International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:1, 2009

Identifying New Sequence Features for Exon-Intron Discrimination by Rescaled-Range Frameshift Analysis

Sing-Wu Liou[1] and **Yin-Fu Huang**[2]
[1]Graduate School of Engineering Science and Technology
National Yunlin University of Science and Technology
Douliou, Yunlin, Taiwan.
g9110808@yuntech.edu.tw

[2]Graduate School of Computer Science and Information Engineering
National Yunlin University of Science and Technology
Douliou, Yunlin, Taiwan.
huangyf@el.yuntech.edu.tw (*corresponding author*)

***Abstract*—For identifying the discriminative sequence features between exons and introns, a new paradigm, rescaled-range frameshift analysis (RRFA), was proposed. By RRFA, two new sequence features, the frameshift sensitivity (FS) and the accumulative penta-mer complexity (APC), were discovered which were further integrated into a new feature of larger scale, the persistency in anti-mutation (PAM). The feature-validation experiments were performed on six model organisms to test the power of discrimination. All the experimental results highly support that FS, APC and PAM were all distinguishing features between exons and introns. These identified new sequence features provide new insights into the sequence composition of genes and they have great potentials of forming a new basis for recognizing the exon-intron boundaries in gene sequences.**

Keywords: Exon-Intron Discrimination, Rescaled-Range Frameshift Analysis, Frameshift Sensitivity, Accumulative Sequence Complexity.

I. Introduction

Exons are born of the obligation for coding the necessary proteins; they have to preserve the protein-coding information, the signals for mRNA transport and localization; In contrast, introns are junk-DNA[29] for being removed during the pre-mRNA maturating processes and they are often characterized by a highly recurrent use of specific triplets[8]. Thus, exons are definitely non-randomness[6] and that distinguishes them from introns[13].

Discrminant analysis (DA) on exon-intron junctions (i.e., the gene splice site) had been studied for a long time. A multi-source integrated method for splice site recognition recruited consensus patterns, free energy and statistical differences of bases usage to discriminate exons from introns[18]; a heuristic informational approach, the discriminant index (DI), combines two k-tuple reference profiles for intron/exon discrimination with the required parameters, the experimentally determined k and window width[5]; the IDQD[17] compared compositional features between exons and introns; the BRAIN[22] inferring Boolean formulae from the training set and combined with a discriminant analysis procedure; a linear discriminant function combining statistical information of specific triplets at specific regions around splice sites was proposed[24].

Yet, the performances of the above-mentioned DA methods are closely related to some result-sensitive parameters, which are usually solved by repetitive trials by choosing the values obtaining good results as thresholds or default values. In this paper, a parameter-free and no domain-knowledge involved methodology, rescaled-range frameshift analysis (RRFA), is proposed, which solves the exon-intron discrimination problem by formulating the dedicated sequence features into quantitative estimators.

Two sequence features, the frameshift sensitivity (FS) and accumulative penta-mer complexity (APC), were firstly identified and experimentally validated in six model organisms (*Human*, *Mouse*, *Rat*, *Arabidopsis*, *C. elegan* and *Drosophila*). Moreover, FS and APC are further integrated into a more effective discriminator, persistency on anti-mutation (PAM), which motivated the exon-clamp hypothesis (ECH). The highly supported ECH in all the six model organisms demonstrated the general applicability of the identified new features.

II. Methods

A. Rescaled full-length comparison (RFLC)

The compositional heterogeneity between exons and introns had been linked up with the composition of triplet repeats[14]; recent evidence also suggested that sequence conservation associated with splice sites may extend relatively far away from intron-exon boundaries[15], [11], [25]; it had also been shown the structural coupling at the exon-intron junctions might not

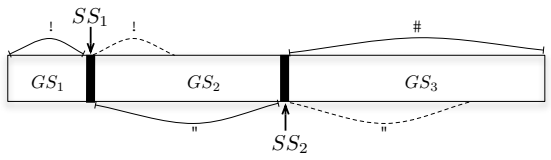

Figure 1. RFLC (rescaled full-length comparison): three contiguous gene segments (exons or introns) of different lengths, GS_1, GS_2 and GS_3, conjoined by two splice sites (SS_1 and SS_2). Supposing the lengths (indicated by solid lines) of GS_1, GS_2 and GS_3 are l, m and n, respectively, where $l < m < n$. RRFA uses the *minimum full-length* strategy to decide the flank regions of splice sites for frameshift analysis. Thus, the range for comparison (indicated by dashed lines) in flank regions of SS_1 is l bps, while for SS_2, it will be rescaled to m bps.

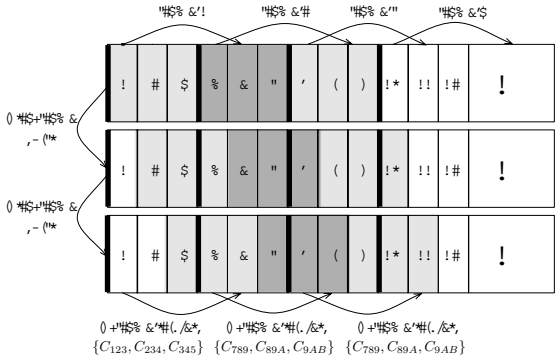

Figure 2. FRPM (frame-relayed pattern model): by adopting the concept of *frameshift* in reprogramming of mRNA translation, any single *frame* has to accommodate a triplet (i.e., 3 bps) and to provide the space for ± 1 *frameshift* (i.e., 2 extra bps); accordingly, the *penta*-mer is defined as a frame, i.e., the size of a frame is 5 bps. In addition, for keeping both the specificity of intra-triplet and the inter-triplet repeats, the frames are *overlapped*; thus, it can be found there are three triplets within each frame; and the last and first 2 bps are overlapped in contiguous frames.

be restricted to the flanking 10~30 nucleotides[3]. Thus, it is possible to discover new distinguishing features between exons and introns by full-length comparison. The full length of the shorter segments in the conjoined exon-intron sequences is adopted as the ranges for comparing sequence compositions, it is called rescaled full-length comparison (RFLC) as illustrated in **Figure** 1.

B. FRPM: Frame-relayed pattern model

Exons and introns are often characterized by a highly recurrent use of oligonucleotides[8]; it had also been suggested counting the frequency of oligomers is an effective measure for discovering patterns in sequences[7]. Moreover, the triplet composition in sequences has great implications in analyzing pre-mRNA sequences[28]; therefore, triplet is chosen as the *unit* oligomer. For retrieving triplet-related information in sequences, a frame-relayed pattern model (FRPM), as shown in **Figure** 2, was constructed and an illustrative example of FRPM is depicted in **Figure** 3.

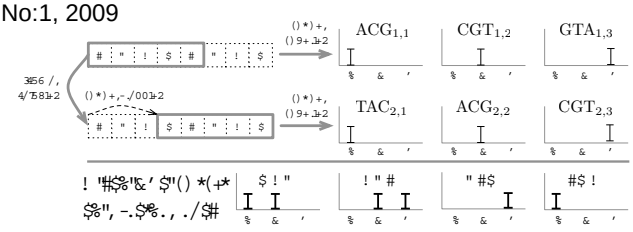

Figure 4. Frame-relayed triplet counting (FRTC): the x-axis is *in-frame* position and the y-axis is the triplet counting. For the example sequence ACGTACGT, the three triplets in the first frame are {$ACG_{1,1}$, $CGT_{1,2}$, $GTA_{1,3}$} and {$TAG_{2,1}$, $ACG_{2,2}$, $CGT_{2,3}$} will be found in the second frame, where the subscriptions are frame-ID and in-frame positions, respectively. An the distribution of triplet repeats can be obtained by combining all the triplet countings within the *relayed* frames. The contiguous frames are overlapped for preserving the detailed information about the distribution of triplet repeats.

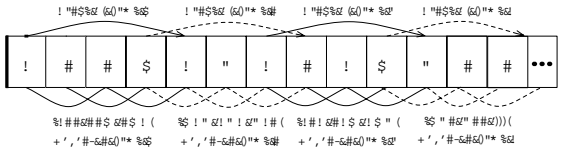

Figure 3. The triplet composition of example sequence ATTGACATAGCTT in FRPM model.

C. The frameshift sensitivity (FS)

1) Triplet repeats and frameshift: Triplet repeats were shown to be closely related with splice regulation[20]; the GGG repeats usually involve in the definition of exon-intron borders[16] and several novel motifs containing GCT are abundant in exons and introns[12], [27]. Moreover, small in-frame shifts leads about 50% of coding transcripts to be targeted by nonsense-mediated decay(NMD)[4]. All of these imply that the exons are very sensitive to frameshift. Therefore, the variation of triplet repeats while frameshifting is formulated to be a discriminator between exons and introns.

2) Triplet fidelity (TF): The distribution of triplet repeats for sequences modeled by FRPM is computed by frame-relayed triplet counting (FRTC), as illustrated in **Figure** 4. FRTC summaries all the distribution of triplet repeats, which form the basis for computing the triplet fidelity (TF). The calculation of TF is depicted in **Figure** 5. An example of TF calculation is illustrated in **Figure** 6.

3) The FS estimator: By summing up all the TF in one sequence, it is the *frameshift sensitivity* (FS) as defined in Equation 1, where k is the number of frames. The triplet mass (tm) is the number of hydrogen bonds in the canonical Watson-Crick base pairing, that is, for $A = T$ pair, the tm is 2, while for $G \equiv C$ pair, the tm is 3; the value of cm ranges from 0.66 ($\frac{6}{9}$) to 1 ($\frac{9}{9}$), where the denominator 9 stands for maximum number of hydrogen bonds in one triplet-pair (i.e., $GGG \equiv CCC$).

$$FS = \sum_{i=1}^{64} \frac{tm_i * TF_i}{k} \tag{1}$$

$$\Delta y = Max(c_1, c_2, c_3) - Mid(c_1, c_2, c_3)$$
$$\Delta x = Max(c_1, c_2, c_3) - Min(c_1, c_2, c_3)$$
$$TF_i = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

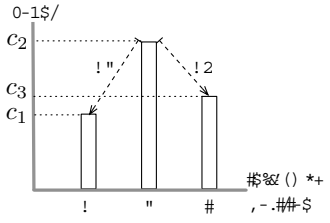

Figure 5. The triplet fidelity (TF): let c_1, c_2, and c_3 be the counts of a specific triplet at in-frame position 1, 2 and 3, respectively, TF equals to $\sqrt{(\triangle x)^2 + (\triangle y)^2}$, where $\triangle x = Max(C_x) - Mid(C_x)$, $\triangle y = Max(C_x) - Min(C_x)$ and $C_x = \{c_1, c_2, c_3\}$. That is, the more skewed the distribution is, the higher the TF value is.

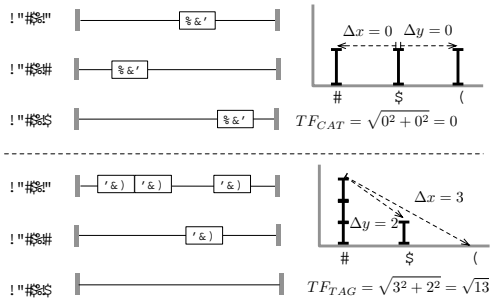

Figure 6. The triplet fidelity (TF): frameshifting causes the growth and decline of triplet repeats, the variation of the number of repeats after frameshifting is defined as the triplet fidelity (TF), which is calculated according to the differences between the maximum number of repeats and the other two repeat numbers. The top half give an example of *zero*-TF, in which the same number of repeats appear in the original sequence ($Shift_0$), the sequence right-shifting one base($Shift_1$) and the sequence right-shifting two bases($Shift_2$). The bottom half give an example of *high*-TF, in which the differences between the number of repeats in $Shift_0$, $Shift_1$ and $Shift_2$ are relatively large comparing to the example in the top half.

D. The accumulative penta-mer complexity (APC)

The low complexity may be preconditioned by strong inequality in biased nucleotide composition by tandem or dispersed repeats[19]; triplet repeats are one of the classic low-complexity sequence patterns in exons and introns[9]. Thus, in this research, the estimation of sequence complexity is accomplished by the occurrences of triplet repeats in frames.

1) The penta-mer complexity (PC): The *frame* is treated as complexity units (CUs), and thus, the CU is equivalent to the *penta*-mer complexity (PC). The values for the PCs are defined as the probability of finding such a 5-bps pattern in penta-mer sequences, therefore, $\parallel NNNNN \parallel = 4^5$,$\parallel AAAAB \parallel = \parallel AAAAB \parallel = \parallel ABABA \parallel = 4 * 3$, where $N = \{a, c, g, t\}$, $A \in N$ and $B \in N - \{A\}$. The formulas for calculating PCs are defined in Equation 2; and a classification scheme according to the FS values (i.e., treating a *penta*-mer segment as a short complete sequence) for the *penta*-mer frames is illustrated in **Figure** 7.

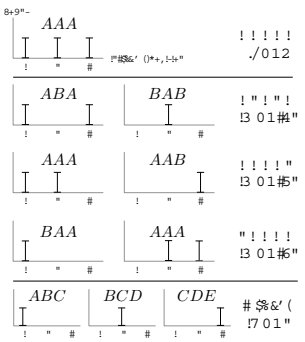

Figure 7. Classifications of PCs: According to the triplet repeats in each penta-mer frame, the PCs are classified into three categories according to the number of triplet repeats in the *penta*-mer frames. For maximum 3-repeats, it is a low-complexity penta-mer (LCP); for maximum 2-repeats, it is a medium-complexity penta-mer (MCP) and the others (i.e., no repeats in the penta-mer frame) are all high-complexity penta-mer (HCP).

$$\begin{cases} PC_{LCP} & = \parallel AAAAA \parallel / \parallel NNNNN \parallel \\ PC_{MCP} & = \frac{\parallel AAAAB \parallel + \parallel BAAAA \parallel + \parallel ABABA \parallel}{\parallel NNNNN \parallel} \\ PC_{HCP} & = 1 - (PC_{LCP} + PC_{MCP}) \end{cases} \quad (2)$$

2) The APC estimator: The total sequence complexity is estimated by accumulating the PCs of sequences, therefore, the accumulative PC (APC) is measured by accumulating the k PCs in the sequence (k is the number of frames) as defined in the equation 3. The APC (accumulative penta-mer complexity) in Equation 3 is devised to quantify the sequence complexity by aggregation of penta-mer complexity. The pm_i is the penta-mer mass, which is the total hydrogen bonds of the penta-mer (similar to the definition of triplet mass cm_i).

$$APC = \sum_{i=1}^{k} (pm_i * PC_i)/k \quad (3)$$

E. Persistency on anti-mutation (PAM)

Both FS and APC are estimators used to measure specific properties of a sequence; for balancing the their effects, they are integrated into Equation 4, which follows the principle for integrating the $precision$ and $recall$ into the $F1$ measurement; and the integrated measurement is termed as PAM (persistency on anti-mutation).

$$PAM = \frac{2 * FS * APC}{FS + APC} \quad (4)$$

F. The exon-clamp hypothesis (ECH)

The exons and introns are playing their roles in very different ways: although introns lack important biological functions that explain their flexible sequence compositions. While the exons have to preserve the protein-coding information; and hence, they should have rigorous sequence composition. Then, the *exon-clamp* hypothesis (ECH) is motivated, which claims

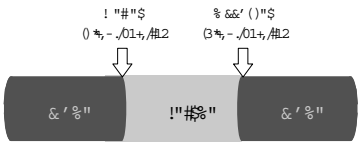

Figure 8. The tri-segment structure of sample sequences in data sets: the exon-(D)-intron-(A)-exon conjoined triple segments.

Table I
NUMBER OF EXTRACTED EXON-DONOR-INTRON (ALSO INTRON-ACCEPTOR-EXON) SEQUENCES

Species	Human	Mouse	Rat	C.ele.	Arab.	Dros.
#samples	35,700	8,466	1,023	99,408	64,898	62,251

exons are more persistent in anti-mutation than introns, i.e., the PAM values of exons will be greater than those in adjacent introns for all donors/acceptors (5SSs/3SSs). The ECH is defined in Equation 5, the result of ECH test is either 1 or 0.

$$ECH = \begin{cases} 1, if\ PAM_{exon} > PAM_{intron} \\ 0, else. \end{cases} \quad (5)$$

For a data set comprising n exon-intron junctions, the *support* of ECH is the total number of positive results in the ECH tests as defined in Equation 6.

$$Support = \frac{\sum_{i=1}^{n} ECH_i}{n} \quad (6)$$

III. RESULTS AND DISCUSSIONS

A. Data sets for ECH

The main source of experimental exons and introns were extracted from the complete pre-mRNA sequences in Xpro[10] with preferred tri-segment structure (see **Figure** 8), a database of eukaryotic protein-encoding genes. The number of samples in the six data sets are tabulated in **Table** I. In addition, two public data sets are also used to verify effectiveness of the devised PAM; one is HS3D[21], which comprises 2,796/2,880 true human exons and introns of length 70 bps; the other one is SpliceDB (SDB)[2], which provides 19,073 and 19,160 exons and introns of length 40 bps flanking at donor and acceptor sites, respectively.

B. Supports of FS, APC and PAM

By applying a vis-a-vis comparison on adjacent exons and introns around splice sites of six model organisms, the supports of FS, APC and PAM in discrminating adjacent exons and introns flanking donor and acceptor sites were shown in **Figure** 9 and **Figure** 10, respectively.

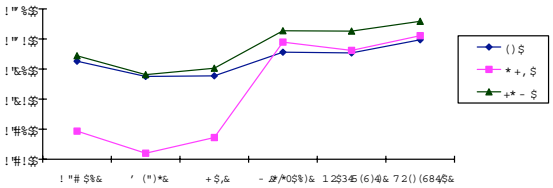

Figure 9. Supports on discriminating adjacent exon and intron flanking donor sites.

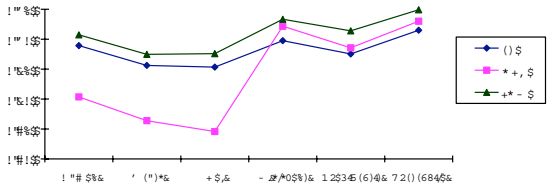

Figure 10. Supports on discriminating adjacent exon and intron flanking donor sites.

Clearly, PAM performed best; it got highest supports in all the six organisms. FS is comparable with PAM, yet, APC demonstrated a special discriminative power in lower eukaryotes. From the results, it showed that FS, APC and, especially, the PAM are all effective and reliable estimator in discriminating exons from introns.

C. Supports of PAM on public data sets

The proposed PAM was applied to two public datasets, SpliceDB and HS3D; these two datasets preserve only small range of flanking exons and introns of splice site. The lengths of sequences in SpliceDB and HS3D are 40 bps and 70 bps, respectively. The experimental results are listed in **Table** II. The supports using only the restricted 40 and 70 bps are lower than the results using full-length comparison as shown in **Figure** 9 and **Figure** 10. The results promote the feasibility of adopting the strategy of full-length comparison.

D. Supports of PAM on two-sided ECH

A more rigorous test on the effectiveness of PAM was performed by comparing both the two-sided flanking exons with the intermediate intron at the same time, The two-sided ECH was defined as Equation 7; and the results were listed in **Table** III.

$$ECH = \begin{cases} 1, if\ PAM_{5SS\_exon} > PAM_{5SS\_intron}\ \&\& \\ \quad PAM_{3SS\_exon} > PAM_{3SS\_intron} \\ 0, else. \end{cases} \quad (7)$$

Table II
PAM TEST ON PUBLIC DATA SETS.

Dataset	Len.	#D	#A	#Sup_D	#Sup_A	Sup_D%	Sup_A%
SpliceDB	40	19,073	19,160	13,782	15,592	**72%**	**81%**
HS3D	70	2,796	2,880	2,285	2,392	**82%**	**83%**

Table III
SUPPORTS ON TWO-SIDED ECH.

Species	#(D+A)	#Sup_D	#Sup_A	#Sup_D+A	%Sup_D+A
Human	35,700	31,133	3,2391	28,393	80%
Mouse	8,466	7,117	7,405	6,316	75%
Rat	1,023	871	896	777	76%
C. ele.	99,408	90,826	92,776	86,473	**87%**
Arbi.	64,898	58,261	59,314	55,104	**85%**
Dros.	62,251	57,859	59,086	55,612	**89%**

The results in **Table** III highly supports the two-sided ECH, especially in lower eukaryotes. They have some implications for sequence compositions of gene, which may form a new infrastructure for splice site recognition.

E. Potentials of the discovered new features

From the (deterministic) experimental results, it is clear that the FS-/APC-/PAM-values in exons are higher than these in introns; thus, the discovered new sequence features and the proposed ECH were all verified. The exon-intron junctions (EIJ) define the structure of eukaryotic protein-coding genes[23], and the researches of EIJ recognition mainly focus on the sequence composition in short-range flank regions (from 40 bps to 70 bps). Although they had been extensively studied[1], the short consensuses still bring on the far outnumbered *pseudo* splice sites[26], which hinder the development of effective and reliable recognition methods. With the rapid increase of genome sequence data, the discovered new discriminative sequence features between exons and introns are valuable information for devising new splice site recognition methodologies.

IV. CONCLUSIONS

In this research, two distinguishing sequence features between full-length exons and introns, , frameshift sensitivity (FS) and accumulative penta-mer complexity (APC), were identified by the proposed rescaled-range frameshift analysis (RRFA). Both FS and APC are arithmetic equations without any result-sensitive parameters and prior-knowledge involved; and from the experimental results, both of them were validated to be effective in discriminating exons from introns. Furthermore, by integrating FS and APC into a new discriminator, the persistency on anti-mutation (PAM), the results showed PAM is more effective than using FS or APC individually for exon-intron discrimination. The devised estimator FS, APC and, especially, the PAM reveal the distinguishing sequence properties between exons and introns, which provide valuable information for analyzing gene sequences.

REFERENCES

[1] M Burset, I A Seledtsov, and V V Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*, 28(21):4364–4375, 2000.

[2] M Burset, I A Seledtsov, and V V Solovyev. Splicedb: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res*, 29(1):255–259, 2001.

[3] V R Chechetkin and V V Lobzin. Study of correlations in segmented dna sequences: application to structure coupling between exons and introns. *J Theor Biol*, 190(1):69–83, 1998.

[4] Tzu-Ming Chern, Erik van Nimwegen, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, and Mihaela Zavolan. A simple physical model predicts small exon length variations. *PLoS Genet*, 2(4):e45, 2006.

[5] J M Claverie and L Bougueleret. Heuristic informational analysis of sequences. *Nucleic Acids Res*, 14(1):179–196, 1986.

[6] Alexei Fedorov, Serge Saxonov, and Walter Gilbert. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res*, 30(5):1192–1197, 2002.

[7] J W Fickett and C S Tung. Assessment of protein coding measures. *Nucleic Acids Res*, 20(24):6441–6450, 1992.

[8] C Frontali and E Pizzi. Similarity in oligonucleotide usage in introns and intergenic regions contributes to long-range correlation in the caenorhabditis elegans genome. *Gene*, 232(1):87–95, 1999.

[9] A Gabrielian and A Bolshoy. Sequence complexity and dna curvature. *Comput Chem*, 23(3-4):263–274, 1999.

[10] Vivek Gopalan, Tin Wee Tan, Bernett T K Lee, and Shoba Ranganathan. Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Res*, 32(Database issue):D59–63, 2004.

[11] Matthew P Hare and Stephen R Palumbi. High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol Biol Evol*, 20(6):969–978, 2003.

[12] Jennifer L Kabat, Sergio Barberan-Soler, Paul McKenna, Hiram Clawson, Tracy Farrer, and Alan M Zahler. Intronic alternative splicing regulators identified by comparative genomics in nematodes. *PLoS Comput Biol*, 2(7):e86, 2006.

[13] M Kozak. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev*, 47(1):1–45, 1983.

[14] S W Liou and Y F Huang. Investigating the intrinsic differences in flank regions of exon-intron junction sites. In *BMEI (2)*, volume 2, pages 96–101. IEEE Computer Society, 2008.

[15] Jacek Majewski and Jurg Ott. Distribution and characterization of regulatory elements in the human genome. *Genome Res*, 12(12):1827–1836, 2002.

[16] A J McCullough and S M Berget. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol*, 17(8):4562–4571, 1997.

[17] G Mengeritsky and T F Smith. New analytical tool for analysis of splice site sequence determinants. *Comput Appl Biosci*, 5(2):97–100, 1989.

[18] K Nakata, M Kanehisa, and C DeLisi. Prediction of splice junctions in mrna sequences. *Nucleic Acids Res*, 13(14):5327–5340, 1985.

[19] Y L Orlov and V N Potapov. Complexity: an internet resource for analysis of dna sequence complexity. *Nucleic Acids Res*, 32(Web Server issue):W628–33, 2004.

[20] Joanna L Parmley and Laurence D Hurst. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol*, 24(8):1600–1603, 2007.

[21] Pasquale Pollastro and Salvatore Rampone. Hs3d: Homo sapiens splice site data set. *Nucleic Acids Research, Annual Database Issue*, 2002.

[22] S Rampone. Recognition of splice junctions on dna sequences by brain learning algorithm. *Bioinformatics*, 14(8):676–684, 1998.

[23] P A Sharp. Splicing of messenger rna precursors. *Science*, 235(4790):766–771, 1987.

[24] V V Solovyev, A A Salamov, and C B Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res*, 22(24):5156–5163, 1994.

[25] Rotem Sorek and Gil Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*, 13(7):1631–1637, 2003.

[26] H Sun and L A Chasin. Multiple splicing defects in an intronic false exon. *Mol Cell Biol*, 20(17):6414–6425, 2000.

[27] Rodger B Voelker and J Andrew Berglund. A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. *Genome Res*, 17(7):1023–1033, 2007.

[28] Erik Willie and Jacek Majewski. Evidence for codon bias selection at the pre-mrna level in eukaryotes. *Trends Genet*, 20(11):534–538, 2004.

[29] G K Wong, D A Passey, Y Huang, Z Yang, and J Yu. Is junk dna mostly intron dna? *Genome Res*, 10(11):1672–1678, 2000.