多样性指标用于基因中剪切位点的识别*

张利绒 罗辽复**

（内蒙古大学物理系，呼和浩特 010021）

摘要 根据基因剪切位点处的碱基保守性特征，和附近位点的碱基组成和关联特征，应用多样性指标和二次判别分析，对几类模式生物的基因结构进行统一的分析和预测，能够较好地识别外显子和内含子及其边界. 计算结果表明，对于4类物种，线虫（*C. elegans*），拟南芥（*A. thaliana*），果蝇（*D. melanogaster*）和人类（human），核苷酸水平的识别精度为92.5%～97.1%，外显子水平的识别敏感性为83.7%～94.5%，特异性为87.8%～97.1%. 预测能力优于GeneSplicer等剪切位点检测软件.

关键词 剪切位点，多样性增量，二次判别法，外显子，内含子

学科分类号 Q61

真核生物基因识别的一个困难问题是Exon/Intron边界剪切位点的识别[1]. 虽然，和真核细胞pre-mRNA的剪切机制相联系，剪切位点处存在序列保守性，例如内含子5′端（供体位点）和3′端（受体位点）绝大部分是GT和AG，一小部分为GC和AG以及其他方式，但是，只根据GT-AG或GC-AG规则判断剪切位点会得到许多虚假的剪切位点. 为了进一步判断哪些是真实的剪切位点，必须综合其他信息. 针对不同的基因组，近年已经发展了多种剪切位点的识别和预测程序. 最近，文献[2]介绍了GeneSplicer，并和几种被认为最好的剪切位点检测器进行比较，如NetPlantGene[3]，NetGene2[3]，HSPL[4]，NNSplice[5]，GENIO，SpliceView等，证明GeneSplicer具有最高的识别效率. 为了把多种信息综合在一起，以上这些程序中使用了神经网络（NN），隐马氏模型（HMM），最大依赖分解（MDD）等算法. 我国学者在前些年也提出过一些算法[6,7]. 原则上，基因识别的途径有两种，一是基于同源识别的extrinsic方法，二是基于序列本身特性的intrinsic方法. 从信息来源讲也有两类，一是对内容敏感的content sensor，二是对信号敏感的signal sensor[8]. 任何一种成功的预测软件都必须综合多方面信息和兼顾上述两种识别途径[9,10]. 算法是软件的核心，一个好的算法也必须有这种综合能力. 本文将提出一个新的基于多样性指标的剪切位点识别方法. 多样性指标本来是生物相似性关系的一种定量表示[11,12]，本文将它适当改造推广，和二次判别法[13]结合起来，用于剪切位点的识别和预测. 它能自动地综合序列本身特性和类似序列的比较，综合剪切位点附近的保守信号和边界两旁序列的碱基组分特征和关联性[14,15]. 这个方法还具有简明易操作，学习参数很少的特点. 本文将使用这个统一的方法对5个基因组进行剪切位点识别，能得到比GeneSplicer等通用软件更好的识别效果.

1 数据与方法

1.1 数据

文中的数据于2002年12月27日从http://mcb.harvard.edu/gilbert/EID下载[16]，所涉及的基因都是实验上已经确定的. 从中挑选出5类物种的基因作为研究对象，并删去含有未知碱基及非A、T、C、G的基因，删去含有partial CDS及基因编码区总长度非3倍数的基因，删去序列重合的基因，得到基因数：线虫（*C. elegans*），185＋22；酵母（*S. pombe*），203＋20；拟南芥（*A. thaliana*），749＋143；果蝇（*D. melanogaster*），1 196＋65；人类（human），1 231＋103（＋号后为含有非标准剪切，即含剪切位点非GT/AG的基因）. 这是本文研究的基因集合，除了非标准剪切集外，随机分为训练集和检验集. 5类物种3个集合中所包含的基因数目、外显子和内含子数目如表1.

* 国家自然科学基金资助项目（90103030）.

** 通讯联系人.

Tel：0471-4992676，E-mail：lfluo@mail.imu.edu.cn

收稿日期：2003-06-11，接受日期：2003-07-31

Table 1 The number of gene，exon and intron in training set，test set and non-standard splicing set for five species

		Gene	Exon	Intron
	Train	95	551	456
C. elegans	Test	90	587	497
	Non_ standard	22	168	146
	Train	121	396	275
S. pombe	Test	82	292	210
	Non_ standard	20	66	46
	Train	387	2 094	1 707
A. thaliana	Test	362	2 188	1 826
	Non_ standard	143	1 072	929
	Train	619	1 890	1 271
D. melanogaster	Test	577	1 832	1 255
	Non_ standard	65	330	265
	Train	636	3 488	2 852
Human	Test	595	3 347	2 752
	Non_ standard	103	829	726

1.2 识别算法中的参数定义

基因剪切位点邻近的碱基存在较强的保守性，碱基与碱基之间存在关联性. 同时，两侧的外显子和内含子具有不同的序列统计特征. 这些是识别剪切位点的基础信息.

首先，选择供体端 ×GT × × × × 的 7 个位点，受体端 × × × × × × AG 的 8 个位点，研究这些位点的碱基保守性. a. 统计除 GT/AG 位的 s 个（s = 5，6）位置上 4 种碱基的数目，定义 $4s$ 维矢量（x_1，x_2，…，x_{4s}），反映碱基保守性；b. 统计 s 碱基片段紧邻和非紧邻位点碱基二联体的数目，定义 $m = C_s^2 \times 16$ 维矢量（x_1，x_2，…，x_m），反映碱基关联特性；c. 统计 s 碱基片段中任意 3 个位点上碱基三联体的数目，定义 $m = C_s^3 \times 64$ 维矢量（x_1，x_2，…，x_m），反映碱基的三重关联.

其次，截取 exon/intron 边界前 L_1 个和后 L_2 个碱基（包括 GT 和 AG），分别称为 L_1 或 L_2 序列，统计序列中的碱基三联体数，定义 64 维矢量（x_1，x_2，…，x_{64}），反映供体位点前或后和受体位点前或后序列的组分特征. 本文取 $L_1 = L_2 = 48$.

对于一对 m 维矢量 X：[x_1，x_2，…，x_m] 和 Y：[y_1，y_2，…，y_m]，其多样性的定义为

$$D(X) = D(x_1, x_2, \cdots, x_m) = M\log_b M - \sum_{i=1}^{m} x_i \log_b x_i \quad (1)$$

$$D(Y) = D(y_1, y_2, \cdots, y_m) = N\log_b N - \sum_{i=1}^{m} y_i \log_b y_i \quad (2)$$

$$D(X+Y) = D(x_1 + y_1, x_2 + y_2, \cdots, x_m + y_m) = (M+N)\log_b(M+N) - \sum_{i=1}^{m} (x_i + y_i)\log_b(x_i + y_i) \quad (3)$$

其中

$$M = \sum_{i=1}^{m} x_i, \quad N = \sum_{i=1}^{m} y_i$$

X 与 Y 的多样性增量为

$$\Delta(X,Y) = D(X,Y) - D(X) - D(Y) \quad (4)$$

根据上述定义，在训练集中建立剪切位点邻近（保守位点）单碱基、碱基二联体和碱基三联体的标准多样性源，以及 L_1 序列和 L_2 序列的碱基三联体组分的标准多样性源. 对于一个待测序列，首先根据 GT-AG 或 GC-AG 规则找到可能的剪切位点，然后在可能的剪切位点邻近计算单碱基、碱基二联体和碱基三联体的多样性，以及待测序列可能剪切位点前后的组分多样性. 根据多样性增量的定义，得到判别分析的 8 个变量：①待测序列可能剪切位点邻近单碱基多样性与训练集单碱基标准多样性源之间的多样性增量；②待测序列可能剪切位点邻近碱基二联体多样性与训练集碱基二联体标准多样性源之间的多样性增量；③待测序列可能剪切位点邻近碱基三联体多样性与训练集碱基三联体标准多样性源之间的多样性增量；④待测序列的 L_1 序列碱基三联体多样性与训练集 L_1 序列的碱基三联体标准多样性源之间的多样性增量；⑤待测序列的 L_1 序列碱基三联体多样性与训练集 L_2 序列的碱基三联体标准多样性源的多样性增量；⑥待测序列的 L_2 序列碱基三联体多样性与训练集的 L_1 序列的碱基三联体标准多样性源的多样性增量；⑦待测序列的 L_2 序列碱基三联体多样性与训练集 L_2 序列的碱基三联体标准多样性源的多样性增量；⑧待测序列的 L_1 序列的碱基三联体多样性与它的 L_2 序列的碱基三联体多样性之间的多样性增量. 前 3 个多样性计算中（1）式的求和分别对 $4s$，$C_s^2 \times 16$，$C_s^3 \times 64$ 个变量进行，后 5 个多样性计算中（1）式的求和都对 64 种三联体进行（图 1）.

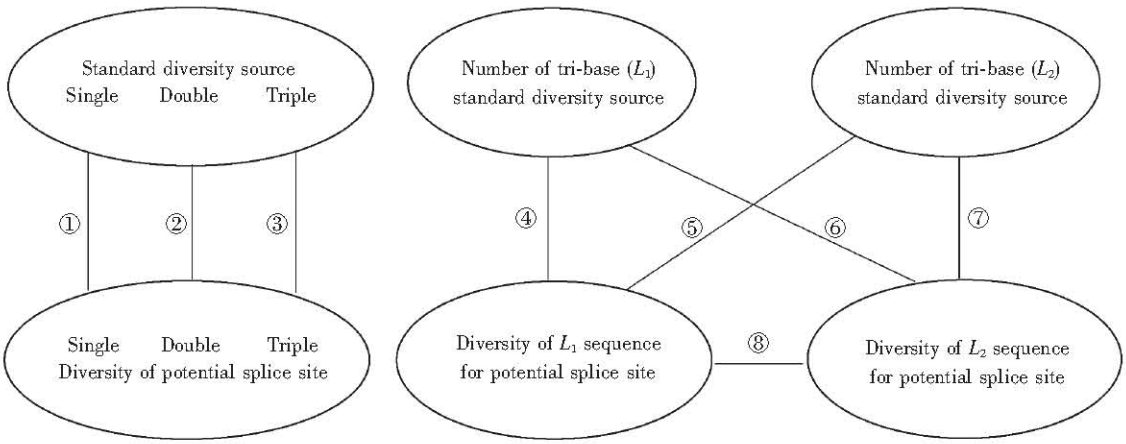

Fig. 1 The eight discriminant variables obtained by calculating the increment of diversity

1.3 基于多样性指标的二次判别分析识别算法

为了提高判别效率，我们在可能的剪切位点集合中只考虑多样性增量的值满足④<⑤或⑥>⑦的样本. 进一步，在训练集中将它们分为真实的剪切位点集合（positive 集）和虚假的剪切位点集合（negative 集）. 在通常理论中，为了判别一个样本在正集或负集，只需计算它的多样性以及它和两个集合的多样性增量，看哪一个多样性增量小，就把它归为这个集合. 而现在我们有 8 个多样性增量，无法照此处理. 本文建议用二次判别法来识别. 为此需计算参数：正集的个数 p，负集的个数 q；正集的各变量平均值 $\vec{\mu}_1$，负集的各变量平均值 $\vec{\mu}_2$；正集的协方差矩阵 Σ_1，负集的协方差矩阵 Σ_2；正集的协方差矩阵行列式的值 $|\Sigma_1|$，负集的协方差矩阵行列式的值 $|\Sigma_2|$. 如

$$\vec{\mu}_1 = \{\mu_1[0], \mu_1[1], \mu_1[2], \mu_1[3], \mu_1[4], \mu_1[5], \mu_1[6], \mu_1[7]\}$$

$$\vec{\mu}_2 = \{\mu_2[0], \mu_2[1], \mu_2[2], \mu_2[3], \mu_2[4], \mu_2[5], \mu_2[6], \mu_2[7]\}$$

（式中 0，…，7 代表 8 种多样性增量）；类似地，协方差矩阵为 8×8 矩阵，每一元素代表一对多样性增量的协方差. 任意一个待确定的可能剪切位点由 X 表示：

$$\vec{X} = \{x[0], x[1], x[2], x[3], x[4], x[5], x[6], x[7]\}$$

判别函数为

$$\xi = \lg \frac{p}{q} - \frac{\delta_1 - \delta_2}{2} - \frac{1}{2} \lg \frac{|\Sigma_1|}{|\Sigma_2|} \qquad (5)$$

其中 δ_i 为

$$\delta_i = (\vec{X} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{X} - \vec{\mu}_i) \quad (i = 1, 2) \qquad (6)$$

在通常的二次判别法中，如果 $\xi > 0$，则 $\vec{X}$ 被判为真，否则判为虚假. 但现在要同时考虑供体端和受体端才能有正确的判别，故令供体端和受体端 ξ 的域值 ξ_D 和 ξ_A 由训练集确定，供体位点 $\xi > \xi_D$ 为真，受体位点 $\xi > \xi_A$ 为真.

2 结 果

ξ_D 和 ξ_A 是本文方法中唯一需定的参数. 定出的值列于表 2 第 2 列. 我们对训练集、检验集和非标准剪切集内的基因剪切位点进行预测，从外显子水平和碱基水平两个角度衡量预测精度. 对于前者，统计基因中两侧剪切位点都预测正确的外显子数 N_1 和仅一侧剪切位点预测正确的外显子数 N_2，令实际外显子数 N_{exon}，预测出的外显子数 N_{pre_exon}，定义敏感性和特异性为

$$Sn = (2N_1 + N_2)/2N_{exon}$$

$$Sp = (2N_1 + N_2)/2N_{pre\_exon} \qquad (7)$$

对于后者，逐个碱基考虑，在外显子位点上预测正确（预测为外显子）的碱基百分数称为 Ac（e），在内含子位点上预测正确（预测为内含子）的碱基百分数称为 Ac（o），全部位点上预测正确的碱基百分数称为 Ac（all）. 对训练集和检验集的预测精度见表 2 和表 3. 本文的预测方法也适用于 GC/AG 型非标准剪切. 对非标准剪切集进行预测时，除了 GT/AG 型剪切的保守位点外，还要考虑 GC/AG 型保守位点. 对非标准剪切集的预测结果见表 4.

Table 2 The accuracy of prediction for splice sites in training set

Species	ξ_D ξ_A	*Sn*/%	*Sp*/%	*Ac*(e)/%	*Ac*(o)/%	*Ac*(all)/%
C. elegans	(-10, -3)	97.2	97.7	98.6	97.2	98.0
S. pombe	(-5, -5)	90.8	93.6	90.8	94.8	91.3
A. thaliana	(-6, -5)	94.0	95.0	98.6	97.5	98.2
D. melanogaster	(-2, -5)	94.9	97.2	97.6	94.9	96.8
Human	(-4, -1)	86.2	89.9	90.9	94.1	93.5

Table 3 The accuracy of prediction for splice sites in test set

Species	ξ_D ξ_A	*Sn*/%	*Sp*/%	*Ac*(e)/%	*Ac*(o)/%	*Ac*(all)/%
C. elegans	(-10, -3)	93.2	96.3	97.0	96.6	96.8
S. pombe	(-5, -5)	79.8	87.9	84.3	89.1	84.9
A. thaliana	(-6, -5)	91.3	93.4	96.8	97.6	97.1
D. melanogaster	(-2, -5)	94.5	97.1	96.7	96.3	96.6
Human	(-4, -1)	83.7	87.8	89.4	93.1	92.5

Table 4 The accuracy of prediction for splice sites in non-standard splicing set

Species	ξ_D ξ_A	*Sn*/%	*Sp*/%	*Ac*(e)/%	*Ac*(o)/%	*Ac*(all)/%
C. elegans	(-10, -3)	78.0	79.4	97.9	96.9	97.5
S. pombe	(-5, -5)	48.5	50.0	73.1	82.4	76.6
A. thaliana	(-6, -5)	75.5	78.5	93.7	91.7	92.7
D. melanogaster	(-2, -5)	66.2	72.4	84.1	93.1	87.0
Human	(-4, -1)	69.8	74.5	83.9	94.0	92.1

3 讨 论

依据剪切信号 GT-AG 或 GC-AG 处的碱基保守性特征和附近位点的碱基组成和关联特征，应用多样性指标和二次判别分析，对 5 类模式生物的基因结构进行统一的分析和预测，能够较好地识别外显子和内含子及其边界. 8 个多样性增量能较好地反映出剪切位点邻近各个位点上碱基的保守性和关联性，还能体现出外显子和内含子序列构成的统计差别. 同时，在本方法中一个待测序列剪切位点的确定，不仅决定于这个序列本身，还决定于和标准多样性源中蕴含的统计信息的比较，这体现了 extrinsic 方法和 intrinsic 方法的结合. 另外，由于 8 个多样性增量间不是独立的，它们间有复杂的关系，我们发现用 Fisher 线性判别法不能得出好的预测，而采用二次判别分析，就能有效地预测出绝大多数实际的剪切位点. 对几个不同的物种，信号提取方式，参数计算方法以及位点的筛选等都保持一致，这既说明了物种在剪切过程中剪切机制的基本一致，也指出了这个方法的有效性. 而理论中的调节参数只有两个，也显示出本方法的吸引力. 从结果看也比较满意，所研究的 5 个物种除 *S. pombe* 外，核苷酸水平的识别精度为 92.5% ~97.1%，外显子水平的识别敏感性为 83.7% ~94.5%，特异性为 87.8% ~97.1%. *S. pombe* 较低，可能与样本数较少有关，因为标准多样性源中含有足够多的样本才能给出正确的统计信息.

文献［2］曾用 GeneSplicer 对 *A. thaliana* 和 human 的剪切位点进行预测，把 GT 和 AG 位点中真正的剪切位点称为 true，其他为 false. 预测依赖于未被预测到的 true 位点百分数，对它设定后，预测精度可用 false 集合中被预测为 true 的位点百分数表示. 后者愈小，精度愈高. 为了和 GeneSplicer 做比较，用我们的方法，但按照完全相同的口径采用分 3 组交叉检验的程序进行预测，对 *A. thaliana* 和 human 的预测结果由表 5 和表 6 给出. 表中最后

一列给出文献［2］的预测结果. 如不计两文所用基因数量的不同引起的差别（该文为 *A. thaliana* 1 323 和 human 1 115），本文的多数结果（对 *A. thaliana* 为 15∶1，对 human 为 12∶4）优于 GeneSplicer. 再考虑到文献［2］的程序在输入剪切位点信息时所用的邻近碱基数量远比本方法多（窗口大），可以认为本文方法的预测能力是比较高的. 如果统计检验方法进一步改善，对参数 ξ_D 和 ξ_A 做进一步的研究和调节，扩大窗口，并对一些物种扩大样本数，当可获得更满意的预测结果.

Table 5 False negative and false positive rates for acceptor and donor site detection on three disjoint partitions of a 749 gene *A. thaliana* data set

	True site missed/%	False positive/%				
		Part 1	Part 2	Part 3	Average	Contrast
Acceptor site (ag) detection	3	8.21	4.79	6.94	6.65	11.7
(3 533 true, 91 525 false)	5	3.45	2.55	2.66	2.89	4.9
Positive: Part 1 1 251	7	2.11	1.81	1.85	1.92	3.3
Part 2 1 201	8	1.85	1.50	1.59	1.65	2.9
Part 3 1 081	10	1.43	1.17	1.14	1.25	2.4
Negative: Part 1 30 206	15	0.82	0.71	0.69	0.74	1.6
Part 2 32 373	20	0.61	0.55	0.46	0.54	1.1
Part 3 28 946	30	0.31	0.31	0.31	0.31	0.7
Donor site (gt) detection	3	5.16	3.59	3.03	3.93	4.7
(3 533 true, 141 850 false)	5	3.14	2.11	1.85	2.37	2.8
Positive: Part 1 1 251	7	2.05	1.51	1.52	1.69	1.9
Part 2 1 201	8	1.87	1.33	1.36	1.52	1.7
Part 3 1 081	10	1.45	1.16	1.19	1.27	1.4
Negative: Part 1 47 315	15	0.98	0.76	0.76	0.83	0.9
Part 2 49 494	20	0.73	0.58	0.52	0.61	0.6
Part 3 45 041	30	0.42	0.34	0.34	0.37	0.4

Table 6 False negative and false positive rates for acceptor and donor site detection on three disjoint partitions of a 1 231 gene human data set

	True site missed/%	False positive/%				
		Part 1	Part 2	Part 3	Average	Contrast
Acceptor site (ag) detection	3	5.16	3.84	4.55	4.52	9.3
(5 604 true, 511 333 false)	5	2.52	1.98	2.28	2.26	5.8
Positive: Part 1 1 843	7	1.55	1.28	1.55	1.46	4.7
Part 2 1 841	8	1.25	1.05	1.37	1.22	4.3
Part 3 1 920	10	0.86	0.80	1.09	0.92	3.7
Negative: Part 1 159 645	15	0.45	0.43	0.58	0.49	2.6
Part 2 196 136	20	0.30	0.26	0.33	0.30	1.9
Part 3 155 552	40	0.09	0.09	0.10	0.09	0.8
Donor site (gt) detection	3	10.18	12.20	7.84	10.07	14.7
(5 604 true, false 765 291)	5	7.05	8.04	5.92	7.00	6.4
Negative: Part 1 1 843	7	5.55	5.87	4.41	5.28	4.8
Part 2 1 841	8	5.11	5.45	4.02	4.86	4.1
Part 3 1 920	10	4.33	4.30	3.28	3.97	3.5
Negative: Part 1 233 681	15	2.89	2.52	1.90	2.44	2.5
Part 2 295 501	20	1.91	1.62	1.18	1.57	1.8
Part 3 236 109	40	0.32	0.32	0.30	0.31	0.7

在前人大多数剪切位点预测工作中都没有考虑非标准剪切，本文的预测方法原则上也适用于非标准剪切. 表 4 给出了对非标准剪切集的预测结果. 我们发现核苷酸水平的预测精度较标准剪切检验集有所降低，但降低幅度不大，而在外显子水平的降低幅度较大. 由于非标准剪切数量不够，没有单独进行训练，预测中都用标准剪切的训练集给出的参数，另外我们只计入了 GC/AG 型而未能对 GC/AG 以外的非标准剪切位点进行搜索，所以预测精度降低是可以理解的. 而在核苷酸水平考察，对非标准剪切的预测结果达到了较高的预测精度，却是令人鼓舞的. 这说明非标准剪切和标准剪切下的 exon/intron 结构的碱基统计性质基本相同.

参 考 文 献

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, **409** (6822): 860 ~ 921

2 Pertea M, Lin X Y, Salzberg S L. Genesplicer: a new computational method for splice site prediction. Nucleic Acids Res, 2001, **29** (5): 1185 ~ 1190

3 Hebsgaard S M, Korning P G, Tolstrup N, *et al*. Splice site prediction in *A. thaliana* DNA by combining local and global sequence information. Nucleic Acids Res, 1996, **24** (17): 3439 ~ 3452

4 Hubbard T, Birney E, Bruskiewich R, *et al*. HSPL: Splice Site Prediction in Human DNA. Cold Spring Habor Meeting on Genome Sequencing and Biology. Cold Spring Harbor, 1999

5 Reese M E, Eeckman F H, Kulp D, *et al*. Improved splice site detection in Genie. J Comput Biol, 1997, **4** (3): 311 ~ 323

6 郑 毅，丁达夫. 果蝇内含子 3′剪切位点的选择机制. 生物物理学报，1994，**10** (3): 459 ~ 464
Zheng Y, Ding D F. Acta Biophy Sin, 1994, **10** (3): 459 ~ 464

7 闻 芳，卢 欣，孙之荣，等. 基于支持向量基 (SVM) 的剪切位点识别. 生物物理学报，1999，**15** (4): 733 ~ 738
Wen F, Lu X, Sun Z R, *et al*. Acta Biophy Sin, 1999, **15** (4): 733 ~ 738

8 Mathe C, Sagot M F, Schiex T, *et al*. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res, 2002, **30** (19): 4103 ~ 4117

9 Guigo R, Agarwal P, Abril J F, *et al*. An assessment of gene prediction accuracy in large DNA sequences. Genom Res, 2000, **10** (10): 1631 ~ 1642

10 Brendel V, Kleffe J, Carle-Urioste J C, *et al*. Prediction of splice sites in plant Pre-mRNA from sequence properties. J Mol Biol, 1998, **276** (1): 85 ~ 104

11 Laxton R R. The measure of diversity. J Theor Biol, 1978, **71** (1): 51 ~ 67

10 徐克学. 生物数学. 北京：科学出版社，1999. 277
Xu K X. Biomathematics. Beijing: Science Press, 1999. 277

12 Zhang M Q. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci USA, 1997, **94** (2): 565 ~ 568

13 Cai D, Delcher A, Kao B, *et al*. Modeling splice sites with Bayes networks. Bioinformatics, 2000, **16** (2): 152 ~ 158

14 Arita M, Tsuda K, Asai K. Modeling splicing sites with pairwise correlations. Bioinformatics, 2002, **18** (1): 1 ~ 8

15 Saxonov S, Daizadeh I, Fedorov A, *et al*. EID: the Exon-Intron Database——an exhaustive database of protein-coding intron-containing genes. Nucl Acids Res, 2000, **28** (1): 185 ~ 190

Recognition of Splice Sites in Genes by Use of Diversity Measure Method*

ZHANG Li-Rong, LUO Liao-Fu**

(*Department of Physics, Inner Mongolia University, Hohhot* 010021, China)

Abstract The conservation of nucleotides at splicing sites and the characteristics of base composition and base correlation in the adjacent segment sequences have been investigated by use of the method of diversity measure combined with quadratic discriminant analysis. About 4 000 genes in five model genomes have been studied. The splicing sites and the exon/intron boundaries are recognized and predicted. The preliminary calculation shows that, through this simple and unified approach the prediction accuracy on the nucleotide basis is from 92.5% to 97.1% for *C. elegans*, *A. thaliana*, *D. melanogaster* and human. The prediction sensitivity and specificity on the exon basis are 83.7% ~ 94.5% and 87.8% ~ 97.1% respectively for these genomes. Non-canonical splicing has also been analyzed. The prediction capacity of the present method is comparable with GeneSplicer and other current splice site detectors.

Key words splice site, increment of diversity, quadratic discriminant analysis, exon, intron

* This work was supported by grants from The National Natural Sciences Foundation (90103030).

** Corresponding author. Tel: 86-471-4992676, E-mail: lfluo@mail.imu.edu.cn

Received: June 11, 2003 Accepted: July 31, 2003