PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(9): 1006~1015
www.pibb.ac.cn

线虫核糖核蛋白基因内含子与相应编码序列的相互作用 *

赵小庆 李 宏 ** 包通拉嘎
(内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 对线虫核糖核蛋白基因内含子序列与相应编码序列采用 Smith-Waterman 方法做局域比对分析，探讨两者之间的相互作用机制．发现内含子中部序列确实存在与相应编码序列的相互作用区域．第一内含子的最佳匹配分布在内含子 15%～55%的区域内，第二内含子的最佳匹配分布在内含子 30%～80%的区域内．对于长内含子，在与外显子序列比对时，最佳匹配分布在内含子 5%～20% 区域内，在与整个编码序列比对时，出现了两个峰区，一个位于内含子 15%～30%区域内，另一个位于内含子 54%～78%区域内．推测第一个峰区与外显子内部序列有关，第二个峰区与外显子 - 外显子结合区域的序列有关．还发现编码序列上存在多个与内含子序列的相互作用域和一些禁配区域分布．推测这些禁配区域与蛋白质结合区域有关．结论印证了内含子序列与相应编码序列协同进化的观点．

关键词 线虫核糖核蛋白基因，内含子，编码序列，局域比对，相互作用
学科分类号 Q61 **DOI**: 10.3724/SP.J.1206.2010.00186

内含子是从 mRNA 前体分子中切除的非编码序列．研究表明，内含子在真核蛋白质编码基因(简称基因)中所占的比例很高，甚至超过 99%，是真核基因的重要组成部分．内含子的生物学功能及其进化特性是近代生物学急需解决的问题[1]．近年来，内含子的“神秘面纱”正被逐步揭开，也从基因组的“垃圾”(junk)片段跃为重要生物学功能的载体[2]．已经发现，内含子中不仅含有微 RNA (microRNA)、核仁小 RNA (snoRNA)等多种非编码 RNA[3]，也包含涉及基因转录、mRNA 加工 (尤其是可变剪接)、输运、mRNA 和蛋白质的空间结构等众多基因表达调控元件[4]．这也可能是 Gabriel 等[5]发现果蝇内含子长度与蛋白质的进化速率成负相关的原因之一．内含子长度受到许多因素的影响[6-7]，如转座子插入、基因调控元件出现、RNA 基因或涉及基因调控 RNA 嵌入、删除事件大小和频数、降低转录耗能选择、维持相对较小活性染色体域倾向、减小外显子之间的 Hill-Roberton 干涉效应[8-14]等．内含子的获得和丢失影响基因内重组和 ncRNA 的变异，是影响真核生物基因进化的主要方面之一，是获得真核新物种的一种动力[15-16]．通过对不同生物正向同源基因内含子 - 外显子的比较分析，发现内含子变异性超过外显子．从内含子序列保守性的角度来看，已经知道具有普适功能的保守序列是内含子序列的两端．Halligan 等研究表明，内含子 5′端约 8 bp 和 3′端约 30 bp 为保守性较强的区域[17]，它的功能主要是完成组织剪接或可变剪接．但已有的研究结果并没有回答内含子中部大量的所谓“非保守”序列具有什么样的普适性和功能这个关键问题．到目前为止，对内含子功能的研究基本限于其被剪切之前的生物学功能，对内含子被剪切之后的生物学功能研究涉及很少．我们认为，内含子的功能远远超出了我们的想象．这不仅仅表现在基因的进化、基因的转录和剪接这些现象中，而同样重要的是在剪接后 mRNA 自然结构的

* 国家自然科学基金资助项目(30660044).
** 通讯联系人.
Tel: 0471-6678889, E-mail: ndlihong@imu.edu.cn
收稿日期：2010-04-17，接受日期：2010-06-14

形成、 mRNA 输运出核、翻译起始和调控等过程中，内含子均起着重要的作用．本文将通过研究内含子与相应编码序列的相互作用关系来推证两者的协同进化关联和内含子被剪切之后可能存在的生物学功能．

理论上，内含子与外显子都是真核生物基因组的重要组成成分，转录前，二者同为基因组序列，共同维护染色体各种生物活性和最适空间结构，故它们之间必存在为实现某种生物学功能的各种协同进化元件，转录后，同一基因中的内含子和其相邻或相近外显子都为这个基因的不同时空准确高效表达而服务，这需要内含子与外显子(或编码序列)协同作用来完成．因此，我们推断内含子与外显子可能存在某种协同进化机制，以保证这些生物学功能代代相传及准确、高效地表达．但现在对这方面认识仍处在模糊的探索阶段，故深入研究两者的相互作用及协同进化机制对进一步挖掘内含子的功能具有重要的生物学意义．

1　数据与方法

1.1　基因序列数据

线虫核糖核蛋白大小亚基基因序列取自 Ribosomal Protein Gene Database (RPG)(http://ribosome.miyazaki-med.ac.jp/)，共 88 条基因序列，剔除 2 条没有内含子及 1 条没有给出明确编码序列的基因后，共得到 85 条基因序列，含有 180 条内含子和 265 条外显子．

1.2　比对方法

将内含子序列转化成互补序列后，与相应的编码序列(CDS)进行局域比对分析，就可以用相同碱基对来代替互补碱基对，得到内含子与其相应编码序列最佳匹配区域．利用 water 局域比对软件(Smith-Waterman local alignment http://mobyle.pasteur.fr/cgi-bin)获得比对序列和被比对序列之间的一个最佳匹配区域．在局域比对计算中采用 Ednafull 矩阵，选取的参数如下：每个空隙罚分(Gap penalty)为 50.0，空隙中每延伸一个碱基位点罚分(Extend penalty)为 5.0，得到两者仅有一个可信的最佳局域配对区域，也是 2 条序列相互作用几率最大的区域．例如图 1 所示．

由于实际比对序列长度各不相同，为得到最佳比对相对位置分布，将比对序列标准化成长度为 100 的目标序列．方法如下：

设有 m 条基因序列，第 i 条序列中含有比对序列(记为 A_i)和被比对序列(记为 B_i)．将比对序列 A_i 与被比对序列 B_i 做局域比对，得到一个最佳局域匹配区域，m 组局域对比共有 m 个匹配结果．设 n_{i0}、n_{i1}、n_{i2} 分别为第 i 条比对序列的长度、最佳匹配区域的起始位点和终止位点；N_{i1}、N_{i2} 分别为标准化后第 i 条比对序列最佳匹配区域的起始位点和终止位点 (i=1，2，…, m–1, m)．做如下变换：

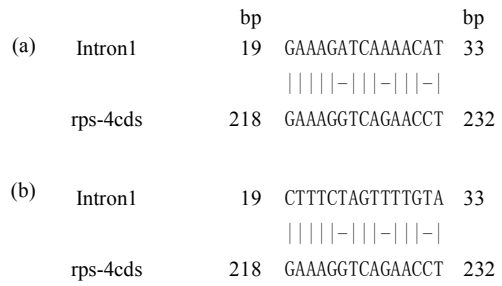

Fig. 1　Sketch map for Smith-Waterman local alignment and authentic matching between intron and corresponding CDS

(a) Smith-Waterman local alignment. (b) Authentic matching.

$$N_{i1}=\begin{cases}[\dfrac{100n_{i1}}{n_{i0}}] & \dfrac{100n_{i1}}{n_{i0}}\text{为整数}\\ [\dfrac{100n_{i1}}{n_{i0}}]+1 & \dfrac{100n_{i1}}{n_{i0}}\text{为非整数}\end{cases} \tag{1}$$

$$N_{i2}=\begin{cases}[\dfrac{100n_{i2}}{n_{i0}}] & \dfrac{100n_{i2}}{n_{i0}}\text{为整数}\\ [\dfrac{100n_{i2}}{n_{i0}}]+1 & \dfrac{100n_{i2}}{n_{i0}}\text{为非整数}\end{cases} \tag{2}$$

公式中，中括弧为取整数运算．这样就将不同长度的 m 条比对序列转化成长度为 100 的标准序列了．

对 m 条标准比对序列的每个位点给予赋值．定义 f_{ij} 为第 i 条比对序列第 j 位点的赋值函数(j=1，2，…, 99, 100)，见公式(3)．凡是在最佳匹配区域内的位点赋予有效值 1，对区域外的位点赋予 0 值．

$$f_{ij}=\begin{cases}1 & N_{i1}\leqslant j\leqslant N_{i2}\\ 0 & j<N_{i1}\text{ 或 }j>N_{i2}\end{cases} \tag{3}$$

定义比对序列第 j 位点的匹配频数 P_j 值(Matching frequency)．它是 m 条比对序列第 j 位点上的赋值函数之和，

$$P_j=\sum_{i=1}^{m}f_{ij} \tag{4}$$

用 P_j 反映比对序列第 j 位点与被比对序列相互作用的几率，P_j 越高，说明该位点参于匹配的几率越大．这样在一组比对中，得到匹配频数 P_j 与比对序列相对位置 j 的分布．

1.3　对比过程

分 3 种情况进行比对．a．根据内含子在基因中的不同位置，将内含子分为第一内含子，第二内含子和其他内含子(线虫中含有多于 3 个内含子的基因数量较少，将其合为一类) 三类，将编码序列分为整个编码序列和外显子序列两类．以三类内含子序列为比对序列，两类编码序列为被比对序列分别做局域比对，得到 6 类内含子的匹配频数与其相对位置分布．b．由于内含子长度差异很大，根据 Halligan[17]研究结论，以 80 bp 为界将内含子分为长内含子和短内含子两类，以长、短两类内含子序列为比对序列，两类编码序列为被比对序列，分别进行比对，得到 4 类内含子匹配频数与内含子相对位置的关系．c．反过来，以整个编码序列为比对序列，3 类内含子、长短内含子和总体内含子共 6 类内含子序列为被比对序列，得到 6 类编码序列配对频数与其相对位置的关系．

2　结　　果

2.1　3 类内含子序列与相应编码序列的局域比对

Halligan 等研究表明，在内含子 5′端约 8 bp 和 3′端约 30 bp 的区域，是内含子剪接或内含子可变剪接的功能保守区域，故我们将内含子分为 3 个部分来分析．第一部分为 5′端剪接区，第二部分为内含子中部非保守区域，第三部分为包含多嘧啶层和 3′端剪接区[17]．以 3 类内含子序列为比对序列，整个编码序列为被比对序列分别做局域比对，得到 3 组内含子匹配频数同其相对位置的分布(图 2).

分析这 3 个分布发现，3 类内含子在中部区域与编码序列的匹配程度均高于两端 5′端剪接区和多嘧啶层 3′剪接区．这表明内含子的中部序列与编码序列存在较强的相互作用，这种相互作用反映了两者之间的一种协同性．比较图 2 中 3 种分布，第一内含子与编码序列的匹配主要在其内含子 5′端约 15%～55% 的范围内，而第二内含子匹配的最佳区域比较宽，约在 30%～80%的区域，其峰值靠近 3′端．其他内含子曲线涨落较大，这是由于线虫基因内含子平均个数为 2.7，其他内含子数目少造成的，但仍显示出中部匹配较强的特征．在第一内含子 3′端长度约 40%的区域内，与编码区的匹配明显降低，而第二和其他内含子 3′端与编码区的匹配是逐渐下降的，说明第一内含子的序列构成和作用与其他内含子确有差别．

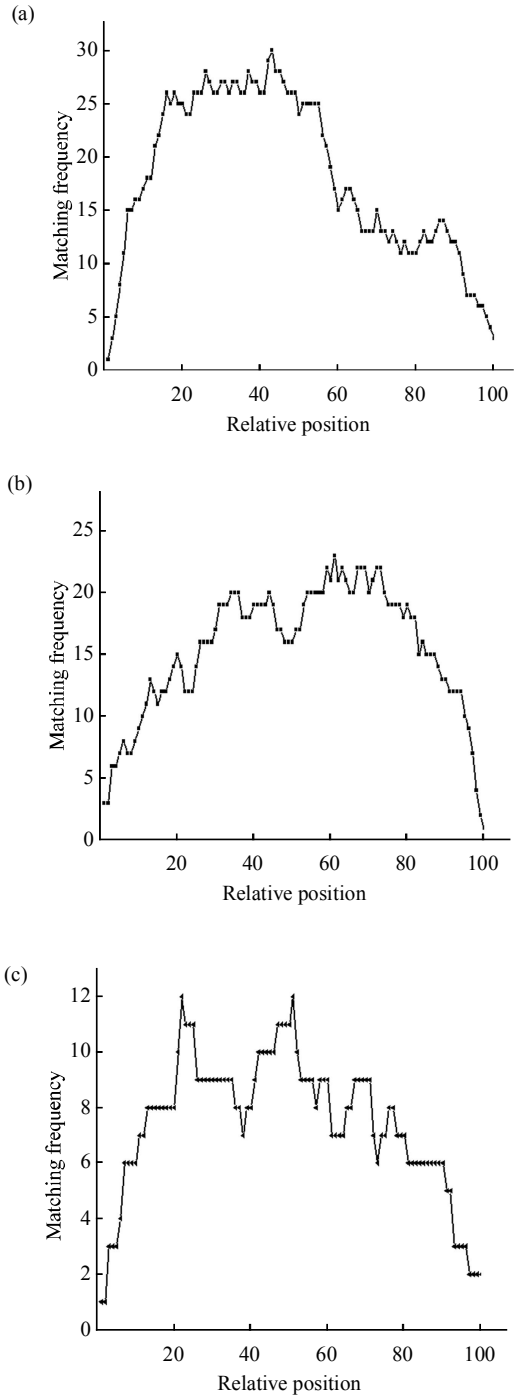

Fig. 2　Matching frequency according to intron relative position for the local alignment between intron and corresponding CDS

(a) First intron. (b) Second intron. (c) Other intron.

2.2　3 类内含子序列与相应外显子序列的局域比对

将编码序列上的外显子作为被比对序列，与 3 类内含子做比对分析见图 3．结果表明：三类内含子的匹配频数分布与上一节的基本相似，即中部区域与编码序列的匹配程度均高于两端的剪接区，但是仍有一些差别存在．这表现在：第一内含子分布，没有像图 2 那样在 5′端区有一段明显的高匹配区域．我们可以这样理解它们的差别：图 2 中，第一内含子的高匹配区域包含与外显子内的最佳匹配和外显子连接处的最佳匹配，而图 3 的第一内含子分布缺少外显子连接处的最佳匹配，才造成高匹配分布呈逐渐下降的趋势．这暗示与外显子内的匹配区比与外显子连接处的最佳匹配区更靠近 5′端．在大于第一内含子 60%的区域，分布基本没有改变．第二内含子中部有一个稳定的高匹配分布，但没有像图 2 那样在第二内含子后半部分出现一个高峰．

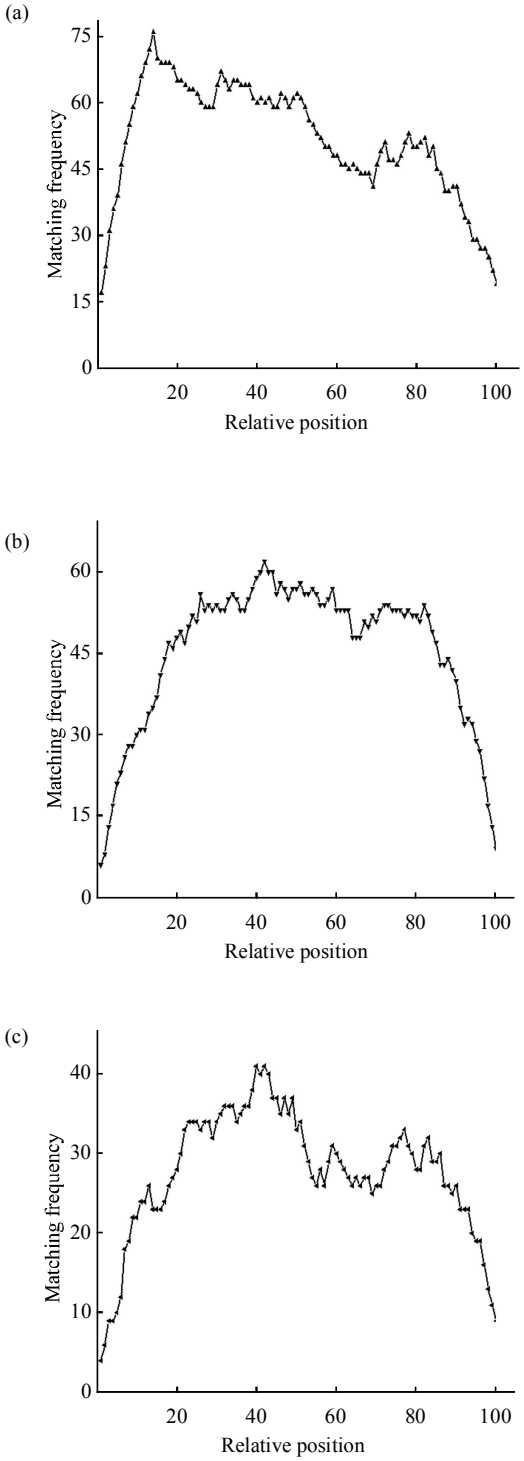

Fig. 3　Matching frequency according to intron relative position for the local alignment between introns and corresponding exons

(a) First intron. (b) Second intron. (c) Other intron.

2.3　长短内含子与相应外显子序列的局域比对

内含子序列与相应编码序列普遍存在一个高匹配区域，说明内含子序列和编码序列在这个区域具有较强的协同进化关系．为了进一步确定这个区域的大小，不区分内含子的位置，将内含子分为长内含子和短内含子两类，短内含子的平均长度为(53.0±9.7) bp，长内含子的平均长度为(177.3±73.1) bp．以长内含子或短内含子为比对序列，仍以整个编码序列和外显子序列作为被比对序列，分别进行局域比对．得到长短内含子与外显子序列和整个编码序列比对的分布．结果见图 4 和图 5．

图 4 分别给出长和短内含子序列与外显子序列的比对结果．发现，对短内含子序列与前面的分布基本一致，最佳匹配区域位于内含子中部，平均约在内含子第 8 个碱基位点到第 35 碱基位点之内，其长度约 27 个碱基．对长内含子分布却有很大的区别，这表现在最佳匹配频数显著区域位于长内含子的前半部分约在 5%～20%的范围内，即平均约在长内含子第 9 个碱基位点到第 36 碱基位点之内，其长度约 27 个碱基．表明与外显子协同进化的序列主要集中在长内含子前半部分，即紧邻 5′剪接位点区域；协同进化的序列长度仅限于一定的范围之内，初步估计大约在 26～40 bp 之内．在长内含子的中部和 3′区域，与外显子序列的配对频数很低．说明就长内含子序列段而言，与外显子序列发生作用的区域主要分布在长内含子的前端．

图 5 分别给出长和短内含子序列与整个编码序列的比对结果．结果显示：对短内含子分布，在中部仍有一个高匹配区域．对于长内含子，其分布与图 4 不同，分布出现两个峰区，分别在 5′剪接区和 3′剪接区附近．通过分析，发现前面的峰区与

图 4 的峰区性质是一样的，反映了与外显子内的最佳匹配．因为整个编码序列包含了外显子的链接部分，所以后面的峰区反映了内含子与外显子交接区域的相互作用．

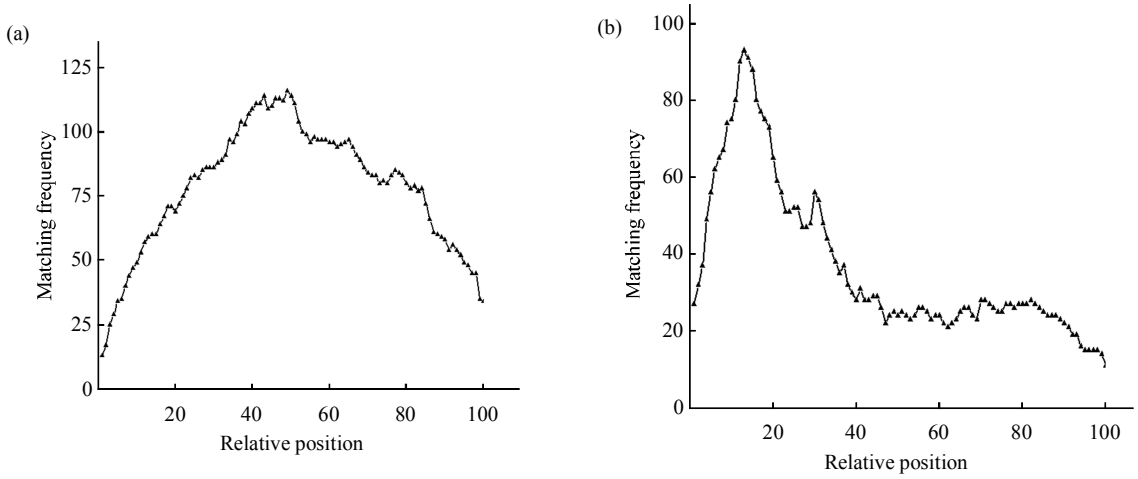

Fig. 4　Matching frequency according to intron relative position for the local alignment between long or short introns and corresponding exons

(a) Short intron. (b) Long intron.

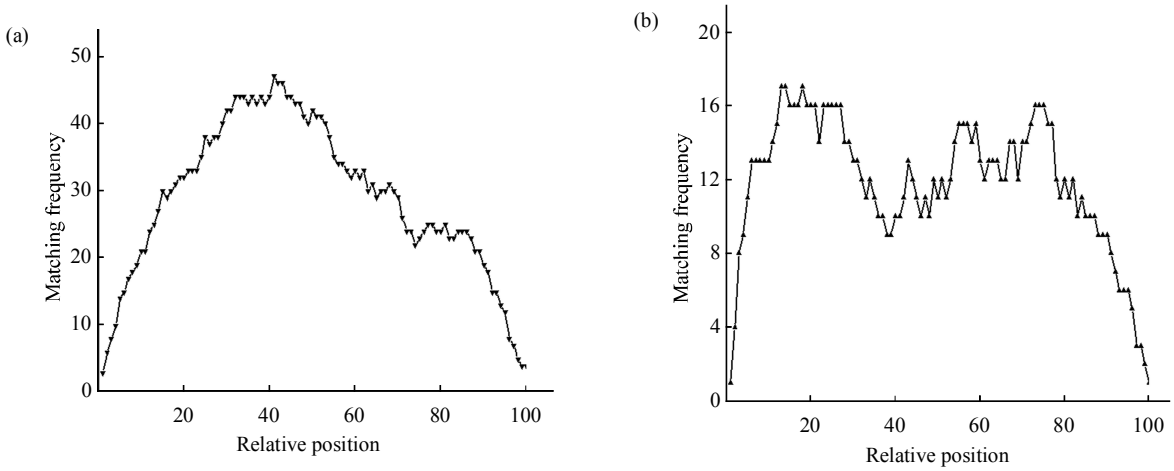

Fig. 5　Matching frequency according to intron relative position for the local alignment between long or short introns and corresponding CDS

(a) Short intron. (b) Long intron.

综合 **2.1**～**2.3** 的分析可以得到这样的结论：内含子中部序列与编码序列有较高的匹配频数，中前部分序列主要与外显子序列相互作用，中后部分主要与外显子链接区域相互作用．对于短内含子，两类相互作用区重叠在一起，因此无法区分这一特征．

2.4　编码序列与相应的内含子序列的局域比对

前面 **2.1**～**2.3** 的分析是从内含子序列的角度，探讨内含子序列上与编码序列或外显子序列的最佳匹配区域分布．接下来，我们将从编码序列的角度来探讨最佳匹配区域在编码序列上的分布情况．

首先，将整个编码序列作为比对序列，分别与

整体内含子、短内含子和长内含子进行局域比对，结果见图 6．其次，仍将整个编码序列作为比对序列，分别与第一内含子、第二内含子和其他内含子进行局域比对，结果见图 7.

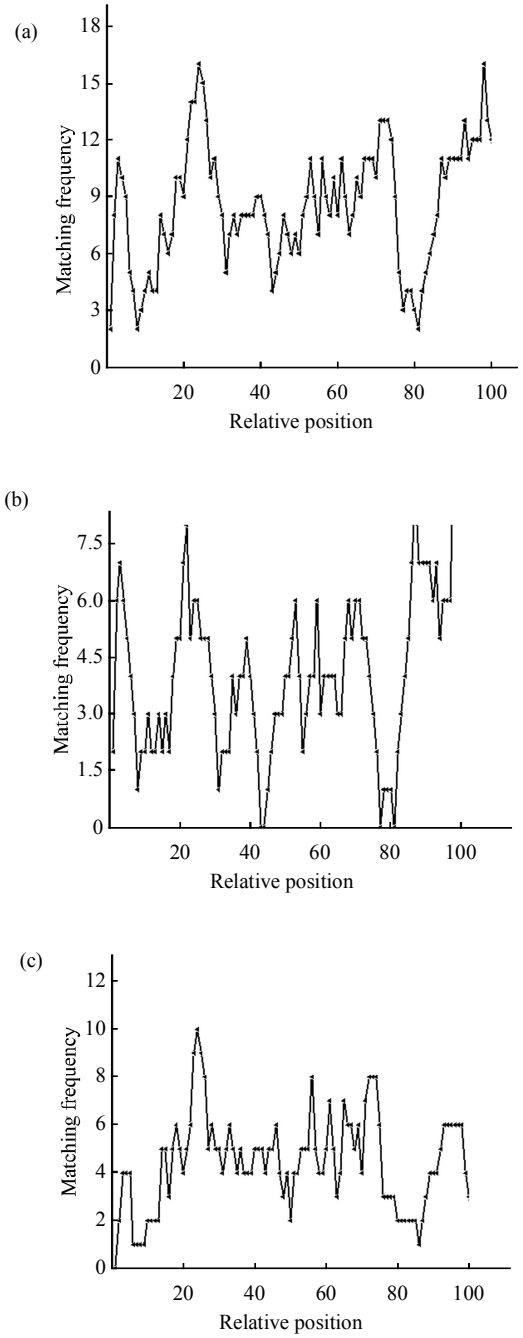

Fig. 6　Matching frequency according to CDS relative position for the local alignment between introns and corresponding CDS

(a) Intron. (b) Short intron. (c) Long intron.

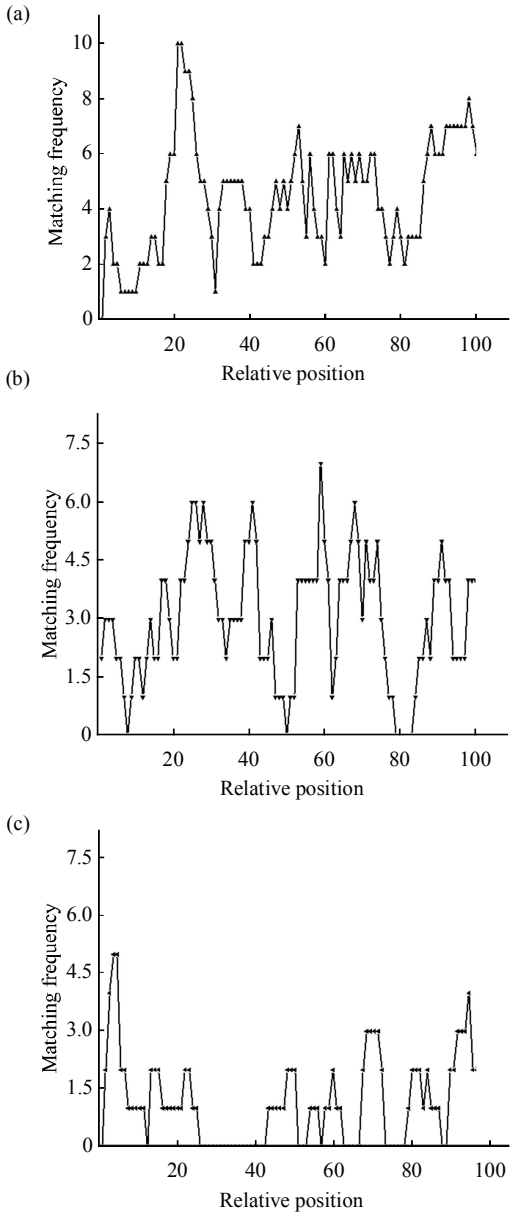

Fig. 7　Matching frequency according to CDS relative position for the local alignment between introns and corresponding CDS

(a) First intron. (b) Second intron. (c) Other intron.

比较图 6 和图 7 中的各个分布，发现整个编码序列与所有类型内含子的比对分布具有一些共同的特征．a．整个编码序列中存在许多与内含子匹配的峰值区域；b．在编码序列约 10%长度处和 80%长度处各有一个很低的匹配区域；c．在编码序列的两端均是与内含子的高匹配区域．d．在编码序

列约 25%长度处均有一个非常显著的高匹配区域.

我们推测，编码序列上确实存在许多与内含子结合的区域，也存在一些与内含子的禁配区域. 在编码序列约 10%长度处和 80%长度处存在的两个禁配区域是值得关注的. 我们认为，这些禁配区域可能是某些蛋白质因子特异性结合区域. 有证据表明，在真核生物细胞无义介导的 mRNA 降解过程中，需要外显子连接蛋白复合体(exon-junction protein complex，EJC)参与[18-19]，而 EJC 正处在外显子连接处附近，这与我们的禁配区域位置一致. 另外，在 mRNA 出核的过程中，Aly 等蛋白质因子也必须结合在编码序列的第一个 EJC 上游附近[20-23]. 因此在编码序列 5′端存在的禁配区域正好说明它们是蛋白质因子的特定结合区域这一现象. 反过来可以推测内含子序列是可以与编码序列结合的，是对本文论点的一个佐证. 3′端禁配区域应该也是蛋白质因子的结合区域，但需要进一步验证. 在编码序列的两端各有一个高配区域，其生物学意义还不太清楚. 我们推测与这两个高配区域结合的内含子可能参与 mRNA 的出核过程. 编码序列上(约 25%长度处)最显著的高匹配区域意义不清楚，是值得深入研究的.

3 讨　论

通过内含子序列与相应编码序列的局域比对分析，发现了内含子中部序列确实存在与编码序列的相互作用区域. 第一内含子和第二内含子与编码序列的匹配区域分布是有差别的，长内含子与外显子 - 外显子连接处和外显子相互作用区域的相对位置是不同的. 编码序列上存在多个与内含子序列的相互作用域和一些禁配区域的分布. 我们推测这些禁配区域是可能的蛋白质结合域. 结论同时印证了内含子序列与编码序列是协同进化的观点.

3.1 断裂基因结合处的信息特征

断裂基因片段(外显子与外显子)结合处的信息特征对真核基因是非常重要的. 通过长内含子分别与整个编码序列和外显子序列比对分析发现，长内含子后部序列与外显子 - 外显子结合处序列紧密相关，而前部序列与外显子内部序列紧密相关. 我们借助内含子从另一个角度发现了外显子 - 外显子结合区域的断裂痕迹或结合信息，这为我们探讨剪接信息特征提供了一种新的思路. 这个结论还给出了长内含子存在的一个理由，因为只有长内含子才能够将内含子不同的功能置放在内含子不同的物理区域，以便各个功能能够顺利表达.

3.2 有关内含子功能的猜想

我们认为内含子中部序列在被剪切后仍具有重要的生物学功能. 一个事实是：原核细胞没有核膜，原核基因没有内含子. 真核细胞有核膜，真核基因有内含子. 内含子与核膜有关吗？没人能够回答这个问题，但无法阻止人们去想这个问题. 我们总觉得，成熟 mRNA 在从核内输运到核外表达的过程中，必然有内含子的参与. 否则，核内大量的内含子序列被弃置不用是不符合生命体经济原则的. 我们认为，剪接后内含子的作用表现在：在出核前，通过内含子序列与成熟 mRNA 的相互作用来调控 mRNA 的结构. 其目的之一是与相关蛋白一起协助 mRNA 出核过程. 从物理学上讲，mRNA 是通过核孔出核的，如果仅靠 mRNA 自身形成的自然结构，出核是不可想象的. mRNA 应该是以一个比较伸展的状态出核的. 如果这个伸展状态仅靠核内的蛋白质来完成，需要的蛋白质数目是可观的，代价太大了，至少目前还没有观测到有许多蛋白质参与这个伸展过程. mRNA 剪接后与周围众多内含子片段弱性结合，调节编码序列的空间结构，使 mRNA 以一个比较伸展的状态出核. 这样内含子在真核细胞中存在就更具有积极的生物学意义. 在出核时，内含子也起到重要的作用. 产生这个观点是基于 Cheng 等的研究结果. Cheng 小组[20-23]给出了一个真核 mRNA 的出核机制，表明酵母的 mRNA 出核机制与人类的不同. 酵母的 mRNA 出核机制是偶联于基因的转录，而人类的 mRNA 出核机制是偶联于基因的剪接. 从另一个角度可解释它们的区别，因为酵母基因组中的基因缺乏内含子(只有 5%的基因有内含子)，而人类基因组中的基因含有丰富的内含子. 所以酵母只能采用与人类不同的机制协助 mRNA 的出核. 我们认为人类的 mRNA 出核机制与内含子直接相关. 出核后，mRNA 序列和结构仍需要内含子参与. 在基因表达过程中，内含子与编码序列的弱性结合还能够有效地保护编码序列不被损坏，并维持一个有利于翻译进行的序列结构，以便调控基因的翻译和延伸速率. 总之，在考虑了内含子的作用之后，对于 mRNA 的出核、翻译和延伸等生物过程机制的理解变得清晰和自然了. 另外，在研究蛋白质与核酸序列的相互作用过程中，人们遇到的一个最大困难

是寻找蛋白质特异性结合位点的问题．本文的结论给出编码序列上存在许多内含子与编码序列相互作用的匹配区域和禁配区域．可以想象，内含子和蛋白质或蛋白质复合体在与编码序列结合的竞争过程中，蛋白质复合体在露出的内含子禁配区域结合的几率胜过内含子．因此，内含子禁配区域就是可能的蛋白质结合位点，这使得寻找蛋白质结合位点变得容易了．我们猜测实际的生命过程大体也是这样．

3.3　内含子中部序列的保守性分析

短内含子中部序列与编码序列有较高的匹配频数，长内含子中前部序列、中后部序列分别与外显子序列和整个编码序列匹配程度较高．我们考虑在不同性质的匹配区域里内含子序列是否具有不同的序列结构特征．做为初步的检验，分别采用二阶信息冗余 $D_2(k)$值和二核苷偏好的平均差异 $\Delta D(m, n)$值来分析判断．

对序列 k，核酸序列的二阶(紧邻碱基)信息冗余定义为：

$$D_2(k)=\frac{1}{\ln 2}\sum_{i,\ j}\frac{(p_{ij}-p_ip_j)^2}{p_ip_j} \tag{5}$$

其中 p_i 或 p_j 为核酸序列中碱基 i 或 j 出现的概率(i, j =A, C, G, T,)，p_{ij} 为二核苷 ij 在序列中出现的联合概率．

对序列 k，二核苷偏好性定义为：

$$D_{ij}(k)=\frac{p_{ij}}{p_ip_j}-1 \tag{6}$$

其中 p_i、p_j 和 p_{ij} 的定义同(5)式．

对两类不同的序列 m 和 n，其二核苷偏好的平均差异定义为：

$$\Delta D(m, n)=\frac{1}{16}\sum_{i,\ j}|D_{ij}(m)-D_{ij}(n)| \tag{7}$$

具体做法是：取出每条短内含子峰区序列(见 **2.3**)，将所有峰区序列顺序连接起来组成一个新的序列，记为序列 s；取每条长内含子前一个峰区序列(见 **2.3**)，后一个峰区序列 (从长内含子的 60%处的碱基到末数第 9 个碱基) 分别将前一个峰区序列和后一个峰区序列连接起来组成两个新的序列,记为序列 m 和序列 n．然后按照公式(5)～(7)进行计算，结果见表 1．

发现长内含子后一个峰区序列的信息冗余与短内含子峰区序列相近，长内含子前一个峰区序列的信息冗余明显高于长内含子后一个峰区序列和短内含子序列，高出约 1 倍．3 类序列的二核苷偏好平均差异也显示出相同的结果．一般 $D_2(k)\geqslant 0$，随机序列此值大于 0．$D_2(k)$值越大，说明序列的碱基组成偏置性越强．结果表明：长内含子前一个峰区序列二核苷有相对强的偏置，明显高于长内含子后一个峰区序列和短内含子序列的碱基偏置．结果显示出内含子序列具有可区分的内部结构．

Table 1　Character comparisons among different intron sequences

$D_2(s)$	$D_2(m)$	$D_2(n)$	$\Delta D(m, n)$	$\Delta D(s, m)$	$\Delta D(s, n)$
0.046	0.095	0.054	0.143	0.142	0.057

The s stands for the short intron sequences, m stands for the first half of long intron sequences and n stands for the second half of long intron sequences. All of the analyzed intron sequences are excluded the conservative segments of 5' end and 3' end.

我们的研究还是初步的，只给出了相互作用区域的相对位置是不够的，应进一步给出作用区域的大小和序列特征．如哪些内含子序列是作用在外显子连接区，哪些是作用在外显子内部．应该从长内含子和短内含子结构的差异入手，进一步探讨内含子长度和序列结构的进化约束．另外需要将两个外显子链接起来与内含子比对，找出确切的禁配区域，以确定蛋白质或蛋白质复合体(EJC)结合的可能位置和序列特异性以及编码序列结合内含子后对其构象的影响等，这些都是下一步工作的重点．本文的研究所选取的样本规模还较小，因为在线一一比对耗时太多．今后将改进比对过程使之能进行集约化比对，对线虫基因组乃至其他基因组中全部基因序列进行分析，期望得到普适的结论．

参　考　文　献

[1] Scott W R, Alexei F, Walter G. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci USA, 2003, **100**(12): 7158−7162

[2] Elodie G, Tomàs M B, Olga F, *et al*. Patterns and rates of intron divergence between humans and Chimpanzees. Genome Biol, 2007, **8**(2): R21.1−R21.13

[3] John S M, Michael J G. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the

development of complex organisms. Mol Biol Evol, 2001, **18**(9): 1611-1630

[4] Ajit N, Shlomo H M, Meliss J M. A quantitative analysis of intron effects on mammalian gene expression. RNA, 2003, **9**(5): 607-617

[5] Gabriel M I, Pierre N, Peter D K, *et al*. Intron size and exon evolution in Drosophila. Genetics, 2005, **170**(1): 481-485

[6] Comeron J M. What controls the length of noncoding DNA?. Curr Opin Genet, 2001, **11**(6): 652-659

[7] Petrov D A. DNA loss and evolution of genome size in *Drosophila*. Genetica, 2002, **115**(1): 81-91

[8] Bartolome C, Masidex, Charlesworth B. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. Mol Biol Evol, 2002, **19**(6): 926-937

[9] Bergman C M, Krettman M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res, 2001, **11**(8): 1335-1345

[10] Maswell E S, Fournier M J. The small nucleolar RNAs. Ann Rev Biochem, 1995, **64**: 897-934

[11] Mattick J S. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep, 2001, **2**(11): 986-991

[12] Petrov D A, Sancster T A, Johnston J S, *et al*. Evidence for DNA loss as a determinant of genome size. Science, 2000, **287**(5455): 1060-1062

[13] Castillo-davis C I, Mekhedov S L, Hartl D L, *et al*. Selection for short introns in highly expressed genes. Nat Genet, 2002, **31**(4): 415-418

[14] Preahumwat A, Devincentts L, Palepoli M F. Intron size correlates positively with recombination rate in *Caenorhabditis* elegans. Genetics, 2004, **166**(3): 1585-1590

[15] Maki Y, Hung D N, Naoya K. Intron dynamics in ribosomal protein genes. PLoS ONE, 2007, **2**(1): e141

[16] Duret L. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet, 2001, **17**(4): 172-175

[17] Halligan D L, Keightley P D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res, 2006, **16**(7): 875-884

[18] Lejeune F, Maquat L E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr Opin Cell Biol, 2005, **17**(3): 309-315

[19] Baker K E. Nonsense-mediated mRNA decay: terminating erroneous gene expression. Curr Opin Cell Biol, 2004, **16**(3): 293

[20] Cheng H, Dufu K, Valencia P, *et al*. mRNA Export [J/OL]. Encyclopedia of Life Sciences [2007-09-28]. http://onlinelibrary.wiley.com/

[21] Cheng H, Dufu K, Lee B, *et al*. Human mRNA export machinery recruited to the 5' end of mRNA. Cell, 2006, **127**(7): 1389-1400

[22] Reed R, Cheng H. TREX, SR proteins and export of mRNA. Curr Opin Cell Biol, 2005, **17**(3): 269-273

[23] Masuda S, Das R, Cheng H, *et al*. Recruitment of the human TREX complex to mRNA during splicing. Genes Dev, 2005, **19**(10): 1512- 1517

Interactions Between Introns and Corresponding Protein Coding Sequences of Ribosomal Protein Genes in *C. elegans*[*]

ZHAO Xiao-Qing, LI Hong[**], BAO Tonglaga

(*School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, China)

Abstract Intron as a kind of non-coding DNA is rich in eukaryote genomes. The functions and involution mechanisms are not very clear besides the splicing. It was thought that introns play a very important role in maintaining and regulating the functional mRNA structure after splicing in the process of mRNA export and translation elongation, *etc*. Moreover, intron sequence and its corresponding coding sequence are existed interaction or co-evolution relations. The relations between intron sequences and its corresponding coding sequences were studied. For the *C. elegans* ribosomal protein genes, 85 genes were selected from RPG (http://www.cbi.pku.edu.cn/chinese/mirrors.html). The intron sequences were divided into first introns, second introns, other introns, short introns, and long introns and the corresponding coding sequences were divided into exons and all protein coding sequences (CDS), then the matching local alignment between introns and the corresponding coding sequences were done with Smith-Waterman local alignment software. The results show that there are really the interaction regions in introns when it is aligned with coding sequences. When intron sequences are aligned with CDSs, the significant interaction regions for the first intron and the other intron are located in about 15%～55% of intron length and it is located in about 30%～80% of intron length for the second intron. The distribution of interaction regions for short introns is similar to the distribution of the first introns. For long introns, there are two significant interaction regions. The first peak region is located about 15%～30% of intron sequence and the second peak region is located about 54%～78% of intron sequence. When long introns are aligned with exons, there is only one peak region. It is located in about 5%～20% of intron upstream region. When CDS are aligned with every kind of introns, it was found that there are many interaction regions and forbidden regions in CDSs. It was also found that there are two common forbidden regions in the CDSs, they are located at the 10% and 80% of coding sequence. The distribution of interaction regions for the first introns is different from the second introns. When compared the distributions of long introns aligned with CDS and aligned with exons, it can be concluded that the segment of the first peak region are acted on the inner exon segment, the segment of the second peak region are acted mainly on the exon-exon junction regions. Furthermore, there are many peak regions and forbidden regions which are distributed in protein coding sequences. It is speculated that the forbidden regions may be the combined regions of protein complex. In a word, all of the intron sequences besides the 5' end and 3' end correlate closely with their corresponding coding sequences or the two kinds of sequence segments are existed co-evolution relation.

Key words *C. elegans* ribosomal genes, introns, protein coding sequences, local alignment, interaction

DOI: 10.3724/SP.J.1206.2010.00186

*This work was supported by a grant from The National Natural Science Foundation of China (30660044).

**Corresponding author.

Tel: 86-471-6678889, E-mail: ndlihong@imu.edu.cn

Received: April 17, 2010 Accepted: June 14, 2010