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

Sing-Wu Liou and Yin-Fu Huang

1Graduate School of Engineering Science and Technology
National Yunlin University of Science and Technology
Douliou, Yunlin, Taiwan.
g9110808@yuntech.edu.tw

2Graduate 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 for being removed during the premRNA maturing processes and they are often characterized by a highly recurrent use of specific triplets. Thus, exons are definitely non-randomness and that distinguishes them from introns.

Discriminant 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; the BRAIN 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.

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.

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; recent evidence also suggested that sequence conservation associated with splice sites may extend relatively far away from intron-exon boundaries; it had also been shown the structural coupling at the exon-intron junctions might not
regions of splice sites for frameshift analysis. Thus, the range for comparison is $l < m < n$. RFPA 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.

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 overlapped in contiguous frames.

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 intratetraplet 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.

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.

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 ($\frac{2}{3}$) to 1 ($\frac{1}{2}$), where the denominator 9 stands for maximum number of hydrogen bonds in one triplet-pair (i.e., $GGG \equiv CCC$).

$$FS = \sum_{i=1}^{64} \frac{tm_i \ast TF_i}{k}$$ (1)
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 $\sqrt{(\Delta x)^2 + (\Delta y)^2}$, where $\Delta x = Max(C_1) - Mid(C_1)$, $\Delta y = Max(C_3) - Min(C_3)$ and $C_3 = \{c_1, c_2, c_3\}$. 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, $||NNNNN|| = 4^5$, $||AAABAB|| = ||AAAAAB|| = ||ABABA|| = 4 \cdot 3$, 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.

\[
\Delta x = Max(c_1, c_2, c_3) - Min(c_1, c_2, c_3)
\]

\[
\Delta y = Max(c_1, c_2, c_3) - Mid(c_1, c_2, c_3)
\]

\[
TF_i = \sqrt{(\Delta x)^2 + (\Delta y)^2}
\]

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 $cm_i$).

\[
APC = \sum_{i=1}^{k} (pm_i \cdot PC_i)/k
\]

E. Persistence 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 (persistence on anti-mutation).

\[
PAM = \frac{2 \cdot FS \cdot APC}{FS + APC}
\]

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
Table I

<table>
<thead>
<tr>
<th>Species</th>
<th>Human</th>
<th>Mouse</th>
<th>Rat</th>
<th>C.ele.</th>
<th>Arab.</th>
<th>Dros.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#samples</td>
<td>35,700</td>
<td>8,466</td>
<td>1,023</td>
<td>99,408</td>
<td>64,898</td>
<td>62,251</td>
</tr>
</tbody>
</table>

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.

\[
ECH = \begin{cases} 
1, & \text{if } PAM_{exon} > PAM_{intron} \\
0, & \text{else.} 
\end{cases} \tag{5}
\]

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.

\[
Support = \frac{\sum_{i=1}^{n} ECH_i}{n} \tag{6}
\]

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 around splice sites of six model organisms, the supports of FS, APC and PAM in discriminating adjacent exons and introns flanking donor and acceptor sites were shown in Figure 9 and Figure 10, 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 discriminating adjacent exons and introns flanking donor and acceptor sites were shown in Figure 9 and Figure 10, respectively.

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.

\[
ECH = \begin{cases} 
1, & \text{if } PAM_{SS,exon} > PAM_{SS,intron} \land \ PAM_{SS,exon} > PAM_{SS,intron} \\
0, & \text{else.} 
\end{cases} \tag{7}
\]

Table II

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Len</th>
<th>#D</th>
<th>#A</th>
<th>#Sup_D</th>
<th>#Sup_A</th>
<th>Sup_D%</th>
<th>Sup_A%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpliceDB</td>
<td>40</td>
<td>19,073</td>
<td>19,160</td>
<td>13,782</td>
<td>15,592</td>
<td>72%</td>
<td>81%</td>
</tr>
<tr>
<td>HS3D</td>
<td>70</td>
<td>2,796</td>
<td>2,880</td>
<td>2,285</td>
<td>2,392</td>
<td>82%</td>
<td>83%</td>
</tr>
</tbody>
</table>
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