-test and γ-test) and the physical meaning of its parameters are discussed in detail.

It is demonstrated by a detailed analysis of the fluctuating velocity signals that the fully developed turbulent Taylor-Couette flow is accurately described by the Hierarchical Structure model. The hierarchical symmetry parameter β is found to be Reynoulds number independent, which indicates that the mechanism connecting the multi-scale and multi-amplitude fluctuations are universal in the flow. The β-test reveals also that the Taylor-Couette flow is more intermittent than the free jet. Furthermore, we found that the most intermittent structure undergo a transition around $R = 10^5$, which correspond to a visually observed breaking of the Taylor vortices. The close connection between the statistical signature and the evolution of fluid structures is encouranging.

Next, we consider the dynamics of two typical spatial-temporally extended systems4, the coupled map lattice system and the sandpile SOC system, in the framework of the Hierarchical Structure model. It is demonstrated that both systems display multiscaling behavior which are well described by the SL Hierarchical Structure model. The hierarchical similarity parameter β gives a quantitative proof that the SOC system is more intermittent than the globally coupled map lattice.

In Chapter five, we applied the hierarchical structure analysis to the statistical study of genomic sequences. It is found that the nucleotide density distribution along the DNA chain satisfies the hierarchical similarity over the scale of the typical size of a gene (from $100bps$ to $8000bps$), indicating the self-organization of genomic sequences over the long evolution. It is also found that the prokaryodic genomes have hierarchical similarity pa-

rameter β larger than Eukaryotic genomes, indicating a greater degree of self-organization in more evolued biological systems. We have also developed a multi-scale method for the grammer analysis of the genomic language, and show that the multi-length word use frequency exhibit hierarchical similarity.

Finally, we present a new method for identifying coding and non-coding DNA sequences – the so-called Multivariate Entropy Distance (MED) method. It is demonstrated that the MED method leads to a very efficient classification of prokaryodic DNA sequences (up to an accuracy above 98%). It is suggested that the MED method forms a variant of methods applicable to the analysis of other complex systems.

Key words: Hierarchical Structure model, Multi-scale analysis, Scaling laws, Complex systems, Turbulence, DNA sequence, Self-organized criticality.

目 录

第一章　引言

§1.1　从湍流看复杂系统

毫无疑问，自然科学，特别是物理学的最新发展，主要集中在对具有复杂行为的系统的研究上。区别于经典物理学的研究对象，这类系统内部包含着众多的尺度和自由度，它们非线性地耦合在一起，使得系统表现出复杂或混沌的时空行为。

经典的物理系统通常认为是由很好的确定性的物理规律（牛顿定律，最小作用原理等）控制，表现为系统的动力学过程由一些简洁的微分（积分）方程所描述，因而是比较确定的，其行为也是相对简单的。

对于复杂系统，事情变得不那么简单。一方面对于很多复杂系统（如生物，社会，经济等系统）我们暂时还难以写出确定性的方程；另一方面，即使可以有相应的方程，其形式也必然是非线性的，方程本身就导致了处理上的困难。

湍流作为这类复杂系统的典范之一，一个多世纪里始终没有找到解析和定量的描述方法。其中的原因很多，但湍流运动的复杂性显然是占主导地位的。由于描述湍流运动的 NS 方程是个强非线性，高自由度的偏微分动力系统，因而对其解析求解几乎是不可能的。直接数值模拟则因为必须涉及巨大的自由度而受到计算机条件的限制，模式理论也因为对湍流运动物理机制的缺乏理解而显得并不成功。

湍流系统具有复杂系统几乎所有的特征，当然也有其本身的一些特殊性质。泛泛地讲，湍流是一个高自由度，强非线性的多体系统。它所对应的应该是甚高维数的随机集合上的运动（从这个意义上讲，湍流不同于当今非线性动力学所能够讨论的低维混沌系统）。这决定了湍流运动表现出很强的随机性，与 NS 方程的确定性很不相称。正是由于这些随机性，导致湍流场的一些物理量如速度，局部能量，涡量等随着时间和空间的分布存在强弱不同的涨落，其行为不能被简单地预测。湍流场的涨落具有典型的多尺度结构，其尺度从与能量耗散相关的 Kolmogorov 微尺度跨越到与流动边界同量级的积分尺度；各种不同尺度，不同强度，不同振幅的旋涡不断产生，演化和衰减；不同尺度旋涡之间既独立又关联，使湍流成为一个典型的大数自由系统，具有时空多层次结构特性。

无论如何，可以由确定性的方程所描述的系统应该具有某种规律性（也就是结构性），湍流也不例外。湍流并非单纯的随机运动，它是有结构的。湍流场的结构也是多尺度的，

既包括大尺度的拟序涡旋结构，如发卡涡等，也包括充分发展湍流的涡丝结构等。Kolmogorov 提出的湍流理论，完全基于随机的湍流运动，其结果并不能对实际湍流行为进行准确的描述，原因就是湍流内部包含有组织性和结构性的运动，主要体现在湍流的间歇性特性上。

从对湍流这一典型复杂系统的描述我们可以看出，复杂系统场既表现了充分的随机性，又展现了一定的结构性，是二者的有机结合。复杂系统的结构性使人们相信其规律性，复杂系统所表现出的明显的随机性特点，使得统计的描述在这里变得必不可少。复杂系统具有自组织性，但又远离平衡态，具有动态的特征。因此当我们观测复杂系统的某些物理量时，会发现多尺度，多强度（多层次）的涨落是它们的基本特征。

尽管上面我们对复杂系统进行了许多泛泛的讨论，在这篇论文里面，作者绝无意从哲学上去探讨复杂性问题，而希望通过对一些具体的复杂系统（如流体湍流，自自组织临界性系统，耦合映象格子，生物体 DNA 序列等）的分析，找到一些分析这些系统的比较有效的方法，为加深人们对复杂现象的理解和认识，探索这些复杂现象背后的物理机制提供一些新的思路。

§1.2 标度律和层次结构模型

无论现象是多么复杂，人们坚信 [41] 复杂系统的背后必然有某些简单的规律性存在，统计标度律就是其中一例。

我们认为，在具有动力学过程的复杂系统（而不是纯粹的复杂几何系统）里，统计标度律是对系统多尺度自组织状态的定量刻划。当多尺度复杂系统进入一定的稳定状态时，各尺度的动力学自由度之间通常有一种耦合。但是由于相互作用的自由度数量巨大（大小尺度分得很开），这种耦合表现出极大的随机性和无规则性。在湍流中，当雷诺数很大时，湍流场的两个特征尺度，即决定流动激发的积分尺度和决定流动耗散的粘耗尺度相差很大。对均匀各向同性湍流场而言，在这两个特征尺度之间，没有其他特征长度。在这种情况下，物理统计量如速度结构函数，局部能量耗散率等随尺度的变化通常呈幂次律。对幂次律的定量刻划就是标度指数，该指数描述了当时矩的变化速度。如果知道了这些指数，就能根据大尺度速度脉动矩预测任何较小尺度的脉动矩。

然而，复杂系统并非完全完全的随机系统，它的内部存在着组织性和结构性。在湍流及混沌系统的研究中人们将这种随机性中表现出的结构称为间歇结构。以湍流为例，

Kolmogorov 1941 年的理论（ $K41$ 理论）[58] 认为，充分发展湍流的惯性区内，速度差的 p 统计阶矩与尺度 l 之间存在简单的标度律关系， $\langle \delta v_l^p \rangle \sim l^{\zeta_p}$ ，其中 $\zeta_p = p/3$ ，是 p 的线性函数，这是将湍流看成完全随机场得到的结果。后来的实验测量表明标度指数对 $K41$ 的理论预测值有系统的偏离 [1]，特别是高阶矩的标度指数（ $p > 3$ ）。看起来 ζ_p 是 p 的非线性函数，而不是线性函数。这种现象后来被称为反常标度律现象，通常认为这是由于湍流场的间歇结构引起的。

一旦反常标度律出现了，一个很自然的问题就是：我们怎样去刻划这种间歇性的动力学。对这个问题的一个最新回答就是由 She 和 L*évê*que 提出的层次结构模型（ Hierarchical structure model ） [91, 92] 。

层次结构描述的是不同强度，不同时空尺度的涨落之间的关系，引进一系列所谓层次量，每个刻划相应强度的涨落结构。层次越高（ p 越大），与层次量 $\epsilon_l^{(p)}$ 相联系的涨落的强度越强。反映强弱涨落的 $\epsilon_l^{(p+1)}$ 和 $\epsilon_l^{(p)}$ 之间由一递推不变关系（层次相似律），(2.12) 式，联系起来。这一递推不变性关系引入参数 β 和 γ 。 β 度量流场的间歇性，而 γ 则度量了最高激发态（最强间歇结构）的奇异性。通过一系列的物理上的考虑，层次结构得出充分发展湍流速度结构函数的标度指数公式，

$$\zeta_p = p/9 + 2[(1 - (2/3)^{p/3}] \tag{1.1}$$

以及层次相似性参数 $\beta = (\frac{2}{3})^{(1/3)}$ 和最高激发态标度指数 $\gamma = 1/9$ 。

湍流的层次结构模型从其发表以来受到了广泛的注意。许多理论，实验和数值模拟工作都提出了支持该模型的证据。人们发现，该模型不仅真实流体湍流 [9, 86, 87, 107] ，数值模拟湍流 [22] ，磁流体湍流 [82] 里成立，而且在被动标量场 [88, 61] ，有限扩散凝聚（ DLA ）模型 [84] ，自然图象的灰度场 [99] ，混沌动力系统 [105] 等领域均成立。反映了层次相似律的普适性。在本文中我们将探讨层次相似律在刻划 Taylor-Couette 湍流，生物体 DNA 序列，以及时空广延的动力系统中的涨落现象中的作用。

第二章　湍流场中的标度律和层次结构

自从 Kolmogorov 的开创性工作以来 [58]，湍流中的标度律问题始终是流体力学的研究热点之一 [36]。在这一章里，我们将对这一问题的背景作一个简单的介绍。本章的内容将是我们以后各章分析的基础和工具。

§2.1　湍流中的标度律

尽管通常认为湍流是由强非线性的 NS 方程控制的复杂流动系统，并且这种流动中的大尺度结构的动力学行为还受到具体流动条件（如边界条件等）的影响，但是人们始终有这样一种愿望，就是去寻找各种湍流系统中的“普适性”的问题。第一次，也或许是最著名的尝试是 20 世纪 40 年代初由苏联科学家 A.N. Kolmogorov[58] 完成的。通过定义所谓的速度结构函数，即尺度相距为 l 的两点的速度差：

$$\delta v_l = v(x+l) - v(x), \tag{2.1}$$

Kolmogorov 猜想惯性区内速度结构函数的 p 阶矩与尺度 l 之间存在简单的标度律关系：

$$S_p(l) = \langle \delta v_l^p \rangle \sim l^{\zeta_p}, \tag{2.2}$$

这里 $l_0 \geq l \geq \eta$, l_0 是湍流能量注入的尺度，η 是能量耗散尺度。l_0 和 η 间的尺度范围称为惯性区。

如果我们假设在雷诺数趋向于无穷的极限下，惯性区内单位质量的平均能量耗散率 ϵ 为常数，并且结构函数 $S_p(l)$ 只与尺度 l 和 ϵ 有关，那么简单的量纲分析可以导出

$$S_p(l) = C_p \epsilon^{\frac{p}{3}} l^{\frac{p}{3}}, \tag{2.3}$$

与 (2.2) 比较可知，这意味着 $\zeta_p = p/3$.

在均匀各向同性并且能量耗散率有限的假设下，Kolmogorov 从 Navier-Stokes 导出了一个精确结果，Kolmogorov 方程：

$$\langle \delta v_l^3 \rangle - 6\nu \frac{d}{dl} \langle \delta v_l^2 \rangle = -\frac{4}{5} \langle \epsilon \rangle l \tag{2.4}$$

惯性区内， (2.4) 中包含粘性的那项可以被忽略掉，于是

$$\langle \delta v_l^3 \rangle = -\frac{4}{5}\langle \epsilon \rangle l \tag{2.5}$$

(2.5) 式就是著名的 Kolmogorov 关于 3 阶速度结构函数的 $-4/5$ 定律，与量纲分析的结果 (2.3) 完全一致。

对于 $p = 2$ ， (2.3) 式的傅立叶变换就是另一个著名的定律， Kolmogorov 关于湍流速度能谱的 $-5/3$ 定律：

$$E(k) = C_K \epsilon^{\frac{2}{3}} k^{-\frac{5}{3}} \tag{2.6}$$

这里， C_K 是个普适的常数。

虽然有对 $K41$ 理论支持的证据 [43, 75] ，但是后来的实验测量表明标度指数对 $K41$ 的理论预测值有系统的偏离 [1, 76] ，特别是高阶矩的标度指数 ($p > 3$) 。看起来似乎 ζ_p 是 p 的非线性函数，而不是线性函数。这种现象后来被称为反常标度律现象。人们又发现，这种标度律偏离 $K41$ 理论的现象是由于湍流的间歇性引起的 [36] 。对湍流间歇性的研究是过去几十年湍流研究的热点。

为了将间歇性的效应考虑进来， Kolmogorov 在 1962 提出了对 $K41$ 理论的修改 [59] ，现称为 $K62$ 理论。该理论引进所谓修正的自相似性假设（ Refined Similarity Hypothesis ， RSH ）：高雷诺数，惯性区内，速度结构函数 δv_l 具有普适的概率分布，其统计性质与能量耗散率 ϵ_l 无关。这里的 ϵ_l 是指半径为 l 的小球内的平均能量耗散率，于是

$$S_p(l) = C_p l^{p/3} \langle \epsilon_l^{p/3} \rangle. \tag{2.7}$$

大家马上就会注意到，（ 2.3 ）式和（ 2.7 ）式的差别仅仅在于（ 2.3 ）式中的 ϵ 在（ 2.7 ）式中由 ϵ_l 所取代。这意味着能量耗散率从一个全局常数变成了局部的涨落量。如果 $S_p(l)$ 对 l 成标度律关系，那么容易证明：

$$\langle \epsilon_l^p \rangle \sim l^{\tau_p}, \tag{2.8}$$

而且

$$\zeta_p = p/3 + \tau_{p/3}. \tag{2.9}$$

显然， ζ_p 的具体形式与 τ_p 联系起来了。

后来，人们提出了很多方法（包括 $K62$ 理论）来从理论上决定 τ_p 的具体形式。最广泛接受的是所谓的多分形模型 [69, 79, 73] 。这种模型将惯性区的级串过程看作是一个

随机倍增过程（ RMP ） W ， W 的统计特征决定了惯性区的标度指数 ζ_p 。这种方法的主要缺点是随机倍增过程的物理和流体力学意义不明显。因此模型里的参数没办法定下来，只能靠人为的调节。

§2.2 扩展的自相似性 -ESS

尽管标度律问题不仅有趣而且非常重要，但是标度律的测量却远远不是一件容易的事。在许多场合，结构函数对尺度在双对数图上显示出非线性的关系，使得人们很难得出令人信服的标度指数。这个问题最终由于扩展自相似性的发现而得以进展。湍流的扩展自相似性 ESS(Extended Self Similarity) 是指 1993 年意大利学者 Benzi 及其合作者所发现的相对标度律现象 [11, 12, 13] ，即：任意 p 阶的速度结构函数对 q 阶 (通常取为 3 阶) 结构函数成幂次函数关系，即

$$\langle \delta v_l^p \rangle \sim \langle \delta v_l^q \rangle^{\zeta_{p,q}} \tag{2.10}$$

上式中 $\zeta_{p,q}$ 一般称为相对标度指数。此相对幂次律直到尺度 l 非常接近 Kolmogorov 耗散尺度时仍保持得很好，即使在雷诺数不很大，没有明显的惯性区的情况下也是如此，而通常此时结构函数自身随尺度的变化已不能保持幂次律关系。从 Kolmogorov 的 $-4/5$ 定律 (2.5) 我们知道， 3 阶结构函数与尺度 l 成正比关系。因此当 $q = 3$ 时， $\zeta_{p,q}$ 可以近似看作 ζ_p 。

ESS 可理解为：在一定范围内，标度律可能已经不再适用，但不同阶的速度结构函数都表现出对幂次律同样的偏离，而它们的相对函数依赖关系却保持不变。更物理地讲，就是没有任何特征量来控制这一系列的结构函数对惯性区标度律的偏离。

ESS 的发现对标度指数的实验测量具有重要意义， Benzi 等人的研究表明，在雷诺数不很高的情况下，即使没有明显惯性区，测量到标度指数 $\zeta_{p,q}$ 与惯性区的标度指数 ζ_p 非常接近。因此测量阶与 3 阶结构函数的相对标度指数可作为一种新的获得标度指数的方法，这一点亦得到实验的证明。但在有些流动中，相对标度指数可能具有更广泛的物理含义。

随后的工作中，同一研究小组又发现适用范围更广的广义扩展的自相似性 (Generalized ESS ，简称 GESS)[13, 14] 。即在从积分尺度到非常小的耗散尺度的整个长度尺度范围内，即使 ESS 不满足，当我们将研究对象从速度结构函数转化为相对于某一阶结构函

数的规一化结构函数时，可以得到归一化函数之间也存在着幂次律关系，即

$$G_{(p,q)}(r) \sim G_{(p',q')}(r)^{\rho(p,q;p',q')} \tag{2.11}$$

这里 $G_{(p,q)}(r)$ 是无量纲的结构函数，定义为：$G_{(p,q)}(r) = \langle \delta v_l^p \rangle / \langle \delta v_l^q \rangle^{p/q}$. 这是一个比 ESS 更加有趣的性质。

“扩展”一词在这里意味着或许还存在着其他形式的标度律，是物理量与物理量之间的标度律。这些标度律在物理量随尺度变化不成标度关系的时候仍然成立。

由 ESS 引起的对标度指数的更加准确的测量，使得人们更加相信，湍流中的标度律与 $K41$ 理论相去甚远。并且，没有迹象表明，随着雷诺数的增加，标度律将向 $K41$ 理论靠拢。这促使人们去从更加物理的角度寻求对反常标度律的深入理解。

实际上， Bershadskii[16] 在研究反常扩散过程时，曾经独立地发现这种相对标度律现象。

§2.3 充分发展湍流的层次结构模型

1994 年， She 和 Lévêque 提出了一种对湍流间歇性的新的描述方法 [91, 98]，得到了与实验及计算结果非常吻合的理论结果。与其他方法想比较，这是一种更为直接的方法。

让我们来看看 She 和 Lévêque 是怎样考虑的。首先定义所谓的层次量：

$$\epsilon_l^{(p)} = \frac{\langle \epsilon_l^{p+1} \rangle}{\langle \epsilon_r^p \rangle} = \frac{\int \epsilon_l^{p+1} P(\epsilon_l) d\epsilon_l}{\int \epsilon_l^p P(\epsilon_l) d\epsilon_l} = \int \epsilon_l Q(\epsilon_l) d\epsilon_l, p = 0, 1, 2, ..., \infty \tag{2.12}$$

这里， $P(\epsilon_l)$ 是 ϵ_l 概率分布函数。 $Q(\epsilon_l) = \epsilon_l^p P(\epsilon_l) / \int \epsilon_l^p P(\epsilon_l) d\epsilon_l$ 是 p 阶加权概率分布函数，它反映了高阶矩的概率分布特性。层次量实际上就是相邻两阶矩之比。由于 $\epsilon_l \geq 0$，很容易证明层次量形成一个不减序列，就是说： $\epsilon_l^{(0)} \leq \epsilon_l^{(1)} \leq ... \leq \epsilon_l^{(\infty)}$ 。

上面的这些定义和性质同样可以自然地推广到物理量 $|\delta v_l|$ 。

基于一些物理上的考虑，层次结构模型提出如下四个基本假设 H1~H4：

H1. 在湍流中，不论雷诺数多大，包括速度 v 和能量耗散率 ϵ_l 在内的所有物理量都是有界的。尽管目前还没有就这一点在三维 Navier-Stokes 方程的一般情形从数学上给出严格证明，但该假设在物理上显然是合理的。

根据这个假设，马上可以有：物理量的 p 阶矩是有界的 $\epsilon_l^{(p)} \leq \epsilon_l^{max}$ ， $\delta v_l^{(\infty)} \leq \delta v_l^{max}$ 等等。

$$\lim_{p\to\infty} \epsilon_l^{(p)} = \frac{\langle \epsilon_l^{p+1} \rangle}{\langle \epsilon_l^{p} \rangle} = \epsilon_l^{(\infty)} \tag{2.13}$$

并且

$$\lim_{p\to\infty} \delta v_l^{(p)} = \frac{\langle \delta v_l^{p+1} \rangle}{\langle \delta v_l^{p} \rangle} = \delta v_l^{(\infty)} \tag{2.14}$$

$\epsilon_l^{(\infty)}$ 和 $\delta v_l^{(\infty)}$ 刻划所谓的最强间歇结构。

H2. 除了积分尺度 l_0 或耗散尺度 η 外，没有其他的特征长度尺度。在尺度 l_0 和 η 之间，就是所谓的惯性区内，物理量如速度结构函数，局部平均能量耗散率等与尺度之间存在标度律关系。这个假设与 $K41$ 理论基本上是一致的。于是有： $\epsilon_l^{(\infty)} \propto l^{\lambda}$.

H3. 与假设 2 类似，对 ϵ_l 和 $|\delta v_l|$ ，除了 $\epsilon_l^{(\infty)}$ 和 $\delta v_l^{(\infty)}$ 外没有其它的特征脉动幅度。

H4. 在充分发展湍流的惯性区内， $\epsilon_l^{(p)}$ 之间通过如下关系联系起来：

$$\epsilon_l^{(p+1)} = C_p (\epsilon_l^{(p)})^{\beta} (\epsilon_l^{(\infty)})^{1-\beta} \tag{2.15}$$

式中 $0 < \beta < 1$ 是与 p 和 l 无关的普适常数，称为间歇参数。 C_p 仅依赖于 p ，与 l 无关。这是层次结构理论最直接的一个假设，也是层次结构模型的创新所在，其正确与否只能靠实验来考察。

从 (2.8) 式， (2.15) 式和假设 **H2** 立刻可以得到，对每个 $p = 0, 1, 2, ...$ ，有下式成立：

$$\tau_{p+2} - (1-\beta)\tau_{p+1} + \beta\tau_p + \frac{2}{3}(1-\beta) = 0. \tag{2.16}$$

实验的证据表明， τ_p 是 p 的非线性函数，这使我们可以猜想 τ_p 具有如下的形式： $\tau_p = -2/3p + 2 + f(p)$ ，式中 $f(p)$ 必须满足 $f(\infty) = 0$ 。于是， (2.16) 式可以写成，

$$f(p+2) - (1-\beta)f(p+1) + \beta f(p) = 0, \tag{2.17}$$

上式是个二阶齐次差分方程，可能有多个解 [66] 。其中一个非平凡的解就是 $f(p) = \alpha\beta^p$ 。即 τ_p 可以写成，

$$\tau_p = -2/3p + 2 + \alpha\beta^p \tag{2.18}$$

对充分发展湍流，容易导得：

$$\tau_p = -\frac{2}{3}p + C(1-\beta^p) \tag{2.19}$$

式中 $\beta = 2/3$ ， $C = 2$ 代表最高激发态的余维数。根据式 (2.9) ， (2.19) 意味着对速度场，结构函数的标度指数为，

$$\zeta_p = p/9 + C[(1 - (2/3)^{p/3}] \tag{2.20}$$

这就是关于充分发展湍流的 SL 标度律公式。

关系式 (2.15) 从其发表以来受到了广泛的注意。许多理论，实验和数值模拟工作都提出了支持该模型的证据。人们发现，该模型不仅真实流体湍流 [9, 86, 87, 107] ，数值模拟湍流 [22] ，磁流体湍流 [82] 里成立，而且在被动标量场 [88, 61] ，有限扩散凝聚 (DLA) 模型 [84] ，自然图象的灰度场 [99] ，混沌动力系统 [105] 等领域均成立。反映了层次相似律的普适性。在本文中我们将探讨层次相似律在刻划 Taylor-Couette 湍流，生物体 DNA 序列，以及时空广延的动力系统中的涨落现象中的作用。

§2.4 β - 检验和 γ - 检验

为了便于对实验测量的真实脉动信号中进行层次结构分析，最近 She 等人发展了一套完整的测量层次相似性参数 β 和最高激发态标度指数 γ 的方法，分别称为 β - 检验和 γ - 检验 [94] 。下面对这两个检验的具体过程作简单介绍。

层次结构模型引入了一系列的层次量，每一层次量描述强度与之对应的涨落。 p 越大，与 $\epsilon_l^{(p)}$ 相联系的涨落的强度越强。反映强弱涨落的 $\epsilon_l^{(p+1)}$ 和 $\epsilon_l^{(p)}$ 之间由一递推不变关系， (2.12) 式，联系起来。这一递推不变性关系引入参数 β 和 γ 。 β 度量流场的间歇性，而 γ 则度量了最高激发态（最强间歇结构）的奇异性。

对于速度场，我们可以引入新的变量重写 (2.15) 式：

$$H_{p+1}(l) = A_p H_(p)(l)^{\beta} H_{\infty}(l)^{1-\beta}, \tag{2.21}$$

式中 $H_p = \epsilon_l^{(p)}$ ， $H_{\infty} = \epsilon_l^{(\infty)}$ 。注意，这里的参数 β 与 (2.12) 式中的不同，但二者间有简单的关系。

你也许会注意到，式中含有 $H_{\infty}(l)^{1-\beta}$ 项，需要涉及的 p 趋向无穷时的行为，这在实际的计算中是没法处理的。于是我们得想别的办法。实际上，我们可以绕过这一项的计算，采取方便一点的办法。只要我们考虑方程在时的比，就可以消除 $H_{\infty}(l)^{1-\beta}$ 项：

$$\frac{H_{p+1}(\ell)}{H_2(\ell)} = \frac{A_p}{A_1} \left(\frac{H_p(\ell)}{H_1(\ell)} \right)^{\beta}. \tag{2.22}$$

如果，在双对数坐标上 $H_{p+1}(l)/H_2(l)$ 与 $H_p(l)/H_1(l)$ 成线性关系，那么我们就说 β - 检验被通过了。最小二乘法拟合所得到的线的斜率就是 β 值。

当 β - 检验通过后，如果进一步假设：

$$F_\infty \sim S_3^\gamma, \tag{2.23}$$

那么，利用性质 $\zeta_3 = 1$ ， SL 标度律的标度指数的全集可以写成 [91] ：

$$\zeta_p = \gamma p + C(1 - \beta^p), \tag{2.24}$$

这里， $C = (1 - 3\gamma)/(1 - \beta^3)$ 。现在，简单的代数运算给出

$$\zeta_p - \chi(p;\beta) = \gamma(p - 3\chi(p;\beta)), \tag{2.25}$$

其中， $\chi(p;\beta) = (1 - \beta^p)/(1 - \beta^3)$ 。分别以 $p - 3\chi(p;\beta)$ 和 $\zeta_p - \chi(p;\beta)$ 为横纵坐标作图，如果横纵坐标间有线性关系存在，则我们断定假设 (2.23) 被满足了。线的斜率就是 γ 值。这个过程被称为 γ - 检验。

尽管严格地讲 γ 所度量的是很高阶矩的性质， $p \to \infty$ [参见 (2.14)] 。我们发现，实际上，对实验和数值模拟的数据，适当高的 p (通常 $3 \le p \le 10$) 就可以很好地定义最高激发态。这主要是因为那些对 $H_p(l)(3 \le p \le 10)$ 做贡献的结构构成了的具有统计意义的强度涨落的大部分。因此， γ 刻划的是那些有限（但很长）的速度记录中的结构。

§2.5 对数 -Poisson 统计

在层次结构模型提出后不久， Dubrulle[31] 和 She&Waymire[92] 各自独立地发现通过一个称为对数 Poisson 的简单随机级串过程，假设 H4 可以精确地得到实现。

湍流级串动力学包含扰动从大尺度向小尺度的传播过程；当湍流达到统计定常的充分发展阶段时，也包含小尺度向大尺度的稳定的反馈作用，两者并存，维持了惯性区的完整存在。通常引入一个随机线性映射或随机倍增过程 $W_{ll'}$ ，在统计意义上将小尺度 (l') 脉动与大尺度 (l) 脉动联系起来：

$$\delta v_{l'} \overset{law}{=} W_{ll'}\delta v_l \tag{2.26}$$

称事件 $W_{ll'}$ 为级串结构。这些事件中的一个子集 $W_{ll'}^{(MI)}$ 具有最大的幅值，且幅值随尺度的变化率最大，这就是最强间歇 (most intermittent, 简称 MI) 结构。

应该指出，控制不同长度尺度之间脉动发展的统计规律，无论在定性上还是定量上都具有复杂得多的形式。用对级串作统计描述，只是为了把级串过程的 (统计) 对称性用一种方便的数学形式 (线性映射) 表示出来。这种方法研究级串现象显然不同于传统的傅立叶分解或分析空间流态变形的方法，它明显的好处是便于进行物理描述。

Dubrulle[31] 和 She & Waymire[92] 独立地发现，如果将随机线性映射 W 选择为对数 -Poisson 的形式，那么 (2.12) 就可以得到精确的实现。

特别地，我们令

$$\delta v_l = (\frac{l}{l_0})^{\gamma}\beta^n \delta v_{l_0}, \tag{2.27}$$

式中 n 是个平均值为 λ 的独立泊松（ Poisson ）随机变量， n 的概率分布为：

$$P(n) = e^{-\lambda}\frac{\lambda^n}{n}, n = 1, 2, ... \tag{2.28}$$

由 (2.27) 可以推导出，

$$\langle \delta V_l^p \rangle = \sum_n \beta^{np} P(n) \langle \delta V_{l_0}^p \rangle \tag{2.29}$$

因为 $\beta < 1$ ，所以最强振幅出现在 $n = 0$ 的地方。其他小振幅的事件可以通过对强振幅事件乘以整数倍的 β 因子获得。因此， She 和 Waymire 称这种级串为量子化的级串过程 [92] 。这是层级结构模型在理论方面的一个重要结果。

§2.6 关于 β 和 γ 的讨论

根据层次结构模型，湍流的标度律是由最强间歇结构的特征决定的。由于标度指数一般是非线性的，湍流场一般表现为多尺度结构，其物理图象可描述如下：湍流场是由不同尺度、不同脉动幅度及不同相干程度 (和 / 或空间填充度) 的层次结构组成的。大尺度 (如积分尺度) 上的脉动由于非线性相互作用传播到小尺度，动力学状态变化的特征时间随空间尺度减小而减小，脉动速度的典型值 (例如均方根值) 亦如此。然而，相对于脉动速度的典型值，在级串过程中渐渐出现稀少的大振幅事件，脉动振幅愈大，它们占据的空间集合的维数愈小，因此它们空间形态的相干性就愈强。其中最大振幅的事件对应于最强间歇结构或最高激发态，它们占有的空间集合维数最小，具有最强的空间相干性。在统计定常情况下，不同尺度和幅度的脉动由相似律联系在一起，形成了一个整体性的层次结构，使湍流在统计意义上成为一个完整的自组织体系。可认为层次结构是由最高激发态和层次相似参数 (间歇参数) 共同决定的。

层次结构模型的参数有其具体的物理意义。参数 β 度量湍流场的间歇性特征。在 $\beta \to 1$ 的极限下，没有间歇性存在。 K41 理论框架下的湍流就属于这个极限。很明显，$\beta \to 1$ 时，由 (2.24) 式可知 $\zeta_p \to p/3$ 。相反，另一个极限是 $\beta \to 0$ 。在这个极限下，只有最高激发态存在，这是有序的极端。这种极限的一个例子就是 Frisch 的黑 - 白 β 模型 [37] ，那里只有白结构对能量耗散起作用。黑 - 白 β 模型的一个特别的例子是随机分布的 Burgers 激波，其能谱为 k^{-2} ，并且对所有 $p \geq 1$ ， $\zeta_p = 1$[96, 90] 。对于真实流体的湍流， $0 < \beta < 1$ ，正是介于 Burgers 湍流和 K41 湍流两个极端之间。因此，真实湍流是有序和随机状态的混合。

γ 度量的是最高激发态 (最奇异结构，H_∞) 的奇异性。如果最高激发态是激波结构，γ 将等于 0 (激波间断面两侧的速度差正比于 l^0 ，或者说与 l 无关)。这是非常强的奇异性。若奇异性的例子是形如布朗运动的速度廓线，那里 $\gamma = 1/2$ 。 γ 越大，最高激发态的奇异性越弱。 Kolmogorov 的耗散结构， $\gamma = 1/3$ ，因此比 Burgers 激波结构的奇异性小，间歇性也小。

γ 与最高激发态相联系可以由下面的过程看出来。利用 SL 标度律公式 (2.24) 和 $\zeta_3 = 1$ 的约束关系，我们有，

$$\zeta_p = \gamma p + \frac{1 - 3\gamma}{1 - \beta^3}(1 - \beta^p) \tag{2.30}$$

式中 $0 < \beta < 1$ ，而 $1/3 > \gamma > 0$ 。很明显，对所有的 p ， ζ_p 随着 p 的增加而增加。这与湍流的对数 - 正态 (log-normal) 模型不同。在那个模型里，当 p 很大以后， ζ_p 会随着 p 的增加而减小，与著名的 Novikov 不等式相矛盾 [36] 。

我们可以求 (2.30) 式的一阶导数：

$$\frac{d\zeta_p}{dp} = \gamma - \frac{1 - 3\gamma}{1 - \beta^3} \cdot \ln\beta \cdot \beta^p \tag{2.31}$$

可以看出，当 $p \to +\infty$ 时， $\frac{d\zeta_p}{dp} \to \gamma$ ，也就是说，

$$\gamma = \lim_{p \to +\infty} \frac{d\zeta_p}{dp} \tag{2.32}$$

所以，当很大时， ζ_p 几乎随 p 线性地增加，线性关系的斜率就是 γ 。现在， γ 同最高激发态联系起来了。图 2.1 中，我们给出了 ζ_p 的一阶导数 $\frac{d\zeta_p}{dp}$ 随 p 的变化关系。

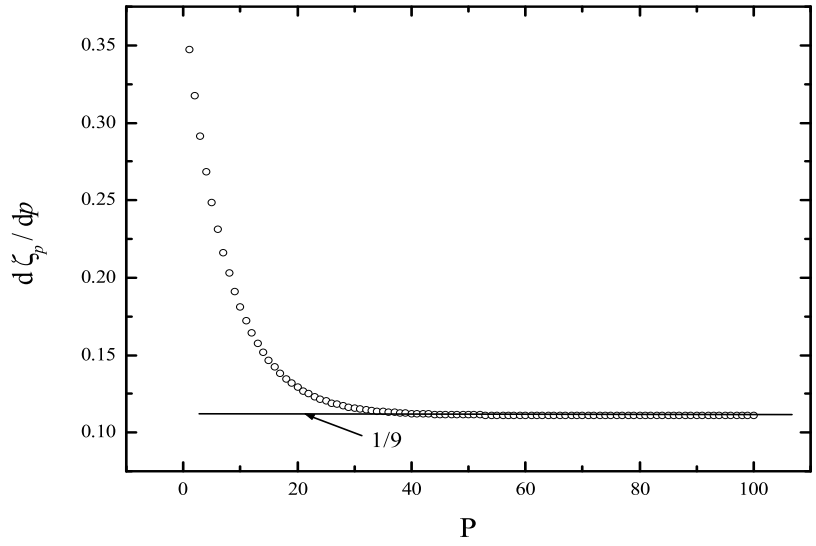

图 2.1: 参数为 $\beta = 0.874$， $\gamma = 1/9$ 时 SL 标度律的一阶导数 $d\zeta_p/dp$，横线为 $d\zeta_p/dp = 1/9$

第三章　Taylor-Couette 流中的标度律和结构

最近这些年关于不同几何条件的湍流的标度律问题有不少研究（参见文献 [3] 及其引文）。我们这里讨论两个同心圆桶间的流体湍流，Taylor-Couette 流的问题。我们讨论的仅限于外桶静止，而内桶旋转的情况，内外桶的半径比为 0.724[64] 。我们分析了一系列雷诺数下（最大雷诺数达 540000 ）测量得到的速度信号。这种封闭系统里的湍流比自由射流和尾流湍流具有更加复杂的结构。我们将证明这种差别可以由结构函数的标度指数和 She-Lévêque 层次结构模型 [91] 的参数 β 反映出来。

在湍流的核心区域，湍流级串所产生的小尺度结构和由壁面附近产生的结构混合起来，形成流体堆（ bulk ）结构。这种小尺度结构与流体堆混合的后果之一，就是即使是在相当高的雷诺数下，也可能没有惯性区存在。在我们考虑的 Taylor-Couette 系统里 [64]，基于 Taylor 微尺度的 Taylor 雷诺数是比较大的 ($R_\lambda \approx 240$)，然而，速度信号的功率谱并没有表现出任何标度特征，也就是说，没有惯性区存在。这也许是因为从桶壁发展起来的流动结构有很强的各向异性，带有多个尺度的信息，扫过探头。即使是在这种情况下，扩展的自相似性 ESS[11, 12] 仍然在比较宽的尺度范围内成立 [64] 。这表明，一旦湍流场比较充分地发展起来了，自相似性就成了这个多尺度场的内在特性。

§3.1　Taylor-Couette 流简介

Taylor-Couette 系统在过去几十年流体力学基本概念发展的过程中起过重要的作用 [27, 64] 。这种系统一般由两个同心圆桶组成，其中或者内桶，或者外桶，甚至两桶同时旋转。两桶之间的流体运动称为 Taylor-Couette 流。

Taylor-Couette 系统的重要几何参数包括：内 (a) 外 (b) 桶半径之比 $\eta = a/b$ ，，相似比 $L = \widetilde{L}/(b-a)$ ，$\widetilde{L}$ 是流体柱长度。基于内外桶半径的雷诺数：$R_i = a(b-a)\Omega_i/\nu$ 和 $R_o = b(b-a)\Omega_o/\nu$ ，$\Omega_i(\Omega_o)$ 是内（外）桶旋转的角速度，ν 是流体的运动粘性系数。此外，系统顶端的几何形状也至关重要。

过去对 Taylor-Couette 流的研究工作主要集中在探讨适当高雷诺数下的流动状态上。通常流动状态随雷诺数的变化是这样的（这里我们只讨论内圆桶旋转的情况）：从低雷诺数的层流开始，随着雷诺数 R_i 的增加，流动发生第一次失稳，导致产生定常的完全

柱对称的切向轴向涡结构。这种状态同 Rayleigh-Benard 对流中产生的环状对流态很类似，但是轴向和径向两个方向不再等价。

当雷诺数 R_i 增加到接近临界雷诺数 R_{ic} 的时候，流动出现了波状涡结构。里面，涡的波状扰动以内桶旋转速度的某个比例的速度向桶周围传播。如果继续增加 R_i ，将出现调制波状涡旋状态，系统出现准周期频率谱，也就是说，出现两个不可约的频率 ω_1 和 ω_2 。然而，随着雷诺数的变化，这两个频率不会发生锁频现象，即使在很小的有理数比的情况下也是如此。继续增加雷诺数 R_i 将出现湍流态。

本文所使用的实验速度信号数据是由美国 University of Texas, Austin 的 G.S. Lewis 和 H.L. Swinney 采集提供的。实验所采用的流动的几何参数在表 3.1 中列出。

a	b	η	$\widetilde{L}$	L
15.999 cm	22.085 cm	0.724	69.5cm	11.4

表 3.1: 测量本文所使用的速度数据的 Taylor-Couette 系统的几何参数。

速度信号是通过热膜测量的，经过校正，消除热膜尺寸和频率响应效应 [63] 。所有数据都对内桶转速归一化过，并且经过数字滤波滤去噪声。速度信号的平均速度约 $42.2 \times \Omega/(2\pi)$ 厘米 / 秒，采样频率为内桶旋转频率的 2500 倍。

雷诺数	粘性系数	旋转频率	采样频率	样本量
12,200	10 cSt	2 Hz	5,000	10,000,000
24,400	10 cSt	4 Hz	10,000	12,000,000
48,800	10 cSt	8 Hz	20,000	12,000,000
34,800	3.5 cSt	2 Hz	5,000	12,000,000
69,000	3.5 cSt	4 Hz	10,000	12,000,000
133,000	1.83 cSt	4 Hz	10,000	12,000,000
266,000	1.83 cSt	8 Hz	20,000	12,000,000
540,000	0.914 cSt	8 Hz	20,000	20,000,000

表 3.2: Taylor-Couette 流速度实验数据的描述。其中第三项是内桶旋转频率。

§3.2 流场的峭度和偏度

Taylor-Couette 流的运动状态不同于通常意义下的湍流图象，这主要体现在两点。第一个是能谱问题。即使在很高的雷诺数下，速度场的能谱也并不表现出明显的标度区。也就是说高雷诺数下的湍流场也没有明显的惯性区存在 [64] 。图 3.1 给出了 3 个不同雷诺数下的湍流能谱。

反映 Taylor-Couette 流奇异性的第二个证据在于速度梯度的峭度和偏度对雷诺数的依赖关系。峭度和偏度在 $R > 100,000$ 后都随着 R 的增加而减小。这与人们在射流湍流，尾迹湍流等许多其他流动中观察到的现象相反 [97] 。这种差别值得进一步的研究。我们猜想这可能是由圆桶的旋转所引起的复杂流动结构间的相互作用引起的。这些复杂的流动结构大多起源于边界层和 Taylor 涡结构。图 3.2 中给出了速度梯度的峭度和偏度对雷诺数的依赖关系。

§3.3 概率分布函数的光滑化处理

我们在计算标度指数的时候，采用从速度增量的截断的概率分布函数 PDF 而不是直接用几何平均的方法，并且研究标度指数随着样本量增加时的收敛性。这样处理可以得到比以前的分析 [64] 更加精确的标度指数值。我们考虑的主要统计量是速度结构函数 $S_p(l) = \langle |\delta v_l|^p \rangle$ ，其中 $\delta v_l = v(t+l) - v(t)$ 是相距 l 的两点的速度差。 v 和 l 都取在平均流场方向。

在标度分析时， l 的取值范围应该选在 $S_p(l)$ 变化满足标度律 $S_p(l) \propto l^{\zeta_p}$ 的区域。然而，在我们分析的系统里，即使在最高的雷诺数下，也没有这种标度律的存在 [64] 。但是，存在一段扩展的自相似律 ESS[11] 成立的尺度范围，也即 $S_p(l)$ 对 $S_3(l)$ 成标度律的范围。图 3.3 是一个典型的 ESS 结果。另人吃惊的是即使只用 1/10 的数据，结果也有标度律成立。最高雷诺数 $R = 540000$ 的情况下， ESS 成立的尺度 l 的范围大约跨越 1.5 个量级。如果这些尺度是取在惯性区内的，那么三阶矩将也跨越 1.5 个量级。但是我们这里，三阶矩与尺度 l 之间根本没有标度律可言。 ESS 中尺度 l 的概念是不明显的。

当湍流处于统计定常态的时候，通常人们都用时间平均来代替系综平均，这么做的合理性是由各态历经假设所保证的。因此， $S_p(l)$ 可以通过时间积分来得到，

$$S_p(l) = \frac{1}{T}\int_0^T (\delta v_l(t))^p dt. \tag{3.1}$$

这种做法在计算高阶矩（例如 $p \geq 10$ ）的时候是有问题的，因为一个不和合理的大振幅事件就可以改变整个结果。如何区分一个假的大涨落事件和一个真的大涨落事件？我们认为大的涨落随时间仍然是平缓变化的，因此概率分布函数也应该是平缓变化的，即使是在 PDF 的尾巴上也是如此。我们通过如下方式除去不太合理的事件。首先作速度增量 δv_l 的直方图，归一化后变成概率分布 $P(\delta v_l)$ 。图 3.4 中给出了两个尺度上的典型的 PDF 。尾巴上的离散的点代表那些在上千万的样本中只出现几次的事件，主要是由于有限样本数效应引起的。那些只出现一两次的事件是不具有统计意义的。我们的噪声去除过程是去掉一些概率最低的事件：假设 p_{min} 是由于有限样本效应导致的最小离散概率密度， $p_{min} = c/N$ ，式中 N 是总样本数， c 是依赖于由于计算直方图引进的对 δv_l 的离散化的常数，那么我们的噪声去除过程就是将阈值 $5 \times p_{min}$ 以下的概率密度函数设为 0 。图 3.5 中给出了三种不同样本数下使用噪声去除程序的概率分布函数。

现在，结构函数通过直接对概率分布函数 PDF 积分来得到：

$$S_p(l) = \int_{-\infty}^{\infty} |\delta v_l|^p P(\delta v_l) d(\delta v_l). \tag{3.2}$$

图 3.6 中给出了 3 个不同的 p 在去除噪声前后 ESS 结果的比较。可以看出，去除噪声前，由于噪声的影响，各阶矩存在较大的涨落；而去除噪声后涨落变的小多了。这表明，光滑后的流体结构的统计系综表现出扩展的自相似结构，直到很高阶矩（ $p = 18$ ）依然如此。本节中所有结构函数的计算都是在去除噪声的基础上进行的。

§3.4 不同雷诺数下测得的参数

下面，我们考虑速度结构函数的标度指数 ζ_p 随统计样本数的收敛性问题。Anselmet 等人 [1] 曾经指出过检验这种收敛性的重要性。标度标度指数 ζ_p 的收敛性与矩的收敛性不同 – 标度指数可以比矩收敛的更快，因为它们刻划的是矩的对数的变化。关于标度指数的误差分析问题还没有理论的处理。对样本容量的依赖性是测量标度指数不确定性的可能的方法。图 3.7 给出了一个典型的结果，这里我们分别用 1/10 ， 1/3 和整个样本数的样本计算了相对标度指数。选取的数据是雷诺数为 $R = 540,000$ 。可以看出， ζ_p 随着样本数的增加而减小。 ζ_p 越小意味着对 $K41$ 理论值的偏离就越大，也就是间歇性越

显著。这是可以理解的，因为随着样本数的增加，更大振幅（通常认为更间歇）的涨落就被包括进来了。我们发现，对 $p \leq 10$ ， ζ_p 收敛很快 (图中的点重叠起来了)，但是对 $p > 10$ ，还需要更长的信号以保证指数的收敛性。

我们也对其他雷诺数 $1.2 \times 10^4 \leq R \leq 5.4 \times 10^5$ 的数据作了类似的分析。对每个雷诺数的数据，我们都先在尺度范围 $2^1\delta$ 到 $2^{10}\delta$（ δ 是数据点间的距离， $\delta = 0.17$ ）内计算速度差的概率分布函数。然后，去除噪声，计算矩。用 ESS 计算相对标度指数时，尺度范围取 $2^3\delta \leq \ell \leq 2^8\delta$ 。每个数据计算时都先用部分样本再用整个样本，以研究其收敛性。

图 3.8 给出四个不同雷诺数下的结构函数标度指数 $\zeta_p \leq 10$ 。雷诺数 $R < 10^5$ 时的 ζ_p 比雷诺数 $R > 10^5$ 时的 ζ_p 大。对于两个大雷诺数的情况，标度指数在统计不确定性误差范围内（比图中的点小）可以认为是相同的，表明标度指数随雷诺数的变化已经收敛。尽管在没有更大雷诺数的数据时这一点还不能被保证。图 3.8 中的大雷诺数下的 ζ_p 比 She 和 L*évê*que 最初预测的均匀各向同性的湍流 (如射流和尾迹湍流 [12]) 的标度指数 [91] 小。

图 3.8 中所有雷诺数下的标度指数都和推广的 She-L*évê*que 标度律 (2.30) 式符合的非常好。式中， β 和 γ 都取至下面的测量结果。

图 3.9 中给出了 Taylor-Couette 流四个雷诺数下的 β - 检验结果。

参数 β 度量湍流场的间歇性特征。我们的测量得到的 β (0.83) 值表明 Taylor-Couette 具有强间歇性的结构，既有混乱的一面又有结构性的一面。这意味着强弱涨落的结构相互关联，并且非常相似。 β 不依赖于雷诺数的特征表明流动的间歇性特征在我们所研究的雷诺数范围内不随着雷诺数的变化而变化。

四个雷诺数下 γ - 检验的结果在图 3.10 中给出。每个雷诺数下都有线性关系成立，表明数据通过 γ- 检验。注意，为了进行 γ 检验，数据必须先通过 β - 检验，得到 β 值。 γ 检验要同时用到上面测量得到的标度指数 ζ_p 值和 β 值。

γ 度量的是最高激发态（最奇异结构， H_∞ ）的奇异性。如果最高激发态是激波结构， γ 将等于 0（激波间断面两侧的速度差正比于 l^0 ，或者说与 l 无关）。这是非常强的奇异性。若奇异性的例子是形如布朗运动的速度廓线，那里 $\gamma = 1/2$ 。 γ 越大，最高激发态的奇异性越弱。 Kolmogorov 的耗散结构， $\gamma = 1/3$ ，因此比 Burgers 激波结构的奇异性小，间歇性也小。

我们测得， γ 值从 $R < 10^5$ 时候的 0.14 变到 $R > 10^5$ 时的 0.11 ，表明最高激发态结构在 $R \simeq 10^5$ 时经历了一个转变，从 $R = 10^5$ 左右开始变得更加奇异了。我们猜想这种转变对应于流场中 Taylor 涡结构的丧失。实验观测显示，在雷诺数增加到大约 10^5 的时候， Taylor 涡结构开始破裂，不再具有明显的结构 [64] 。新的结构更加无序，混乱，更加湍流化。因此更加奇异， γ 减小。表 3.3 中列出了我们在各种雷诺数下测得的 γ 值。

Reynolds number	R_λ	λ_T (cm)	β	γ
12,000	34	0.55	0.83	0.12
24,000	48	0.40	0.83	0.14
34,000	57	0.34	0.83	0.14
48,000	67	0.29	0.83	0.14
69,000	80	0.24	0.83	0.13
133,000	110	0.18	0.83	0.10
266,000	160	0.13	0.83	0.10
540,000	220	0.09	0.83	0.10

表 3.3: 不同雷诺数下测量得到的 β 和 γ 。 Taylor 微尺度 λ_T 和基于 Taylor 微尺度的雷诺数 R_λ 分别由下面两个公式获得： $\lambda_T = 47.0R^{-0.473}$ cm ， $R_\lambda = 0.324R^{0.495}$ 。 Taylor 微尺度 λ_T 比耗散尺度 η 大很多，后者在 $R = 12000$ 和 $R = 540000$ 时分别为 0.075 和 0.0057 。

本章小结：我们在层次结构的框架下分析了不同雷诺数下 Taylor-Couette 流的速度信号。结构函数不是通过直接黎曼求和，而是通过去除噪声后概率分布函数的方式得到。通过扩展的自相似律 ESS 的方法求的标度指数 ζ_p ，并且对 $p \leq 10$ 的情况，我们发现 ζ_p 在我们的样本数下已经收敛。

Taylor-Couette 流的速度数据可以由层次结构模型很好地描述，以很好的线性度通过 β- 检验和 γ- 检验。该模型存在一种不变性，这种不变性定义了一种变换群 [93] ，并且可以由对数 -Poisson 级串过程精确实现 [31, 92] 。

β- 检验得到的值与雷诺数无关，表明联系大小尺度（ ℓ ）涨落，不同强度（ p ）涨落的机制是普适的，与雷诺数无关。 β 值 0.83 比在自由射流模拟，和尾迹湍流中测得的 $\beta \approx 0.87$ 小。我们无法说明 Taylor-Couette 流与自由射流和尾迹湍流有何不同，但这一点肯定值得进一步的研究。

我们还发现，参数这里测得的 γ 值在 $R = 10^5$ 左右时从 0.14 变到 0.11 ，表明最高激发态结构从 $R = 10^5$ 左右开始变得更加奇异了。

关于此研究的进一步的实验应该采集更长的速度信号（比我们用的 2×10^7 个点更长）。更长的信号将包括更多的大振幅事件，使得概率分布的尾部更可信，以便可能去研究更大的 p（ $p > 10$ ）时， ζ_p 的收敛性。而且，应该多测不同空间点，特别是近壁区的速度信号。近壁区由于湍流偏离均匀各向同性， γ 值有可能发生很大变化。最后，应该利用 β- 检验和 γ- 检验研究和刻划其他几何条件下的间歇结构。

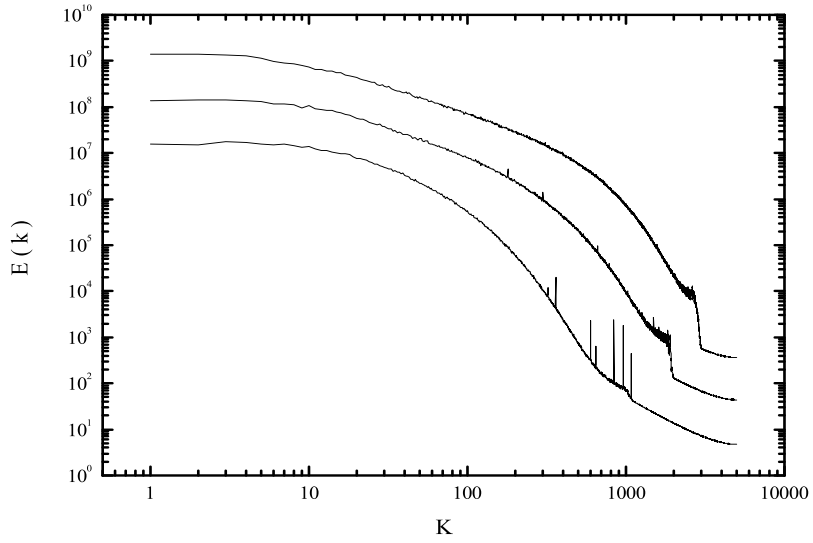

图 3.1: 三个不同雷诺数 $R = 12k$(下), $R = 266k$(中) 和 $R = 540k$(上) 下的流速度场的能谱。为个能够分辨清楚，我们已经将曲线在纵轴方向作了平移。

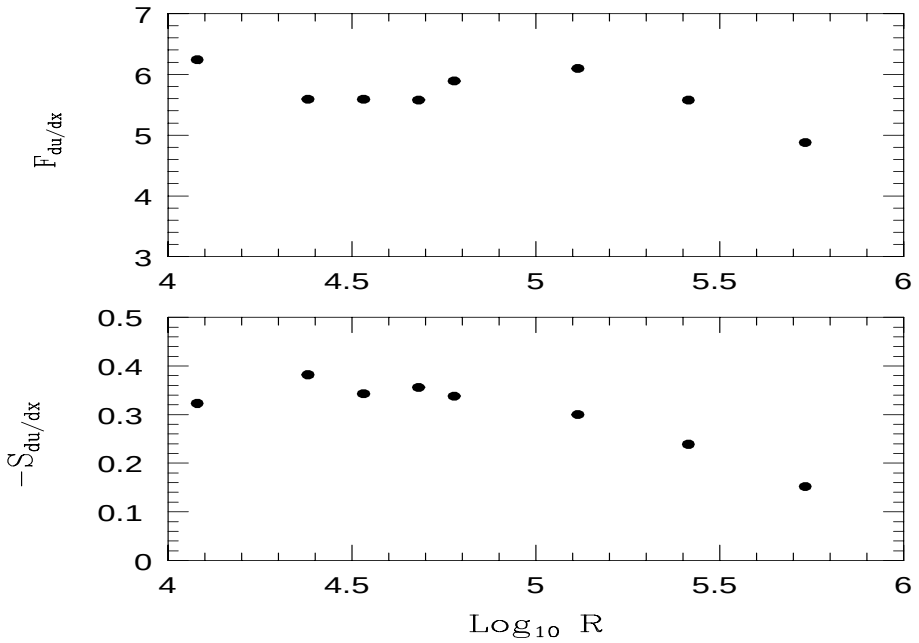

图 3.2: 速度导数的偏度 S_3 和峭度 F 随着雷诺数的变化关系。注意在 $R > 100,000$ 时，二者都有下降的趋势。

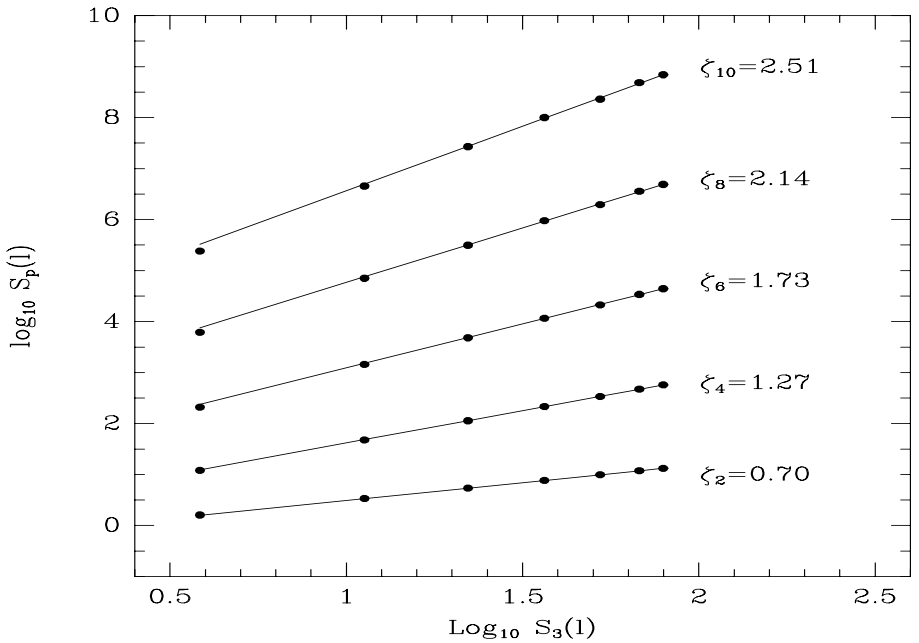

图 3.3: 扩展的自相似性 ESS：尺度范围跨越 1 个多量级（ $2^3\delta \le l \le 2^8\delta$ ， δ 是两个数据点之间的距离， 0.17 mm 。我们使用的数据量为 2×10^6 个数据点，只占整个数据量的 1/10 。雷诺数 $R = 540,000$ 。

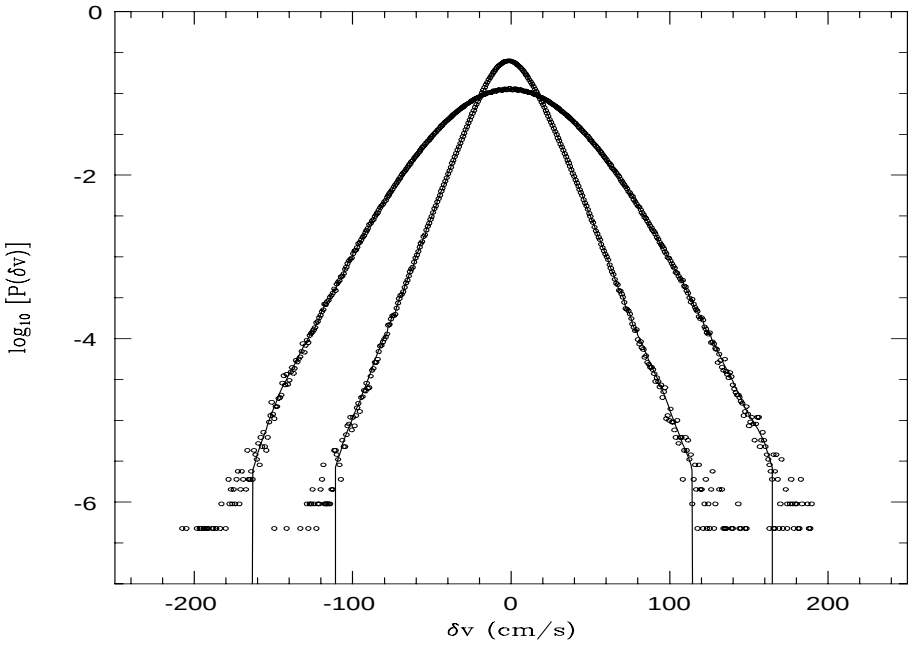

图 3.4: 尺度为 $2^3\delta$ 和 $2^8\delta$ 时的速度增量的概率分布函数 PDF，其中 δ 是两个数据点之间的距离，0.17 mm 。点和线分别是去除噪声前后的 PDF 。我们使用的数据量为 2×10^7 个数据点。雷诺数 $R = 540,000$ 。

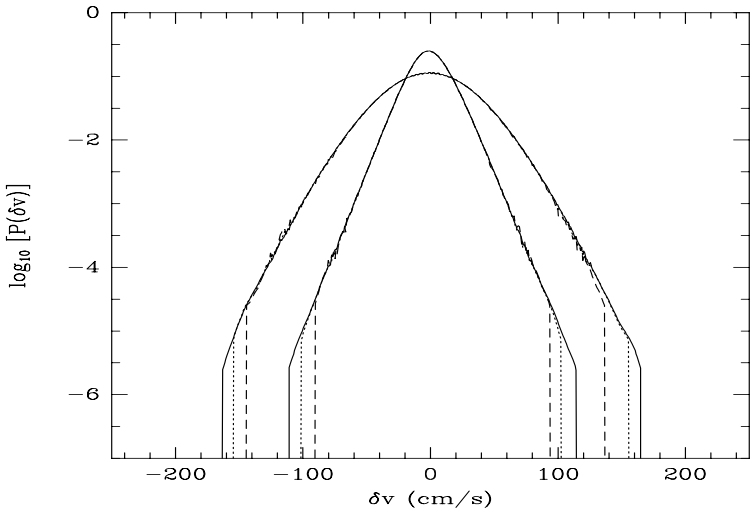

图 3.5: 三个不同样本量 2×10^6（虚线），6×10^6（点线），2×10^7（实线）下的概率分布函数 PDF（已经去除噪声）。可以看出，样本量逐渐增加时，PDF 的尾巴伸展的更长了。雷诺数 $R = 540,000$ 。

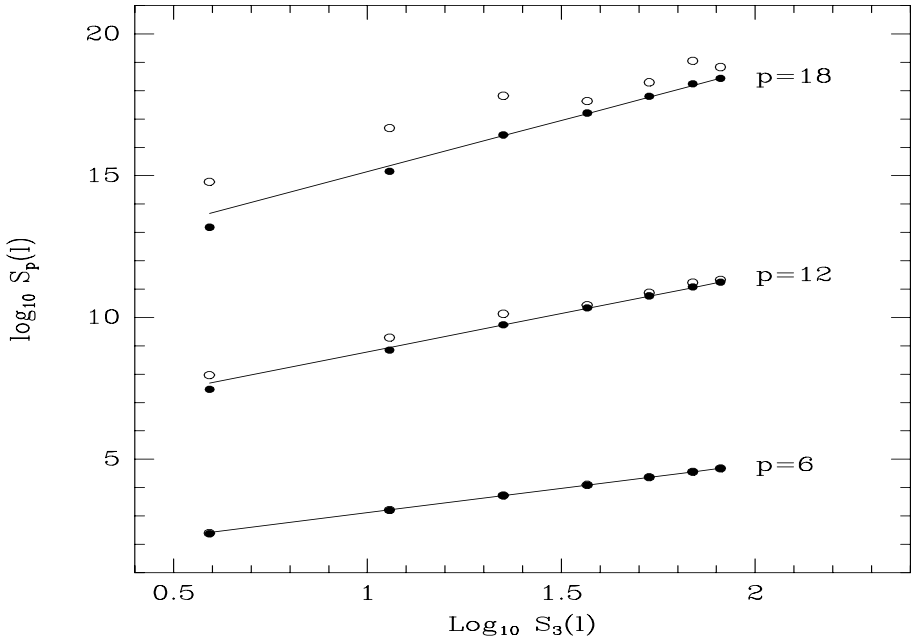

图 3.6: 使用去除噪声前后的 PDF 进行相对标度律 ESS 计算的比较。实心符号为去除噪声后的情况，空心符号为去除噪声前的情况。去除噪声后，矩相对于 $S_3(l)$ 的涨落变小了。线是求相对标度指数时所用的最小二乘法拟合的结果。我们使用的数据量为 2×10^7 个数据点。雷诺数 $R = 540,000$ 。

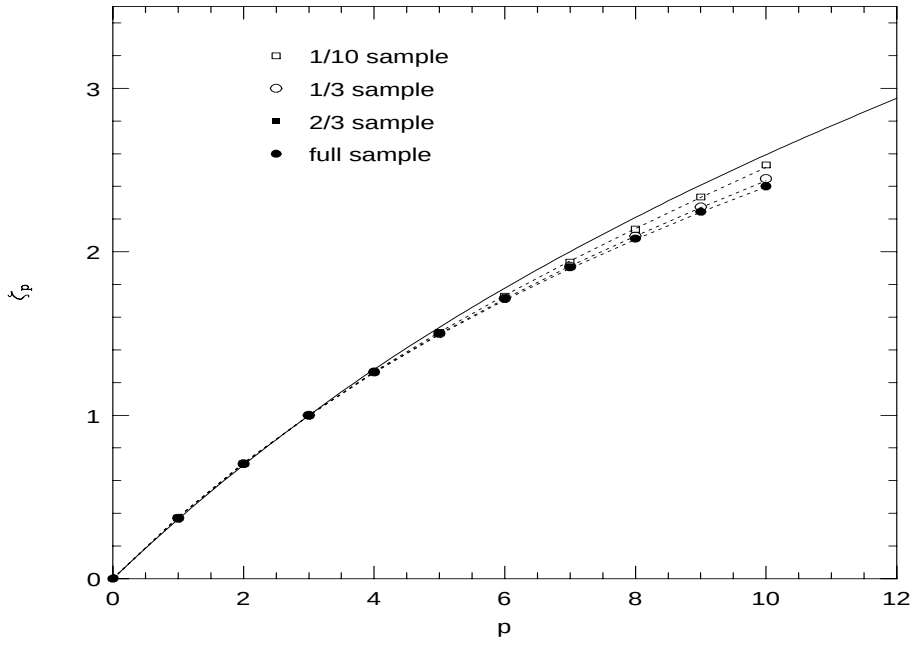

图 3.7: 雷诺数 $R = 540,000$, 样本量分别为 1/10 , 1/3 和整个样本 (2×10^7 个数据点) 时的结构函数标度指数 ζ_p 。虚线是用测量得到的 β 和 γ , 由 She-Lévêque 预测的标度指数。实线是用 $\beta = (2/3)^{1/3}$, $\gamma = 1/9$ 由 She-Lévêque 模型预测的充分发展湍流的标度指数。

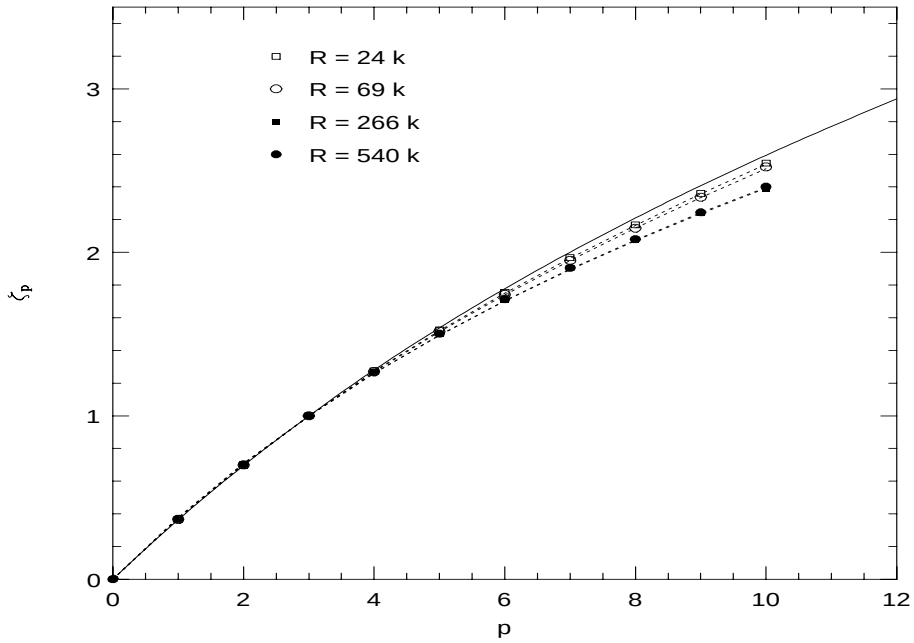

图 3.8: 雷诺数 R 分别为 2.4×10^4（空心方框），6.9×10^4（空心圆），2.66×10^5（实心方框）和 5.4×10^5（实心圆）时的结构函数的标度指数 ζ_p。后两个雷诺数下的 ζ_p 在图中的尺度下几乎不可区分了。虚线是用测量得到的 β 和 γ，由 She-Lévêque 预测的标度指数。实线是用 $\beta=(2/3)^{1/3}$，$\gamma=1/9$ 由 She-Lévêque 模型预测的射流，尾流等充分发展湍流的标度指数。

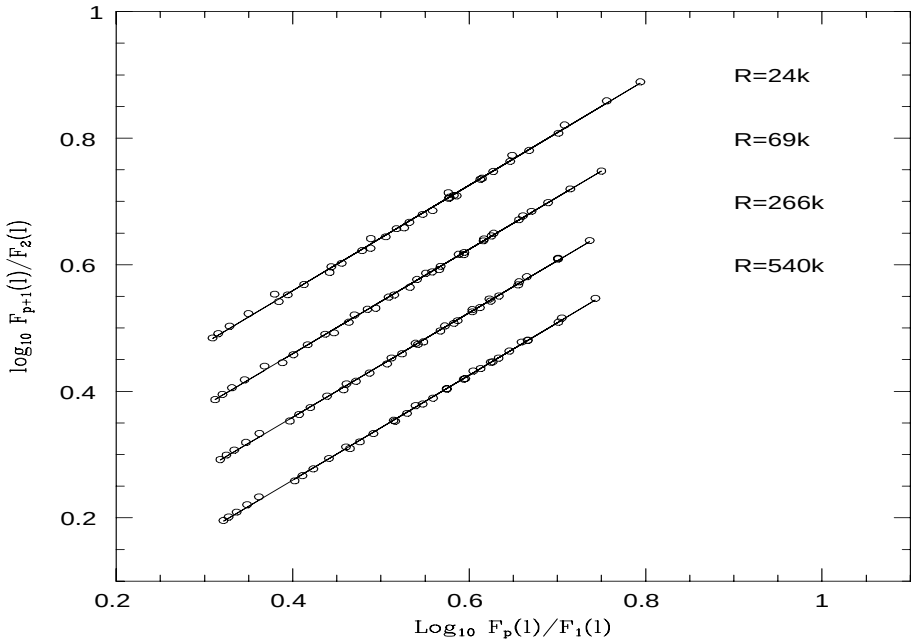

图 3.9: 四个雷诺数 R 下的 Taylor-Couette 流数据很好地满足 β- 检验：每个数据都有线性关系存在。而且 β 的数值，也就是线的斜率在四个雷诺数下都是相等的，0.83，比自由射流的 $\beta\approx0.87$ 稍小。为了看的方便，我们已经将不同雷诺数下的点在纵轴方向进行了平移。

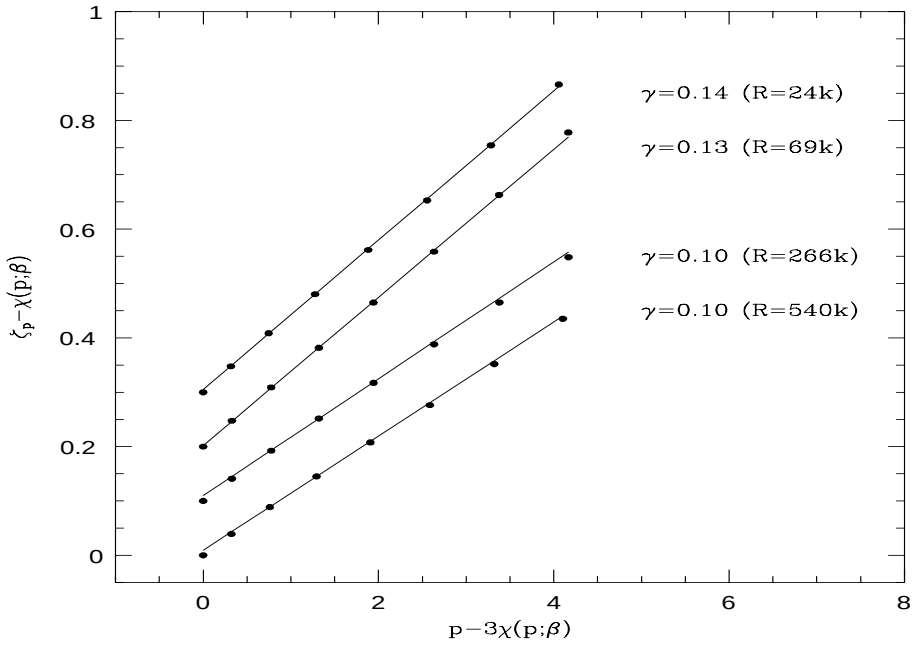

图 3.10: 这些点的线性关系表明 Taylor-Couette 数据通过了层次结构的 γ- 检验。注意对于 $R < 10^5$，$\gamma \approx 0.14$ 而对于 $R > 10^5$， $\gamma \approx 0.10$ 。为了看得清楚，我们已经将不同雷诺数下的点在纵轴方向进行了平移。

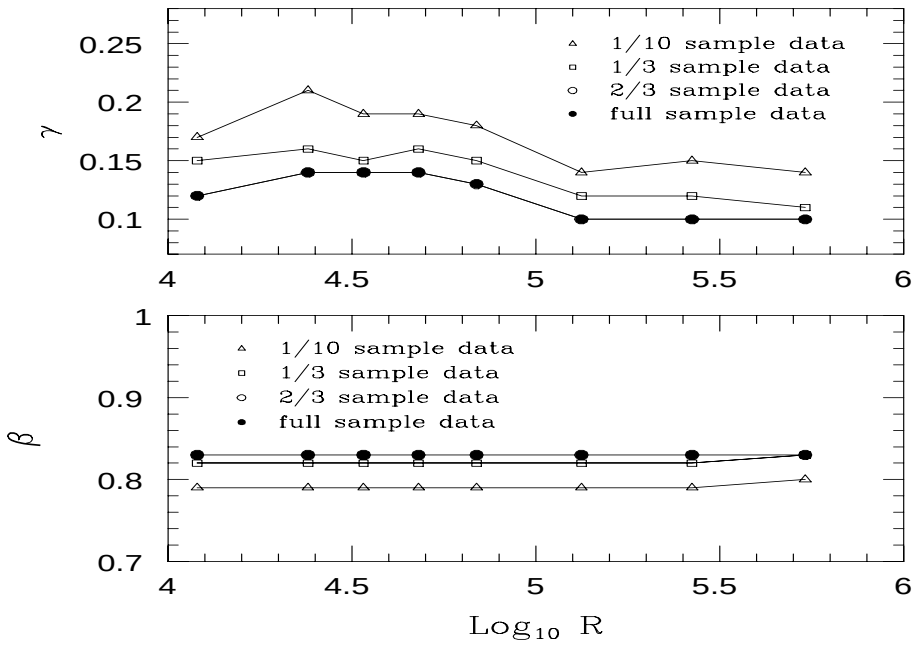

图 3.11: 四种不同样本数下测得的 β 和 γ 随着雷诺数的变化。样本数为 2/3 时的结果同全部数据的结果基本重叠，表明两种检验的收敛性很好。

第四章　时空广延系统中的层次结构

过去的二十年里，凝聚态物理学家们对于时空广延系统的动力学行为倾注了大量的研究 [33] 。这不仅因为这些系统存在许多新颖有趣的复杂行为，有重要的科学研究价值，而且因为对这些系统的研究具有重要的实际应用价值。

有两类系统被研究的十分广泛。一类是由 Kaneko 提出的所谓耦合映像格子(Coupled Map Lattices， CML ）系统 [53, 50, 104] ，另一类是由 Bak， Tang 和 Wiesenfeld 提出的自组织临界性（ Self-Organized Criticality， SOC ）系统 [4, 7, 100] 。本章中，作者将利用第二章所提出来的层次结构模型对这两类系统进行多尺度分析。我们的结果表明，这两类系统均可以由层次结构所描述，这种描述对我们认识系统的动力学行为和其中的标度律现象随系统尺寸大小的变化会有所帮助。

§4.1　耦合映象格子中的相似律

耦合映象格子最初是由（ couple map lattice， CML ）Kaneko[52] 引进的，后来被许多作者所发展 [50, 104] 。耦合映象格子由坐落在规则格点上的大量非线性映射构成，通常格点之间由某种局部的相互作用（如扩散）联系起来。因此系统是时间空间离散化的，但状态变量是连续的。我们来看一个具体的 2 维耦合映象格子的例子。考虑一个 2-维格子，格点坐标为 (i, j)(i=1, N; j=1, N) 。每个格点上定义一个状态变量 $u_{i,j}(n)$ ， n 是时间变量。状态变量的动力学由下式给出 [51] ：

$$x_{t+1}(i) = (1-\eta) f(x_t(i)) + \frac{\eta}{2d} \sum_{\mathbf{i}'} f(x_t(\mathbf{i}')), \tag{4.1}$$

，这里 **i'** 是 **i** 的 2D 个相邻格点。函数 $f(x)$ 是个局部映射，通常是诸如 logistic 映射，帐篷映射等的混沌映射。系统 (4.1) 的动力学行为非常丰富，在过去的一些年里受到过广泛的研究 [50] 。

我们将要研究的是一种由 (4.1) 式推广的系统，通常称为全局耦合映象 [54].

$$x_{t+1}(i) = (1-\eta) f(x_t(i)) + \frac{\eta}{N} \sum_{j=1}^{L} f(x_t(j)), \tag{4.2}$$

这个系统由作为 (4.1) 的平均场推广而提出的，受到了广泛的研究 [77] 。细心的读者马

上会发现，方程 (4.2) 和 (4.1) 的唯一差别在于最后一项的求和，(4.2) 式是对相邻格点，而 (4.1) 式是对全部格点。

现在，我们来考虑 (4.2) 中尺寸为 L 的系统的平均场的涨落：

$$h_{(L,t)} = (1/L)\sum_{i=1}^{L}(f(x_t(i))) \tag{4.3}$$

这个量类似于湍流场的能量耗散率 ε_l 的场。这里，我们分别选择局部映射 $f(x)$ 为 logistic 映射：

$$f(x) = 1 - ax^2 \tag{4.4}$$

和帐篷映射：

$$f(x) = 1 - a|x| \tag{4.5}$$

图 (4.1) 给出了两段尺度 $L = 1200$ 时典型的平均场涨落信号。

我们计算了不同尺度 $L = 1200$ 下的涨落 $h_{(L,t)}$ 的概率分布（ PDF ）。图 (4.3) 中显示的是 $L = 150$, $L = 500$ and $L = 2000$ 时 $h_{(L,t)}$ 的 PDF 。图中使用的是半对数坐标，横坐标是线性的，纵坐标是对数坐标。计算每个 PDF 使用的样本数都是 10^5 个。

下一步需要计算的是涨落 $h_{(L,t)}$ 的 p 阶矩：

$$\varepsilon_L^p = \langle h_{(L,t)}^p \rangle = \int h_{(L,t)}^p P(h_{(L,t)}) dh_{(L,t)} \tag{4.6}$$

式中 $P(h_{(L,t)})$ 代表 $h_{(L,t)}$ 的 PDF ， $\langle .. \rangle$ 代表时间平均。我们曾在双对数坐标上检查过下面的标度关系：

$$\varepsilon_L^p \sim L^{\tau_p} \tag{4.7}$$

并没有明显的标度区存在。也就是说没有绝对标度律存在，结果在图 4.4 中给出。相反，当我们考虑以下形式的相对标度律 ESS 的时候，

$$\varepsilon_L^p \sim (\varepsilon_L^3)^{\tau_{(p,3)}}, \tag{4.8}$$

我们观察到了标度区的存在。图 4.5 给出了 ESS 结果，可以看出 ε_L^p 对 ε_L^3 在双对数坐标上的线性关系，表明存在很好的标度律。我们还计算了不同系统参数下的相对标度指数，结果在表 4.1 中统一给出。

很明显，测得的这些指数并不是普适的，随着系统的参数变化而变化。而我们知道这些参数同系统的混沌性质密切相关 [77] 。

local map	parameter	$\tau_{1,3}$	$\tau_{2,3}$	$\tau_{4,3}$	$\tau_{5,3}$	$\tau_{6,3}$	$\tau_{7,3}$	$\tau_{8,3}$	β	γ
$f(x)=1-ax^2$	$a=1.80,\eta=0.05$	0	0.29	1.97	3.14	4.48	5.99	7.64	0.83	2.23
	$a=1.80,\eta=0.1$	0	0.50	1.66	2.44	3.36	4.40	5.53	0.92	2.28
	$a=1.99,\eta=0.1$	0	0.37	1.81	2.75	3.82	4.98	6.24	0.79	1.54
$f(x)=1-a\|x\|$	$a=1.80,\eta=0.05$	0	0.33	1.94	3.11	4.49	6.05	7.78	0.88	2.87
	$a=1.80,\eta=0.1$	0	0.32	1.98	3.20	4.66	6.33	8.19	0.89	3.43
	$a=1.99,\eta=0.1$	0	0.33	1.94	3.10	4.46	5.97	7.64	0.86	2.53

表 4.1：不同参数下耦合映像格子模型测量得到的相对标度指数及 β 和 γ 的值。相对标度指数是利用 ESS 性质求得的。

又一次，我们可以检验层次相似律的存在性，并测量层次相似性参数。让我们首先定义：

$$H_p(L)=\frac{S_{p+1}(L)}{S_p(L)} \tag{4.9}$$

于是，在双对数坐标上画出 $H_p(L)/H_1(L)$（横坐标）和 $H_{p+1}(L)/H_2(L)$（纵坐标），就构成一个 β- 检验过程。图 4.6 给出了 β- 检验的结果。我们选取的局部映射是帐篷映射。内图给出的是局部映射为 logistic 映射时的 β- 检验。测得的 β 分别是 0.83 和 0.88，稍有不同。我们，还测量了其他参数下的涨落 $h_{(L,t)}$ 的 β- 检验，结果罗列在表 4.1 中。

我们可以利用 (2.25) 式进行所谓的 γ- 检验。注意，由于这里绝对的标度律不成立，所以这里测得的 γ 将是相对于 3 阶矩的值。图 4.7 中观察到的线性关系表明 γ 被很好地满足。

我们在不同的参数下研究了 GCM 模型的平均场涨落的标度行为，发现虽然没有绝对的标度律，但相对标度律 ESS 却很好地存在。相对标度指数 $\tau_{p,3}$ 不是普适的，倚赖于模型参数 a 和 η。这些参数与系统的混沌性质有关。因此相对标度指数也与系统的混沌性质有关。平均场涨落可以被层次结构模型很好地描述，以很好的线性度通过 β- 检验和 γ- 检验。参数 β 和 γ 的值也与系统的混沌性质相联系

§4.2 沙堆模型的多尺度分析

自组织临界性（SOC）的概念最初是由 Bak，Tang 和 Wiesenfeld 提出的，用来描述那些出现在缓慢驱动的耗散系统里的时空相关性现象 [4, 5]。尽管后来又有许多新

的模型被提出， Bak 等人的沙堆一直被人们看作是关于 SOC 概念的范式，通常被称为 BTW 模型。

尽管表面上看起来似乎特别简单， 2 维 BTW 模型的标度律问题至今还没有很好地弄清楚。特别是诸如刻划“沙崩”分布的标度指数等还没有精确的结果。人们作了许多数值实验，但并没提供多少有用的信息。关于 SOC 概念的提出和发展，有兴趣的读者可以参阅文献 [100] 。

最近， De Menech 等人将矩分析引进来分析 BTW 模型 [29] ，后来别的作者又将其用于分析其他的 SOC 模型，该方法在 SOC 模型分类的研究中显示了一定有效性的。本节，我们将在层次结构模型的框架下重新考虑 BTW 模型的矩分析，我们发现，不同尺寸的系统内“沙崩”大小涨落可以由层次结构很好地描述。

§4.2.1 模型介绍

这里，我们以 2 维为例子，介绍 BTW 模型的基本思路 [5] 。考虑一个尺寸为 L 的 2 维的正方形格子，每个格点上有一个非负整型变量 $Z(i,j)$ ，它代表某些物理量，如局部能量，应力，沙柱的高度等。如果某个格点上的 $Z(i,j) > Z_c$ ，则称该格点为不稳定的，这里 Z_c 是某一临界值，通常取 $Z_c = 2D$ ， D 是模型的维数。一个不稳定的格点将发生弛豫，其值将减去 Z_c ，周围 2D 个临近点每个点上的值加 1 。数学上，我们可以通过如下的规则扰动格子系统：

$$z(i-1,j) \to z(i-1,j) - 1 \tag{4.10}$$

$$z(i-1,j) \to z(i-1,j) - 1 \tag{4.11}$$

$$z(i,j) \to z(i,j) + 2 \tag{4.12}$$

如果 $Z(i,j) > Z_c$ ，那么， $Z(i,j)$ 将以以下方式弛豫：

$$z(i,j) \to z(i,j) - 4 \tag{4.13}$$

$$z(i,j \pm 1) \to z(i,j \pm 1) + 1 \tag{4.14}$$

$$z(i \pm 1,j) \to z(i \pm 1,j) + 1 \tag{4.15}$$

通过这种方式，临近的格点都被扰动起来，各种大小的“沙崩”事件都可能发生。这些“沙崩”事件具有很多有趣的性质，如大小“沙崩”发生的频率分布满足标度律关系。在临界稳定态（最初称为最小稳定态），相应的畴域大小的概率分布应该满足标度律。

图 (4.8) 和 (4.9) 中分别给出了 2D 和 3D 情况下沙堆模型模拟的典型的最小稳定态构型。

在自组织临界性的沙堆模型中，系统的输入是恒定的常数（加一粒沙子），而输出是大小 - 频率遵从标度律分布的一系列“沙崩”事件。

§4.2.2 矩分析结果

本节我们将利用 Chessa 等人 [24] 引进的矩分析方法来刻划的一些基本标度属性。然而，我们的重点将放在层次结构分析上，而不是仅仅求几个标度指数。

我们对系统 4.10，4.13 进行了 2D 数值模拟。模拟取格点尺寸 32×32 到 1024×1024。在每个尺寸的模拟中，我们都收集了 10^5 个样本点。图 4.10 给出了系统尺寸为 32×32，64×64 和 128×128 时的畴域大小分布。可以看出，这些分布具有明显的标度律。我们这里测得的标度指数大约是 -1.19，与 Bak 等人的最初结果 -1.0 有微弱差别，但是与 Manna[70] 后来进行的大规模数值模拟的结果 -1.22 非常接近。

这里，畴域大小的概率分布 $P(s)$ 的 q 阶矩定义为：

$$\langle s^q\rangle = \int s^q P(s)ds. \tag{4.16}$$

我们希望下面的标度关系成立，

$$\langle s^q\rangle \sim L^{\sigma_q} \tag{4.17}$$

然而，我们这里的数值模拟并没有发现很好的绝对标度律，这与最近的研究结果不同。当然，一旦考虑相对标度律情况就变了。图 4.11 中我们给出了 ESS 的结果。可以看出线性度还是挺好的。我们选取的参照是 $p=0.3$ 的情景。因此，得到的标度指数也是相对于 $p=0.3$ 的指数而言的。

图 4.12 中画出了矩分析得到的标度指数随 p 的变化。因为我们使用的是相对标度律，所以我们得到的标度指数与其他作者们给出的具体值差别较大 [68]。但是曲线的总体形状是类似的。在 p 比较小的时候（$p<1$），曲线的非线性非常明显；当 p 很大时，ζ_p 随 p 基本上是线性变化的，其变化率，也就是曲线的斜率，在 p 很大时大约 6.3。在层次结构的框架下，这个斜率就是最高激发态的标度指数 γ。后面的分析将会证明这点。

β - 检验和 γ - 检验分别在图 4.13 和 4.14 中给出。

我们发现，在进行 γ- 检验时，最小二乘法可以得到斜率 $\gamma=6.35$，与标度指数 ζ_p 曲线在 p 很大时的渐进斜率非常接近（理论上 γ 就是 p 趋向于无穷大时的渐进斜率），

这说明我们这里的层次结构分析， β- 检验， γ- 检验的整个过程是完全自恰的。

我们重新考虑了二维 BTW 模型的矩分析，发现，这里的一些物理量如畴域大小涨落可以由层次结构很好地描述。与最近的研究结果不同，我们这里并没有发现很好的绝对标度律，而只有相对标度律。相对标度指数 ζ_p 在 p 小于 1.2 时是 p 的非线性函数， p 较大时近似为线性函数。这表明这里的统计是非高斯的，而不是高斯的。我们的研究显示，对于沙堆模型，其系统性质随着系统尺寸大小的变化不是简单的外推关系，而是与 Log-Poisson 乘积过程类似的变化过程。

自从 SOC 概念提出以来，人们设计了许多不同的模型。这些模型通常都具有类似的但却不完全相同的标度行为。那么人们很自然会提出这样的问题，就是这些不同的模型是否属于同一普适类，也就是这些模型的标度行为究竟是真的不同，还是实际上相同的？许多人尝试过这方面的研究 [81, 101, 25, 10, 24, 68] ，但是通常得出相互矛盾的结果。例如，文献 [81, 101] 中作者通过重整化方法预言 SOC 的 BTW 模型和 Manna 模型 (随机沙堆模型)[71] 属于同一普适类，而文献 [10, 68] 却通过数值模拟认为二者属于不同的普适类。我们认为，在以上的文献中只是通过简单的标度指数来讨论普适类问题似乎有点牵强，是远不够得出正确结论的。而这里所讨论的层次结构模型可能是探讨这一问题的合适方法之一。层次结构的参数具有明显的物理意义，而且这些参数对标度指数的变化并不敏感。层次结构所描述的是系统的整体结构。

需要指出的是，在本章中我们所讨论的层次结构模型与前面一章所讨论的概念有所不同。层次结构模型最初的提出是为了描述湍流场中的间歇性现象，实际上的应用对象是多分形场。而这里我们所讨论的是系统的行为随着系统尺寸变化的关系。这个关系并不是是简单的线性关系或者单标度关系，而是某种意义上的多标度关系。也就是说，当系统尺寸变化时，系统里各种不同强度的涨落的变化是不同的，但都满足标度律关系，可以由不同的标度指数刻划。这组标度指数与涨落强度（体现在 p 上）之间是一种非线性的关系。于是整个系统行为表现出类似多标度的性质。系统行为随着系统尺寸变化而变化的过程是个类似于 log-Poisson 级串过程的过程。所以该过程可以由层次结构所描述。表现为整个层次结构的两个检验 β- 检验和 γ- 检验被很好地满足。

时空广延系统（特别是耦合映像格子和自组织临界性的沙堆模型）里层次相似律的成立绝对不是一件平凡的事。理解时空广延系统系统的动力学行为一直是科学和工程领域内，甚至社会科学领域的科学家们所关注的问题。不仅因为这些系统存在许多新颖有

趣的复杂行为，有重要的科学研究价值，而且因为对这些系统的研究具有重要的实际应用价值。这些模型大多数可以看作是实际系统的抽象模型。例如，许多研究者 [18] 都曾经考虑过将类似 (4.2) 的 CML 看作是湍流的简单动力系统模型。 Bak 等人也曾试图用封闭边界条件的 SOC 模型作为湍流的玩具模型 [6] 。因此，对这类系统的研究将有助于人们认识真实的复杂系统。对于此类系统的层次结构研究，将增加我们对层次结构的物理含义本身的进一步理解。

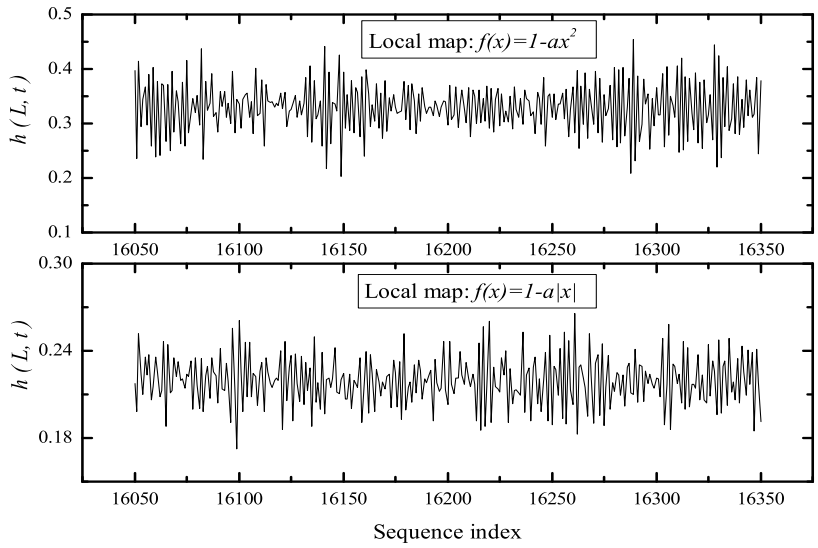

图 4.1: 尺度 $L = 1200$ 时的平均场涨落 $h_{(L,t)}$ 的信号。（ a ）局部映射是 $f(x) = 1 - ax^2$ ，参数分别为： $a = 1.80$ ， $\eta = 0.10$ ；（ b ）局部映射是 $f(x) = 1 - a|x|$ ，参数与（ a ）相同。

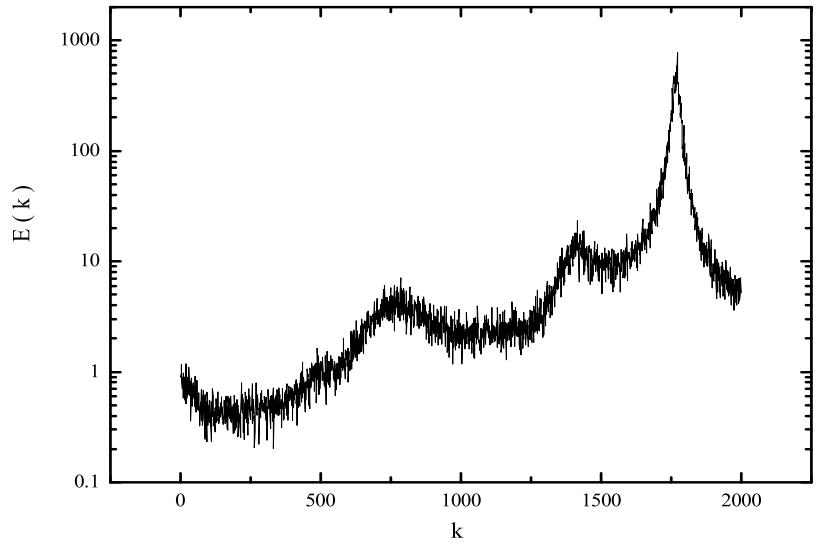

图 4.2: 平均场涨落 $h_{(L,t)}$ 的功率谱。尺度为 $L = 1200$, 局部映射为 $f(x) = 1-a|x|$, 参数是 $a = 1.80$, $\eta = 0.10$ 。注意纵坐标取的是对数尺度。

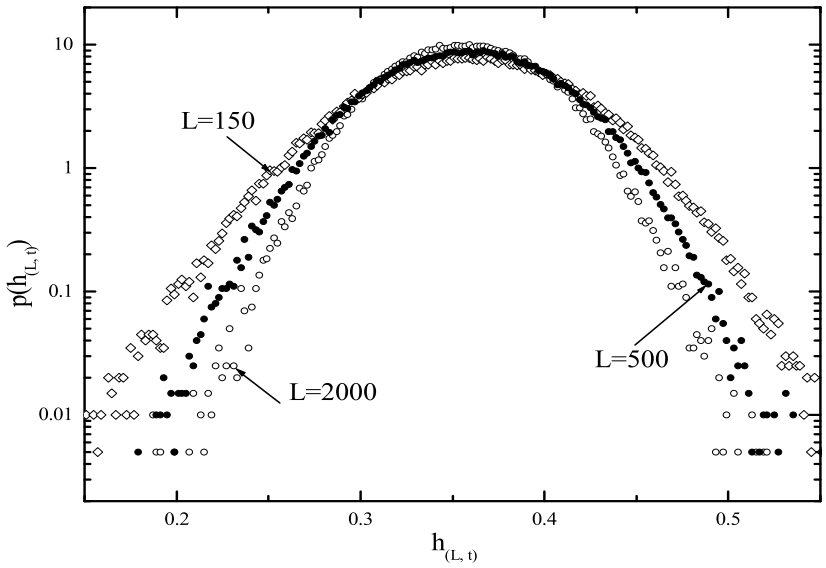

图 4.3: 尺度分别为 $L = 150$， $L = 500$ 和 $L = 2000$ 的涨落 $h_{(L,t)}$ 的概率分布函数。局部映射是 $f(x) = 1 - a|x|$，参数为 $a = 1.80$，$\eta = 0.10$。

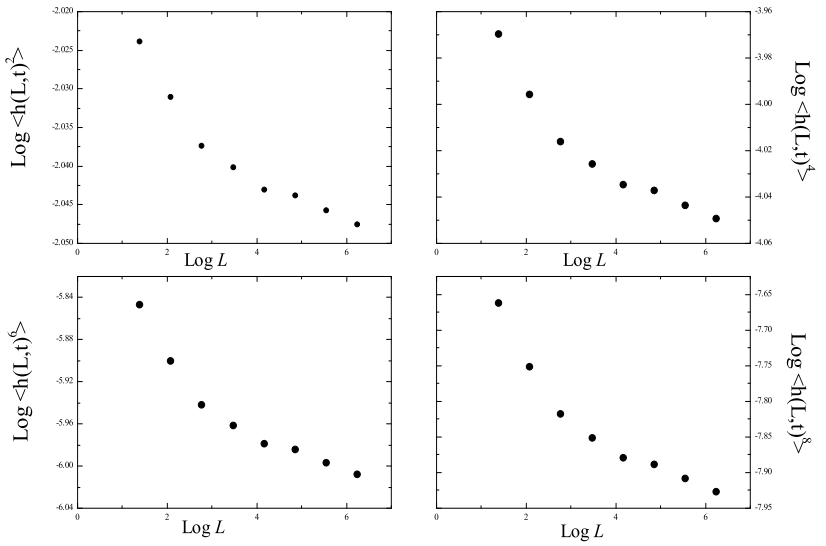

图 4.4: 平均场涨落 $h_{(L,t)}$ 的绝对标度律检验。局部映射为 $f(x) = 1 - ax^2$，参数 $a = 1.80$，$\eta = 0.10$。图中给出的分别是 $p = 2$， $p = 4$， $p = 6$，和 $p = 8$ 的情况。

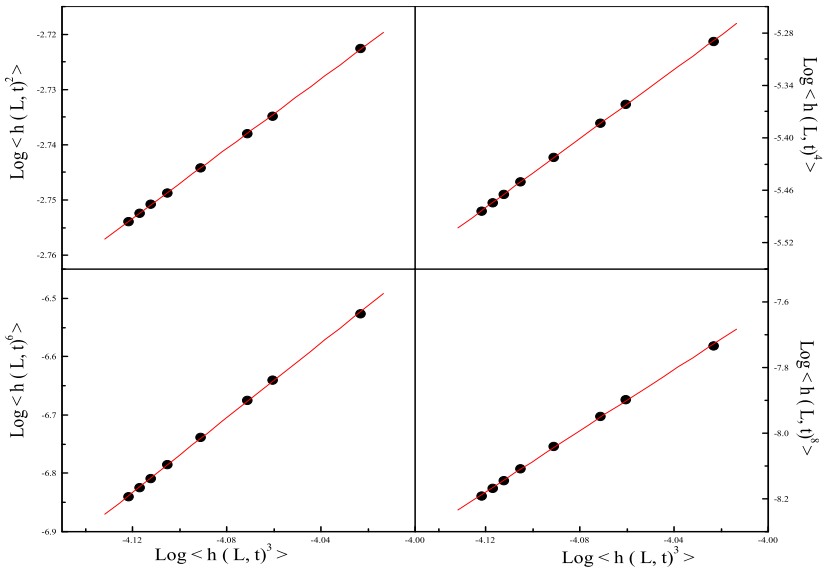

图 4.5: 相对标度律 ESS 检验。选取的参考矩是 3 阶矩。四个子图分别是 $p = 2$， $p = 4$， $p = 6$，和 $p = 8$ 的情况，局部映射是 $f(x) = 1 - a|x|$ 。参数分别为： $a = 1.80$， $\eta = 0.10$ 。

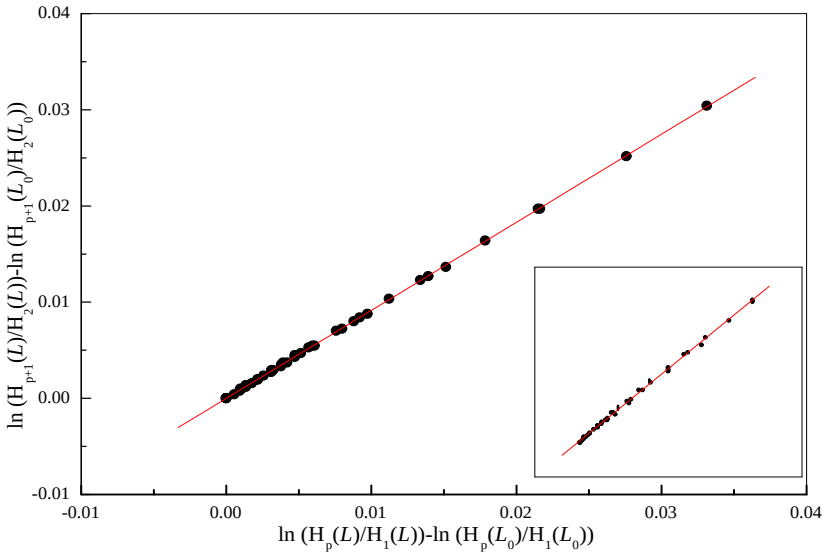

图 4.6: 平均场涨落 $h_{(L,t)}$ 的 β- 检验。局部映射是 $f(x) = 1 - a|x|$， 参数是 $a = 1.80$， $\eta = 0.10$ 。内图：局部映射为 $f(x) = 1 - ax^2$ 时的 β- 检验，参数是 $a = 1.80$， $\eta = 0.10$ 。

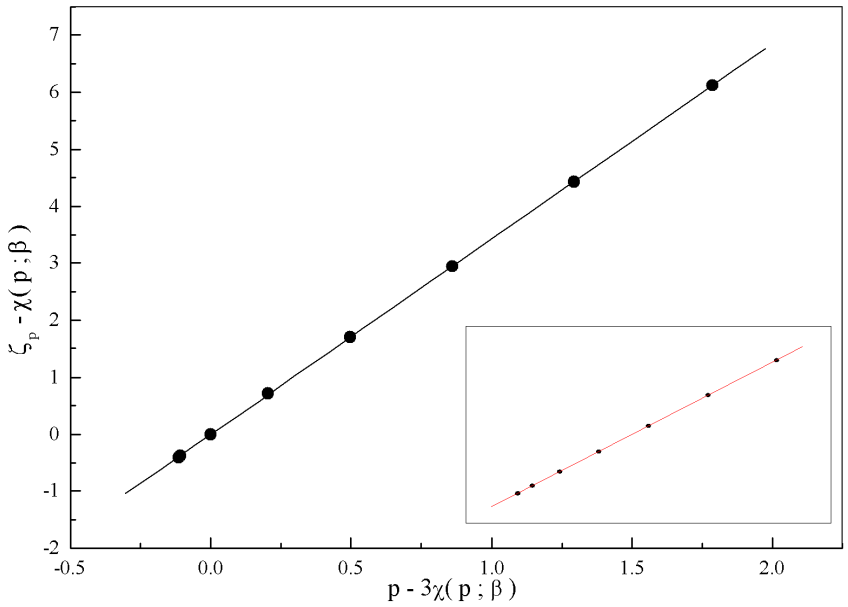

图 4.7: 平均场涨落 $h_{(L,t)}$ 的 γ- 检验。局部映射是 $f(x) = 1 - a|x|$ ，参数是 $a = 1.80$ ， $\eta = 0.10$ 。内图：局部映射为 $f(x) = 1 - ax^2$ 时的 γ- 检验，参数是 $a = 1.80$ ， $\eta = 0.10$ ，与 β- 检验参数时相同。

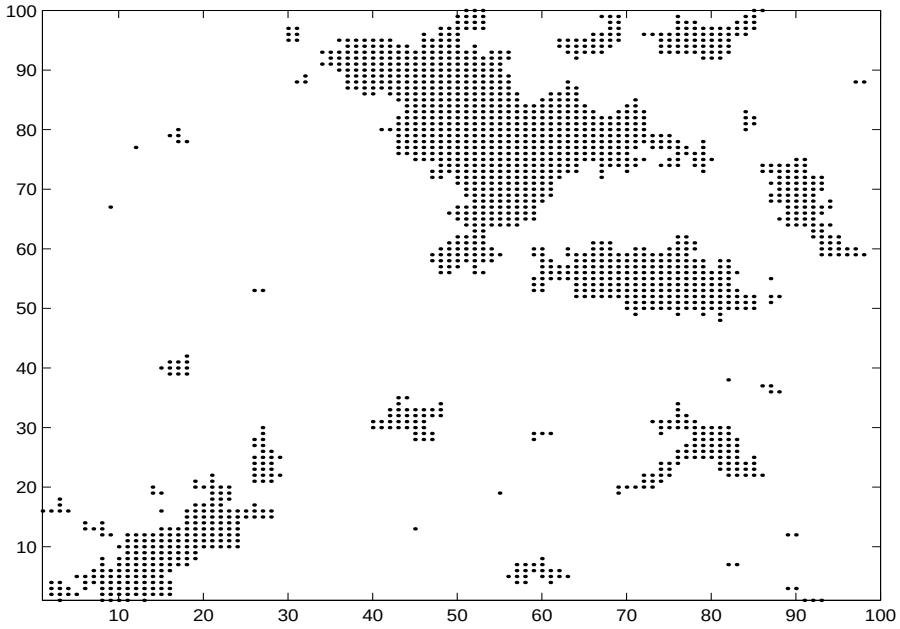

图 4.8: 典型的二维 BTW 模型的最小稳定构型。模拟时系统尺寸大小为 100×100 。扰动不同的点时，系统发生的“沙崩”范围的大小不同，但基本上“沙崩”发生的频率与其大小成标度律关系。

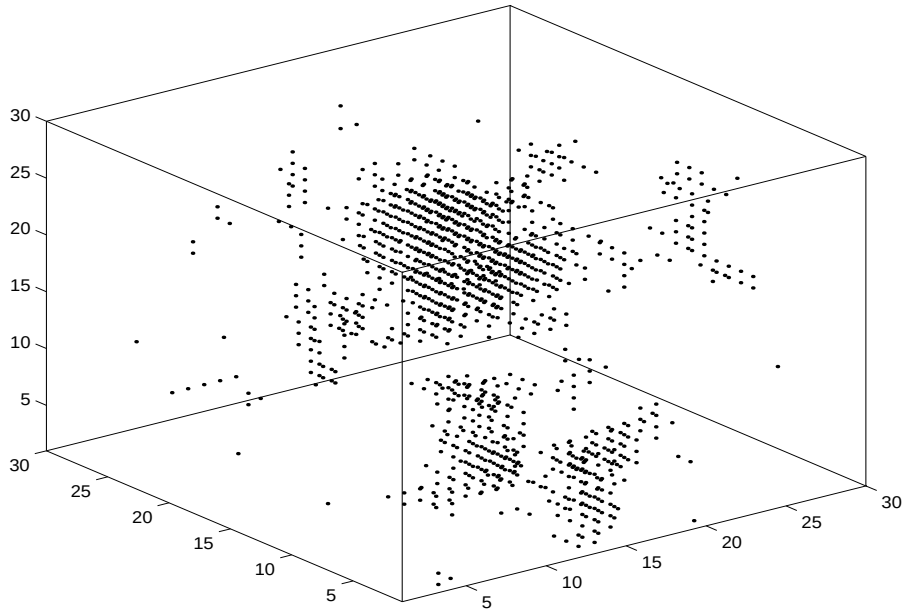

图 4.9: 典型的三维 BTW 模型的最小稳定构型。模拟时系统尺寸大小为 $30 \times 30 \times 30$

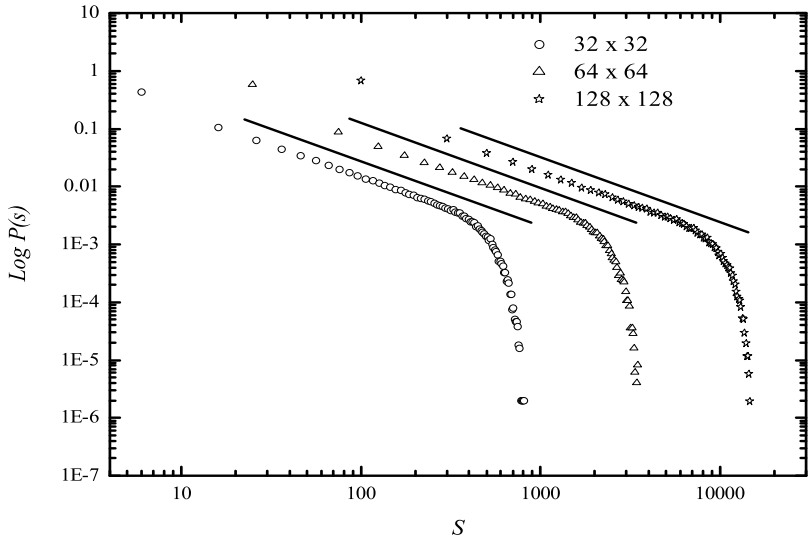

图 4.10: 二维 BTW 沙堆模型的畴域大小 s 的概率分布函数 $P(s)$ 。模拟时取临界值 $Z_c = 4$ ，系统大小分别为 32×32 ， 64×64 和 128×128 。为了表明 $P(s)$ 对 s 的标度律关系，我们取了双对数坐标，并且将曲线在纵轴方向进行了平移。

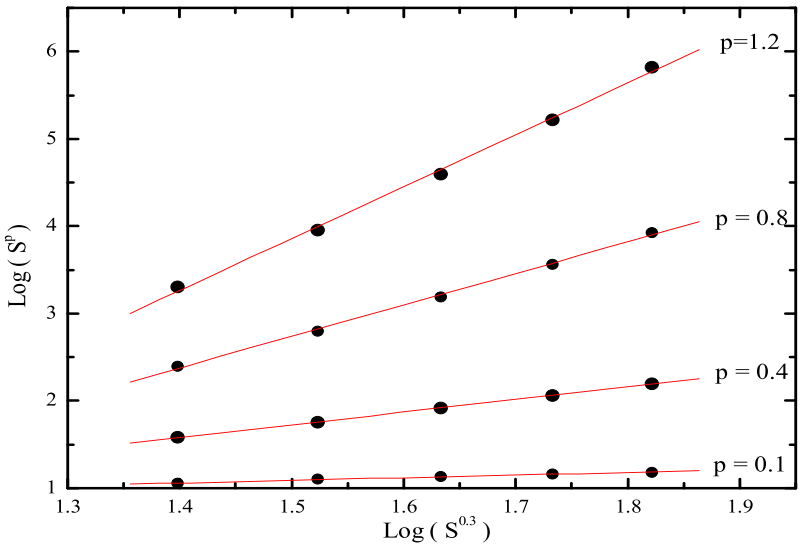

图 4.11: 二维 BTW 沙堆模型畴域尺寸大小分布的相对标度律 ESS 检验。我们选取的参考矩为 0.3 阶矩。图中的四个矩分别为 $p = 0.1$， $p = 0.4$， $p = 0.8$ 和 $p = 1.2$，直线是最小二乘拟合的结果。

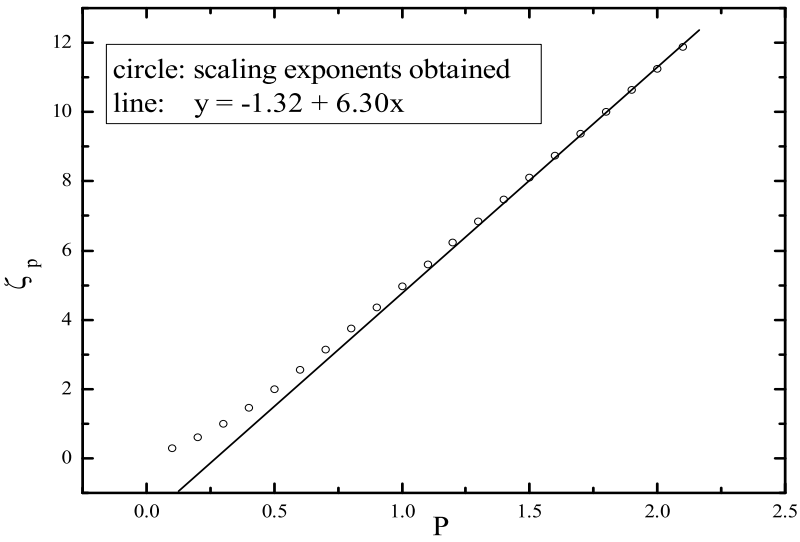

图 4.12: 二维 BTW 沙堆模型畴域尺寸大小分布相对标度律检验的相对标度指数。图中实线方程为：$y = -1.32 + 6.30x$。可以看出线的斜率 6.30 与图中 γ- 检验的 γ 值 6.35 非常接近，与理论估计完全一致。

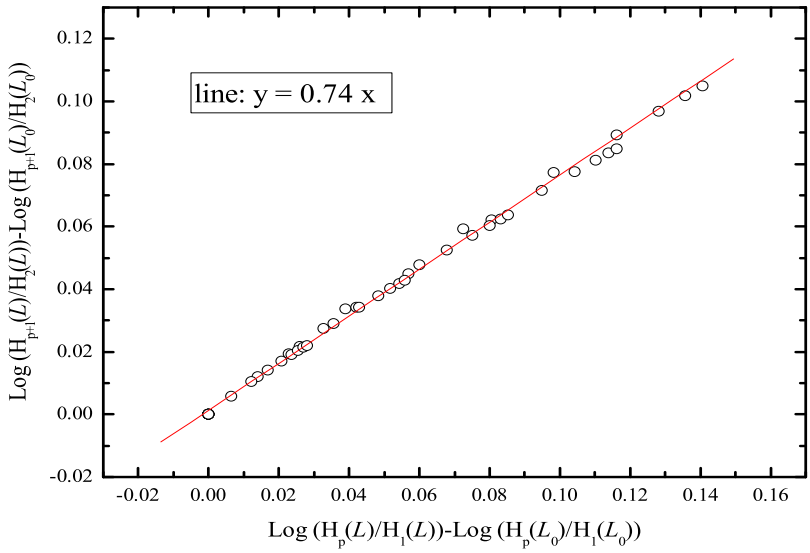

图 4.13: 二维 BTW 沙堆模型畴域尺寸大小分布的 β- 检验。图中实线方程为： $y = 0.74x$ 。

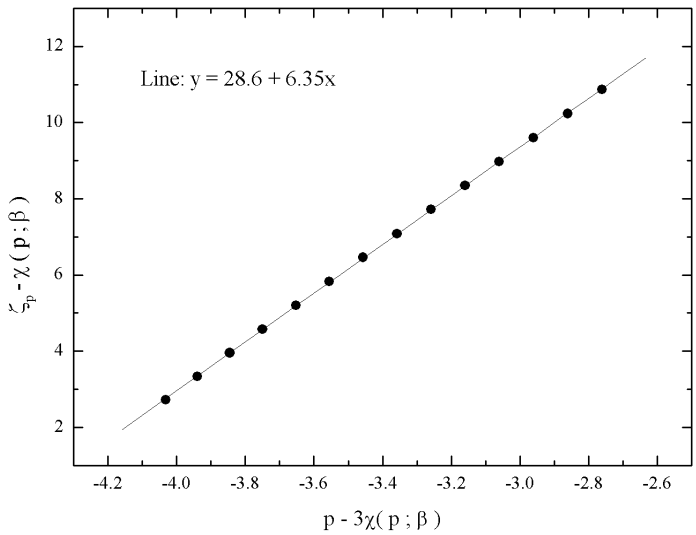

图 4.14: 二维 BTW 沙堆模型畴域尺寸大小分布的 γ- 检验。图中实线方程为： $y = 28.6 + 6.35x$ 。

第五章　DNA 序列的多尺度分析

生命系统作为最显著的复杂系统之一，曾经受到过广泛而深刻的研究。在许多情况下，生物系统是由大量的观测事实所得出的经验关系所描述的。人们花费许多精力去研究这些经验关系背后的机制，取得了不少成绩。其中一个典型的例子就是遗传物质， DNA 和它的双螺旋结构的发现。尽管今天的生命科学涉及到诸如物理学，化学和许多别的学科，越来越多的迹象表明更多的数学的，定量的分析对研究，理解生命系统复杂性来说是必不可少的。在这一章中，我们将讨论如何描述 DNA 序列 – 遗传信息的携带者 – 中隐藏的复杂信息问题。

§5.1　DNA 序列的复杂性

过去的生物学研究认为，生物体的遗传信息通常储存在它的脱氧核糖核酸（ deoxyribonecleic acid ， DNA ）序列中（除了在某些病毒中， RNA 取代了 DNA ）。在繁衍过程中，生物体并不复制它本身，而是给其下一代提供能够构造一个新生物体的携带遗传信息的遗传物质，这种遗传物质就是 DNA 。 DNA 是由 4 种不同的碱基分子构成的线性高分子长链。这 4 种碱基分子分别是：腺嘌呤 A, 胞嘧啶 C, 鸟嘌呤 G 和胸腺嘧啶 T 。它们由糖 - 磷酸盐的骨架连接起来。 A 和 G 是嘌呤， C 和 T 是嘧啶。这两种分子的主要区别在于它们不同的化学结构 [62].

DNA 的空间结构通常认为是双螺旋，由两条互补的单链组成：一条链上的 A 总是与另一条链上的 T 相对应， C 与 G 相对应。因此两条链可以相互转换，通常研究一条链就足够了。碱基分子在 DNA 序列上的排列顺序决定了序列的功能结构。

随着人类基因组测序工作接近尾声和大量模式生物全基因组 DNA 序列的获得 [32, 40] ，从新世纪初开始，人类已经有可能对包含生物全部遗传信息的基因组 DNA 序列进行全面研究，从这种仅表现出四个碱基一维排列的完整生命信息数据中解读出最基本的生物学规律，揭开生命体形成、发育、疾病、衰亡等过程的全部奥秘。各种生物全基因组 DNA 链不仅包含着制造生物体全部蛋白质的信息（即基因），而且包含有按照特定时空模式把这些蛋白质装配成为生物体的四维调控信息。如何找到这些信息的编码方式、调节规律，这是分子生物学的发展给科学带来的新课题。对 DNA 序列的分析将有助于我

们揭示许多新的生物学的规律。

通常，人们研究 DNA 序列的方法是统计分析。长（短）程相关 [65, 80, 102]，信息和熵的含量 [56, 45, 26]，序列复杂性 [48, 15] 以及语言学特征 [35, 72] 等都是有趣的研究对象。随机行走模型 [106]，标度指数方法 [67]，分数阶矩分析 [74]，小波分析 [2] 等各自在刻划 DNA 某些特征方面都取得了一定的成功。 DNA 序列对比算法和软件也在种系发育及进化研究中起了重要作用 [103]。另外，一些非线性研究方法、密码学方法也被引入进行序列特征及 Junk DNA 等研究 [46]。

在以上的研究中，有些方法主要着眼于序列中的一些局部的信号，另外一些方法则主要着眼于序列中全局性相关。

§5.2 碱基密度分布的多尺度分析

生命体是一典型的复杂系统，其中既包含规律性，又包含丰富的随机性 [44]。我们认为生命复杂性的分子基础应是基因组 DNA 序列中四种碱基的有序排列与无序排列在多种尺度上交错分布的复杂性。人类基因组及其它各种生物完整基因组的序列测定为我们研究生物体系复杂性、发展相应的复杂性理论提供了直接的数据材料，而对完整基因组有效的理论分析和比较则有望建立起关于生物起源、进化和生物调控的基本框架。

根据生物学上的中心法则，遗传信息的传递，生命系统的复杂功能等都与一维 DNA 序列上碱基的排列方式密切相关 [62]。 DNA 序列上碱基密度涨落分布展现了有趣的不均匀性。我们希望通过对碱基密度涨落的多尺度描述，对生命系统复杂功能的定量刻划有所帮助。

过去，物理学家们对 DNA 序列的研究相当一部分集中在讨论序列中是否存在长（短）程相关的问题上。1992 年，基于随机行走模型， Peng 等人 [80] 认为尽管 DNA 序列中的编码部分不存在长程相关，非编码部分确实存在长程相关。这一报道引起了人们对 DNA 序列中的相关性问题的广泛兴趣。接下来的工作得到了一些相互矛盾的结果。 Voss[102] 通过对 DNA 序列的 FFT 分析，认为即使在编码区也存在长程相关，与 Prabhu 等人的结果 [83]，以及 Chatzidimitrion-Dreismann 的结果 [23] 相一致。 Nee[78] 和 Karlin[55] 却认为在编码和非编码部分都不存在长程相关。虽然其后又有大量的工作，关于 DNA 序列中的相关性问题仍然没有定论。有兴趣的读者可以参阅文献 [65]。

正如 Li[65] 所注意到的，过去对长程相关的研究主要基于对单碱基间的统计。然而，

这种点对点的统计可能不是揭示大尺度，全局性性质的有效手段。也就是说，过去的统计研究并未给我们认识 DNA 序列的总体统计性质，全局相关性提供多少素材。

本节应用新近发展的复杂系统层次结构理论 [91, 92] ，对大肠杆菌 *E.coli* 基因组的 DNA 序列进行层次结构分析，发现大肠杆菌基因组碱基密度分布具有多尺度结构，且满足层次相似律。进一步发现，不同碱基组合的密度分布的层次结构参数具有明显的生物学含义，尤其是强氢键作用的 G 、 C 分布相对于弱氢键作用的 A 、 T 分布具有不同的层次相似参数。这一研究提示了一种刻画基因组结构复杂性的新的定量方法，对于揭示各种生物体基因组自组织结构及物种之间定量关系有一定的意义。

让我们首先简单描述一下我们这里所使用的 DNA 序列数据。所有的数据都是使用匿名 ftp 服务从 NCBI 在中国的镜像站点，北京大学生物信息中心网站上下载的，地址是：ftp.cbi.pku.edu.cn/pub/ncbi/ 。我们感谢北京大学生物信息中心所提供的帮助。在表 5.1 中，我们罗列了一些本文分析所用到的数据。其中比较特殊的一个是大肠杆菌 *E.Coli* 全基因组数据。

E.Coli 是最早建议为全基因组测序的生物体，其 K-12 菌株全基因组测序于 1997 年完成 [17], 全长 4,639,221 碱基对，已标注编码基因或可能的编码基因有 4288 个，这是实验研究得最充分的模式生物体之一，是我们下面分析的主要侧重点。

首先，在 DNA 序列中引入大小为 L 的窗口，计算窗口内某种碱基（如 A ）的个数，对窗口尺度进行归一，得到窗口内该碱基的密度 ρ_L 。顺着 DNA 序列平移窗口，将得到 ρ_L 沿序列的变化曲线数学上， ρ_L 的计算可表示为：

$$\rho_L = \frac{1}{L}\sum_{i=1}^{L}\delta_{u_i,A} \tag{5.1}$$

这里， $\delta_{u_i,A}$ 是 Kronecker 符号，当 $u_i = A$ 时， $\delta_{u_i,A} = 1$ ，否则， $\delta_{u_i,A} = 0$. 放大窗口宽度 L, 则可以得到另一组 ρ_L ，依次类推。实际上，每次放大窗口宽度都对应于一次对 ρ_L 的粗粒化过程。这个过程类似于湍流标度律研究中处理局部能量耗散率 ε_r 的过程。图 5.1 中，我们给出了窗口尺寸为 $L = 100$ 时一段典型的碱基 A 密度涨落信号。数据取至大肠杆菌 *E.coli* 全基因组。

因为下面计算的需要，我们测量了不同窗口尺寸 L 下 ρ_L 的概率分布。图 5.2 给出了 $L = 512$ 和 $L = 1024$ 时的概率分布函数。纵坐标用的是对数尺度。

现在计算涨落量 ρ_L 的 p 阶矩：

$$\varepsilon_L^p = \langle \rho_L^p \rangle = \int \rho_L^p P(\rho_L) d\rho_L \tag{5.2}$$

这里 $P(\rho_L)$ 代表 ρ_L 的概率分布函数。我们检验了标度关系： $\varepsilon_L^p \sim L^{\tau_p}$ ，发现这一绝对标度律并不成立。也就是说没有绝对的自相似性存在。

相反，当我们考察相对标度律 ESS[11] 时，我们得到了正面的结果。图 (5.3) 中给出了四个双对数坐标下的 $\langle \rho_L^p \rangle$ 对 $\langle \rho_L^3 \rangle$ 的结果，可以看出线性关系是很明显的。表明相对标度律很好地成立。 ESS 得到的标度指数 $\tau_{p,3}$ 罗列在表 5.2 中。不难看出， $\tau_{p,3}$ 随 p 的变化是非线性的，证明这里的统计行为是非高斯的，而不是高斯的。

现在，我们的分析程式来到了所谓的 β- 检验 (2.22) 过程。我们在图 5.4 上用双对数坐标画出对的关系。图中的线性关系使我们相信系统通过了 β- 检验。这里我们所考虑的尺度范围是： L 从 128 到 8092 ， p 从 1 到 5 。线性拟合得到的 β 值分别是：对于碱基 A ， $\beta = 0.93$ ；对于碱基 C ， $\beta = 0.89$ 。二者都比 She-Lévêque 对于均匀各向同性的湍流速度场的 β 值 0.874[91] 大，表明两种系统的间歇性，相关性不同。

我们又测量了碱基 T 和 G 的 β 值，得到一个有趣的结果： T 的密度涨落测得的 β 大约是 0.94 ，而 G 的密度涨落测得的 β 大约是 0.90 。于是，我们可以看出 $\beta_A \approx \beta_T$ ， $\beta_C \approx \beta_G$ 。也就是说， A, C, G 和 T 分成两组： A/T 组和 C/G 组。这难道仅仅是巧合吗？不是，绝对不是。实际上，我们在分析其他生物基因组数据的时候发现了同样的事情。关于分组现象的意义我们放在本节的后面讨论。

图 5.5 是 γ- 检验图。线性拟合给出 A 的 γ 为 5.18 ，与 C 的 γ 很不同， C 的 γ 是 3.04 。同样可以测量 G 和 T 的 γ ，得到 $\gamma_A \approx \gamma_T$ ， $\gamma_G \approx \gamma_C$ 。又一次发现了分组现象。

关于大肠杆菌 *E.coli* 全基因组的层次结构分析得到的所有参数都列在表 5.2 中。

我们还分析了人类基因组各条染色体序列的碱基密度分布，其中第 22 条染色体的 β- 检验结果在图 5.6 中给出。

利用 2.22 式给出的 β- 检验，我们对一些原核和真核生物的 DNA 序列进行了比较系统的研究，我们发现， β 值随着物种的不同有比较大的变化。在表 5.1 中，我们给出了细菌及人类染色体的碱基 A 密度变化的分析结果。可以看出，虽然 β 值存在一定的涨落，但基本上分成两组，细菌基因组和人类染色体，二者之间 β 值的差别可以在图 5.7 中清楚地表现出来。

从以上分析结果可以看出， A 、 T 、 G 、 C 四种碱基的密度变化基本上可以分成

两组， A 、 T 和 G 、 C ，各组内部两种碱基间的差别要远远小于两组之间的差别。由于 GC 富含区也就是基因富含区，因此， AT 、 GC 的差别也就是编码区和非编码区的差别，即非编码区呈现出很强的长程相关 [21] 。 G 与 C 间通过三根氢键配对， A 与 T 间通过两根氢键配对，它们具有相对不同的化学功能，因而具有自组织性上的差异。同时，由于碱基配对的互补性，分组现象可能是 DNA 双链对称性的一种自然体现 [15] 。对 *E.coli* 基因组四种碱基层次结构分析表明，从典型的基因内（ $100bp$ ）到普通基因组（ $8000bp$ ）之间的尺度上，层次相似律是比较好地存在的。反映出存在基因在基因组中分布的特征复杂性结构的可能性。

实际上，层次相似律的存在性直接反映了 DNA 序列中碱基分布的强不均匀性 [15, 38, 85] 。正是这种不均匀性使得序列的统计特征明显的偏离高斯性。尽管人们对这种不均匀性产生的原因及其生物学含义目前还不了解很多 [15, 85] ，但我们有理由相信，这种不均匀性与生物形成与演化等重要问题密切相关 [38] 。从以上的分析结果可以看出：大肠杆菌的基因组序列具有多尺度结构，并且是强间歇的。在这样的序列中，当我们考察某些物理参量时，发现在从基因内（ $100bp$ ）到基因组（ $8000bp$ ）之间的尺度上，存在扩展的自相似律，同时层次相似律也得到很好的验证。这反映了碱基分子密度分布在长期的进化过程中演化到一定的统计自组织状态。层次相似参数将是对其统计状态的定量刻画。我们发现，原核生物基因组序列的平均 β 值比其它自然现象如充分发展湍流的大，反映了序列在某种意义上为多样性更丰富的多尺度结构，而人类基因组序列的平均 β 值比湍流的小。由于生命过程是无数低概率事件的集合，而层次结构所描述的正是低概率、强涨落事件的多尺度标度性，因此有望成为定量描述生物完整信息的模型。由于非编码区具有许多基础的生物学功能，如染色体复制、分裂、重组，染色体稳定性以及与细胞核的相互作用，所以为了防止或减少错误的发生，非编码区通过自然选择，保留了高丰度率和低信息量的特性。层次结构参数可作为基因组复杂系统自组织的特征量度，用以系统研究不同物种序列的层次结构；对基因组碱基所具有的高度非均匀分布，该理论可用以系统研究编码与非编码区的层次结构。

§5.3 DNA 序列的词汇使用频率图分析

对 DNA 序列的语言学研究曾经引起过不少人的注意 [35, 42, 49, 57, 72] 。 DNA 序列语言学的研究一般分两个不同的方向，一个就是利用传统的 Zipf 试验或者 Shannon 试

验等研究 DNA 序列语言与自然语言的差别和联系 [72, 35]，另外一个就是从词汇使用的角度来考察 DNA 序列的语言学 [42, 49, 57]。后一种方法显得更加物理化一些，也是本文将要使用的方法。

过去对 DNA 序列词汇使用频率的研究多半集中在仅仅对某一特定长度的词汇使用情况做统计分析，或者孤立地对某些长度的词汇使用做研究，忽略了各种长度词汇使用之间的联系，因而难免丢掉某些重要信息，如序列词汇使用上的相关性等。作者这里的工作将试图弥补这一缺点，着重讨论 DNA 序列中各种长度词汇使用之间的关系。

在对 DNA 序列词汇使用频率的进行研究时，第一步要做的当然是建立序列的词汇使用频率字典。频率字典可以是真正的，也可以是重构的，还可以是经过变换的 [42]。这里，我们将要讨论的都是真实的频率字典。

为了使人们能对长 DNA 序列有个直观的，整体的印象， Hao 等人提出了我们称为词汇频率地形图 (WFL) 的 DNA 序列二维表示方法 [49]。基本思路就是将 DNA 序列中所有可能出现的字长为 K 的词汇排列在一个的二维方阵上，每一点代表一个词汇，然后将该词汇在 DNA 序列中出现的频率用相应深度的颜色表示出来，这样就得到了一个类似于地形图的二维图，图上每点 (对应于一个单词) 的颜色代表该单词在 DNA 序列中出现的频率。关于 WFL 的详细介绍请参阅文献 [49]。 Hao 等人曾经讨论过许多生物体全基因组 DNA 序列的频率地形图，得到了不少有用的结果，我们也曾经将该方法推广来研究编码和非编码序列的差别，图 5.8 和 5.9 分别给出了大肠杆菌 *E.coli* 全基因组编码区和非编码区字长为 9 时的词汇频率地形图，可以看出二者差别非常大，说明这种方法是一种比较有效的方法。

受前面的这些杰出的工作的启发 [42, 49, 57]，我们将在序列的单词使用频率空间去研究序列的结构。因为以前的工作都集中在某一个独立的词汇长度空间，我们的研究将更加注意各种长度单词间的关系。

让我们首先描述一下如何构造一个单词频率字典 (Word Frequency Dictionary, WFD)。对于一个由给定字母表 $ALP = \{A, C, G, T\}$ 生成的 DNA 符号序列，我们将其分成一些给定长度 K (下面的演示图中我们取长度为 $K = 6$) 的子序列 $S_k = P_{i_1}P_{i_2}..P_{i_k}$，$P_{i_k} \in ALP$。后面的讨论中我们称这些子序列为“单词”。

... A T G C G A C T T C A A G A T C G A C G T ...

我们所关心的是每个单词在给定的原始序列中究竟出现了多少次。简单的组合学知识告

诉我们，对于由 A,C,G,T 四个符号组成的长度为 K 的字符串，所有可能的组合方式为 4^K 种。为了避免计算种引进非常耗费计算时间的字符串比较，我们给每个单词赋予一个唯一的正整数。方法如下：首先，将 A,C,G,T 映射到一个四进制整数系统，

$$\{A,C,G,T\} \longmapsto \{0,1,2,3\}$$

然后，对于每个字符串 S_k，有唯一的整数 I 与之对应 [49]:

$$I = \sum_{j=1}^{K} Q_{i_j} 4^{K-j}, \tag{5.3}$$

式中，Q_{i_j} 是与 P_{i_j} 相对应的四进制数。容易证明 I 是不大于 4^K-1，不小于 0 的整数。于是我们就在长度为 K 的所有单词和 0 到 4^K-1 之间的所有整数之间建立了一一对应关系。现在，给定的 DNA 序列被转换成了一个整数序列。只需要计算一下每个整数在这个序列中出现的次数，再除以序列的总长度，我们就可以得到与其相应的单词在该序列中出现的频率。为了便于不同生物体之间的比较，与文献 [49] 的做法类似，我们将得到的频率对长度为 100 万的序列归一。

下一步，我们将所有长度为 K 的单词排成一个一维的序列，规则是按单词所相应的整数 I 从小到大的顺序，也就是这些单词出现在字典上的次序。我们在下文中称该序列为一个尺度上的词汇频率字典。通过构造各种长度尺度的字典，我们就可以得到一个如图 5.10 所示的词汇频率字典系列。注意图 5.10 的一个有趣的性质是，每一个左边的字典 K 都可以通过对其右侧字典 $K+1$ 的相邻四个频率求和而得到。如果你对第 K 个字典进行 $K-1$ 次类似的求和，你就可以得到原始序列的 A,C, G 和 T 四种字符的各自含量。但是所有这些过程都不能向反方向进行，尽管曾经有人尝试过这么做 [20, 42] 。

图 (5.11) 中我们给出四个尺度为 6 的典型的词汇频率字典。使用的序列分别是大肠杆菌 *Escherichia Coli* 全基因组序列，*Methanococcus jannasckii* (*M.jan*) 全基因组序列，人类第 21 条染色体序列以及一个随机序列 [1]（注意，因为各个信号的涨落存在跨量级的差别，我们没有将图 5.11 的四个子图的 Y 坐标设在同样的标度上，以便可以看清楚各个图内部的涨落。）

除了涨落大小的巨大差别外，我们发现四个信号在其空间结构上存在巨大的差别，表明四个序列在词汇使用上的巨大差别。细心的读者会发现，每个信号都存在关于中间点的对称性。我们将在下面对此给予解释。图 5.20 中，我们收集了一些生物序列的词汇

[1]我们构造随机 DNA 序列的方法是将大肠杆菌 *E.coli* 的 DNA 序列上碱基分布的顺序随机打乱，这样做的目的是保持序列内的四种碱基的含量完全不变。

频率字典。为了比较信号的涨落信息，我们在图中给出了不同信号的概率分布。很明显，随机序列的涨落最小，而 M.jan 和人类第 21 条染色体的涨落最大。在后两种情况下，概率分布显示出明显的非高斯性。

尽管我们也认为研究全基因组可能会得到更多的信息，但是，我们认为上面这种方法本身可以应用于任何长度的序列。实际上，文献 [42] 中，作者就利用类似的方法研究过生物体的 16S RNA 序列，得到了很多有用的结果。

现在，我们将利用 Fourier 分析为工具来刻划我们得到的频率字典的一些基本性质。对于给定离散序列 $\{u_k\}$ 的 Fourier 变换 (FT) 可以写成 [21]：

$$q_f = \sum_{k=0}^{N-1} u_k exp(ikf2\pi/N) \tag{5.4}$$

相应的功率谱写成：

$$E(f) = |q_f|^2 + |q_{N-f}|^2 \tag{5.5}$$

虽然 f 可以从 0 取至 $N-1$，但是由于 Fourier 变换在 $N/2$ 的共轭对称性，通常，只有前半部分可以使用。

我们测量了不同字长的 WFD 的功率谱，对应于图 5.11 中信号的结果在图 5.13 中给出。这些功率谱图的最显著的特点就是谱峰结构，反映了原序列中的周期性结构。为了更容易地辨认出峰值对应的波数，我们已经将横坐标设为对数尺度。我们发现，最高的 5 个峰之间的距离相等，意味着线性尺度下，这些距离是成常数倍增加的。在这里，这个常数是 4。在其他不太明显的峰值之间也有相同的情况发生。也就是说：如果长度为 K 的字典有某个波数为 n 的明显周期结构，那么它必然同时存在相同振幅的波数为 $n4^m$ 的明显周期，其中 $m = 0, 1, ..., K-1$，除非 n 比 4^{K-1} 还大。下面的讨论中我们将称 n 基本波数 (basic wave number, BWN)，相应的 $n4^m$ 称为导出波数 (derived wave number, DWN)。很自然，这种 4 倍周期的现象是由于两个因素造成的。一个是构成 DNA 序列字母表只有 4 个字母，导致 4 倍周期的出现。另一个原因是因为 DNA 序列中 A，C，G 和 T 的含量并非完全相等，造成明显的峰值结构。这个解释可以由以下事实所支持，就是打乱顺序后的 *E.coli* 基因组 (碱基含量保持不变) 序列仍然存在明显的峰值结构 (见图 5.13(d))，而一个四种碱基含量相等的随机序列只有四倍周期而没有明显的峰值结构 (这里没有给出此图)。

现在，让我们转向研究不同功率谱之间的差别。从图 5.13 可以看出，*E.coli* 的最突出的基本波数是 3，而其几个的基本波数是 1，这实际上反映了其字典的结构有所不同。

波数 1 反映了整个字典存在明显的关于中心的完全对称结构，而波数是 3 表明这种字典不是完全关于中心对称，对称中心有了偏移。关于词汇频率字典的的详细分析值得进一步的探讨。

在上面的分析中，我们主要集中在对单个词汇频率字典（ WFD ）的分析上。下面我们将引进多尺度的概念，讨论各个不同单词长度的字典之间的关系问题。

首先，我们引进尺度 l ，定义为字典大小的倒数，即 $l = 4^{-K}$ 。为什么要用字典大小的倒数呢？我们是出于这样的考虑：当我们增加所考虑的单词的长度时，我们实际上是看到序列更加精细的结构。如果考虑长度为 1 的单词，则我们只能得到关于序列很粗糙的信息， A ， C ， G ， T 的含量；考虑长度为 2 的单词时，不仅可以得到各个碱基的含量信息，而且可以知道 AA ， AC ， AG ， AT ， CA ，. . . 等的含量信息，就是它们的使用频率。依次类推，考察的单词长度越长，反映的是越精细的结构，所以我们取为字典大小 4^K 的倒数，而不是字典大小本身。

然后，我们考虑不同字长的字典上各种单词使用频率 f_l 的涨落之间的关系。我们定义尺度 l 上该涨落的 p 阶矩为：

$$S_p(l) =< f_l^p >= \int_{-\infty}^{+\infty} (f_l)^p p(f_l) df_l \tag{5.6}$$

于是我们可以检验关系式： $S_p(l) \sim l^{\zeta_p}$ ，看看有无标度律成立。容易证明，如果所考察的 DNA 序列是一个完全随机，而且四种碱基均匀分布的序列，那么该标度律一定成立，并且标度指数将是 $\zeta_p = p$ 。我们发现，对于真实的 DNA 序列，标度指数并不是，而是比略小。图 5.14 中给出了三阶矩随尺度的变化。我们发现，对于我们的人工随机 DNA 序列， $\zeta_3 \approx 2.99$ ，与理论预测的 3 非常接近。而对于大肠杆菌全基因组序列 $\zeta_3 \approx 2.65$ 。最有意思的是对于人类第 21 条染色体 DNA 序列，没有明显的标度律存在，对其他真核生物染色体的结果也如此。以上结果表明，细菌和真核生物 DNA 序列在词汇使用方面存在着很大的不同。细菌更接近随机序列，虽然与随机序列大不相同。实际上，很容易想象，词汇频率字典直接反映了 DNA 序列上碱基的分布规律，是碱基分布不均匀性的自然体现。

同样，图 DNA5.15(a) 是大肠杆菌 *E.coli* 序列词汇使用频率字典的 β- 检验，显然，线性度很好，表明该检验被通过了。 β 值为 0.93 ，与随机序列的 1.0 相差较大。与标度律检验结果类似，当我们考虑真核生物的序列时，我们观测到 β- 检验失败的情形，见图 5.15(b) 。图中所用的数据是人类第 21 条染色体序列。我们用其他真核生物体的序列进

行 β- 检验时，也发现检验的点比较分散，意味着没法通过此检验。我们这里进行 β- 检验时，都取了尺度范围是 $K=3$ 到 $K=8$， p 取从 $p=0.5$ 到 $p=5$ 。

我们注意到，对于大肠杆菌 *E.coli* 来说，基因组的编码部分远远大于非编码部分，实际上， *E.coli* 大约 87.8% 的序列都是编码序列 [17] 。而对于人类基因组序列来说，绝大部分都是所谓的 “ *junk* ” DNA ，是非编码的部分，只有大约 $3\sim5\%$ 的序列是有编码功能的 [28] 。这令人马上想到，大肠杆菌和人类第 21 条染色体在 β- 检验方面反映出来的差别可能就是编码与非编码部分差别的自然体现，这是一个值得进一步探讨的问题。

我们尝试着在词汇使用频率空间研究了 DNA 序列的结构行为，以帮助我们更好地研究 DNA 序列之间的差异。对于频率字典的 Fourier 分析表明，尽管不同生物的 DNA 序列之间的确存在许多共同点，例如谱空间的四倍周期现象，但是也确实存在许多明显的不同点，例如看起来似乎这些差别同样存在于编码和非编码区之间。

细菌类 DNA 序列不同长度的词汇使用频率字典之间存在很好的标度律关系。并且，这些不同尺度，大小涨落之间的关系可以由层次结构模型来描述。 β- 检验和 γ- 检验都被通过。对于真核生物的 DNA 序列，标度律关系不再成立，层次结构模型的 β- 检验也难以被通过。以上结果说明从语言学的角度来看，原核生物和真核生物的序列在词汇使用频率这一点上存在较大的差别，这些差别反映了二者进化过程的机制有所差别。

由于非编码区具有许多基础的生物学功能，如染色体复制、分裂、重组，染色体稳定性以及与细胞核的相互作用，所以为了防止或减少错误的发生，非编码区通过自然选择，保留了高丰度率和低信息量的特性，因而非编码区的词汇使用相对单一。而编码区则由于生物学功能多样性的需要而采用高信息量的词汇排列方式。

§5.4 DNA 序列的多变量熵距离分析

在本文主要内容的最后一部分，让我们抛开多尺度分析，从另外的角度来看看该如何去刻划一个比较复杂的系统。我们将仍然以对 DNA 序列的研究为例，讨论我们的一些新的想法。

DNA 序列研究的一个有趣的方向就是寻找能够对结构相似的序列进行识别的方法，包括所谓的基因识别 [47] ，也就是从未注释过的序列中寻找基因编码区。大部分基因识别的研究都是首先要找到编码区序列里所隐藏的结构信息。马尔可夫近似方法 [19, 60] ，语言学方法 [30] ，决策树方法 [89] 以及 Z 曲线方法 [106] 都是其中比较成功的例子。然

而，大部分方法都是要不着眼于过于局部的信息，要不就是眼于过于全局的信息，因而都不能提供足够的信息去进行高精确度的预测。

一个完整的基因组包含着几十万，几百万甚至几千万的碱基对。编码区几乎包括了合成蛋白质的所需的全部信息，而剩余的部分（非编码区）具有其他一些不那么特定和准确的生物学功能。长久以来人们一直认为编码区与非编码区是有着本质不同的 [62]，问题是如何揭示刻划这些不同。如果能找到这样的方法，那么就可以用它来寻找新的编码序列。我们认为这种区别在二者通过遗传密码所生成的伪氨基酸序列上有所反映。在原核生物的基因组里存在着成千上万的真实的氨基酸序列，这些序列终将被翻译成各种各样的蛋白质。我们相信这与非编码区的伪氨基酸序列是有所不同的。也就是说，有两种不同的氨基酸序列，一种是真的，另一种是假的。只有前者具有生物学的意义。

在图 5.16 中，我们给出了由大肠杆菌的编码（上图）和非编码序列（下图）所生成的两组氨基酸序列。信号代表氨基酸的种类沿着序列的变化。注意，我们这里共有 21 个氨基酸指标，前面 20 个代表 1-20 种氨基酸，第 21 个代表终止密码子。

第一眼很难看出两个信号有和区别。二者都表现出复杂的结构。然而，下面我们引进的多变量熵密度方法却可以作为系统地区分这两种序列的比较敏感的方法。

我们提出利用由一组熵密度组成的熵密度廓线 EDP 来刻划有限长的 DNA 序列（一般几百个氨基酸）的方法，形成所谓的多变量熵描述的方法。下面我们将证明，整个编码与非编码区序列定义了两个相互分开的平均熵密度廓线。而且，我们发现适当长度 >300bps 以上的编码和非编码序列基本上聚集在各自中心点的周围。利用单个序列对两个中心的距离的差别这一性质，我们可以将两种序列几乎完全分开。这一点在原核生物的 DNA 序列中已经得到了充分的证明。

我们的结果表明 EDP 可以反映编码序列的基本属性，携带使序列成为有生物学含义（例如决定其所翻译的蛋白质的三维结构）的基本信息。这一结果对进一步研究其他生物学结构，如 α 螺旋，β 折叠片等二级结构具有一定的意义。

§5.4.1 MED 方法

MED 方法使用归一化的熵密度廓线（ EDP ）来刻划一个氨基酸序列。熵密度廓线定义为

$$S_i = -\frac{1}{H} p_i \log p_i, \quad \text{i=1, ... ,21} \tag{5.7}$$

这里 p_i 代表第 i 种氨基酸出现的频率， H 是给定氨基酸序列的 Shannon 信息熵，由下式定义。

$$H = -\sum_{i=1}^{21} p_i log p_i \tag{5.8}$$

在实际的计算过程中， p_i 就是第 i 种氨基酸出现的次数除以氨基酸序列中总的氨基酸的个数。注意，变量 $\{S_i\}$ 的选择是与有生物学内容的 p_i（氨基酸的丰度）紧密联系起来的。对于非编码序列，我们也做同样的事情，只不过由于它不编码氨基酸，这里的氨基酸只是一种伪氨基酸。 p_i 曾经被拿来刻划氨基酸序列，和进行基因预测 [95, 34] 。这里的我们的创新之处在于使用全部 20 个氨基酸（加上一个终止密码子）构成一个对有限长度的序列的多变量的描述。这种多变量的描述的好处在于它能够一系列复杂序列中的分组性质。编码，非编码序列就在这种性质上有差别，因此需要使用这种多变量的描述。尽管我们无法严格证明，但是下面的事实说明对一定长度以上（长于 100 氨基酸，或者说 300 个碱基对）的序列，熵密度廓线 $\{S_i\}$ 具有足够的信息去决定相应的 DNA 序列是否是有生物学含义的编码序列。

熵密度廓线（ EDP ） $\{S_i\}$ 是随序列不同而变化的。下面关键的一步就是发现对于编码和非编码序列，熵密度廓线分别有各自的 “中心”。换言之，尽管编码非编码内部各序列之间 EDP 有些涨落，但两个系综各自还是存在一个平均的熵密度廓线。为了说明这个中心的确存在，我们给出了图 5.18 。因为无法在 21 维空间显示这个中心，这里取的是其在两维平面上的投影。这个平面的两维坐标由 Ala 和 Cys 两种氨基酸构成，分别为横纵坐标。可以看出，虽然存在一定的重叠，但是，编码和非编码区确实存在相互分开的趋势，也就是说，二者有不同的中心存在。可以想象，当我们考虑高维空间的时候，图 5.18 中的两个部分会更清楚地分开，这就是为什么我们要引进高维空间描述的原因。图 5.17 中我们给出了大肠杆菌编码和非编码区的中心，即平均熵密度廓线（ EDP ）

令 $\{\overline{S_i}\}^c$ 和 $\{\overline{S_i}\}^{nc}$ 分别代表由大量编码序列的 EDP 得到的编码序列的平均 EDP 和由大量非编码序列的 EDP 得到的非编码序列的平均 EDP ，即两种序列的中心点。那么对于每个序列，我们可以定义一个多变量的熵距离（ multivariate entropy distance ，MED ）：

$$D_\alpha = \sqrt{\sum_{i=1}^{21} (S_i - \overline{S_i}^{\alpha})^2} \tag{5.9}$$

实际上就是在 21 维熵密度空间的每个序列与中心点的距离。 α= “ c ” 或 “ nc ” 分别代表距编码中心 $\{\overline{S_i}\}^c$ 和非编码中心 $\{\overline{S_i}\}^{nc}$ 的距离。如果在 21 维的熵空间内两组相点是分开的，那么，编码区的相点距编码区中心相点的距离应该小于它们距非编码区中心相点的距离。同样，非编码区的相点距非编码区中心相点的距离应该小于它们距编码区中心相点的距离。于是，我们就可以将编码序列与非编码序列区分开来。在图 5.19 中，我们给出了 500 个编码序列与 500 个编码序列距两个中心的距离比较。纵横坐标分别是相点距编码和非编码中心的距离。图中斜线为角平分线。可以看出，两组序列有很好的聚类现象，编码非编码序列被分开。因此可以说明，编码序列的熵密度廓线与非编码序列的熵密度廓线大不相同。

§5.4.2 Genbank 数据分析

在上一节的试验过程中，一直隐含着这样的一个假设，就是编码和非编码区的平均熵密度廓线（ EDP ）是已知的。如果要拿这种方法来做实际的编码区非编码区识别，我们就会遇到这样的问题：编码区和非编码区的中心是未知的，需要从有限的序列中去 “学习” 到编码和非编码区的平均熵密度廓线（ EDP ），于是，中心随着序列的变化的敏感性问题变的十分重要。我们的研究表明，只用 10 到 50 个已知序列来求出一个大概的中心也能达到很高的精度，也就是说，中心随着序列的变化是很稳定的。这种稳定性再次证明编码和非编码在高维熵空间是相互分开的。

我们以方法来分析了一些原核生物的 DNA 数据，结果罗列在表 5.3 中。

MED 方法的结果有其生物学的合理性。氨基酸的熵密度与其在序列中出现的频率及占有的百分比是紧密相连的。对于那些出现较多的氨基酸，出现频率与其在一维序列中的平均距离也是紧密相关的。而这一平均距离很可能与序列的二级结构和其他性质密切相关。因而， MED 方法给出的是对序列的整体属性的描述，而不是仅仅抓住序列某种比较微妙的特征。初步的研究显示，不同的基因组序列其编码，非编码中心是不同的。关于此性质的进一步研究有可能导致一种系统地区分不同基因组的方法。

本文提出的思想和方法也可以被用来尝试研究其他复杂系统。

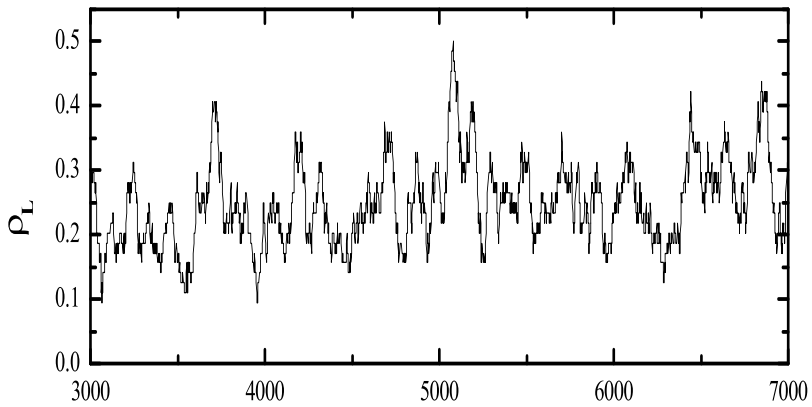

图 5.1: 一段典型的碱基 A 密度涨落 ρ_L 沿序列的变化，窗口尺寸为 $L = 100$ 。数据取自大肠杆菌 *E.coli* 全基因组 DNA 序列。

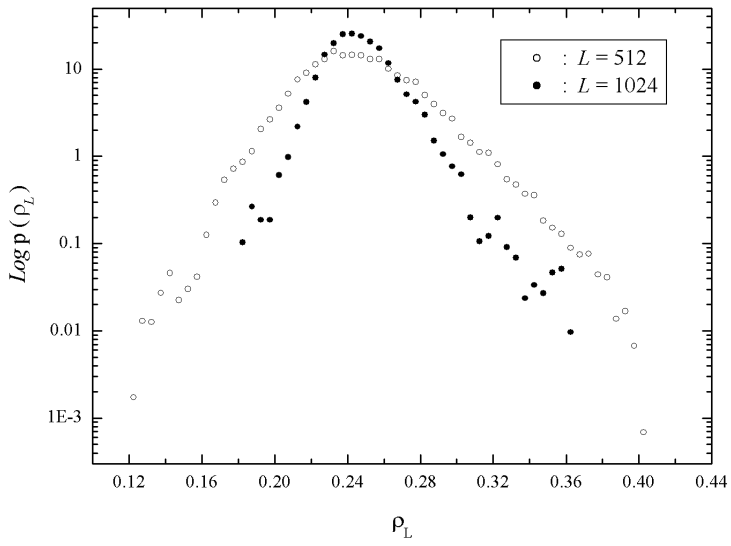

图 5.2: 大肠杆菌 *E.coli* 全基因组碱基 A 密度涨落 ρ_L 的概率分布函数，窗口尺寸分别为 $L = 512$（空心圆点）和 $L = 1024$（实心圆点）。

英文简称	数据大小	*beta* 值	英文简称	数据大小	*beta* 值
A.ero	1.67M	0.92	*A.ful*	2.18M	0.90
M.jan	1.66M	0.90	*M.the*	1.75M	0.90
P.abyssi	1.77M	0.87	*P.yro*	1.74M	0.89
U.ure	1.55M	0.90	*B.bur*	0.91M	0.87
B.sub	4.21M	0.86	*C.pneu*	1.23M	0.93
C.tra	1.04M	0.92	*C.jej*	1.64M	0.89
E.coli	4.64M	0.93	*H.inf*	1.83M	0.90
H.pyl	1.67M	0.89	*N.men*	2.27M	0.84
M.gen	0.58M	0.90	*M.pneu*	0.82M	0.87
M.tub	4.41M	0.88	*Synecho*	3.57M	0.87
T.mar	1.86M	0.86	*T.pal*	1.14M	0.87
H.sp Chr1	23.08M	0.82	*H.sp Chr2*	19.43M	0.82
H.sp Chr3	8.68M	0.84	*H.sp Chr4*	11.49M	0.81
H.sp Chr5	13.21M	0.80	*H.sp Chr6*	41.99M	0.81
H.sp Chr7	78.56M	0.81	*H.sp Chr8*	8.72M	0.82
H.sp Chr9	4.85M	0.80	*H.sp Chr10*	4.60M	0.79
H.sp Chr11	7.93M	0.84	*H.sp Chr12*	22.01M	0.80
H.sp Chr13	1.99M	0.84	*H.sp Chr14*	19.77M	0.81
H.sp Chr15	2.09M	0.79	*H.sp Chr16*	16.22M	0.86
H.sp Chr17	28.37M	0.84	*H.sp Chr18*	3.52M	0.82
H.sp Chr19	14.56M	0.81	*H.sp Chr20*	22.25M	0.83
H.sp Chr21	17.70M	0.83	*H.sp Chr22*	33.34M	0.83
H.sp ChrX	60.75M	0.79	*H.sp ChrY*	5.92M	0.84

表 5.1：本文分析所使用的部分 DNA 数据及其大小列表。每部分的第三列是碱基 A 密度涨落层次结构分析所测得的 β 值。微生物的数据全部为全染色体数据，人类除了第 21 和 22 两条染色体外都是部分染色体数据。为了比较的方便，我们在进行 β- 检验的时候，已经将尺度范围全部定在 128 和 1024 之间。

	$\tau_{2,3}$	$\tau_{3,3}$	$\tau_{4,3}$	$\tau_{5,3}$	$\tau_{5,3}$	β	γ
A	0.35	1	1.95	3.18	4.68	0.93	5.18
T	0.35	1	1.95	3.17	4.67	0.94	5.73
G	0.36	1	1.89	3.00	4.31	0.90	3.23
C	0.36	1	1.91	3.05	4.41	0.89	3.04

表 5.2: *E.coli* **基因组碱基密度分布层次结构分析所测得的参数。**

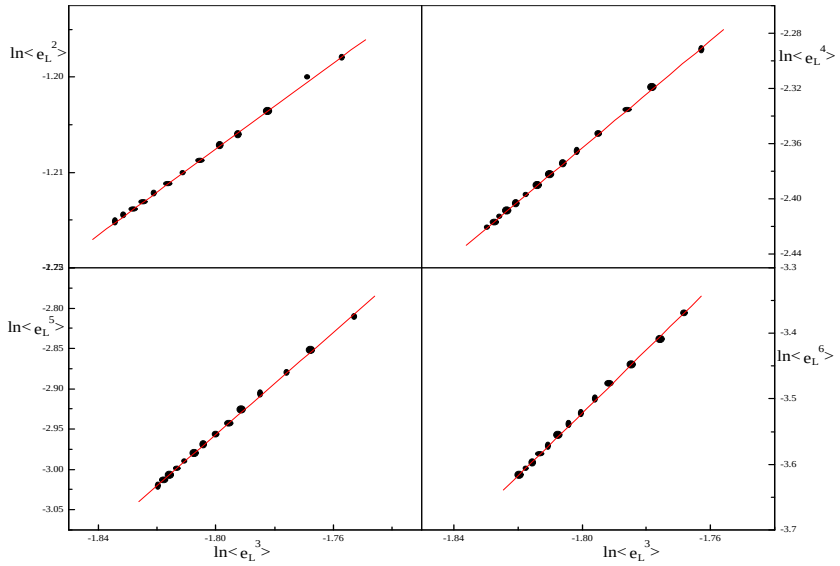

图 5.3: 碱基 A 密度涨落的相对标度律 ESS 分析，纵横坐标分别为对数坐标下的 $\langle \rho_L^p \rangle$ 和所用数据是大肠杆菌 *E.coli* 全基因组序列，矩分别为 $p = 2$， $p = 4$， $p = 5$ 和 $p = 6$。最小二乘法拟合得到的相对标度指数分别为 $\tau_{2,3} = 0.35$， $\tau_{4,3} = 1.95$， $\tau_{5,3} = 3.18$ 和 $\tau_{6,3} = 4.68$，显示了随的 p 非线性变化关系。

英文简称	Sn	Sp	Sq	Sr	Ac	CC
A.ful	99.2	98.4	98.7	99.3	97.8	97.8
A.quae	97.2	99.1	99.3	97.5	96.6	96.6
B.bur	95.8	98.2	98.4	96.4	94.4	94.4
B.sub	97.6	99.0	99.3	98.2	97.1	97.1
C.jej	97.4	97.8	98.1	97.7	95.5	95.5
C.pneu	95.9	97.3	97.8	96.8	93.9	93.9
C.tra	95.0	98.1	98.4	96.0	93.8	93.7
E.coli	98.4	99.0	99.2	98.8	97.6	97.6
H.inf	97.4	97.5	98.1	98.0	95.5	95.5
H.pyl	96.4	99.3	99.5	97.2	96.2	96.2
N.men	98.3	94.9	96.7	98.9	94.4	94.4
M.gen	95.7	96.7	97.3	96.4	93.1	93.1
M.jan	98.3	98.7	99.0	98.7	97.3	97.3
M.pneu	95.1	98.1	98.6	96.2	94.0	94.0
M.the	98.8	99.5	99.6	99.1	98.4	98.4
P.abyssi	98.8	99.1	99.2	99.0	98.1	98.1
Synecho	98.0	99.6	99.7	98.5	97.9	97.9
T.mar	99.5	99.3	99.4	99.6	98.9	98.9
R.pxx	95.2	99.5	99.7	97.1	95.7	95.7
R.pNGR	98.8	97.2	98.4	99.3	96.9	96.9
T.pal	96.3	97.4	97.8	96.8	94.1	94.1
U.ure	96.0	98.2	98.5	96.7	94.7	94.7

表 5.3: MED 方法识别各种不同生物体编码区的准确性参数。 Sn 和 Sq 分别度量了算法对编码和非编码序列的敏感性 (sensitivity)； Sp 和 Sr 分别度量了算法对编码和非编码序列的特定性 (specificity)； AC 和 CC 是综合参数。关于这些参数的具体意义，参见文献 [8] 这里所选取的序列长度都大于 100 个氨基酸。注意，进行此表的计算时，我们只考虑了熵密度空间的前 20 个维度。若将第 21 个维度 (终止密码子) 也考虑进来，得到的参数更高，平均综合得分 CC 达 98 以上。

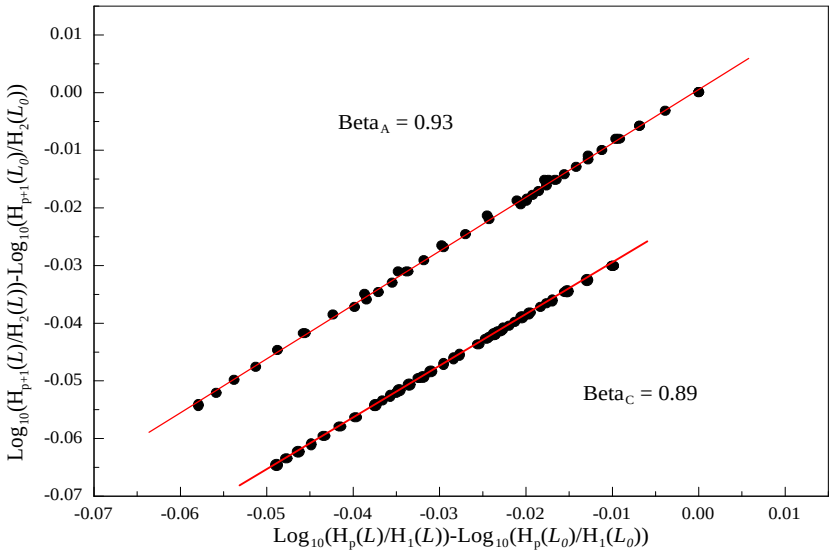

图 5.4: 图中的线性关系表明 *E.coli* 全基因组碱基密度涨落场 ρ_L 通过了 β- 检验。上下两组分别是碱基 A 和碱基 C 的情况。最小二乘拟合（图中直线）给出两个 β 值分别为 0.93（ A ）和 0.89（ C ）。

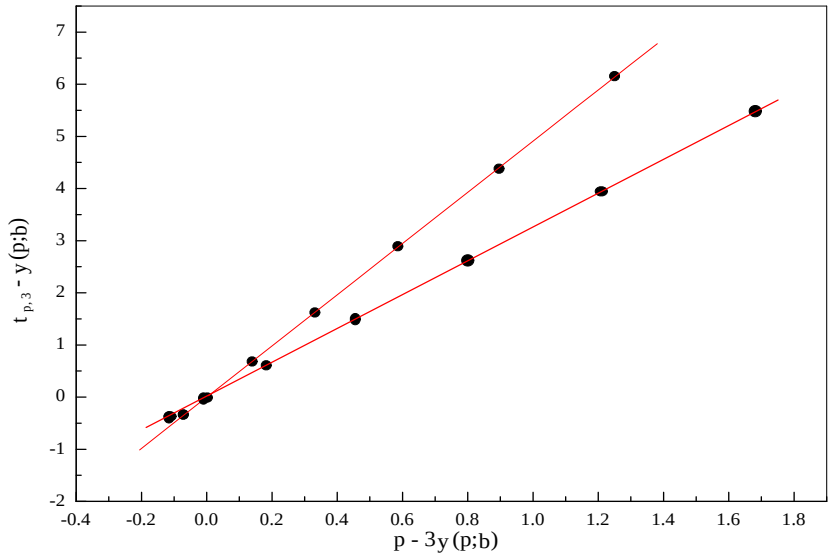

图 5.5: 与图 5.4 相应的 γ- 检验。线性拟合得到 A 和 C 的 γ 值分别为 $\gamma \approx 5.18$ 和 $\gamma \approx 3.04$ 。我们还测量了和的 γ 值，分别为 $\gamma \approx 3.23$（ G ）和 $\gamma \approx 5.73$（ T ）。

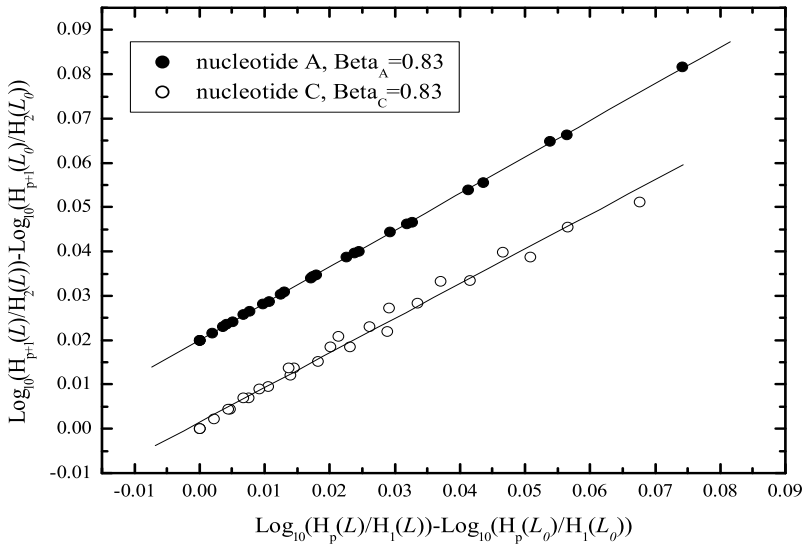

图 5.6: 人类第 22 条染色体碱基密度涨落场的 β- 检验。上下两组分别是碱基 A 和碱基 C 的情况。最小二乘拟合（图中直线）给出两个 β 值分别为 0.83（ A ）和 0.77（ C ）。

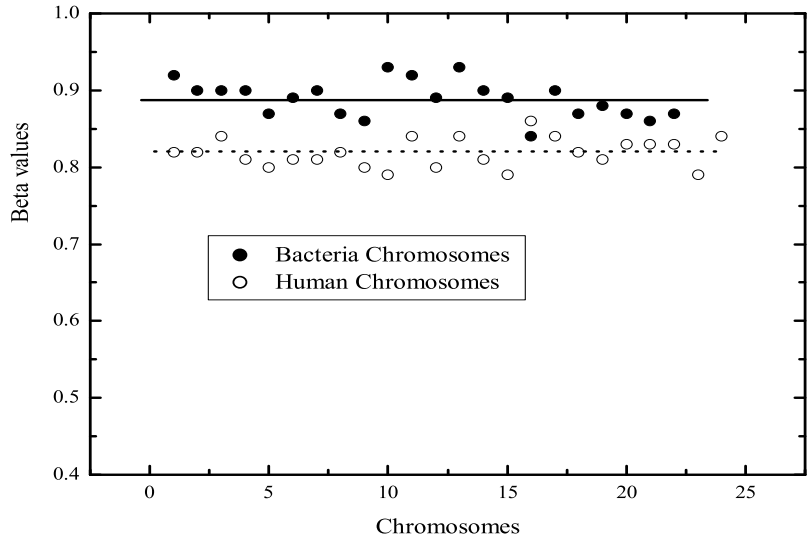

图 5.7: 不同染色体碱基 A 密度涨落场的 β 值随着物种的变化。图中每个点代表一条染色体（对于大多数原核生物实际上一种生物只有一条染色体），空心圆点代表人类 24 条染色体（ 22 条常染色体加上 X， Y 染色体），实心圆点代表 22 种微生物，实线（实心圆点）和点线（空心圆点）分别是两组的平均 β 值，分别是 0.89 和 0.82 。

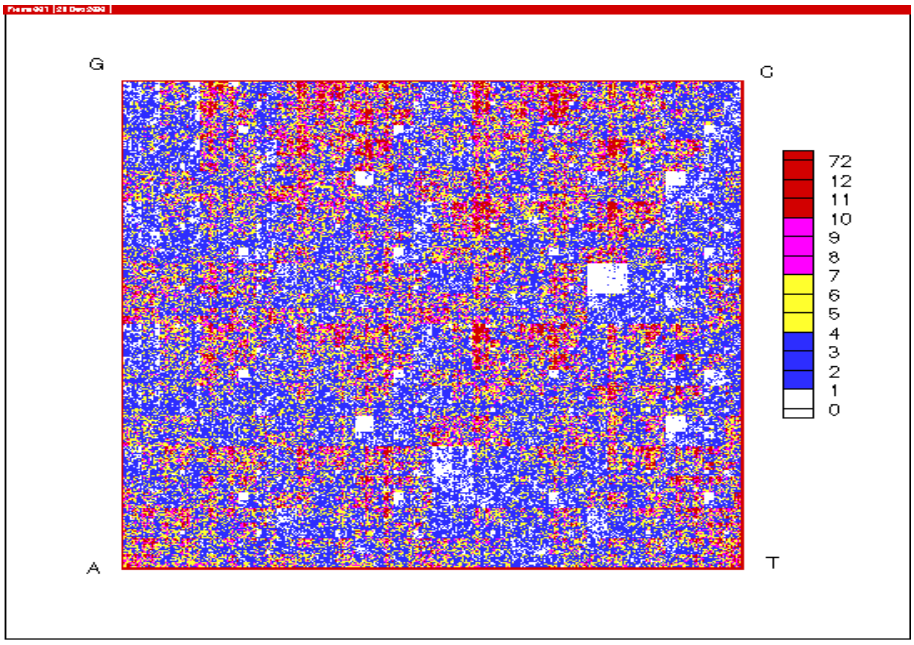

图 5.8: 大肠杆菌 *E.coli* 全基因组编码区字长为 9 时的词汇频率地形图。

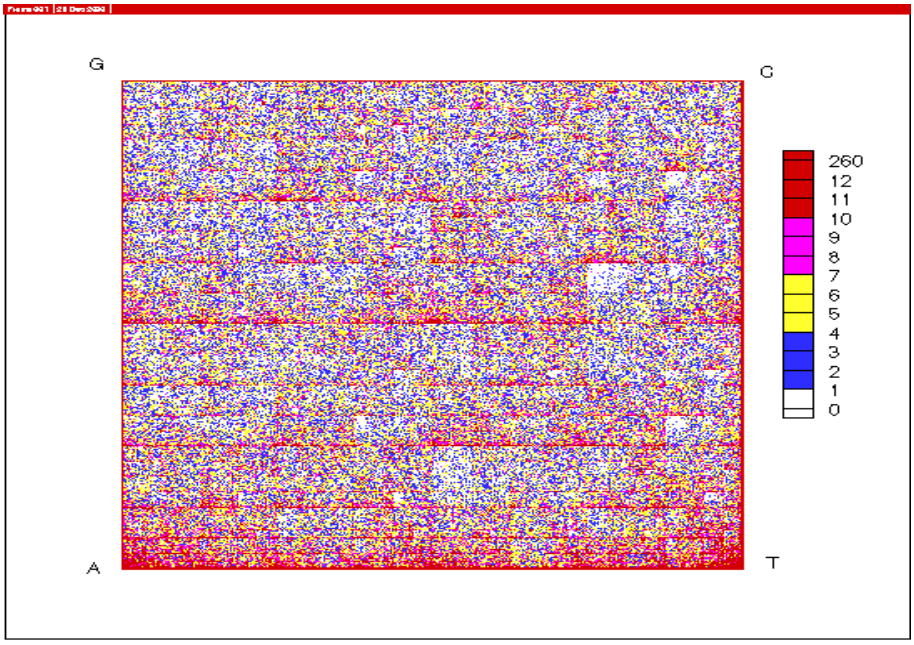

图 5.9: 大肠杆菌 *E.coli* 全基因组非编码区字长为 9 时的词汇频率地形图。

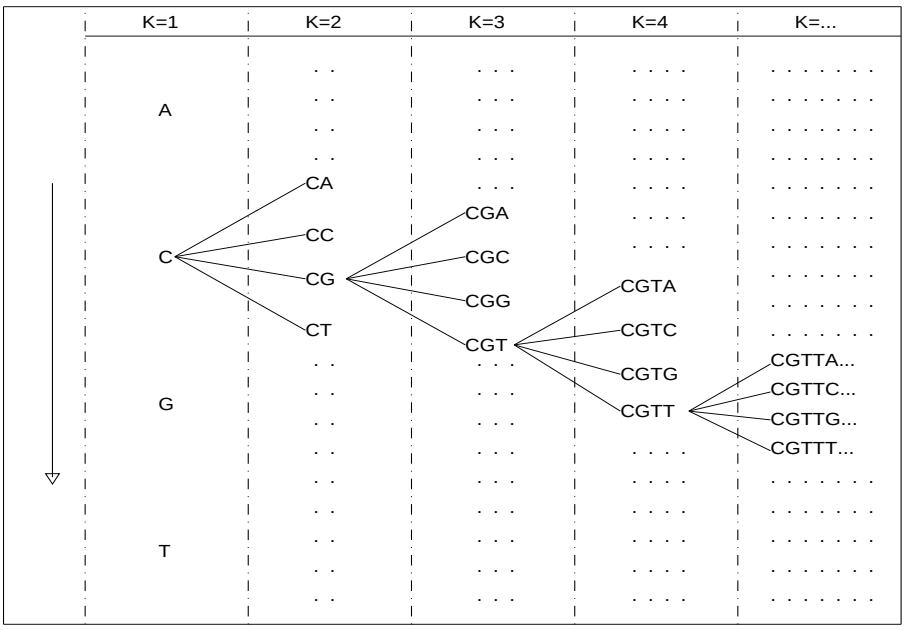

图 5.10: 不同字长 K 的字典组成的字典层次，每个层次的字典内的单词排序均为从上向下，层次从左向右递增。注意第 K 个层次的字典可以通过对第 $K+1$ 层次的字典中相邻四个字求和得到，但反过来不行。

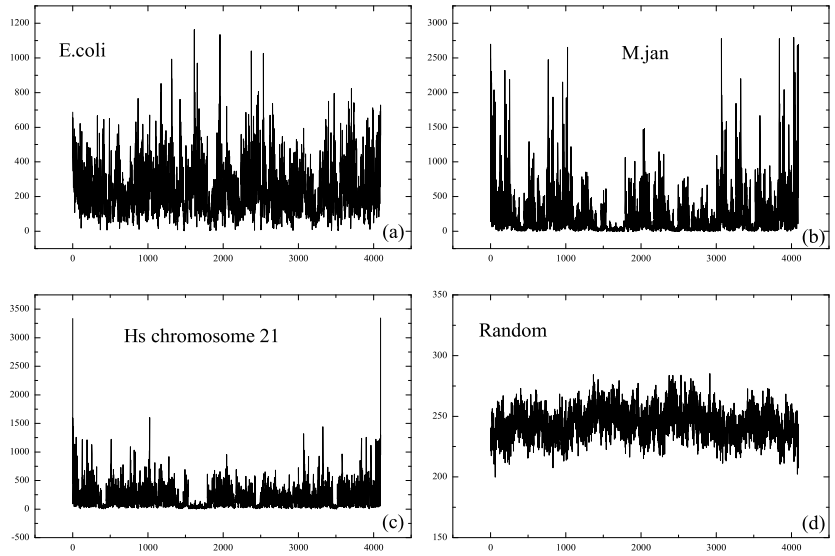

图 5.11: 四种不同序列 *E.coli* 全基因组序列， *M.jan* 全基因组序列，人类第 21 条染色体序列，以及随机序列的字长为 6 的词汇频率字典。注意我们没有将四个信号的纵轴设为统一的标度，主要是因为各自涨落相差太大的缘故。

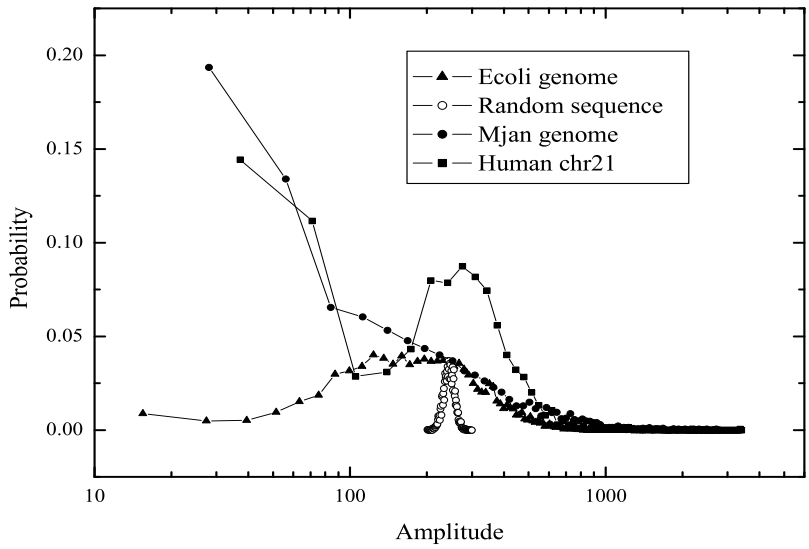

图 5.12: 图 5.11 中四个信号涨落分布的概率密度函数。可以看出随机序列具有最小的涨落，而人类第 21 条染色体序列的涨落最大。此图的纵坐标已经取了对数。

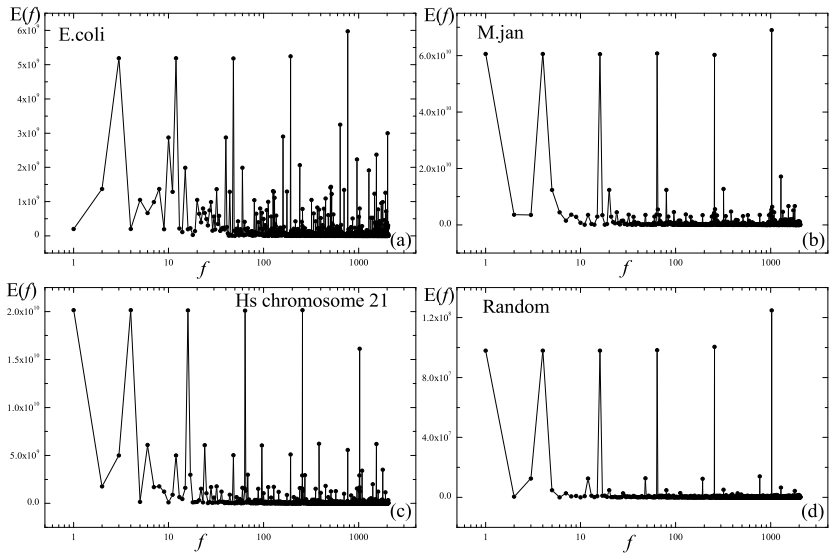

图 5.13: 图 5.11 中四个涨落信号的功率谱分析比较。图中的峰值结构对应于 WFD 信号中的周期行为。为了便于观察，我们将横坐标取了对数。

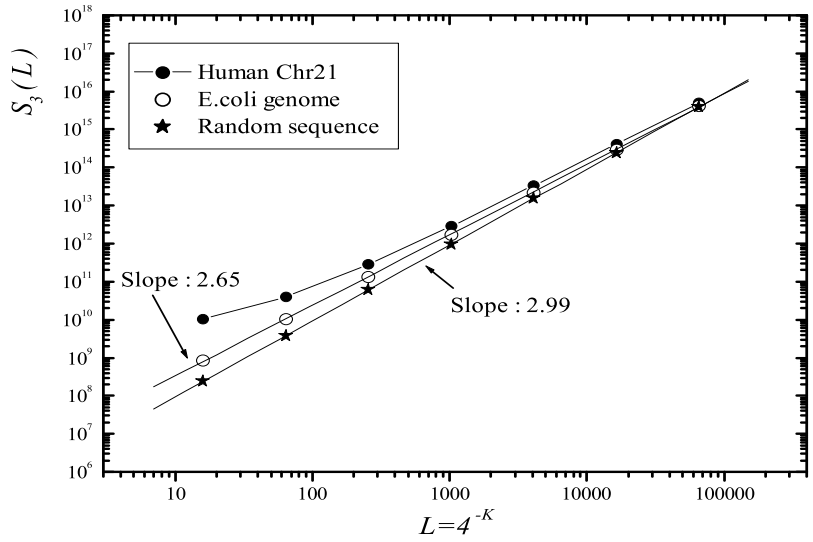

图 5.14: 多层次词汇频率字典的标度律分析。图中取的是 3 阶矩对尺度 $L = 4^{-K}$ 的标度律关系。最小二乘拟合得到 *E.coli* 全基因组序列和随机序列的标度指数分别为 2.65 和 2.99 。

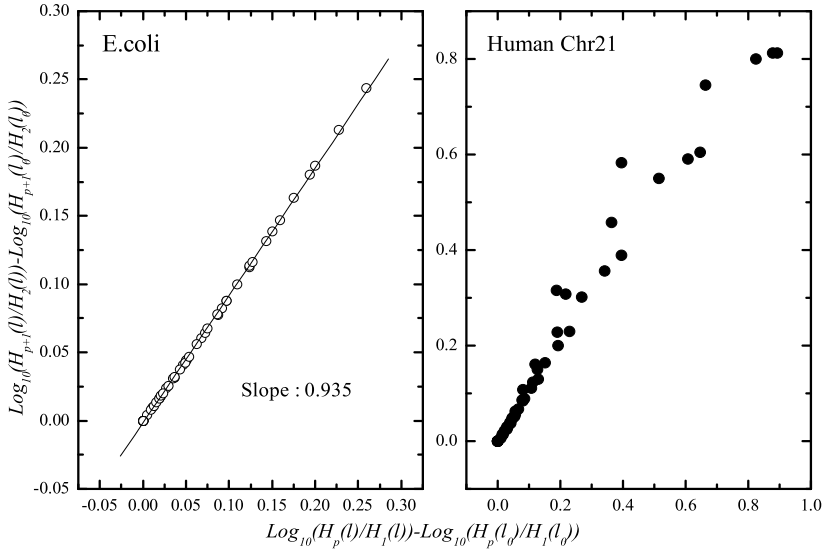

图 5.15: 多层次词汇频率字典的 β- 检验，左右两图使用的数据分别是 *E.coli* 全基因组序列和人类第 21 条染色体序列。对于左图，我们拟合得斜率为 0.935 ；右图的散乱的点表明 β- 检验没有通过。

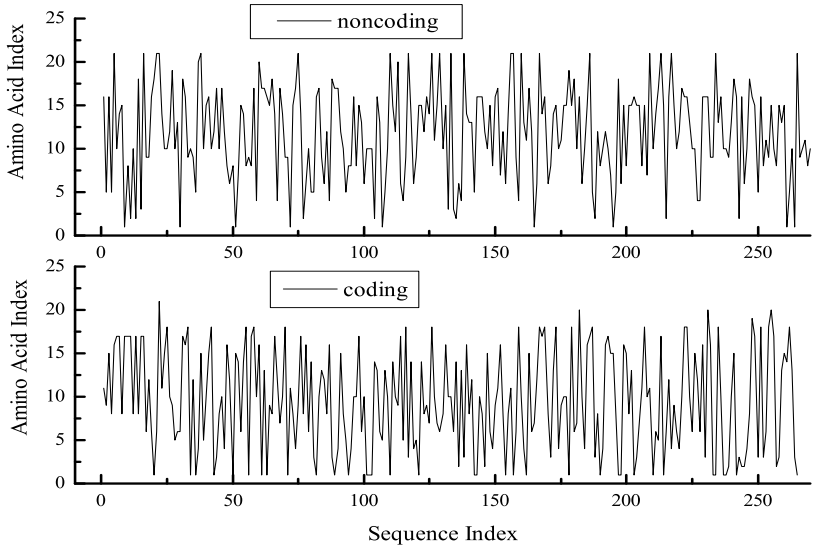

图 5.16: 大肠杆菌 *E.coli* 非编码（上图）和编码（下图）序列产生的伪氨基酸和氨基酸序列。纵坐标为氨基酸代号，横坐标为序号

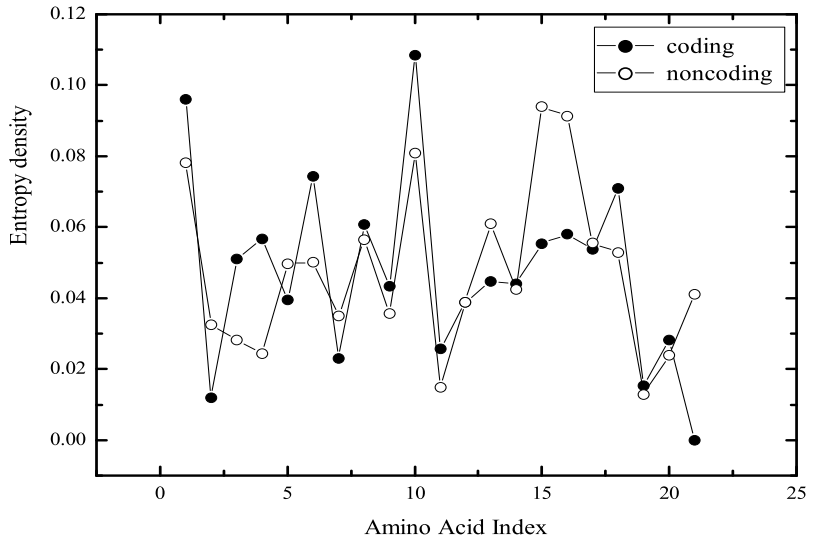

图 5.17: 大肠杆菌 *E.coli* 编码（黑点）和非编码区（圆圈）的平均熵密度廓线（ EDP ）。注意第 21 个指标代表终止密码子。

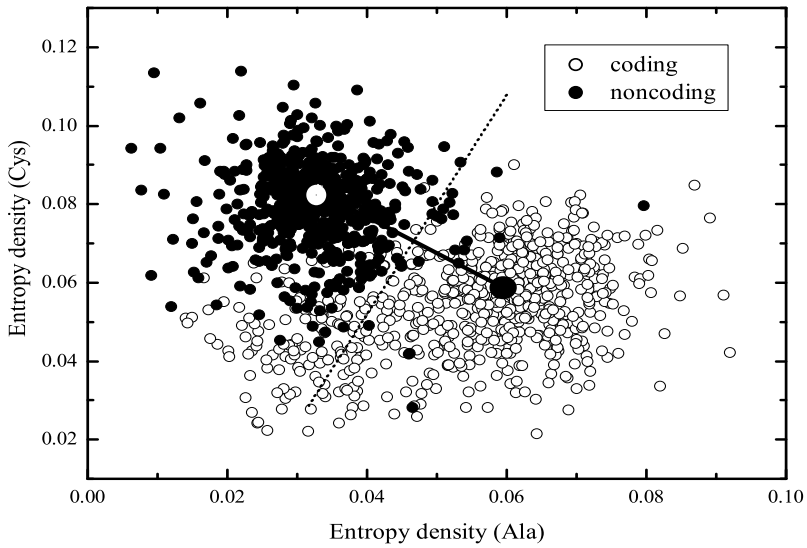

图 5.18: 二维熵密度空间的相点分布，横纵坐标分别为为氨基酸 Ala 和 Cys 的熵密度。数据取自大肠杆菌 *E.coli* 全基因组的编码和非编码部分，各 500 个相点，长度均大于 100 个氨基酸。图中大的黑点和白点分别是编码和非编码区中心。实线是两个中心的连线，虚线是其垂直平分线。

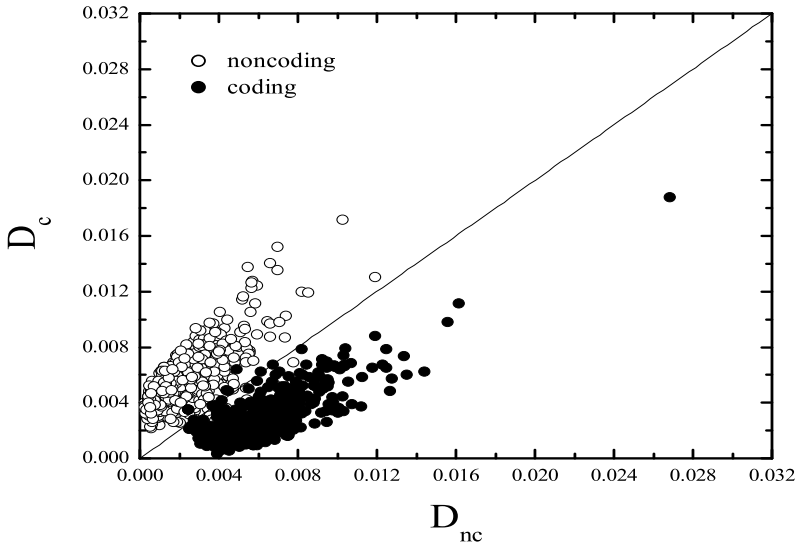

图 5.19: 大肠杆菌 *E.coli* 编码（黑点）和非编码区（圆圈）序列分别距两个中心的距离比较。纵坐标为距编码中心的距离，横坐标为距非编码中心的距离，斜线为角平分线。这里所取的编码和非编码的长度均大于 100 个氨基酸，双方各 500 条序列

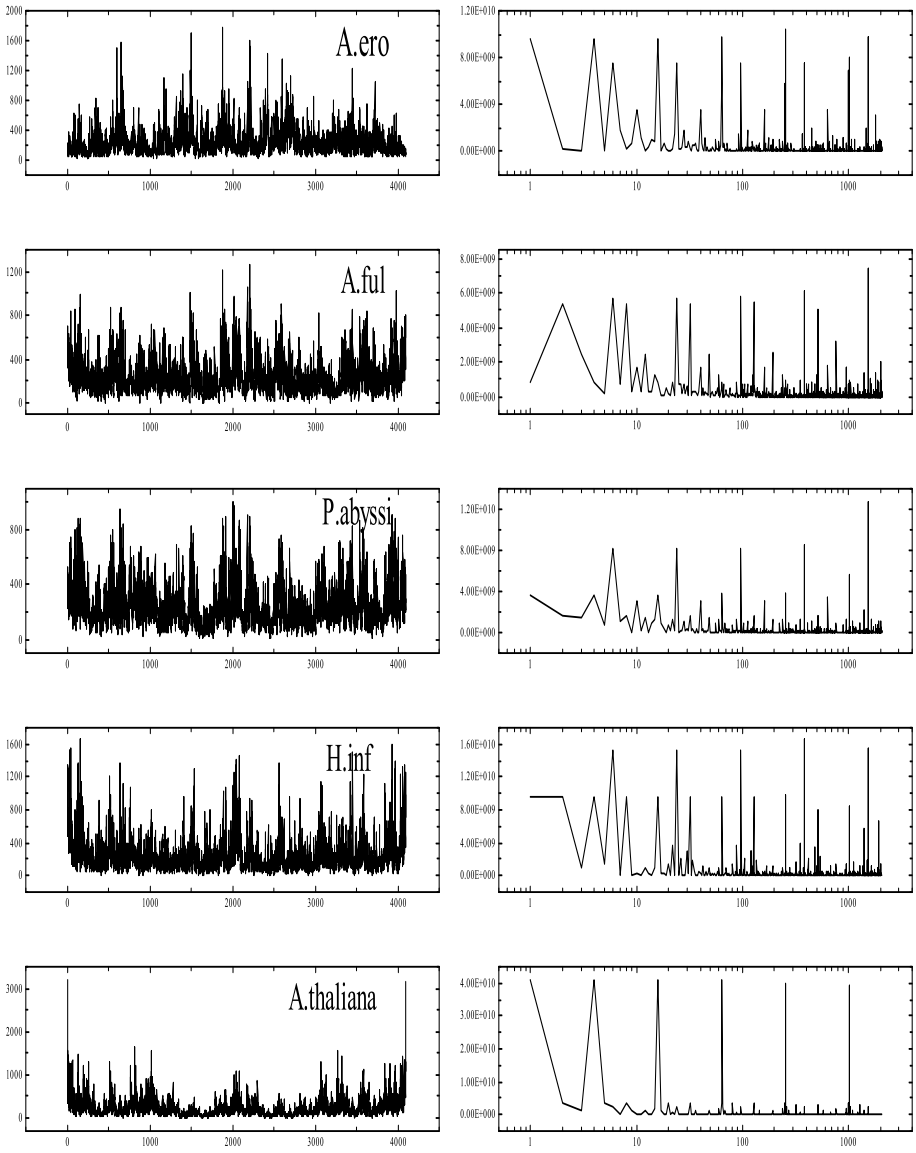

图 5.20: 一些典型生物 DNA 序列的词汇频率字典及其功率谱分析。

第六章　简单的总结

到现在为止，本文的主要结果都已经陈述完毕，我们可以简单地总结一下了。

在这篇论文里，我们以最近提出的层次结构模型 [91, 92] 为基础，发展了对 Taylor-Couette 流动系统，时空广延系统 (包括耦合映像格子和沙堆模型) 以及生物体的 DNA 序列等几个典型的复杂系统的多尺度分析方法。我们分析的对象既有理论模型 (耦合映像格子和沙堆模型) 又有实验数据 (Taylor-Couette 流速度信号， DNA 序列) 。

首先，在本文的第二章里，我们以湍流为例比较详细地介绍了标度律研究的基本概念 (结构函数，标度律等) 和充分发展湍流的层次结构模型的基本思路及其基本分析工具 (β- 检验和 γ- 检验) ，并简要地讨论了层次相似性参数 β 和最高激发态的标度指数 γ 的物理意义。

接下来，在第三章里，通过对一个典型的复杂流动系统 -Taylor-Couette 流的速度信号的详细分析，展现利用层次结构模型分析具体系统的的实际过程，揭示层次结构模型在描述和刻划现象，发现新规律方面的可能应用潜力。这一章的主要结果如下：我们分析了不同雷诺数下 Taylor-Couette 流的速度信号，引进从概率分布函数去除噪声的方法，通过扩展的自相似律 ESS 的方法求得标度指数 ζ_p ，并且对 $p \leq 10$ 的情况，我们发现 ζ_p 在我们的样本数下已经收敛。 Taylor-Couette 流的速度数据可以由层次结构模型很好地描述，以很好的线性度通过 β- 检验和 γ- 检验。 β- 检验得到的值与雷诺数无关，表明联系大小尺度 (l)，不同强度 (p) 涨落的机制是普适的，与雷诺数无关。 β 值 0.83 比在自由射流模拟，和尾迹湍流中测得的 $\beta \approx 0.87$ 小，显示了 Taylor-Couette 流系统具有更强的间歇性。我们还发现，这里测得的参数 γ 值在 $R = 10^5$ 左右时从 0.14 变到 0.11 ，表明最高激发态结构从 $R = 10^5$ 左右开始变得更加奇异了，这与实验中所观察到的有序 Taylor 涡的破碎相对应。这些统计分析令人鼓舞地给出了流体结构的演化图象。

从第四章开始，我们走出物理的湍流系统，向其他复杂系统推广以层次结构分析为主线的多尺度分析。我们主要考察了两个典型的时空广延动力系统，耦合映像格子(CML) 和自组织临界性 (SOC) 概念的经典模型， BTW 沙堆模型的统计动力学行为和系统尺寸之间的关系。我们发现：在不同的参数下 GCM 模型的平均场涨落满足相对标度律 ESS 关系，相对标度指数 $\tau_{p,3}$ 不是普适的，倚赖于模型参数 a 和 η 。这些参数与系统的混沌性质有关。因此相对标度指数也与系统的混沌性质有关。平均场涨落可以被层次结

构模型很好地描述，高信度地通过 β- 检验和 γ- 检验。同样参数 β 和 γ 也和系统的混沌性质有关。对二维 BTW 模型的矩分析表明，畴域大小涨落与系统尺寸的标度律关系是非平凡的。标度指数 ζ_p 在 p 小于 1.2 时是 p 的非线性函数， p 较大时近似为线性函数，表明这里的统计是非高斯的，而不是高斯的。层次相似性参数 β 比在其他系统测到的都小，反映了 SOC 动力学过程的强间歇性。层次结构模型给出了对两类系统的定量的刻划。

第五章关于 DNA 序列的分析使我们相信，在对此类复杂系统的处理过程中，多尺度，多参数的概念对我们是大有益处的。高维熵空间序列分析的初步成功证明了复杂系统研究中引进新的处理方法的重要性。本章的主要内容分为三个部分：第一部分是关于 DNA 序列碱基密度分布的层次结构分析。我们发现生物体基因组序列中碱基密度分布具有多尺度结构，并且是强间歇的。在从基因内（ $100bp$ ）到基因组（ $8000bp$ ）之间的尺度上，存在扩展的自相似律，同时层次相似律也得到很好的验证。这反映了碱基分子密度分布在长期的进化过程中演化到一定的统计自组织状态。层次相似参数将是对其统计状态的定量刻划。我们发现，原核生物基因组序列的 β 值比真核生物 (如人类) 的大，反映了序列在某种意义上为多样性更丰富的多尺度结构。通过对 DNA 序列词汇使用频率的研究我们发现，细菌类 DNA 序列不同长度的词汇使用频率字典之间有一定的相关性，存在很好的标度律关系，然而在真核生物的 DNA 序列中却不存在这种关系；原核生物 DNA 序列不同尺度的字典中词汇使用频率的涨落满足层次相似律，真核生物中又不存在这种关系， β- 检验难以被通过，说明从语言学的角度来看，原核生物和真核生物的 DNA 序列在词汇使用频率这一点上存在较大的差别。我们认为，这种差别实际上反映 DNA 序列编码和非编码部分在生物进化过程的演化机制的差别。

在以上几种典型的复杂系统中，层次结构模型都较好地成立，表明该模型一定程度上的普适性。说明层次结构模型可以被用来刻划这一类系统中的多尺度涨落。尽管这种刻划是比较唯象的，但其仍将有助于我们理解产生这些多尺度涨落的物理机制。层次结构模型的参数具有一定的物理意义，但不应该孤立地去理解它们。广义地讲，参数 β 度量了系统的间歇性和自组织性；而 γ 则度量了系统的最高激发态结构的奇异性。然而，一旦涉及到具体的系统，必须联系实际去理解这些参数，特别是讨论的物理量已经与原始的层次结构模型的物理量不同的时候。层次结构的普适性鼓励我们将其推广到更一般的复杂系统中去。

第五章第三部分的研究方法与前面几章略有不同。这里我们引进了一个刻画 DNA 序列的新方法 – 多变量熵距离法 (MED 法) 。我们证明， MED 方法对 DNA 编码序列的发现具有非凡的效率 (平均综合得分 >98%)。我们提出， MED 方法可能是对复杂系统描述的普遍适用的一种方法。

本文一直围绕着层次结构分析这一主要线索，提倡对复杂系统多尺度，多层次的分析方式，最后又试图跳出层次结构的框架，提出对类似于生物系统的复杂系统采取多参数的高维空间描述的想法，显示了作者对探讨复杂系统认识的一步步的深入过程。

参考文献

[1] F. Anselmet, Y. Gagne, E. J. Hopfinger, and R. A. Antonia, *High order velocity structure functions in turbulent shear flows, J. Fluid Mech.*, 1984, 140: 63-89

[2] A. Arneodo, E. Bacry, P.V. Graves and J.F. Muzy, *Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis, Phys. Rev. Lett.*, 1995, 74: 3293-3296

[3] A. Arneodo *et al.*, *Structure functions in turbulence, in various flow configurations, at Reynolds number between 30 and 5000, using extended self-similarity, Europhys. Lett.*, 1996, 34:411-416

[4] P. Bak, C. Tang and K. Wiesenfeld, *Self-Organzied Criticality: An Explanation of 1/f noise, Phys. Rev. Lett.*, 1987, 59: 381-384

[5] P. Bak, C. Tang and K. Wiesenfeld, *Self-Organzied Criticality, Phys. Rev. A*, 1988, 38: 364-374

[6] P. Bak, K. Chen and C. Tang, *A forest-fire model and some thoughts on turbulence, Phys. Lett. A*, 1990, 147: 297-300

[7] P. Bak, *How nature works: The Science of Self-Organzied Criticality*, 1996, New York: Copernicus, 212p.

[8] P. Baldi *et al*, *Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics*, 2000, 16: 412-424

[9] F. Belin, P. Tabeling, H. Willaime, *Exponents of the structure functions in a low temperature helium experiment, Physica D*, 1996, 93: 52-63

[10] A. Ben-Hur and O. Biham, *Universality in sandpile models, Phys. Rev. E*, 1996, 53: R1317-R1320

[11] R. Benzi, S. Ciliberto, R. Tripiccione, C. Baudet and S. Succi, *Extended self-similarity in turbulent flows, Phys. Rev. E*, 1993, 48: R29-R32

[12] R. Benzi, S. Ciliberto, C. Baudet, and G. Ruiz Chavarria, *On the scaling of 3-dimensional homogeneous and isotropic turbulence, Physica D*, 1995, 80: 385-398

[13] R. Benzi, L. Biferale, S. Ciliberto, M. V. Struglia and R. Tripiccione, *Generalized scaling in fully developed turbulence, Physica D* 1996, 96: 162-181

[14] R. Benzi, L. Biferale, S. Ciliberto, M. V. Struglia and R. Tripiccione, *Scaling property of turbulent flows, Phys. Rev. E*, 1996, 53: R3025-R3027

[15] P. Bernaola-Galvan, J.L. Oliver and R. Roman-Roldan, *Decomposition of DNA Sequence Complexity*, 1999, 83: 3336-3339

[16] A. Bershadskii, *Generalized scaling in nonscaling diffusion, Physica A*, 2000, 278: 497-503

[17] F.R. Blattner *et al.*, *The Complete Genome Sequence of Escherichia Coli, Science*, 1997, 277: 1453-1462

[18] T. Bohr, M.H. Jensen, G. Paladin and A. Vulpiani, *Dynamical systems approach to turbulence*, 1998, Cambridge University Press.

[19] M. Borodovsky and J. McIninch, *GENMARK: Parallel gene recognition for both DNA strands, Comput. Chem.*, 1993, 17: 123133

[20] N.N. Bugaenko, A.N. Gorban and M.G. Sadovsky, *Maximum Entropy Method in Analysis of Genetic Text and Measurement of its Information Content , Open Sys. and Information Dyn.* 1998, 5: 265-278

[21] S.V. Buldyrev, A.L. Goldberger, S. Havlin *et al*, *Long-range correlation properties of coding and noncoding DNA sequence: Genbank analysis, Phys. Rev. E*, 1995, 51: 5084-5091

[22] N.-Z. Cao, S.-Y. Chen, Z.-S. She, *Scaling and Relative Scaling in the Navier-Stokes Turbulence, Phys. Rev. Lett.*, 1996,76: 3711-3714

[23] D. Chatzidimitrion-Dreismann, *Long-range correlations in DNA, Nature*, 1993, 361: 212-213

[24] A. Chessa, H.E. Stanley, A. Vespignani, and S. Zapperi, *Universality in sandpiles, Phys. Rev. E*, 1999, 59: R12-R15

[25] A. Corral and A. Díaz-Guilera, *Symmetries and fixed point stability of stochastic differential equations modeling self-organized criticality, Phys. Rev. E*, 1997, 55: 2434-2445

[26] M. Crochemore and R. Verin, *Zones of low entropy in genomic sequences, Computers Chem.*, 1999, 23: 275-282

[27] M.C. Cross and P.C. Hohenberg, *Pattern formation outside of equilibrium, Rev. Mod. Phys.*, 1993, Vol. 65, No. 3

[28] B.D. Davis *et al.*, *The Human genome and other inititives, Science*, 1990, 249: 342-343

[29] M. De Menech, A.L. Stella and C. Tebaldi, *Rare events and breakdown of simple scaling in the Abelian sandpile model, Phys. Rev. E*, 1998, 58: R2677-R2680

[30] S. Dong, and D.B. Searls, *Gene structure prediction by linguistic methods, Genomics*, 1994, 23: 540551

[31] B. Dubrulle, *Intermittency in Fully Developed Turbulence: Log-Poisson Statistics and Generalized Scale Covariance, Phys. Rev. Lett.*, 1994, 73: 959-962

[32] L. Dunham, N. Shimizu, B.A. Roe *et al*, *The DNA sequence of human chromosome 22, Nature*, 1999, 402: 489-495

[33] D. Feng and G. Jin, New perspective on condensed matter physics , Shanghai Scientific & Technical Publishers, 1992, 422p.

[34] J.W. Fickett, *Recognition of protein coding regions in DNA sequences, Nucleic Acids Res.*, 1982, 10: 53035318

[35] F. Flam, *Hints of a language in Junk DNA, Science*, 1994, 266: 1320

[36] U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov*, 1995, Cambridge University Press.

[37] U. Frisch, M. Nelkin, P.-L. Sulem, *A simple dynamical model of intermittent fully developed turbulence, J. Fluid Mech.*, 1978, 87: 719-736

[38] C. Gautier, *Compositional bias in DNA, Curr. Opin. Genet. Dev.*, 2000, 10: 656-661

[39] http://www.ncbi.nlm.nih.gov/

[40] M.R. Gerald, D.Y. Mark, R.W. Jennifer *et al*, *Comparative Genomics of the Eukaryotes, Science*, 2000, 287: 2204-2215

[41] N. Goldenfeld & L.P. Kadanoff, *Simple Lessons from Complexity, Science*, 1999, 284: 87-89

[42] A.N. Gorban T.G. Popova and M.G. Sadovsky, *Classification of Symbolic Sequence over Their Frequency Dictionaries: Towards the Connection Between Structure and Natural Taxonomy, Open Sys. and Information Dyn.*, 2000, 7: 1-17

[43] H.L. Grant, R.W. Stewart, and A Moilliet, *Turbulent spectra from a tidal channel, J. Fluid Mech.*, 1962, 12: 241-268

[44] P. Grassberger, *Toward a quantitative theory of self-generated complexity, Inter. J. Theor. Phys.*, 1986, 25: 907-938

[45] I. Grosse, H. Herzel, S.V. Buldyrev and H.E. Stanley, *Species independence of mutual information in coding and noncoding DNA, Phys. Rev. E*, 2000, 61: 5624-5629

[46] P. Gtaziano, A. Marcella and S. Cecilia, *Linguistic Analysis of Nucleotide Seuences: Algorithms for Pattern Recognition and Analysis of Codon Strategy, Method in Enzymology*, 1996, 266: 281-294

[47] R. Guigo, *Computational gene identification: an open problem, Computers Chem.*, 1997, 21: 215-222

[48] V.D. Gusev, L.A. Nemytikova and N.A. Chuzhanova, *On the complexity measures of genetic sequences, Bioinformatics*, 1999, 15: 994-999

[49] B.L. Hao, H.C. Lee and S.Y. Zhang, *Fractals related to long DNA sequences and complete genomes, Chaos, Solitons and Fractals*, 2000, 11: 825-836

[50] *International Journal of Bifurcations and Chaos*, 1992, Vol 2., No 3.

[51] K. Kaneko, *Period-doubling of kink-antikink patterns, quasiperiodicity in antiferro-like structures and spatial intermittency in coupled logistic lattice: towards a preclude of a "field theory of chaos", Prog. Theor. Phys.*, 1984, 72: 480

[52] K. Kaneko, *Collapse of Tori and Genesis of Chaos in Dissipative System*, 1986, World Scientific Publishing Co. Pte Ltd.

[53] K. Kaneko, *Pattern dynamics in spatialtemporal chaos, Pattern selection, diffusion of defect and pattern competition intermittency, Physica D*, 1989, 34: 1; Spatialtemporal chaos in one-and two-dimensional coupled map lattices, Physica D, 1989, 37: 60

[54] K. Kaneko, *Globally Coupled Chaos Violates the Law of Large Numbers but Not the Central-Limit Theorem, Phys. Rev. Lett.*, 1990, 65: 1391-1394

[55] S. Karlin and V. Brendel, *Patchiness and correlations in DNA sequences, Science*, 1993, 259: 677-680

[56] O.V. Kirillova, *Entropy concepts and DNA investigations, Phys. Lett. A*, 2000, 274: 247-253

[57] O.V. Kirillova, *Comparative statistical analysis of bacteria genomes in "word" context, Physica A*, 2001, 290: 453-463

[58] A.N. Kolmogorov, *Local structure of turbulence in a viscous incompressible fluid at high Reynolds number, Proc. R. Soc. Lond. A*, 1991, 434: 9-13

[59] A.N. Kolmogorov , *A refinement of previous hypothesis concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number, J Fluid Mech.*, 1962, 13: 82-85.

[60] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, *Hidden Markov models in computational biology: application to protein modeling, J. Mol. Biol.*, 1994, 235: 15011531

[61] E. Lévêque, Ruiz-Chavarria G, Baudet C, Ciliberto S, *Scaling laws for the turbulent mixing of a passive scalar in the wake of a cylinder, Phys. Fluids*, 1999, 11: 1869-1879.

[62] B. Lewin, *Genes VII*, 2000, Oxford University Press.

[63] G.S. Lewis, Ph.D. dissertation, 1996, University of Texas at Austin.

[64] G.S. Lewis & H.L. Swinney, *Velocity structure functions, scaling, and transitions in high-Reynolds number Couette-Taylor flow, Phys. Rev. E*, 1999, 59: 5457-5467

[65] W. Li, *The study of correlation structures of DNA sequences: a critical review, Computers Chem.*, 1997, 21: 257-271

[66] S.D Liu, *Most intermittent structures and invariance principle of multifractal field*, 2001, preprint

[67] X. Lu, Z. Sun, H. Chen and Y. Li, *Characterizing self-similarity in bacteria sequences, Phys. Rev. E*, 1998, 58: 3578-3584

[68] S. Lübeck, *Moment analysis of the probability distribution of different sandpile models, Phys. Rev. E*, 2000, 61: R204-R209

[69] B. Mandelbrot, *Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier, J. Fluid Mech.*, 1978, 62: 331-358

[70] S.S. Manna, *Large scale simulation of avalanche cluster distribution in sandpile model, J. Stat. Phys.*, 1990, 59: 509-521

[71] S.S. Manna, *Two-state model of self-organized criticality, J. Phys. A: Math. Gen.*, 1991, 24: L363-L369

[72] R.N. Mantegna, S.V. Buldyrev and A.L. Goldberger *et al*, *Liguistic Features of Noncoding DNA Sequences, Phys. Rev. Lett.*, 1994, 73: 3169-3172

[73] C. Meneveau and K. Sreenevasan, *Simple multifractal cascade model for fully developed turbulence, Phys. Rev. Lett.*, 1987, 59: 1424-358

[74] A.K. Mohanty and A.V.S.S. Narayana Rao, *Factorial Moments Analysis Show a Characteristic Length Scale in DNA Sequences, Phys. Rev. Lett.*, 2000, 84: 1832-1835

[75] A.S. Monin and A.M. Yaglom, *Statistical Fluid Mechanics, Vol. 1* ed. J. Lumley, MIT press, Cambridge, MA, 1971

[76] A.S. Monin and A.M. Yaglom, *Statistical Fluid Mechanics, Vol. 2* ed. J. Lumley, MIT press, Cambridge, MA, 1975

[77] S. Morita, *Scaling law for the Lyapunov spectra in globally coupled tent maps, Phys. Rev. E*, 1998, 58: 4401-4412

[78] S. Nee, *Uncorrelated DNA walks, Nature*, 1992, 357: 450

[79] G. Parisi and U. Frisch, in *Turbulence and Predictability of Geophysical Flows and Climate Dynamics* ed. N. Ghil, R. Benzi and G. Parisi, 1985, P.84

[80] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger *et al.*, *Long-range correlations in necleotide sequences, Nature*, 1992, 356: 168-170

[81] L. Pietronero94, A. Vespignami and S. Zapperi, *Renormalization scheme for self-organized criticality in sandpile models , Phys. Rev. Lett.*, 1994, 72: 1690-1693

[82] H. Politano and A. Pouquet, *Model of intermittency in magnetohydrodynamic turbulence, Phys. Rev. E*, 1995, 52: 636-641

[83] V.V. Prabhu and J.M. Claverie, *Correlations in intronless DNA, Nature*, 1992, 359: 782

[84] D. Queiros-Conde, *Geometrical Extended Self-Similarity and Intermittency in Diffusion-Limited Aggregates, Phys. Rev. Lett.*, 1997, 78: 4426-4429

[85] R. Roman-Roldan R, P. Bernaola-Galvan and J.L. Oliver, *Sequence Compositional Complexity of DNA through Entropic Segmentation Method, Phys. Rev. Lett.*, 1998, 80:1344-1347

[86] G. Ruiz Chavarria, C. Baudet and S. Ciliberto,*Hierarchy of Energy Dissipation Moments in Fully Developed Turbulence, Phys. Rev. Lett.*, 1995, 74: 1986-1989

[87] G. Ruiz Chavarria, C. Baudet, R. Benzi and S. Ciliberto, *Hierarchy of Velocity Structure Functions in Fully Developed Turbulence, J. de Phys. II*, 1995, 5: 485-490

[88] G. Ruiz Chavarria, C. Baudet, S. Ciliberto, *Scaling laws and dissipation scale of a passive scalar in fully developed turbulence, Physica D*, 1996, 99: 369-380

[89] S. Salzberg, *Locating protein coding in human DNA using a decision tree algorithm, J. Comput. Biol.*, 1995, 2: 473485

[90] Z.-S. She, E. Aurell and U. Frisch, *The Inviscid Burgers Equation with Initial Data of Brownian Type, Comm. Math. Phys.*, 1992, 148: 623-641

[91] Z.-S. She & E. Lévêque, *Universal Scaling Laws in Fully Developed Turbulence, Phys. Rev. Lett.*, 1994, 72: 336-339

[92] Z.-S. She, E.C. Waymire, *Quantized Energy Cascade and Log-Poisson Statistics in Fully Developed Turbulence, Phys. Rev. Lett.*, 1995, 74: 262-265

[93] Z.-S. She, *Universal Law of Cascade of Turbulence Fluctuations, Prog. Theor. Phys.*, 1998, S130: 87-102

[94] Z.-S. She and L. Liu, *"Measuring intermittency parameters in turbulence"*, submitted to Phys. Rev. E. (2000); L. Liu and Z.-S. She, *"Quantifying intermittent structures of turbulence"*, submitted to Phys. Fluids (2000).

[95] J.C. Shepherd, *Method to determine the reading frame of a protein from the*

purine/pyrimidine genome sequence and its possible evolutionary justification, Proc. Natl. Acad. Sci., 1981, 78: 15961600

[96] Ya.G. Sinai, *Statistics of shocks in solution of inviscid Burgers equation, Comm. Math. Phys.*, 1992, 148: 601

[97] K.R. Sreenivasan & R.A. Antonia, *The phenomenology of small-scale turbulence, Annu. Rev. Fluid Mech.*, 1997, 29: 435-472

[98] 佘振苏, 苏卫东, 湍流脉动的层次结构描述, 湍流研究最新进展, 科学出版社, 2001

[99] A. Turiel , G. Mato, N. Parga and J.P. Nadal, *Self-Similarity Properties of Natural Images Resemble Those of Turbulent Flows, Phys. Rev. Lett.*, 1998, 80: 1098-1101

[100] D.L. Turcotte, *Self-organized criticality, Rep. Prog. Phys.*, 1999, 62: 1377-1429

[101] A. Vespignami, S. Zapperi and L. Pietronero94, *Renormalization approach to the self-organized critical behavior of sandpile models, Phys. Rev. E*, 1995, 51: 1711-1724

[102] R.F. Voss, *Evolution of Long-Range Fractal Correlation and 1/f Noise in DNA Base sequences, Phys. Rev. Lett.*, 1992, 68: 3805-3808

[103] N. Wang and R. Chen, *Comparative analysis of phylogeny based on intron and exon, Chinese Science Bulletin*, 1999, 44: 2095-2102

[104] W.-M. Yang, *Spatialtemporal Chaos and Coupled Map Lattics, Advanced Series in Nonlinear Science.*, 1994, Shanghai Scientific and Technological Education Publishing House, Shanghai.

[105] J.M. Yoon, PhD Thesis, 1999, University of California, Los Angeles.

[106] R. Zhang and C.-T. Zhang, *Z curves, an intuitive tool for visualizing and analyzing the DNA sequences, J. Biomol. Struct. Dyn.*, 1994, 11: 767-782

[107] 邹正平, 湍流层次结构理论的实验研究, 2001, 湍流及复杂系统国家重点实验室博士后研究报告, 111p.

致谢

我非常感谢我的导师刘式达教授和佘振苏教授在我研究生期间给我的指导和一贯的支持。这三年里他们不仅关心我的学习和工作，而且生活上也给我许多关怀。两位老师严谨的治学态度，活跃的创新思维令我终生难忘。他们勇于探索的精神激励作者去争取更大的进步。

我也要感谢北大力学系的苏卫东教授，地球物理系的刘式适教授。前者在我刚刚进入湍流研究这个新领域的时候给了我大量的帮助和指导。后者给我介绍了非线性科学，特别是非线性波动的的最新知识，使作者受益匪浅。

R. D'Hulst 教授和 G.J. Rodgers 教授给了作者一些关于他们最近工作的单行本，为作者更好地理解自己的工作提供了很好的思路。南京大学生化系的王进教授介绍给作者很多与生物学有关的知识，并且给了作者很多鼓励。

我非常高兴能够遇到美国 Tennessee 大学的吴介之教授， Arizona 大学的周明德教授，以及北大力学系的蔡庆东教授。从他们身上我学到了许多专业知识和做人及做学问的道理。

必须提及，作者这三年在北京大学的学习过程中，得到了湍流及复杂系统国家重点实验室及地球物理学系的大力支持和帮助。作者在湍流室的机群系统上进行了本文的部分计算，一并致谢。

作者还希望向以下各位表示感谢：付遵涛博士，邹正平博士，诸元杰，王龙（我们一起度过了一段难忘的时光）；熊鳌魁博士，程雪玲博士，欧阳正清，杨岩，郭昊（他们同作者进行了许多有益的讨论）；李晨光，李黎明，王在文（他们知道我为什么要感谢他们）以及梁允宽博士（我多年的良师益友）。

最后，我要感谢我的父母，是他们的辛劳培育着我长大，并支持我完成了学业。我为他们感到自豪。

北京大学　　任奎

二 OO 一年五月三十一日