PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(10): 1188~1194
www.pibb.ac.cn

人类基因组中可变和组成性剪接位点的预测 *

张利绒 ** 罗辽复 邢永强 晋宏营
(内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 根据剪接位点的核酸序列保守特征，以及邻近位点的碱基组成和关联特性，结合一对可变剪接位点之间的距离参数和受体端剪接位点前 30 位碱基的 GC 和 TC 含量，利用结合多样性指标的二次判别方法(IDQD)，预测了人类基因组中可变和组成性内含子的供体端和受体端的剪接位点，对可变的供体端和受体端剪接位点，阈值 ξ 选择–2 时，总的预测精度分别为 87.9%和 89.9%，对组成性的供体端和受体端剪接位点，阈值 ξ 选择–1，总的预测精度分别为 92.8%和 94.3%.

关键词 可变剪接，组成性剪接，二次判别方法
学科分类号 Q61

可变剪接是指从一个 mRNA 前体中通过不同的剪接方式(选择不同的剪接位点组合)产生不同的 mRNA 剪接异构体的过程．可变剪接是调节基因表达和产生蛋白质组多样性的重要机制[1, 2]．高通量的基因组测序和表达序列标签(EST)测序，使得利用生物信息学的方法研究可变剪接问题成为可能．最近几年，多个研究小组通过不同的生物信息学方法，从整个人基因组的水平进行分析，结果一致显示约 75%的人类基因有可变剪接形式[3~6].

可变剪接能够产生惊人的多样性，但我们对其调节机制所知不多．剪接位点的选择受到结合到非剪接位点 RNA 元件剪接因子的多重调节．参与可变剪接调节的 RNA 元件包括 ESE、ISE、ESS、ISS．剪接因子包括 SR 和 hnRNP 家族蛋白质等多种因子．剪接反应由剪接体执行，剪接体包括 5 个小核糖核蛋白复合体 U1，U2，U4，U5 和 U6，snRNPs 和 50～100 种非 snRNP 蛋白．剪接体通过 RNA-RNA，RNA- 蛋白质，蛋白质 - 蛋白质等多重相互作用以精确切除每个内含子和以正确次序连接外显子.

内含子中有 4 种短核苷酸序列可以作为信号，指导剪接体在何处，包括：位于内含子起始处的剪接信号，叫 5′端(供体端)剪接位点；位于内含子末端的剪接信号，叫 3′端(受体端)剪接位点；位于内含子末端的分支位点；以及最后的多聚嘧啶区[7]. 由小核 RNA 和蛋白质形成的剪接体(splicesome)，负责识别内含子起始和终止的部位，并从前体 mRNA 转录本中切掉内含子，把外显子连接起来形成 mRNA．而为了有效剪接，绝大部分内含子需要：a．一个保守的 5′剪接位点，A/CAG↓GURAGU；b．一个分支点序列 BPS，YNYURAY，后面跟着一个多聚嘧啶尾，Y10～20；c．一个 3′剪接位点 Y/AG．通常情况下，可变的内含子具有共同的供体(或受体)端，有两个不同的受体(或供体)端，从而形成一对可变内含子.

虽然，已经有许多预测基因结构的软件，而且得到了很好的预测结果，但目前还没有一种有效的方法来预测基因的可变剪接位点．有些研究工作利用表达序列标签(ESTs)或 cDNA 文库来探测可变剪接事件[8~10]，有些是利用机器学习和基因的同源比较来识别[11~14]，Wang 等[15]提出结合模式识别和组分模式的神经网络方法，Xia 等[16, 17]引入基于剪接位点的竞争机制来识别可变的供体和受体剪接位点，供体端竞争位点对(CSSP)的预测总精度为

* 国家自然科学基金资助项目(10447003, 90403010)和内蒙古自治区自然科学基金资助项目(200308020102).
** 通讯联系人.
Tel: 0471-4995052, E-mail: pyzlr@mail.imu.edu.cn
收稿日期：2008-01-30，接受日期：2008-04-08

89.21%，受体端竞争位点对(NCSSP)的预测总精度为 87.98%，而在此基础上对可变剪接位点的预测准确率约 70%，相应的假阳性率约 30%.

本文利用 ASD 数据库[18]，以人类基因组作为研究对象，构建可变的供体和受体端剪接位点数据库，作为对比，构建组成性的供体和受体端剪接位点数据库，利用结合多样性指标的二次判别方法[19~22]来预测人类基因组中的组成性剪接位点和可变剪接位点. 对可变剪接位点，选择 ξ 阈值为–2 时，预测总精度为：供体端 87.9%，受体端 89.9%；对组成性剪接位点，选择 ξ 阈值为–1 时，供体端 92.8%，受体端 94.3%，这个精度在目前已见报道的预测中是较高的.

1　可变剪接的模式

基本上，可变剪接包括 5 种基本类型(图 1)[18]：

a. 内含子的保留(intron retention, IR). 一段核苷酸序列在一个转录体中是外显子的一部分而在另外一个可变的转录体中是一个内含子.

b. 可变的供体剪接位点(alternative donor introns, AD). 具有竞争的供体(5′端)剪接位点，而内含子下游的受体位点相同.

c. 可变受体剪接位点(alternative acceptor introns, AA). 具有竞争的受体(3′端)剪接位点，而内含子上游的供体位点相同.

d. 外显子盒(cassette exons, CE). 在一个转录体中是一个外显子，在另外一个转录异构体中是一个内含子的一部分.

e. 互斥外显子(mutually exclusive exons, MEE). 一组外显子中在其中的一个转录体中只选择其中的一个.

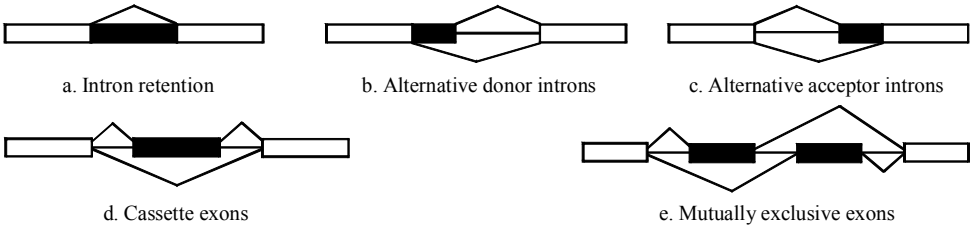

Fig. 1　Main types of alternative splicing

上述 5 种类型中，占可变剪接比例由大到小的顺序依次是：外显子盒、可变 3′剪接位点、可变 5′剪接位点、互斥外显子和内含子保留. 本文选择可变的受体剪接位点和可变的供体剪接位点类型作为研究对象，来预测人类基因组中的可变剪接位点.

2　数　　据

文中数据来自 ASD [18] (Alternative Splicing Database: http://www.ebi.ac.uk/asd/altsplice)中 altsplice 数据库的人类基因组 release 2 数据资料. 从 AltSplice-rel2.events 文件中，选择 INTRON ISOFORM (II-5P)类型构建可变的受体端剪接位点序列，选择 INTRON ISOFORM (II-3P)类型构建可变的供体剪接位点序列，不考虑内含子两端同时可变的情况. 组成性剪接数据来自人类基因组的 AltSplice-rel2.intron 文件.

在人类基因组中，统计分析一对可变的受体端剪接位点和供体端剪接位点之间的距离分布(小于 50 nt)，结果如图 2 和图 3 所示，多数受体端可变

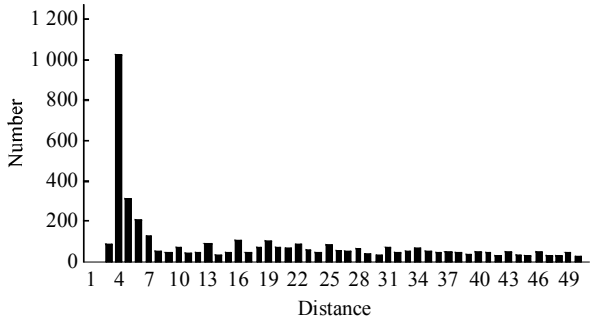

Fig. 2　The distribution of distances between alternative acceptor splice sites for human genome

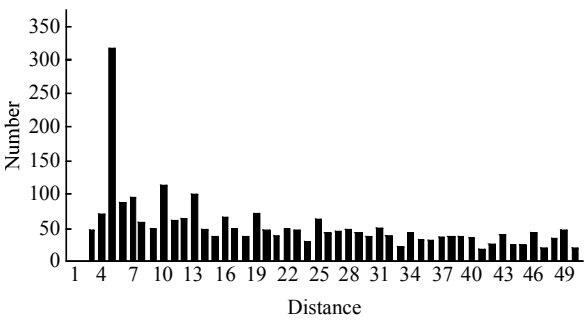

Fig. 3　The distribution of distances between alternative donor splice sites for human genome

内含子情形中一对可变剪接位点的间隔为 4 个碱基，而多数供体端可变内含子的情形中一对可变剪接位点的间隔为 5 个碱基．图 2 和图 3 给出距离小于 50 个碱基的分布．

对 INTRON ISOFORM (II-5P)类型，数据库中对同一个受体位点，给出了两个可变的供体剪接位点，构建供体端可变的数据库时，a．删除前后两个可变的供体位点间距大于 200 nt 的序列，b．删除两个可变供体位点中含非 GT 剪接位点的序列，以数据库中所给的第一个剪接位点为基准，前后各取 200 nt，将长度为 400 nt 的序列作为研究对象．

对 INTRON ISOFORM (II-3P)类型，数据库中对同一个供体位点，给出了两个可变受体剪接位点，构建受体端可变的数据库时，a．删除前后两个可变的受体位点间距大于 200 nt 的序列，b．删除两个可变受体位点中含非 AG 剪接位点的序列，以数据库所给的第一个剪接位点为基准，前后各取 200 nt，将长度为 400 nt 的序列作为研究对象．

由 AltSplice-rel2.intron 数据库，得到所有的内含子数据资料，与 ASD 数据库中所给出的各种可变剪接数据资料比较，排除所有参与可变剪接的内含子，选择其中内含子长度大于 200 nt 的作为研究对象．为了减少组成性内含子的样本数量，对供体端，进一步选择供体端紧邻的外显子长度大于 200 nt，以供体端剪接位点为基准，前后各取 200 nt，构建组成性的供体端剪接位点数据库．对受体端，进一步选择受体端紧邻的外显子长度大于 200 nt，以受体端剪接位点为基准，前后各取 200 nt，构建组成性的受体端剪接位点数据库．

可变剪接位点数据库，由数据库中所给出的真正的剪接位点构成正集(positive)，在长度为 400 nt 的范围内所有其他的含碱基二联体 GT 或 AG 位点作为虚假的剪接位点，构成负集(negative)，对组成性剪接位点亦如此，数据库如表 1 所示．

Table 1 The number of acceptor and donor set for alternative and constitutive splice site in human genome

Class	Splice site	Total number	Number of L < 200 bp (Non-GT/AG)	Positive	Negative
Alternative	Donor	5 924	4 608(483)	4 125	83 355
	Acceptor	6 783	5 837(18)	5 819	161 740
Constitutive	Donor	–	5 984(64)	5 920	116 697
	Acceptor	–	5 160(16)	5 144	142 728

L means the distance between a pair of alternative acceptor or donor sites and numbers in bracket gives splice sites of non-GT/AG.

3 特征与方法

在基因的剪接位点预测方面，我们利用基于多样性指标的二次判别方法(IDQD)，成功地预测了人类等 5 种模式生物的剪接位点[22]．本文，利用 IDQD 方法，分别预测了人类基因组中供体和受体端的可变剪接位点，作为对比，同时预测了人类基因组中供体和受体端的组成性剪接位点．

供体端，选择 GT 前 3 个位点，后 6 个位点(包括 GT)作为保守位点．a．统计长度为 7nt 的保守片段上，固定位点的单碱基、碱基二联体、碱基三联体数目，建立多样性源；b．截取剪接位点前后各 48 个碱基序列，统计前后序列的碱基三联体数目，建立多样性源；c．统计一对可变剪接位点之间的距离．对于一个待测的位点，计算反映其保守特征的 3 个多样性指标，反映剪接位点前后碱基组分特性的 5 个多样性指标，以及待测位点距已知剪接位点的距离指标，共 9 个指标，利用二次判别方法，判断该位点是否为真正的剪接位点．

受体端，选择 AG 前 10 个位点(包括 AG)作为保守位点，a．统计长度为 8 nt 的保守片段上，固定位点的单碱基、碱基二联体、碱基三联体数目，建立多样性源，b．截取剪接位点前后各 48 个碱基序列，统计前后序列的碱基三联体数目，建立多样性源，c．统计一对可变剪接位点之间的距离，d．供体端上游存在分支位点和多嘧啶尾，统计反映其组分特征的参数，分别是受体剪接位点前 30 位碱基的 G+C 含量和 T+C 含量．对于一个待测的位点，计算反映保守特征的 3 个多样性指标，反映剪接位点前后碱基组分特性的 5 个多样性指标，以及待测位点距已知剪接位点的距离指标和前 30 位碱基的 G+C 含量和 T+C 含量，共 11 个指标，利用二次判别方法，判断该位点是否为真正的剪接位点．

为了判别在已知剪接位点附近是否有可变的剪接位点，多样性源由可变剪接位点数据库中的序列

构建．在训练集中，对于任意一个待判别的可能的剪接位点，计算其特征参数构成 9 维(供体序列)或 11 维(受体序列)向量，然后，根据已知的可变剪接位点(正集)和非剪接位点集合(负集)来训练参数，最后，通过判别函数

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

来识别剪接位点．其中 δ_i 为

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

式中 p 和 q 分别表示正负集中的样本数，$\vec{\mu}_1$ 和 $\vec{\mu}_2$ 分别表示正负集中各变量平均值，Σ_1 和 Σ_2 表示相应的协方差矩阵，$|\Sigma_1|$和$|\Sigma_2|$为其行列式值．通常情况下，如果 $\xi>0$，则待测位点$\vec{X}$被判为真，否则判为虚假．但是，两个集合的分界点不一定是 0，因此，对供体端，选取不同的阈值 ξ_D，当 $\xi>\xi_D$，判别为真正的供体位点,否则为假，对受体端，选取不同的阈值 ξ_A，当 $\xi>\xi_A$，判别为真正的受体位点,否则为假．

4 预测结果

利用 3-fold 交叉检验，将数据分为 3 组，其中 1 组检验，其他 2 组训练，预测供体和受体端的剪接位点．定义敏感性指标 S_n 和特异性指标 S_p，以及总精度 TA 和相关系数 MCC 如下：

$$S_n=\frac{TP}{TP+FN}\times100\%$$

$$S_p=\frac{TN}{TN+FP}\times100\%$$

$$TA=\frac{TP+TN}{TP+FN+TN+FP}\times100\%$$

$$MCC=\frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}$$

在可变的供体剪接位点数据库中，预测一对可变剪接位点，阈值不同，得到不同的预测结果，如表 2 所示．综合考虑敏感性和特异性指标，选择 $\xi_D=-2$ 时，一对可变剪接位点的预测敏感性分别为 90.3%和 86.4%，总的敏感性指标为 88.3%，特异性指标为 87.8%，总精度为 87.9%．

Table 2 The result of prediction on alternative donor splice sites in human genome

ξ_D	Sn1	Sn2	Sn	Sp	TA	MCC
−5	96.3	95.5	95.9	76.3	78.1	0.451
−4	95.1	94.0	94.6	80.1	81.4	0.484
−3	93.3	91.3	92.3	83.9	84.7	0.520
−2	90.3	86.4	88.3	87.8	87.9	0.555
−1	85.6	78.2	82.0	91.6	90.7	0.589

在组成性供体剪接位点数据库中，预测其剪接位点，结果如表 3 所示．选择 $\xi_D=-1$ 时，敏感性指标为 97.0%，特异性指标为 92.6%，总精度为 92.8%．

Table 3 The result of prediction on constitutive donor splice sites in human genome

ξ_D	Sn	Sp	TA	MCC
−5	99.6	78.4	79.4	0.384
−4	99.4	82.1	82.9	0.423
−3	99.2	85.7	86.4	0.470
−2	98.5	89.2	89.7	0.526
−1	97.0	92.6	92.8	0.597
0	92.7	95.8	95.7	0.682

在可变的受体剪接位点数据库中，预测一对可变剪接位点，阈值不同，得到不同的预测结果，如表 4 所示．综合考虑敏感性和特异性指标，选择 $\xi_A=-2$ 时，一对可变剪接位点的预测敏感性分别为 84.0%和 78.9%，敏感性指标为 81.5%，特异性指标为 90.5%，总精度为 89.9%．

Table 4 The result of prediction on alternative acceptor splice sites in human genome

ξ_A	Sn1	Sn2	Sn	Sp	TA	MCC
−5	95.2	93.1	94.1	72.2	73.7	0.355
−4	92.8	90.4	91.6	78.9	79.7	0.403
−3	89.5	85.8	87.6	85.1	85.3	0.458
−2	84.0	78.9	81.5	90.5	89.9	0.515
−1	76.4	68.6	72.5	94.7	93.3	0.567

在组成性受体剪接位点数据库中，预测其剪接位点，结果如表 5 所示．选择 ξ_A=–1 时，敏感性指标为 94.1%，特异性指标为 94.3%，总精度为 94.3%．

Table 5 The result of prediction on constitutive acceptor splice sites in human genome

ξ_A	Sn	Sp	TA	MCC
–5	99.4	71.3	72.3	0.280
–4	98.9	78.2	78.9	0.329
–3	98.2	84.5	85.0	0.392
–2	96.8	90.1	90.3	0.474
–1	94.1	94.3	94.3	0.574
0	87.6	97.1	96.8	0.661

5 讨 论

依据剪接位点处的保守性特征和前后序列的组成性特征，以及待测序列距已知剪接位点的距离参数和受体端序列的 GC 含量、TC 含量，识别人类基因组中可变内含子和组成性内含子的供体和受体剪接位点，结果较好．

利用 IDQD 方法预测剪接位点时，当所取的阈值不同时，得到不同的预测精度，阈值相对小时，几乎可以预测出所有的真实的剪接位点，但同时也包含了较多的虚假的剪接位点，所取阈值较大时，虚假的剪接位点数减少，同时又失去了部分真实的剪接位点．综合考虑敏感性和特异性指标，对可变剪接位点，选择–2 为阈值，预测结果较好，而对组成性剪接位点，选择阈值为–1 时预测结果较好．这时对可变的供体端和受体端剪接位点，总的预测精度分别为 87.9%和 89.9%，对组成性的供体端和受体端剪接位点总的预测精度分别为 92.8%和 94.3%．据我们所知，对于可变剪接位点的预测，目前文献资料报道的精度皆在 80%以下，上述结果显示了 IDQD 方法的预测能力．

为了进一步讨论各种剪接位点对 ξ 值的依赖关系，我们对供体端和受体端真正的剪接位点和满足 GT/AG 碱基对所有的虚假剪接位点，计算其 ξ 值，得到可变剪接位点、组成性剪接位点、以及虚假的剪接位点序列的 ξ 值分布，如图 4 所示．

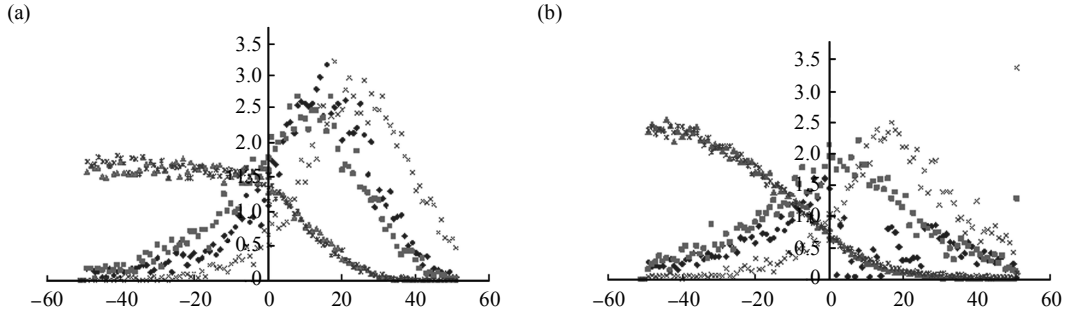

Fig. 4 The distribution of ξ value for true and false donor splice site (a) and true and false acceptor splice site (b) in human genome

x-axis indicates ξ(×10) and *y*-axis gives the percentage of splice sites with ξ in each 0.1 interval. AS1 , AS2 and CS mean true alternative splice site 1 and 2 and constitutive splice site respectively. AF and CF are false splice sites in alternative and constitutive splice site databases respectively. ◆ : AS1; ■ : AS2; ▲ : AF; × : CS; ✱ : CF.

结果显示，无论是可变剪接数据库，还是组成性剪接位点的数据库，其虚假剪接位点 ξ 值绝大部分都小于 0，可变剪接位点和组成性剪接位点的 ξ 值分布，有一定的区别，组成性剪接位点的 ξ 值更大一些．以 ξ 值为依据，可以很好地区分真正的剪接位点和虚假的剪接位点．

有意思的是，利用 ξ 值还可进一步对可变剪接位点和组成性剪接位点进行一定的区分．根据可变剪接位点和组成性剪接位点不同的 ξ 值分布，对供体端，以 ξ_D=1.6 为界，当 ξ_D > 1.6 时，判别为组成性剪接位点，否则为可变剪接位点，对受体端，以 ξ_A=1.2 为界，当 ξ_A > 1.2 时，判别为组成性剪接位点，否则为可变剪接位点，结果列于表 6 中．

Table 6 Prediction results of classification between alternative and constitutive splice sites

	ξ	Sn	Sp	TA	MCC
Donor	1.6	62.8	70.8	66.2	0.333
Acceptor	1.2	61.1	70.9	64.2	0.298

与其他的预测相比，上述结果高于 Wang 等[15]的预测结果．尽管本文的 IDQD 方法主要目标是从大量虚假剪接位点的背景中找出真正的组成性和可变性剪接位点，不是为区别这两类剪接而设计的，但所得结果也可以用来较好地识别二者，而且判别方法仅仅通过一个参数 ξ，既简单又意义明确．

Xia 等[17]研究了一对剪接位点的竞争性并提出 CSSP(竞争位点对)和 NCSSP(非竞争位点对)的识别方法，对此我们也用 IDQD 方法进行了研究．在构建 CSSP(竞争位点对)和 NCSSP(非竞争位点对)数据库的基础上，我们的预测结果如表 7 所示，供体端高于 Xia 等的方法 1.31 个百分点，受体端高于该方法 3.51 个百分点．

Table 7 The performance of splice site competition recognition by IDQD and comparison with other method

		Sn	Sp	TA	MCC
Donor	Our method(ξ=−1.7)	84.43	90.64	90.52	0.339
	Xia *et al*[17]	84.94	89.35	89.21	0.395
Acceptor	Our method(ξ =−1.5)	78.25	91.75	91.49	0.323
	Xia *et al*[17]	78.34	88.31	87.98	0.346

进而，根据 Xia 等的方法，对任意一个已知的剪接位点，和前后 200 bp 内的 GT(供体端)或 AG(受体端)位点构成一对一对的待判别的竞争位点对，若其中任意一对位点被判别为竞争位点对，则该位点为可变剪接位点，否则，为组成性剪接位点．但是，存在一种可能，被认为是竞争的位点未必是真正已知的剪接位点．以受体端为例，在给定预测敏感性下检验所预测的可变剪接位点中真正的竞争位点的精确度．如图 5 所示，PS 给出预测可变剪接位点的敏感性指标，而 TS 给出在预测为可变剪接位点中找到真正的竞争剪接位点的敏感性指标，所以从一对位点的竞争性来判别一个剪接位点的特性并不容易．

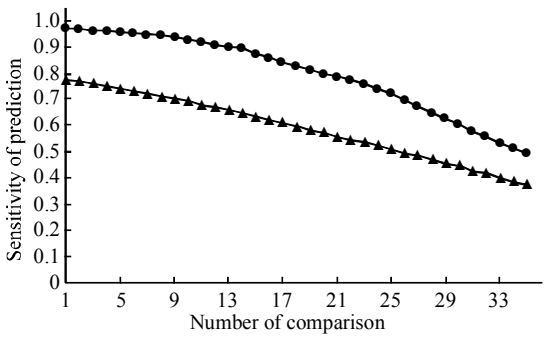

Fig. 5 The examination of truly competitive splice site

PS is sensitivity of predicted alternative splice sites and TS is sensitivity of true competitive sites in predicted alternative splice sites. ●—●:PS; ▲—▲ : TS.

由于生物体在识别剪接位点时是同时识别 GT 和 AG，寻找供体和受体位点，因此，在一段核酸序列上，我们可以寻找可能的供体和受体剪接位点，然后通过位点之间的匹配关系，确定可变剪接位点和组成性剪接位点．若同一个超过阈值的 AG 位点，有两个(或以上)超过阈值的 GT 与之配对，则为可变的受体剪接位点，若只有一个，则为组成性剪接位点；对同一个超过阈值的 GT 位点，有两个(或以上)超过阈值的 AG 与之配对，则为可变的供体剪接位点，若只有一个，则为组成性剪接位点．我们认为，这样结合 ξ 值和供体与受体端剪接位点之间的匹配关系，将能更好地识别可变剪接位点和组成性剪接位点．

参 考 文 献

1 Barmak M, Christopher L. A genomic view of alternative splicing. Nature Genetics, 2002, **30**(1): 13～19

2 Christopher L, Qi W. Bioinformatics analysis of alternative splicing. Briefings in Bioinformatics, 2005, **6**(1): 23～33

3 Kriventseva E V, Koch I, Apweiler R, *et al*. Increase of functional diversity by alternative splicing. Trends in Genetics, 2003, **19**(3): 124～128

4 Brenton R G. Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics, 2001, **17**(2): 100～107

5 Brett D, Pospisil H, Valcarcel J, *et al*. Alternative splicing and genome complexity. Nature Genetics, 2002, **30**(1): 29～30

6 Mironov A A, Fickett J W, Gelfand M S. Frequent alternative splicing of human genes. Genome Research, 1999, **9**(12): 1288～1293

7 Clark F, Thanaraj T A. Categorization and characterization of transcriptconfirmed constitutively and alternatively spliced introns and exons from human. Human Molecular Genetics, 2002, **11**(4): 451～464

8 Modrek B, Resch A, Grasso C, *et al*. Genome-wide analysis of alternative splicing using human expressed sequence data. Nucleic Acids Research, 2001, **29**(13): 2850～2859

9 Kan Z Y, Rouchka E C, Gish W R, *et al*. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Research, 2001, **11**(5): 889～900

10 Gupta S, Zink D, Korn B, *et al*. Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics, 2004, **20**(16): 2579～2585

11 Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics, 2005, **21**(7): 897～901

12 Chen T M, Lu C C, Li W H. Prediction of splice sites with dependency graphs and their expanded bayesian networks. Bioinformatics, 2005, **21**(4): 471～482

13 Chuang T J, Chen F C, Chou M Y. A comparative method for identification of gene structures and alternatively spliced variants. Bioinformatics, 2004, **20**(17): 3064～3079

14 Modrek B, Lee C. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nature Genetics, 2003, **34**(2): 177～180

15 Wang M, Marin A. Characterization and prediction of alternative splice sites. Gene, 2006, **366**(2): 219～227

16 Wen F, Li F, Xia H Y, *et al*. The impact of very short alternative splicing on protein structures and functions in the human genome. Trends in Genetics, 2004, **20**(5): 232～236

17 Xia H Y, Bi J N, Li Y D. Identification of alternative 5′/3′ splice sites based on the mechanism of splice site competition. Nucleic Acids Research, 2006, **34**(21): 6305～6313

18 Thanaraj T A, Stamm S, Clark F, *et al*. ASD: the Alternative Splicing Database. Nucleic Acids Research, 2004, **32** (Database issue): D64～D69

19 Laxton R R. The measure of diversity. J Theoretical Biology, 1978, **70**(1): 51～67

20 Li Q Z, Lv Z Q. The prediction of the structural class of protein: application of the measure of diversity. J Theoretical Biology, 2001, **213**(3): 493～502

21 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

22 Zhang L R, Luo L F. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Research, 2003, **31**(21): 6214～6220

The Prediction for Alternative and Constitutive Splice Sites in Human Genome*

ZHANG Li-Rong**, LUO Liao-Fu, XING Yong-Qiang, JIN Hong-Ying

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

Abstract Based on the conservation of nucleotides at splice sites, the characteristics of base composition and base correlation in the adjacent segment sequences, the distance between alternative donor or acceptor splice sites and the content of GC and TC near splice sites, the donor and acceptor splice sites for alternative and constitutive introns are predicted by use of the method of diversity measure combined with quadratic discriminant analysis. For alternative splice sites the total prediction accuracies are 87.9% and 89.9% for donors and acceptors respectively (with the chosen threshold –2). For constitutive splice sites the total accuracy are 92.8% and 94.3% for donors and acceptors respectively (with the chosen threshold –1).

Key words alternative splice，constitutive splice，quadratic discriminant analysis

---

*This work was supported by grants from The National Natural Sciences Foundation(10447003, 90403010) and The Natural Sciences Foundation of Inner Mongolia (200308020102).

**Corresponding author.

Tel: 86-471-4995052, E-mail: pyzlr@mail.imu.edu.cn

Received: January 30, 2008 Accepted: April 8, 2008