第 26 卷 第 5 期
2011 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.26 No. 5
Oct. 2011

高等真核生物基因组 isochore 边界确定中的对称性

崔家峰
（天津科技大学理学院，天津 300457）

摘 要：高等真核生物基因组的 isochore 结构与许多重要的生物学特征相关，而对其边界的确定则是 isochore 结构分析的重点，同时也是难点. 针对基于 Z 曲线的累积 GC 轮廓图法、基于 Shannon 熵的递归分段算法以及基于二次散度的分段算法三种典型的应用，分析出其分段依据本质上是对于基因组序列求取碱基对换对称性的对称中心. 基于此结果，在寻找 isochore 结构分析度量时，只要度量函数满足一定对换对称性要求，即可达到殊途同归的目的.

关键词：真核生物；基因组；isochore；Z 曲线；Shannon 熵
中图分类号：Q811.4 **文献标志码**：A **文章编号**：1672-6510(2011)05-0072-04

Measures of Symmetry in the Identifying Isochore Boundaries of Eukaryotic Genomes

CUI Jia-feng
(College of Science, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: Many higher eukaryotic genomes are composed of large sequence segments with fairly homogeneous GC content, namely isochores, which have been linked to many important biological functions. It is not only important but also difficult to determine the isohore boundaries. Three methods, GC-Profile algorithm based on Z curve, recursive segmentation algorithm based on Shannon entropy and on quadratic divergence, are analyzed and conclusion is drawn that the common feature of three algorithms is to find the symmetry center of the genome sequence. Based on this result, the same purposes can be achieved as long as the measure functions for the analysis of the isochore structures meet some exchange symmetry requirements.

Keywords: eukaryotic; genome; isochore; Z curve; Shannon entropy

随着 2003 年对人类全基因组测序的完成[1]，越来越多的生物全基因组数据以及各种组学技术把生物学带入了系统科学时代. 哺乳动物基因组最重要的特征之一就是它的 GC 含量在大尺度上的变化，这些变化的尺度从几十万个到几百万个碱基对不等，这就是所谓的基因组的 isochore 结构. 这些碱基组成的变化影响到序列的编码区和非编码区，也反映了基因组结构的基本特征. 哺乳动物基因组的 mosaic 结构是在 20 世纪 70 年代中期对牛的基因组做密度梯度离心实验时揭示的[2]. 根据 Bernardi 的分析结果，有 5 个 isochore 家族：其中两个 GC 含量较低的家族是 L1（GC 含量＜38%）和 L2（38%≤GC 含量＜44%）；另外三个是 GC 含量较高的家族 H1（44%≤GC 含量＜48%），H2（48%≤GC 含量＜52%）和 H3（GC 含量≥52%）. isochore 结构很早就被发现，然而至今关于它的诸多问题仍然在激烈地讨论研究[3]中. 一般认为，isochore 结构与许多重要的生物学特征相关，比如：基因密度、基因长度、密码子使用、重复元件、重组频率以及复制开关等[4-5].

1 isochore 结构研究方法

研究 isochore 结构最简单最直接的方法就是滑动窗口法[6]，即在每一个窗口内计算 GC 的含量，然后再根据某些判据断定局部 GC 含量是否有显著变化. 但是这种方法有一个最致命的缺点，就是窗口的

收稿日期：2011–02–14；**修回日期**：2011–03–25
作者简介：崔家峰（1975—），男，天津人，副教授，cuijf@tust.edu.cn.

大小会影响到局部 GC 含量以及其标准差，使得窗口过大则抹杀了 GC 含量变化的细节，而窗口过小又会产生很大的统计涨落. 利用无窗技术研究 isochore 结构也产生了许多方法，如基于 Shannon 熵的递归分段算法[7]、最小二乘优化分段法[8]、隐马尔科夫方法[9]、基于 Z 曲线的累积 GC 轮廓图法[10]、基于二次散度的分段算法[11]. 其中，最小二乘优化分段法其实是一种基于局部小窗（大小为 10 万个碱基对的不重叠窗口）的算法，但由于成功地避免了把富含 GC 的片段过度划分，着眼于基因组序列的全局差异而不是局部差异，获得了良好的效果.

2　三种典型算法中的对称性

2.1　基于 Z 曲线的累积 GC 轮廓图法

Z 曲线理论最早是由我国学者张春庭[10]院士在 20 世纪 90 年代初提出的，并在很多方面得以应用.

考虑长度为 N 的单链 DNA 序列，A_n,C_n,G_n,T_n 分别表示该序列前 $n(n=0,1,\cdots,N)$ 个碱基中所含四种碱基的数目，显然 $A_n+C_n+G_n+T_n=n$. 若以三个变量 x_n,y_n,z_n 分别表示前 n 个碱基中嘌呤碱基与嘧啶碱基之差、氨基碱基与酮基碱基之差、弱氢键碱基与强氢键碱基之差，即

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

$$x_n,y_n,z_n\in[-N,N],n=0,1,\cdots,N$$

则 x_n,y_n,z_n 可以对应于空间中的一个点 P_n，若把这一系列点 P_n（共 $N+1$ 个点）连接起来构成的曲线称为 DNA 序列的 Z 曲线，很容易证明，Z 曲线和 DNA 序列一一对应，即 Z 曲线包含了 DNA 的全部信息，是 DNA 序列等价表示的一种几何形式.

Z 曲线的三个分量 x_n,y_n,z_n 描述了 DNA 序列的三种独立类型的碱基分布，其中 z_n 分量刻画了沿 DNA 序列 GC 含量的分布. 一般而言，对于 GC 含量丰富的基因组，z_n 近似为 n 的单调递减函数. 利用最小二乘法直线拟合 z_n–n 曲线，得直线 $z=kn$，其中 k 为拟合直线的斜率. 令 $z'_n=z_n-z=z_n-kn$，若以 $\overline{\mathrm{G+C}}$ 表示一段序列区间 Δn 的 GC 平均含量，则

$$\overline{\mathrm{G+C}}=\frac{(\Delta n-\Delta z_n)}{2\Delta n}=\frac{(\Delta n-\Delta z'_n-k\Delta n)}{2\Delta n}=\frac{1}{2}\left(1-k-\frac{\Delta z'_n}{\Delta n}\right)\equiv\frac{1}{2}(1-k-k')$$

其中 $k'=\Delta z'_n/\Delta n$ 是在区间 Δn 上 Z' 曲线的斜率.

很明显，若 Z' 曲线是上升曲线，则表明该区间 GC 含量降低；反之则表明该区间 GC 含量增加. 若 Z' 曲线可以用直线拟合（即 k' 为常数）则说明 GC 含量相对均匀，且该直线线性回归效果越明显，该段序列 GC 含量越均匀. 事实上，Z' 曲线的意义在于，如果将整体 GC 含量的累积效果看作是背景压力，则 Z' 曲线恰好从中去除了这个背景，从而更加充分地表现出局部 GC 含量的均匀特征. 于是，Z' 曲线激增或锐减的地方将可能是局部 GC 含量发生突变的转折点，这个区域就有可能成为 isochore 的边界.

值得注意的是，这样的区域恰好体现了 GC（强氢键碱基）与 AT（弱氢键碱基）对换对称的特性，也就是说，若将这两类碱基对换，则该区域依然可能是 isochore 的边界位点.

2.2　基于 Shannon 熵的递归分段算法

该方法是一种分而治之的算法. 对于由 k 个符号（符号集记为 $\Omega=\{\omega_1,\omega_2,\cdots,\omega_k\}$，例如对于核酸序列来说，$\Omega=\{A,C,G,T\}$）组成的序列，设其长度为 N，计算整条序列的熵 H，每一位点 $n(0<n<N)$ 处的左右子序列的熵 H_l 和 H_r，即

$$\begin{cases} H=-\sum\limits_{j\in\Omega}\dfrac{N_j}{N}\log\dfrac{N_j}{N} \\ H_l(n)=-\sum\limits_{j\in\Omega}\dfrac{N_{j,l}}{n}\log\dfrac{N_{j,l}}{n} \\ H_r(n)=-\sum\limits_{j\in\Omega}\dfrac{N_{j,r}}{N-n}\log\dfrac{N_{j,r}}{N-n} \end{cases}$$

其中，$N_j,N_{j,l}$ 和 $N_{j,r}$ 分别表示整条序列中符号 j 出现次数、左侧及右侧子序列中符号 j 出现次数. 为了刻画序列的异质性，引入最大化 Jensen-Shannon 散度[12]：

$$D_{JS}=\max_n D_{JS}(n)=\max_n\left[H-\frac{n}{N}H_l(n)-\frac{N-n}{N}H_r(n)\right]$$

如果 D_{JS} 足够大，就可以断定该序列是异质的，须进一步分割成片段（segments）进行分别研究. 在分割成的两个片段基础上，采用相同的办法，递归地进行进一步分割，直至满足一定条件（即 D_{JS} 小于某个事先指定的阈值）为止.

本算法的关键是寻找合适的位点 n 使得 $D_{JS}(n)$ 最大（同时要求 D_{JS} 大于事先指定的阈值）. 下面就核酸序列推导该位点满足的条件，这种情况比单纯寻找 isochore 边界更广泛一些，因为寻找 isochore 边界

只涉及到两类碱基 S(即 G、C)和 W(即 A、T).

由于熵 H 在 Jensen-Shannon 散度中与位置 n 无关，于是求解 D_{JS} 的最大值就转变为求解 $K(n)=\frac{n}{N}H_l(n)+\frac{N-n}{N}H_r(n)$ 的最小值了. 为了以下叙述的方便，可以将各 n 的函数对 n 的变差记作导数形式(此时的导数在数学上被称为 Radon-Nikodym 导数).

记符号集 $\Omega=\{A,C,G,T\}$，$(N_{Al},N_{Cl},N_{Gl},N_{Tl})$ 与 $(N_{Ar},N_{Cr},N_{Gr},N_{Tr})$ 表示各碱基在位点 n 处的左右子序列中的计数，$p_l=(p_{Al},p_{Cl},p_{Gl},p_{Tl})$ 与 $p_r=(p_{Ar},p_{Cr},p_{Gr},p_{Tr})$ 表示各碱基在位点 n 处的左右子序列中的比例. 显然有，$p_l=(N_{Al},N_{Cl},N_{Gl},N_{Tl})/n$ 及 $p_r=(N_{Ar},N_{Cr},N_{Gr},N_{Tr})/(N-n)$.

令 $\frac{\mathrm{d}K(n)}{\mathrm{d}n}=0$，即

$$\frac{\mathrm{d}K(n)}{\mathrm{d}n}=\frac{1}{N}H_l(n)-\frac{n}{N}\sum_{j\in\Omega}p'_{jl}\log p_{jl}-\frac{1}{N}H_r(n)-\frac{N-n}{N}\sum_{j\in\Omega}p'_{jr}\log p_{jr}=0$$

因为

$$p'_{jl}=\left(\frac{N_{jl}}{n}\right)'=\frac{1}{n}\left(N'_{jl}-p_{jl}\right)$$

$$p'_{jr}=\left(\frac{N_{jr}}{N-n}\right)'=\frac{1}{N-n}\left(N'_{jr}+p_{jr}\right)\quad(j\in\Omega)$$

于是

$$\frac{\mathrm{d}K(n)}{\mathrm{d}n}=\frac{1}{N}H_l(n)-\frac{1}{N}\left[\sum_{j\in\Omega}N'_{jl}\log p_{jl}-\sum_{j\in\Omega}p'_{jl}\log p_{jl}\right]-\frac{1}{N}H_r(n)-\frac{1}{N}\left[\sum_{j\in\Omega}N'_{jr}\log p_{jr}+\sum_{j\in\Omega}p'_{jr}\log p_{jr}\right]=-\frac{1}{N}\sum_{j\in\Omega}N'_{jl}\log p_{jl}-\frac{1}{N}\sum_{j\in\Omega}N'_{jr}\log p_{jr}=0$$

即有

$$\sum_{j\in\Omega}N'_{jl}\log p_{jl}+\sum_{j\in\Omega}N'_{jr}\log p_{jr}=0$$

由于 $N'_{jr}=-N'_{jl},j\in\Omega$，于是，使得 $K(n)$ 最小(即 D_{JS} 最大)的位点 n 应满足

$$\sum_{j\in\Omega}N'_{jl}\left(\log p_{jl}-\log p_{jr}\right)=0$$

该式表明，当位点 n 的两侧的碱基满足对换对称时，此位点有可能是要找的 isochore 边界分割点(注意上式位点 n 只是序列分割点的必要而非充分条件).

2.3　基于二次散度的分段算法

对于向量 $V=(v_1,v_2,\cdots,v_k)$，定义

$$S(V)=\sum_i v_i^2$$

设 $P=(p_1,p_2,\cdots,p_k)$ 和 $Q=(q_1,q_2,\cdots,q_k)$ 为两个概率分布函数，即 $0\leqslant p_i,q_i\leqslant 1$ 且 $\sum_i p_i=\sum_i q_i=1$. 当 $k=4$ 时，称 $S(P)=\sum_i p_i^2$ 为基因组序参数[11]. 又设 α,β 为权重因子，且有 $0<\alpha,\beta<1,\alpha+\beta=1$. 定义两个分布 P 和 Q 之间的二次散度

$$\Delta S(P,Q)=\alpha S(P)+\beta S(Q)-S(\alpha P+\beta Q)$$

$\Delta S(P,Q)$ 量化了两个分布 P 和 Q 之间的差异. 经过简单推导，易得其几个有用的性质：

(1) $\Delta S(P,Q)=\alpha\beta S(P-Q)=\alpha\beta\sum_i(p_i-q_i)^2$；

(2) 非负性：$\Delta S(P,Q)\geqslant 0$，当且仅当 $P=Q$ 时，$\Delta S(P,Q)=0$；

(3) 对称性：$\Delta S(P,Q)=\Delta S(Q,P)$；

(4) 三角不等式：$\sqrt{\Delta S(P,R)}+\sqrt{\Delta S(R,Q)}\geqslant\sqrt{\Delta S(P,Q)}$，其中 P,Q,R 为三个分布.

由上面性质(1)—(4)易知，$\Delta S(P,Q)$ 是一个距离的度量.

对于长度为 N 的基因组 DNA 序列，符号集 $\Omega=\{A,C,G,T\}$，对于给定的位置 n 满足 $1<n<N$，将整条序列分为两个子序列. 令 $\alpha=\frac{n}{N},\beta=\frac{N-n}{N}$，$(N_{Al},N_{Cl},N_{Gl},N_{Tl})$ 与 $(N_{Ar},N_{Cr},N_{Gr},N_{Tr})$ 表示各碱基在位点 n 处的左右子序列中的计数，$p_l=(p_{Al},p_{Cl},p_{Gl},p_{Tl})$ 与 $p_r=(p_{Ar},p_{Cr},p_{Gr},p_{Tr})$ 表示各碱基在位点 n 处的左右子序列中的比例. 显然有

$$p_l=(N_{Al},N_{Cl},N_{Gl},N_{Tl})/n$$

$$p_r=(N_{Ar},N_{Cr},N_{Gr},N_{Tr})/(N-n)$$

于是

$$\Delta S(p_l,p_r)=\alpha S(p_l)+\beta S(p_r)-S(\alpha p_l+\beta p_r)$$

因为

$$S(\alpha p_l+\beta p_r)=\sum_{j\in\Omega}\left(\frac{N_{jl}+N_{jr}}{N}\right)^2=\sum_{j\in\Omega}\left(\frac{N_j}{N}\right)^2$$

其中，$N_j=N_{jl}+N_{jr}$ 表示整条序列中碱基 j 的数目，与位点 n 无关，故若 n 是使上式取得最大值的位点，则以下结论等价：

(1) $\alpha S(p_l)+\beta S(p_r)$ 最大；

(2) $\Delta S(p_l, p_r)$ 最大；

(3) $\frac{n(N-n)}{N^2}\sum_{j\in\Omega}(p_{jl}-p_{jr})^2$ 最大.

记 $K(n)=\alpha S(p_l)+\beta S(p_r)$，其中 $\alpha=\frac{n}{N},\beta=\frac{N-n}{N}$.

令 $\frac{\mathrm{d}K(n)}{\mathrm{d}n}=0$，即

$$\frac{\mathrm{d}K(n)}{\mathrm{d}n}=\frac{1}{N}S(p_l)+\frac{n}{N}\sum_{j\in\Omega}2p_{jl}p'_{jl}-\frac{1}{N}S(p_r)+\frac{N-n}{N}\sum_{j\in\Omega}2p_{jr}p'_{jr}=0$$

因为 $p'_{jl}=\left(\frac{N_{jl}}{n}\right)'=\frac{1}{n}(N'_{jl}-p_{jl})$ 及 $p'_{jr}=\left(\frac{N_{jr}}{N-n}\right)'=\frac{1}{N-n}(N'_{jr}+p_{jr}),j\in\Omega$，于是

$$\frac{\mathrm{d}K(n)}{\mathrm{d}n}=\frac{2}{N}\sum_{j\in\Omega}p_{jl}N'_{jl}-\frac{1}{N}S(p_l)+\frac{2}{N}\sum_{j\in\Omega}p_{jr}N'_{jr}+\frac{1}{N}S(p_r)=0$$

由于 $N'_{jr}=-N'_{jl},j\in\Omega$，于是，使得 $K(n)$ 最大（即 $\Delta S(p_l,p_r)$ 最大）的位点 n 应满足

$$\sum_{j\in\Omega}N'_{jl}(p_{jl}-p_{jr})=\frac{S(p_l)-S(p_r)}{2}$$

该式表明，当位点 n 的两侧的碱基满足对换对称时，此位点有可能是要找的 isochore 边界分割点（注意上式只是必要条件，而非充分条件）.

3　分段算法对称性要求及特点

通过对上面三种常见的 isochore 边界识别方法的分析，得出以下结论：

(1) 序列出现最大异质的位点恰好是使得在该位点的左右两侧子序列碱基分布满足对换对称性的地方. 当然，一般真实的生物序列不一定恰好满足这种对称性，于是用事先给定的阈值来确认序列可能的 isochore 边界.

(2) 从计算量上来看，二次散度方法更简洁一些，不过基于信息熵的方法最完美，因为其在可能出现最大异质点处的熵不依赖于度量方法.

(3) 一般而言，若所定义的序列异质度量函数关于单个碱基分布对换对称，则在一定程度上混淆了碱基之间的区别，同时也忽略了碱基在序列构成上的相互关联性. 故要想进一步区别不同碱基在构成生物序列上的关联特性，高阶熵的考虑是必须的.

参考文献：

[1] Collins F S, Green E D, Guttmacher A E, et al. A vision for the future of genomics research [J]. Nature, 2003, 422(6934): 835–847.

[2] Macaya G, Thiery J P, Bernardi G. An approach to the organization of eukaryotic genomes at a macromolecular level [J]. Journal of Molecular Biology, 1976, 108(1): 237–254.

[3] Bernardi G, Olofsson B, Filipski J, et al. The mosaic genome of warm-blooded vertebrates[J]. Science, 1985, 228(4702): 953–958.

[4] Bernardi G. The human genome: organization and evolutionary history[J]. Annual Review of Genetics, 1995, 29: 445–476.

[5] Tenzen T, Yamagata T, Fukagawa T, et al. Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex [J]. Molecular and Cellular Biology, 1997, 17(7): 4043–4050.

[6] Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*[J]. Nature, 1997, 390(6657): 249–256.

[7] Li W. Delineating relative homogeneous G + C domains in DNA sequences [J]. Gene, 2001, 276(1/2): 57–72.

[8] Haiminen N, Mannila H. Discovering isochores by least-squares optimal segmentation [J]. Gene, 2007, 394(1/2): 53–60.

[9] Melodelima C, Gautier C, Piau D. A markovian approach for prediction of mouse isochors[J]. Journal of Mathematical Biology, 2007, 55(3): 353–364.

[10] Zhang C T, Zhang R. Analysis of distribution of bases in the coding sequences by a digrammatic technique [J]. Nucleic Acids Research, 1991, 19(22): 6313–6317.

[11] Zhang C T, Gao F, Zhang R. Segmentation algorithm for DNA sequences [J]. Physical Review E, 2005, 72: 041917.

[12] Lin J. Divergence measures based on the Shannon entropy [J]. IEEE Transactions on Information Theory, 1991, 37(1): 145–151.