Hans Journal of Computational Biology 计算生物学, 2016, 6(3), 41-49
Published Online September 2016 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2016.63006

Hans汉斯

Research on Statistical Analysis of Gene Splicing Sites

Hongbin Li*, Guangzhong He

Medical School, Xianyang Vocational and Technical College, Xianyang Shaanxi
Email: *leehbin@126.com

Received: Aug. 5th, 2016; accepted: Aug. 19th, 2016; published: Aug. 26th, 2016

Abstract

The genes of eukaryotes are composed of several exons and introns. After transcript process, sequences of exons are retained, while sequences of introns are cleaved off. A large number of experiments of molecular biology validate that the splicing sites between exon and intron follow the rule of GT-AG, only a few GT or AG sequences are true splicing sites, and the accuracy of the prediction still needs to be improved. In this study, the training dataset of splicing site of HS3D was downloaded, and a statistical analysis of the sequence near the splicing site of the promoter was carried out. The sequence showed high specificity when the true and false sequence lengths of the left splicing site side and right splicing site side were both more than seven, which was helpful to train the sequences characters so as to accurately identify the true and false splicing sites.

Keywords

Gene, Splice Site, Statistical Analysis

基因剪切位点的统计分析研究

李宏彬*，赫光中

咸阳职业技术学院医学院，陕西 咸阳
Email: *leehbin@126.com

收稿日期：2016年8月5日；录用日期：2016年8月19日；发布日期：2016年8月26日

*通讯作者。

文章引用：李宏彬，赫光中．基因剪切位点的统计分析研究[J]．计算生物学, 2016, 6(3): 41-49.
http://dx.doi.org/10.12677/hjcb.2016.63006

摘　要

真核生物的基因由若干外显子和内含子交替组成，外显子序列在转录后保留，而内含子序列转录过程中被剪切掉。大量分子生物学实验验证基因的剪切位点遵从GT-AG规则，然而只有很少的含GT或AG序列是真剪切位点,目前预测的准确程度仍有待提高。本研究下载了HS^{3}D剪切位点训练数据集，对启动子剪切位点附近的序列进行了统计分析研究。当真、假序列长度在剪切位点左旁和右旁均超出各七个位点时，序列呈现很高的特异性，可以使用这些特异性序列作为特征进行训练，从而准确地识别真假剪切位点。

关键词

基因，剪切位点，统计分析

1. 引言

基因组学是研究基因序列结构、功能分析和如何利用基因的一门学科，而基因剪切位点识别是其中重要的研究方向之一。真核生物基因是由一段段编码区和非编码区碱基序列嵌合而成，编码区又称为外显子(exon)，它们之间的非编码区被称为内含子(Intron)，基因的首尾还各有一段具备一定功能的非编码区，分别称为启动子和终止子。外显子和内含子的大小变化不定，内含子一般要远长于外显子。真核基因先转录为前 mRNA (包含所有外显子和内含子序列)，然后序列中的内含子需要被除去而外显子相互链接为成熟的 mRNA，这个过程称为剪接(Splicing)，如图 1 所示。成熟的 mRNA 每个三联体核苷酸构成一个密码子，将被翻译成一个氨基酸，它们决定了蛋白质的氨基酸线性顺序。因此，如果剪切不够准确，如多出或缺少一个核苷酸，下游经翻译的密码子就会出误，最终生成错误的蛋白质。大量实验数据表明绝大多数剪切位点遵从 GT-AG 规则(极少数个例显示遵从 AT-AC 规则)，外显子-内含子连接区呈现高度保守性，也就是在内含子序列的 5’端(从外显子过渡到内含子)特征为 GT，而在其 3’端(从内含子过渡到外显子)特征为 AG，然而海量基因组测序数据显示满足 GT-AG 规则的序列绝大多数并不是真内含子序列。基因剪接位点常用的研究方法有：人工神经网络[1]、隐马尔可夫模型[2]、动态规划[3]、支持向量机[4]、贝叶斯网络[5]和频谱 3-周期性[6]等。

2. 方法和分析

HS^{3}D (Homo Sapiens Splice Sites Dataset) [7]是意大利 Pollastro 从 GeneBank DNA 序列数据库中提取的基因剪接位点序列数据集，数据集中的每个条目记录剪切位点从上游到下游总长为 140 个字符的 DNA 序列数据，剪接符均遵从 GT-AG 规则，GT 位于位点 71 到 72，AG 位于位点 69 到 70。数据集分为四个部分：真 EI (exon to intron)、假 EI、真 IE (intron to exon)和假 IE，真 EI 和真 IE 记录的数据条目数相对较少，分别为 2796 和 2880 个，而假 EI 和假 IE 数据条目数相对极多，分别为 271,928 和 329,360 个。为了观察真假剪切位点临近序列的差异，本研究依据 HS^{3}D 数据集中的数据，对 140 个位点真、假 EI 和 IE (的四种碱基(A、T、C、G)出现频率进行了比较，分别如图 2、图 3 所示。其中横坐标代表临近序列位置，纵坐标表示某位置的真(左)、假(右) EI、IE 碱基 A (红)、T (绿)、C (蓝)或 G (黄)出现频率(0 到 1 之间)。从图 2、图 3 中可以观察到，总体来说，从剪接点上游位点到下游位点，真剪接位点临近序列的碱基呈现随位置变化的出现频率，特别在剪接特征符左右十个位置(统计数据见表 1)，而假的除特征符 GT 和 AG 左右一两个位置以外，在其余位点呈现近似接近的碱基出现频率。真 EI 序列在剪接位点的一致序列

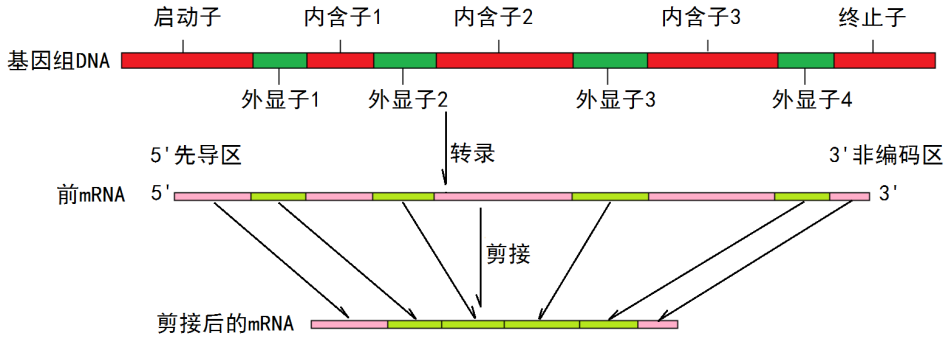

Figure 1. Structure, transcription and splicing of eukaryotic gene
图 1. 真核基因的结构、转录和剪接

Table 1. The base frequency comparison in sites from −10 to 10 between true splicing sequences and false splicing sequences
表 1. 真、假剪切位点序列−10 到 10 各碱基出现频率比较

位置	真 EI				假 EI				真 IE				假 IE			
	A	T	C	G	A	T	C	G	A	T	C	G	A	T	C	G
−10	0.24	0.21	0.27	0.28	0.25	0.26	0.23	0.26	0.08	0.43	0.37	0.12	0.24	0.28	0.25	0.23
−9	0.27	0.18	0.26	0.28	0.26	0.27	0.23	0.24	0.07	0.48	0.34	0.11	0.25	0.28	0.24	0.23
−8	0.23	0.22	0.28	0.26	0.27	0.26	0.23	0.23	0.08	0.44	0.38	0.11	0.24	0.28	0.23	0.24
−7	0.24	0.21	0.28	0.26	0.26	0.27	0.24	0.23	0.07	0.43	0.39	0.11	0.25	0.28	0.23	0.24
−6	0.27	0.19	0.27	0.27	0.27	0.26	0.23	0.24	0.09	0.38	0.41	0.12	0.26	0.28	0.22	0.24
−5	0.26	0.23	0.27	0.24	0.27	0.27	0.25	0.21	0.09	0.39	0.42	0.09	0.25	0.28	0.23	0.24
−4	0.27	0.18	0.30	0.26	0.25	0.26	0.27	0.22	0.06	0.40	0.47	0.07	0.24	0.29	0.24	0.23
−3	0.33	0.11	0.36	0.19	0.26	0.26	0.24	0.23	0.07	0.47	0.39	0.07	0.22	0.29	0.26	0.22
−2	0.58	0.14	0.14	0.14	0.29	0.24	0.25	0.22	0.21	0.22	0.35	0.23	0.24	0.27	0.24	0.24
−1	0.09	0.07	0.03	0.81	0.31	0.34	0.08	0.27	0.04	0.19	0.76	0.01	0.26	0.17	0.34	0.23
	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
1	0.47	0.03	0.03	0.47	0.19	0.26	0.21	0.34	0.23	0.09	0.14	0.54	0.27	0.22	0.23	0.29
2	0.70	0.08	0.08	0.13	0.25	0.30	0.22	0.23	0.23	0.32	0.21	0.24	0.28	0.24	0.24	0.23
3	0.06	0.05	0.05	0.83	0.24	0.28	0.25	0.23	0.22	0.24	0.28	0.25	0.28	0.26	0.22	0.24
4	0.15	0.44	0.18	0.23	0.25	0.29	0.23	0.23	0.22	0.24	0.30	0.24	0.26	0.26	0.23	0.25
5	0.26	0.16	0.21	0.37	0.25	0.26	0.23	0.26	0.25	0.24	0.29	0.22	0.26	0.26	0.21	0.26
6	0.20	0.23	0.28	0.29	0.25	0.29	0.23	0.24	0.22	0.23	0.30	0.25	0.26	0.25	0.22	0.27
7	0.19	0.22	0.32	0.28	0.26	0.27	0.23	0.25	0.25	0.22	0.26	0.28	0.26	0.25	0.22	0.27
8	0.18	0.22	0.27	0.33	0.25	0.28	0.22	0.25	0.24	0.22	0.29	0.25	0.26	0.27	0.22	0.25
9	0.20	0.20	0.29	0.31	0.25	0.28	0.24	0.24	0.24	0.23	0.26	0.27	0.26	0.26	0.22	0.26
10	0.19	0.22	0.28	0.31	0.25	0.29	0.22	0.23	0.24	0.21	0.27	0.29	0.26	0.26	0.23	0.25

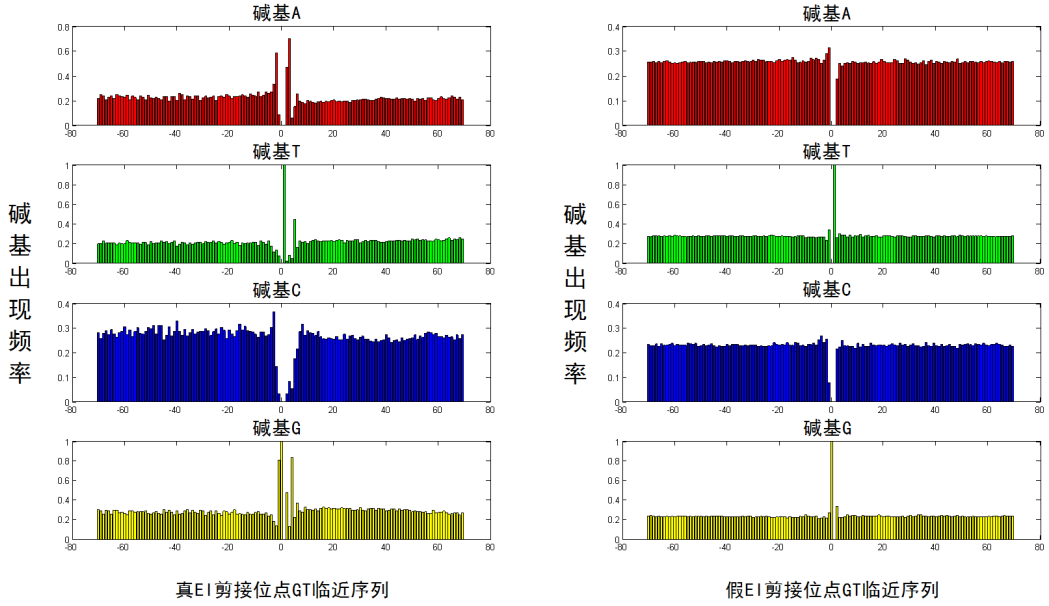

Figure 2. The base statistical frequency comparison of sequence close to splicing site between true EI dataset and false EI dataset
图 2. 真、假 EI 剪接位点临近序列碱基统计频率比较

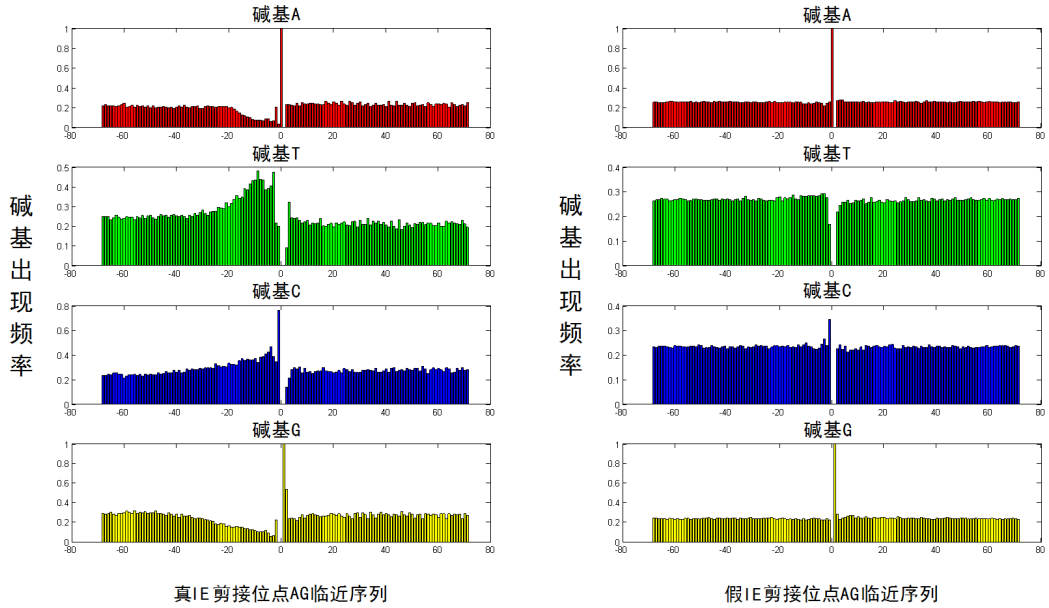

Figure 3. The base appearance frequency comparison of sequence close to splicing site between true IE dataset and false IE dataset
图 3. 真、假 IE 剪接位点临近序列碱基出现频率比较

是 MAG GT RAG (M: A/C, R: A/G)真 IE 序列在剪接位点的一致序列是 YYNC AG RN (Y: T/C, N: A/T/C/G)。

参考真 EI、IE 数据集，本研究对以剪切位特征符(GT、AG)为中心的碱基六聚体出现百分频率进行了统计，若总频率为 1，则真 EI 和真 IE 频度靠前的六聚体 EI、IE 数据集中真剪接位点多聚体出现频率数据分别如表 2 和表 3 所示，这些六联体总频度占真 EI 和真 IE 数据集中样本量的绝大多数，分别为 82.53%

Table 2. The top 24 frequency statistic of 6-mer base close to splicing sites in true EI dataset
表 2. 真 EI 剪接位点六聚体碱基出现频率统计(前 24 个)

真 EI 六聚体出现频率统计								
序号	六聚体	出现频率	序号	六聚体	出现频率	序号	六聚体	出现频率
1	5'-AG**GT**GA-3'	0.1516	9	5'-AG**GT**AC-3'	0.0393	17	5'-AA**GT**GA-3'	0.0132
2	5'-AG**GT**AA-3'	0.1195	10	5'-GG**GT**AA-3'	0.0368	18	5'-AA**GT**AA-3'	0.0129
3	5'-TG**GT**GA-3'	0.0497	11	5'-AG**GT**AT-3'	0.0315	19	5'-CA**GT**GA-3'	0.0129
4	5'-AG**GT**AG-3'	0.0490	12	5'-AG**GT**GC-3'	0.0190	20	5'-CT**GT**GA-3'	0.0125
5	5'-AG**GT**GG-3'	0.0476	13	5'-AG**GT**CA-3'	0.0154	21	5'-AG**GT**GT-3'	0.0104
6	5'-GG**GT**GA-3'	0.0465	14	5'-CT**GT**AA-3'	0.0154	22	5'-AT**GT**AA-3'	0.0097
7	5'-CG**GT**GA-3'	0.0454	15	5'-CG**GT**AA-3'	0.0143	23	5'-CA**GT**AA-3'	0.0097
8	5'-TG**GT**AA-3'	0.0401	16	5'-AT**GT**GA-3'	0.0136	24	5'-TG**GT**AT-3'	0.0093

Table 3. The top 63 frequency statistic of 6-mer close to splicing sites in true IE dataset
表 3. 真 IE 剪接位点六聚体出现频率统计(前 63 个)

真 IE 六聚体出现频率统计								
序号	六聚体	出现频率	序号	六聚体	出现频率	序号	六聚体	出现频率
1	5'-CC**AG**GT-3'	0.0469	22	5'-CT**AG**GT-3'	0.0128	43	5'-AT**AG**GA-3'	0.0073
2	5'-CC**AG**GC-3'	0.0410	23	5'-GC**AG**AC-3'	0.0128	44	5'-TC**AG**AG-3'	0.0073
3	5'-CC**AG**GG-3'	0.0392	24	5'-GC**AG**GC-3'	0.0128	45	5'-TC**AG**CT-3'	0.0073
4	5'-CC**AG**GA-3'	0.0313	25	5'-GC**AG**CT-3'	0.0125	46	5'-GC**AG**CA-3'	0.0073
5	5'-GC**AG**GT-3'	0.0281	26	5'-AC**AG**AT-3'	0.0118	47	5'-AT**AG**GT-3'	0.0069
6	5'-AC**AG**GT-3'	0.0271	27	5'-TC**AG**AT-3'	0.0115	48	5'-CT**AG**GC-3'	0.0069
7	5'-TC**AG**GA-3'	0.0229	28	5'-TC**AG**GC-3'	0.0115	49	5'-CC**AG**TG-3'	0.0069
8	5'-GC**AG**GG-3'	0.0226	29	5'-GC**AG**AA-3'	0.0115	50	5'-CC**AG**AA-3'	0.0066
9	5'-TC**AG**GT-3'	0.0212	30	5'-TT**AG**GA-3'	0.0108	51	5'-TC**AG**TG-3'	0.0063
10	5'-TC**AG**GG-3'	0.0208	31	5'-AC**AG**AA-3'	0.0101	52	5'-CT**AG**AA-3'	0.0063
11	5'-AC**AG**GG-3'	0.0205	32	5'-AC**AG**CC-3'	0.0097	53	5'-AC**AG**TG-3'	0.0059
12	5'-GC**AG**GA-3'	0.0205	33	5'-CC**AG**CC-3'	0.0097	54	5'-CC**AG**CA-3'	0.0059
13	5'-CC**AG**AG-3'	0.0177	34	5'-AC**AG**CT-3'	0.0094	55	5'-AT**AG**GG-3'	0.0056
14	5'-AC**AG**GA-3'	0.0174	35	5'-TT**AG**GT-3'	0.0094	56	5'-TT**AG**AT-3'	0.0056
15	5'-CC**AG**AC-3'	0.0160	36	5'-GC**AG**CC-3'	0.0094	57	5'-TT**AG**GC-3'	0.0056
16	5'-CC**AG**CT-3'	0.0160	37	5'-TC**AG**AA-3'	0.0090	58	5'-TC**AG**CC-3'	0.0056
17	5'-GC**AG**AG-3'	0.0153	38	5'-TC**AG**AC-3'	0.0083	59	5'-CT**AG**GG-3'	0.0056
18	5'-AC**AG**GC-3'	0.0149	39	5'-AC**AG**AG-3'	0.0080	60	5'-CC**AG**TT-3'	0.0056
19	5'-CC**AG**AT-3'	0.0146	40	5'-TT**AG**GG-3'	0.0076	61	5'-GT**AG**GA-3'	0.0056
20	5'-GC**AG**AT-3'	0.0146	41	5'-GC**AG**TT-3'	0.0076	62	5'-AC**AG**AC-3'	0.0049
21	5'-CT**AG**GA-3'	0.0142	42	5'-GC**AG**TG-3'	0.0076	63	5'-AC**AG**TT-3'	0.0049

和 82.65%。统计发现，剪切位特征符左右各有两个碱基核苷酸，应有 4 的 4 次方(256)中可能性，数据集中真，EI 和 IE 剪接六聚体至少出现一次分别占 118 和 181 种可能性，而假 EI 和 IE 剪接六聚体覆盖所有的 256 种可能性，反映剪接位点临近序列的特异性。本研究依次对 EI、IE 数据集中真、假剪接位点多聚体重合度统计，包括 6 聚体、8 聚体(4096 种可能性，特征符前三后三)、10 聚体(65,536 种可能性，特征符前四后四)、12 聚体(1,048,576 种可能性，特征符前五后五)和 14 聚体(16,777,216 种可能性，特征符前六后六)，如图 4 所示。图中反映 6 聚体、8 聚体和 12 聚体，真假剪接聚体的重合度很高(重合是指，某个位于剪接位点的多聚体，若真、假数据集中都至少出现 1 例，则该聚体重合)，若依此聚体序列为特征进行识别会导致极高的错误率，而 14 聚体以上真、假聚体序列的重合度大幅度下降，特异性显著增强，有利于以此为特征进行训练和进行真剪接位点识别判断。

那么总长相同的聚体，不同的选取方式，会对重叠率有什么影响呢？本研究做了一个实验，这些多聚体总长相同，但在剪接特征符 GT 或 AG 前后选取的核苷酸数目不同，然后统计真、假多聚体的重叠率，结果如表 4 所示，其中示例 EI 和 IE 数据集中真、假剪接位点前五后五模式多聚体分布比较分别如图 5 和图 6 所示。十二聚体中，选取方式前 6 后 4 (前 6 GT 后 4)在 EI 真、假数据集中获得最低的重叠率，前 2 后 8 (前 2 AG 后 8)在 IE 真、假数据集中获得最低的重叠率。十四聚体中，选取方式前 8 后 4 (前 8 GT 后 4)在 EI 真、假数据集中获得最低的重叠率，前 3 后 9 (前 3 AG 后 9)在 IE 真、假数据集中获得最低的重叠率。获得最低重叠率的十二聚体位置与十四聚体接近，但十二聚体重叠率远高于十四聚体，如果以十二聚体作为特征进行真、假剪接位点识别，虽然特征碱基数目少、速度快，但会产生过高的错误率，因此真假剪接位点识别训练应选择十四聚体以上作为特征模式。

Table 4. The overlap rate comparison between true and false dataset of 12-mer and 14-mer close to splicing sites by different selection methods
表 4. 剪接临近十二和十四聚体在不同选取方式下真、假数据集的重叠率比较

	选取方式	EI 重叠率	IE 重叠率
十二聚体	前 1 后 9	35.59%	39.12%
	前 2 后 8	36.10%	**38.01%**
	前 3 后 7	35.01%	41.58%
	前 4 后 6	34.90%	43.29%
	前 5 后 5	34.81%	45.61%
	前 6 后 4	**34.31%**	48.68%
	前 7 后 3	36.82%	48.59%
	前 8 后 2	35.11%	52.12%
	前 9 后 1	37.94%	54.75%
十四聚体	前 1 后 11	5.54%	5.92%
	前 2 后 10	4.99%	5.04%
	前 3 后 9	5.08%	**5.00%**
	前 4 后 8	4.99%	5.93%
	前 5 后 7	4.49%	6.73%
	前 6 后 6	4.42%	6.85%
	前 7 后 5	4.70%	6.60%
	前 8 后 4	**4.21%**	6.98%
	前 9 后 3	4.77%	7.81%
	前 10 后 2	5.30%	8.70%
	前 11 后 1	5.67%	10.55%

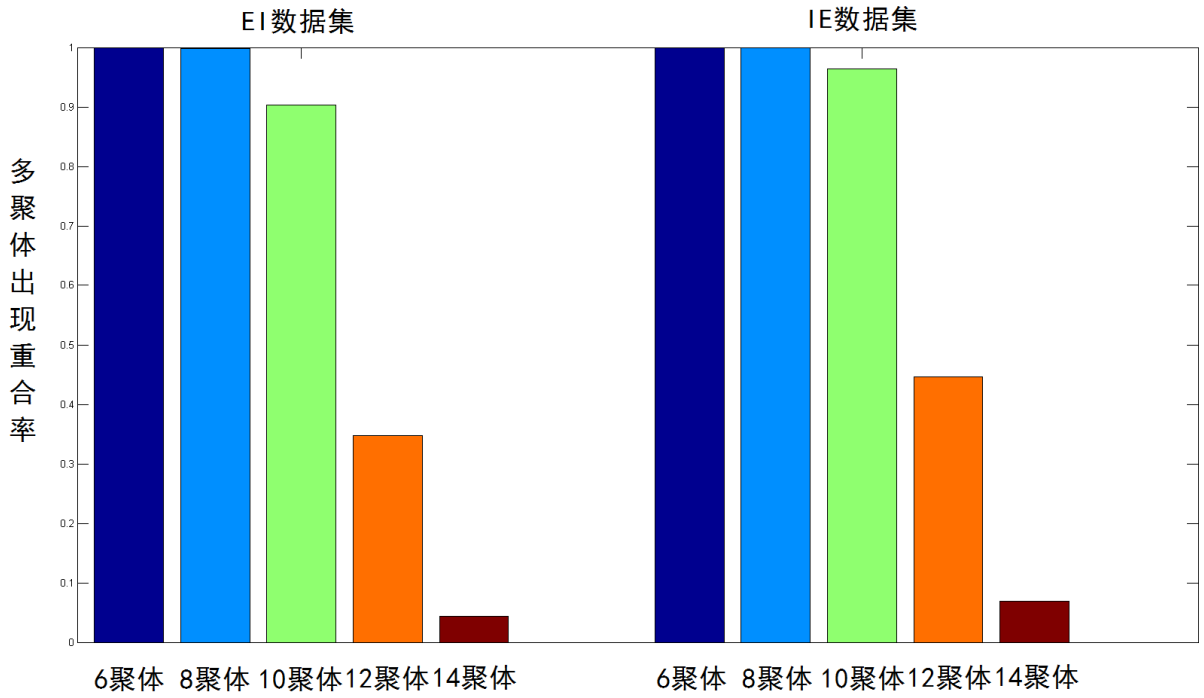

Figure 4. The N-mer overlap rate statistic of true and false splicing site in EI and IE dataset
图 4. EI、IE 数据集中真、假剪接位点多聚体重合度统计

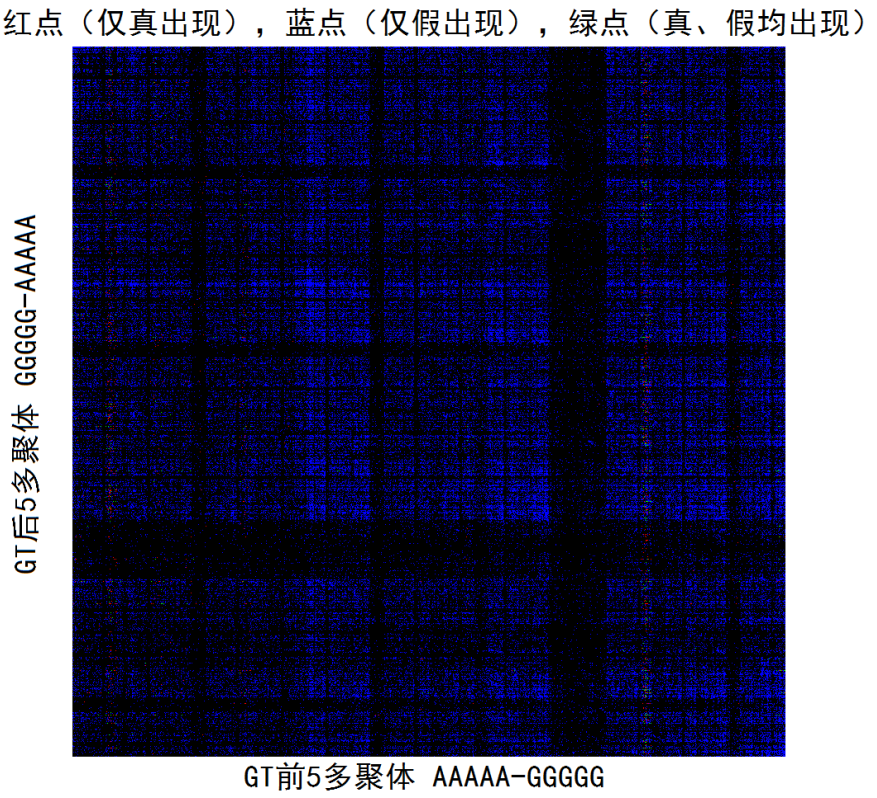

Figure 5. The N-mer (five before and five after GT site) distribution comparison between true and false splicing site in EI dataset
图 5. EI 数据集中真、假剪接位点 GT 前五后五模式多聚体分布比较

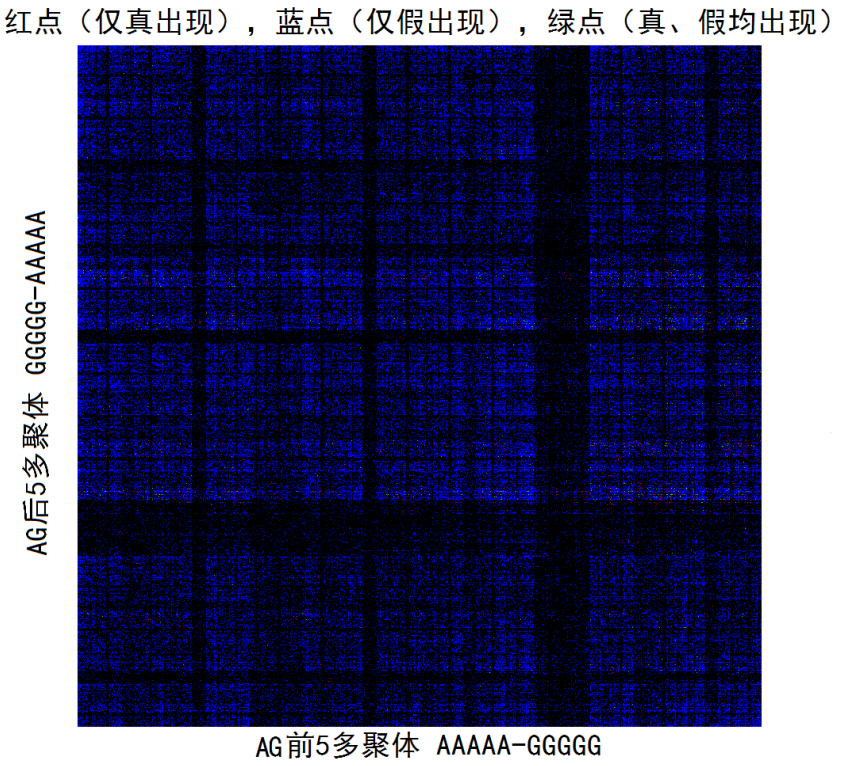

Figure 6. The N-mer (five before and five after AG site) distribution comparison between true and false splicing site in IE dataset
图 6. IE 数据集中真、假剪接位点 AG 前五后五模式多聚体分布比较

3. 结论

本文使用公共数据库 HS^3D 的序列数据对基因剪接位点的序列进行了统计分析。通过统计获得真假剪接位点的碱基出现频率，分析反映真剪接位点临近序列的碱基呈现随位置变化的出现频率，而假的除特征符 GT 和 AG 左右一两个位置以外，在其余位点呈现近似接近的碱基出现频率。通过统计还获得了占数据库中绝大多数的真剪接位点 EI 和 IE 六联体序列。研究还发现 14 聚体以上，真剪切位点临近序列的特异性显著增强，这有利于以此序列为特征进行训练，从而准确地识别真假剪切位点。

致　谢

感谢陕西省科技厅社会发展科技攻关项目基金(2016SF-343)资助。

参考文献 (References)

[1] Sun, J. (1993) Predicting the Splicing Sites of mRNA by Neural Network. *Acta Biophysica Sinica*, **9**, 127-131.

[2] Xia, H., Zhou, Q. and Yanda, L.I. (2002) Application of Hidden Markov Model in the Recognition of Splicing Sites. *Journal of Tsinghua University*, **42**, 1214-1217.

[3] Snyder, E.E. and Stormo, G.D. (1993) Identification of Coding Regions in Genomic DNA Sequences: An Application of Dynamic Programming and Neural Networks. *Nucleic Acids Research*, **21**, 607-613. http://dx.doi.org/10.1093/nar/21.3.607

[4] Zhang, L.R. and Luo, L.F. (2003) Splice Site Prediction with Quadratic Discriminant Analysis Using Diversity Measure. *Nucleic Acids Research*, **31**, 6214-6220. http://dx.doi.org/10.1093/nar/gkg805

[5] Cai, D., Delcher, A., Kao, B. and Kasif, S. (2000) Modeling Splice Sites with Bayes Networks. *Bioinformatics*, **16**, 152-158. http://dx.doi.org/10.1093/bioinformatics/16.2.152

[6] Yin, C. and Yau, S.T. (2007) Prediction of Protein Coding Regions by the 3-Base Periodicity Analysis of a DNA Sequence. *Journal of Theoretical Biology*, **247**, 687-694. http://dx.doi.org/10.1016/j.jtbi.2007.03.038

[7] Pollastro, P. and Rampone, S. (2002) HS^{3}D, a Dataset of Homo Sapiens Splice Regions, and Its Extraction Procedure from a Major Public Database. *International Journal of Modern Physics C*, **13**, 1105-1117. http://dx.doi.org/10.1142/S0129183102003796