离散稳恒信号的多重分形谱的计算及其应用

陈双平[①②]　郑浩然[②]　刘金霞[②]　童　庆[②]　王煦法[②]
[①](中国科学技术大学电子工程与信息系统系　合肥　230027)
[②](中国科学技术大学计算机科学与技术系　合肥　230027)

摘　要：对于未知信号而言，一般将其视为稳恒信源的输出。因而，利用统计的方法计算信源输出信号的多重分形谱，与理论上计算的结果加以比较，据此就可以判断信源模型参数估计的合理性。该文给出了计算信号多重分形谱的一般方法，并且探讨了计算过程中的相关问题。并将该方法应用于染色体中碱基序列的分析中，实验结果表明，在某种程度上，碱基序列可视为某个离散稳恒信源的输出。。

关键词：离散稳恒信源；多重分形谱；脱氧核糖核酸(DNA)序列；统计分析
中图分类号：TN911.7　　**文献标识码**：A　　**文章编号**：1009-5896(2007)05-1054-04

Computation and Applications of Multi-fractal to Discrete Stationary Signals

Chen Shuang-ping[①②]　Zheng Hao-ran[②]　Liu Jin-xia[②]　Tong Qing[②]　Wang Xu-fa[②]
[①](*Department of Electronic Engineering and Information System*, *University of Science and Technology of China* (*USTC*), *Hefei* 230027, *China*)
[②](*Department of Computer Science and Technology*, *USTC*, *Hefei* 230027, *China*)

Abstract: Unknown signals are always be treated as outputs of stationary information sources which are easy to be dealt with. So, it is possible to compute multi-fractal spectrum of the signals, which are compared with theoretical results to verify whether the estimation of parameters for a model is correct or not. This paper describes how to compute the multi-fractal spectrum and other problems concerning the methods. Applying the methods to the analysis of DNA sequences shows that, in a sense, genomic sequence can be viewed as outputs of a stationary information source.

Key words: Discrete stationary information sources; Multi-fractal spectrum; DeoxyriboNucleic Acid (DNA) sequence; Statistical analysis

1　引言

信源输出信号在数学中可以用随机过程加以描述，因此，可以说信源的建模在某种程度上也就是用恰当的随机过程来描述信号[1]。在实际的建模过程中，我们观察到的只有信源的输出信号，对信源的真实模型并不了解。能不能找到一些统计特征能反映信源的本质特征，从而可以利用这些特征来推测信源所属的模型类，这个问题在信号建模和机器学习中一直被长期研究，但是并没有得到彻底的解决。如果这种统计特征存在，那么它可以用于模型的评估，因为模型统计量的理论值可以和统计获得的观测值加以比较，倘若偏差太大，则可以拒绝得到的模型。

稳恒信源是信源研究中最主要的一种信源，因为很多实际信源在较短的一段时间内都可以用稳恒信源作为其数学模型，而且稳恒信源的研究又是非稳恒信源研究的基础[1]。一般而言，将未知信源看成是稳恒的信源可以简化对问题的处理方式，这是因为稳恒信源有一系列很好的性质，如满足渐近等同分割定理等[1]。在实际的应用中，稳恒性通常是一个默认的假设，比如在语音处理中应用得比较广泛的隐马尔可夫模型[2]，就是一个稳恒信源模型。这是因为语音信号在短时期内是近似稳恒的，所以这隐马尔可夫模型对实际应用有足够的精度。基于以上的原因，本文中讨论的信源仅限于稳恒信源。

熵率是稳恒信源的重要特征，它是多重分形谱的特殊一维，即信息维[3]。1958 年，Kolmogorov指出熵相等是动力系统同构的必要条件[1]。从多重分形谱的角度而言，多重分形谱反映了随机过程的动力学特性，因此它可以作为验证信源是否同构的必要条件。稳恒信源熵率的存在性[1]，从另一个角度说明了，稳恒信源的多重分形谱可能是存在的。简单稳恒信源的多重分形谱可以直接用解析法求得，复杂信源的多重分形谱可以用统计物理中的方法来计算，因此多重分形谱的存在性可以通过上述两种方法加以验证[4]。

在以前的相关工作中，我们已经探讨了如何从理论上计算信源的多重分形谱[4]。本文针对的是信源的输出信号，依据信号的统计量，估计信源的多重分形谱。本文探讨了这个

2005-10-08 收到，2006-03-13 改回
中国科学院知识创新工程重要方向基金项目(KSCX2-SW-329)和中国科学技术大学高水平大学建设重点项目资助课题

估计过程中存在的问题，给出了多重分形谱判定的准则和相关的算法。本文还给出了一个实际的例子，计算酵母DNA序列的多重分形谱，实验结果表明DNA序列在某种意义上可以视为一个稳恒信源的输出，因而可以用一个稳恒的随机过程加以模拟。这也从一个角度解释了为什么用隐马尔可夫模型等稳恒随机过程在处理生物数据时能获得巨大的成功[5]。

2　离散信号的多重分形谱

2.1　定义和记号

稳恒信源输出的信号是一个稳恒的随机过程。当信源字母表是离散的，且信号取值时刻也是离散时，此时的稳恒信源就称为离散稳恒信源[1]。设离散稳恒信源的字母表为 $\{a_1,a_2,\cdots,a_K\}$，信源的输出序列用 $\{\cdots,u_{-2},u_{-1},u_0,u_1,u_2,\cdots,u_t,\cdots\}$ 来表示。根据稳恒随机过程的定义，信源输出序列的一切有限维概率分布与时间轴起点的选择无关，即有 $P\left(u_tu_{t+1}\cdots u_{t+N}=A\right)=P\left(u_ju_{j+1}\cdots u_{j+N}=A\right)$，其中 A 为某一特定的字母序列[1]。这一点为稳恒信源的研究带来很大方便。

假设稳恒信源字母序列的长度有限，设为 N，并用 $(u_1,u_2,\cdots,u_N)$ 来表示，那么，可以将该有限长度的序列看成一个随机矢量。该随机矢量的熵可以用联合熵 $H(U_1,U_2,\cdots,U_N)$ 来表示，于是，每个字母的平均熵 $H_N(U)$ 可以表示为 $H_N(U)=(1/N)H(U_1,U_2,\cdots,U_N)$。当 $N\to\infty$ 时，若 $H_N(U)$ 趋于某一极限，则定义该极限为信源的熵率，记为 $H_\infty(U)$，即 $H_\infty(U)=\lim_{N\to\infty}H_N(U)$。对于一般的稳恒信源，可以证明 $H_\infty(U)$ 一定存在[1]。

从混沌和复杂性理论来看，熵率只是多重分形谱中特殊的一维，即信息维[3]。对于稳恒信源长度为 N 的输出随机矢量 $(U_1,U_2,\cdots,U_N)$，其实例 $(u_1,u_2,\cdots,u_N)$ 一共有 K^N 种。将 K^N 种输出 $(u_1,u_2,\cdots,u_N)$ 按照概率的大小分成不同的集合 $E_p=\left\{(u_1,u_2,\cdots,u_N)\,|\,P(u_1,u_2,\cdots,u_N)=p\right\}$，其中 p 为某个大于 0 的数。令 $\left|E_p\right|$ 表示集合 E_p 中元素的数目，即 E_p 的基数，则 $\log_2(|E_p|)/N$ 和 $\alpha=-(\log_2(p)/N)$ 之间具有函数关系，记该函数为 $f_N(\alpha)$，即 $\log_2(|E_p|)/N=f_N(-\log_2(p)/N)$。当 $N\to\infty$，$f_N(\alpha)$ 收敛于函数 $f_\infty(\alpha)$，我们将函数 $f_\infty(\alpha)$ 记为信源的多重分形谱。对于一般的稳恒信源 $f_N(\alpha)$ 的收敛性难以证明，并且 $f_\infty(\alpha)$ 的解析表达式难以求得。但是可以用数值的方法，如统计物理的方法计算，并且由标度不变性表明其收敛性。熵率 $H_\infty(U)$ 满足[6]：$H_\infty(U)=f\left(H_\infty(U)\right)$ 且 $f_\infty\left(H_\infty(U)\right)=1$。

对于一个信源的输出，它通常是一个或者多个序列，对于一个长度为 M 的输出信号，我们得到 $M-N+1$ 个可重叠的长度为 N 的片断，其中 $0<N<M$。由于信源稳恒的特性，我们用 $(u_1,u_2,\cdots,u_N)$ 出现的频率，用式(1)来估计 $(u_1,u_2,\cdots,u_N)$ 的真实概率 $P(u_1,u_2,\cdots,u_N)$，从而用 $\hat{P}$ 来计算多重分形谱：

$$\hat{P}(u_1,u_2,\cdots,u_N)=\frac{(u_1,u_2,\cdots,u_N)\text{的出现次数}}{M-N+1}\tag{1}$$

2.2　按照定义计算时的缺陷

考虑一个简单的离散无记忆信源，二项分布，即 N 次伯努利试验的输出，其字母表为 {0,1}，每次以概率 e 输出 0，概率 $1-e$ 输出 1，其中 $0<e<1$。则按照定义，对于长度为 N 的输出随机矢量 $(U_1,U_2,\cdots,U_N)$，其实例 $(u_1,u_2,\cdots,u_N)$ 一共有 2^N 种。将 2^N 种输出 $(u_1,u_2,\cdots,u_N)$ 按照概率的大小分成不同的集合，对 $p_i=e^i(1-e)^{n-i}$，$E_{p_i}=\{(u_1,u_2,\cdots,u_N)\,|\,P(u_1,u_2,\cdots,u_N)=p_i\}$，即 $E_{p_i}=\{(u_1,u_2,\cdots,u_N)\,|\,u_1,\cdots,u_N$中有$i$个为0$\}$，其中 $0\le i\le n$。则对于 p_i，$\left|E_p\right|=C_N^i=N!/(i!(N-i)!)$。则 $\log_2(|E_p|)/N$ 和 $\alpha=-\log_2(p)/N$ 之间具有函数关系 $f_N(\alpha)$ 可以用图表示出来(见图 1)。并且，$f_\infty(\alpha_\infty)$ 可以解析求得[3]，对于 $\xi\in[0,1]$, $\alpha_\infty(\xi)=-\xi\log_2 e-(1-\xi)\cdot\log_2(1-e)$, $f_\infty(\alpha_\infty(\xi))=-\xi\log_2\xi-(1-\xi)\log_2(1-\xi)$。从图 1 中可以看到，$f_N(\alpha)$ 收敛到 $f_\infty(\alpha)$ 的速度不是很快。对于 $N=80$，f_{80} 和 f_∞ 还是有一定差异，此时如果要计算概率的话，需要计算 2^{80} 个值，这样的计算量是难以承受的。这一点说明，即使获得了理论上的分布 $P(u_1,u_2,\cdots,u_N)$，按照定义计算得到的多重分形谱也难以保证与理论结果吻合。

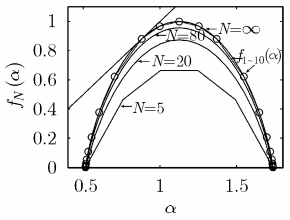

图 1　p=0.3 的二项分布的多重分形谱

更严重的问题是对于只知道输出信号的未知信源，我们采用的是出现频率 $\hat{P}(u_1,u_2,\cdots,u_N)$ 来估计 $P(u_1,u_2,\cdots,u_N)$。按照大数定律，估计的准确度取决于输出信号的数据量。通常情况下，要估计长度为 N 的 K^N 个概率需要的数据量也是随 N 指数增长的，这一点对于实际的信号来说可能是苛刻的。如酵母第 4 条染色体中的碱基数约为 1.5×10^6，碱基有 4 种类型，用该长度的数据估计 4^N 个概率，顶多估计的长度 $N\le\left\lfloor\log_4(1.5\times10^6)\right\rfloor=10$ (等式满足说明数据量已经少于要估计的参数的个数，估计将不再有效)。并且 $\hat{P}$ 还存在误差，如直接采用定义计算多重分形谱，将导致更多的偏差。图 2 演示了这种偏差。当 $N=8$ 时，对 $\hat{P}$ 的估计已经出现偏差，这时多重分形谱 f_8 已经变形。在后面的分析中我们将发现 $N\le7$，对概率的估计数据量是充分的。

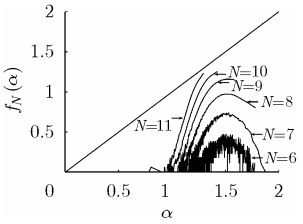

图 2　酵母第 4 条染色体的多重分形谱

除了上面描述的跟信号数据量有关的问题以外，在文献[7]中还描述了计算多重分形谱的其他有关问题，本文并不涉及这些问题。

2.3 离散信号的多重分形谱计算

为避免上面描述的两种偏差，在统计物理中采用Legendre变换[3]，但其计算是比较繁琐的。本文将文献[8]中的方法推广到离散信源的情形，可以用如下条件判断多重分形谱的存在(以保证估计的有效性)，并计算其多重分形谱，从而使得求得的多重分形谱避免上述缺点。令 $\boldsymbol{U}_N=\{(u_1,u_2,\cdots,u_N)\}$，即 $\boldsymbol{U}_N$ 表示稳恒信源长度为 N 的随机矢量 $(U_1,U_2,\cdots,U_N)$ 所有可能的输出，一共 K^N 种。我们定义配分函数 $\chi_N(q)$，它满足下式：

$$\chi_N(q)=\sum_{\boldsymbol{u}\in\boldsymbol{U}_N}\hat{P}_{\boldsymbol{u}}^q=2^{-N\tau(q)} \tag{2}$$

对于确定的 q，$\tau(q)$ 是一个常数，因此变换式(2)得：

$$\log_2\chi_N(q)=\log_2\sum_{\boldsymbol{u}\in\boldsymbol{U}_N}\hat{P}_{\boldsymbol{u}}^q=-N\tau(q) \tag{3}$$

这表明配分函数的对数 $\log_2\chi_N(q)$ 与 N 成线性关系，其斜率为 $-\tau(q)$。式(3)可以作为多重分形谱是否存在的判定标准，对于满足线性关系的 N 的区间 $[n_1,n_2]$，其多重分形谱存在，其中 $n_1<n_2$ 为正整数。

对于线性关系得到满足的区间 $[n_1,n_2]$，$\forall N\in[n_1,n_2]$，推广文献[8]中方法，可以计算其多重分形谱 $\alpha_N(q)$, $f_N(\alpha_N(q))$：

$$\alpha_N(q)=\frac{\sum_{u\in U_N}\hat{P}_u^q\log_2\hat{P}_u}{N\sum_{v\in U_N}\hat{P}_v^q} \tag{4}$$

$$f_N(\alpha_N(q))=\frac{-\log_2\sum_{v\in U_N}\hat{P}_v^q+q\sum_{u\in U_N}\hat{P}_u^q\log_2\hat{P}_u}{N\sum_{v\in U_N}\hat{P}_v^q} \tag{5}$$

最后采用平均或者加权平均等线性回归的方法得到最终的多重分形谱：

$$\alpha_{n_1\sim n_2}(q)=\frac{\sum_{N=n_1}^{n_2}\alpha_N(q)}{n_2-n_1+1} \tag{6}$$

$$f_{n_1\sim n_2}(\alpha_{n_1\sim n_2}(q))=\frac{\sum_{N=n_1}^{n_2}f_N(\alpha_N(q))}{n_2-n_1+1} \tag{7}$$

或者

$$\alpha_{n_1\sim n_2}(q)=\frac{n_2\alpha_{n_2}(q)-n_1\alpha_{n_1}(q)}{n_2-n_1} \tag{8}$$

$$f_{n_1\sim n_2}(\alpha_{n_1\sim n_2}(q))=\frac{n_2f_{n_2}(\alpha_{n_2}(q))-n_1f_{n_1}(\alpha_{n_1}(q))}{n_2-n_1} \tag{9}$$

在这里当 $q=1$ 时，$\alpha_{n_1\sim n_2}(1)=f_{n_1\sim n_2}(\alpha_{n_1\sim n_2}(1))$，即为熵率。由此，我们得到计算信号的熵率的算法1。

3 实验结果分析

算法1的有效性可以从图1中 $f_{1\sim10}$ 和 f_∞ 之间的差别看出来。$f_{1\sim10}$ 是图中带圈的折线，最外围的是 f_∞，两者差异在 10^{-2} 量级以下，大大小于用定义计算的 f_{80}。

本文计算了酵母的全部染色体(数据来自http://www.yeastgenome.org/)的多重分形谱，计算结果表明全部染色体都存在多重分形谱。我们用第4条染色的结果为例说明这个计算过程，酵母的第4条染色体包括1,531,916个碱基。从图3中 $\log_2\chi_N(q)$ 与 N 之间的关系，可以看出线性关系在 $N=1$ 和 $N=7$ 之间满足。$N\geq8$ 以后这种线性关系已经得不到满足。正如前面所分析的，这是因为信号数据不足，导致对概率的估计出现偏差。利用算法1计算[1,7]区间的 α 和 $f_{1\sim7}(\alpha)$ 显示在图4中，此即其多重分形谱。由图4可见，其多重分形谱非常光滑，这一方面表明了算法是有效的，另外一方面，由多重分形谱的存在性，暗示我们DNA序列可能是某个稳恒随机过程的产物，因而这从某种层面熵解释了为什么用隐马尔可夫或者其他稳恒的概率模型来建模是非常有效的。同时，一旦对DNA序列建模成功，就可以利用文献[4]中的方法，计算模型多重分形谱的理论值，以此检查模型的有效性。

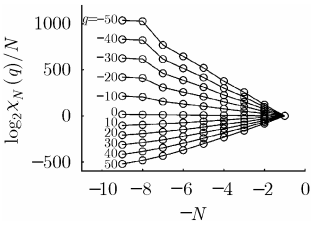

图3 酵母第4条染色体的 $\log_2\chi_N(q)$ 和 N 之间的关系

从图4中的切点处，求得其熵率为1.942294。我们在图5中，直接计算 $N\cdot H_N$ 得到熵率为图中直线的斜率。由图5可见直接计算熵率时，当 $N=9$ 时，线性关系很好的满足，表明熵率估计基本准确。这一点同样体现了数据的多少对计算多重分形谱的影响，由此可见我们可以用图3中变量是否满足线性关系作为熵率存在和估计准确性的依据之一。

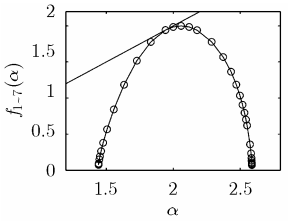

图4 酵母第4条染色体的多重分形谱

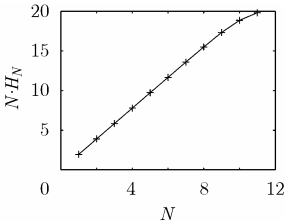

图5 酵母第4条染色体的熵随着 N 增加

算法1 计算离散稳恒信号的多重分形谱

(1)用式(1)计算 $\hat{P}_u=\hat{P}(u_1,u_2,\cdots,u_N)$。

(2)用式(3)计算配分函数，对不同的 q，做出 $\log_2\chi_N(q)$ 与 N 之间的关系图。

(3)寻找 $\log_2\chi_N(q)$ 与 N 图形中线性关系得到满足的区间 $[n_1,n_2]$。

(4)利用式(4)~式(9)计算多重分形谱 $f_{n_1\sim n_2}$。

(5)使用多重分形谱和 $f_{n_1\sim n_2}(\alpha)=\alpha$ 的切点求解熵率。

4 结束语

稳恒信源的多重分形谱体现了信源的统计特征，它为鉴

别信号建模的合理性提供了一个必要条件。本文对信号的多重分形谱计算中存在的问题加以分析，并且给出了判断信号多重分形谱是否存在以及计算信号多重分形谱的方法。用以上方法，我们计算了酵母染色体的多重分形谱，试验结果表明，其多重分形谱存在，这暗示了遗传信息可能是某个稳恒信源的输出。本文方法有助于理解稳恒随机过程与它的输出信号之间的联系。

参 考 文 献

[1] 朱雪龙. 应用信息论基础. 北京: 清华大学出版社, Mar. 2000: 74–108.

[2] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 1989, 77(2): 257–286.

[3] 孙霞, 吴自勤, 黄昀. 分形原理及应用. 合肥: 中国科学技术大学出版社, 2003, 53–88.

[4] 陈双平, 郑浩然, 马猛等. 用统计物理的方法计算信源熵率. 电子与信息学报, 2007, 29(1): 129–132.
Chen Shuang-ping, Zheng Hao-ran, and Ma Meng, *et al.*. Computing the entropy rate of information source with methods of statistical physics. *Journal of Electronics & Information Technology*, 2007, 29(1): 129–132.

[5] Durbin R, Eddy S, Krogh A, and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. London: Cambridge University Press, 1998, chapter 3.

[6] 周炜星, 王延杰, 于遵宏. 多重分形奇异谱的几何特性: i.经典 renyi 定义法. 华东理工大学学报, 2000, 26(4): 385–389.
Zhou Wei-xing, Wang Yan-jie, and Yu Zun-hong. Research on scale inhibition of combination of polyaspartic acid and oxidized starch. *Journal of East China University of Science and Technology*(*Natural Science Edition*), 2000, 26(4): 385–389.

[7] Chen Huiping, Sun Xia, Chen Huixuan, and Wu Ziqin, *et al.*. Some problems in multifractal spectrum computation using a statistical method. *New J. Phys.*, 2004, 60 (1): 84–100.

[8] Mach J, Mas F, and Sagues F. Two representations in multifractal analysis. *Journal of Physics A: Mathematical and General*, 1995, 28(19): 5607–5622.

陈双平：　男，1976 年生，博士后，研究方向为数据挖掘、生物信息学和复杂性.

郑浩然：　男，1967 年生，副教授，研究方向为计算智能、生物信息学、模式识别.

刘金霞：　女，1985 年生，硕士生，研究方向为生物信息学.

童　庆：　男，1977 年生，博士生，研究方向为生物信息学.

王煦法：　男，1948 年生，教授，主要研究方向为计算机网络、计算智能、信号处理和模式识别等.