2012-02

Identical sequence patterns in the ends of exons and introns of human protein-coding genes

COMPUTATIONAL BIOLOGY AND CHEMISTRY, OXFORD, v. 36, n. 1, pp. 55-61, FEB, 2012
http://www.producao.usp.br/handle/BDPI/33546

Computational Biology and Chemistry 36 (2012) 55–61

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Identical sequence patterns in the ends of exons and introns of human protein-coding genes

Raphael Tavares[a,1], Gabriel Renaud[a,1,2], Paulo Sergio Lopes Oliveira[c,3], Carlos G. Ferreira[b,2], Emmanuel Dias-Neto[d,e,4], Fabio Passetti[a,*]

[a] Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rua André Cavalcanti, 37 - CEP 20231-050, Rio de Janeiro, RJ, Brazil
[b] Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rua André Cavalcanti, 37 - CEP 20231-050, Rio de Janeiro, RJ, Brazil
[c] Laboratorio Nacional de Biociências, Caixa Postal 6192 - CEP 13083-970, Campinas, SP, Brazil
[d] Laboratório de Neurociências (LIM-27), Instituto de Psiquiatria, Faculdade de Medicina, Universidade de São Paulo, R. Dr. Ovídio Pires de Campos, 785-Caixa Postal 3671 - CEP 01060-970, São Paulo, SP, Brazil
[e] Lab. of Medical Genomics, Centro Internacional de Pesquisa e Ensino (CIPE), Hospital AC Camargo, Rua Taguá, 440 - CEP 01508-010, São Paulo, SP, Brazil

ARTICLE INFO

Article history:
Received 11 November 2011
Accepted 4 January 2012

Keywords:
Transcriptomics
Bioinformatics
Genome
Sequence analysis

ABSTRACT

Intron splicing is one of the most important steps involved in the maturation process of a pre-mRNA. Although the sequence profiles around the splice sites have been studied extensively, the levels of sequence identity between the exonic sequences preceding the donor sites and the intronic sequences preceding the acceptor sites has not been examined as thoroughly. In this study we investigated identity patterns between the last 15 nucleotides of the exonic sequence preceding the 5′ splice site and the intronic sequence preceding the 3′ splice site in a set of human protein-coding genes that do not exhibit intron retention. We found that almost 60% of consecutive exons and introns in human protein-coding genes share at least two identical nucleotides at their 3′ ends and, on average, the sequence identity length is 2.47 nucleotides. Based on our findings we conclude that the 3′ ends of exons and introns tend to have longer identical sequences within a gene than when being taken from different genes. Our results hold even if the pairs are non-consecutive in the transcription order.

1. Introduction

The removal of introns from pre-mRNA is known as splicing. One of the most important steps during this process for spliceosomal introns is the recognition of the intronic dinucleotides (the GU–AG rule) respectively located at the 5′ and 3′ splice sites acting as a reference for the whole enzymatic machinery (Modrek and Lee, 2002). As a result, the nucleotide composition of the donor and acceptor splice sites have been analyzed since the early 1980s (Breathnach and Chambon, 1981; Mount, 1982; Shapiro and Senapathy, 1987). These studies tried to search for consensus sequences that could help not only to explain the splicing mechanism but also to define the exon/intron organization of a gene (Zhang, 1998; Bernard and Michel, 2009). Early studies constructed sequence profiles for sequences around the donor and acceptor splice sites and, even with very few samples, were able to observe a consensus sequence of [A|C]AG/GT[A|G]AGT for the former and [C|T]N[C|T]AG/G for the latter where "/" represents the splice site (Breathnach and Chambon, 1981; Mount, 1982). As more samples became available, the presence of both consensus sequences was observed in other organisms thus reaffirming its biological significance (Mount et al., 1992; Burset et al., 2000).

Another characteristic of the 5′ and 3′ of intron splice regions that has been reported in the literature is the presence of tandem sequences following the pattern GYN|GYN (where "|" is the exon–intron junction and Y stands for C or T; N stands for A, C, G or T) located in donor splice sites (Hiller et al., 2006) and NAG|NAG pattern in acceptor spice sites (Hiller et al., 2004). Although the similarity between the intronic sequence preceding the acceptor splice site (almost exclusively AG) and the exonic sequence preceding the donor splice site (predominantly AG) has been known for years, few modern studies have looked beyond the trivial comparison that can be made between both. In addition, these sequences are sometimes referred to as "shadow sequences" (Qiu et al., 2004). One study aiming at investigating the model which speculates that the duplication of small genomic sequences is a possible origin of spliceosomal introns conducted such a comparison between both

* Corresponding author. Tel.: +55 21 3207 6546.
E-mail addresses: rtsilva@inca.gov.br (R. Tavares), grenaud@inca.gov.br (G. Renaud), paulo.oliveira@lnbio.org.br (P.S.L. Oliveira), cferreira@inca.gov.br (C.G. Ferreira), emmanuel@usp.br (E. Dias-Neto), passetti@inca.gov.br (F. Passetti).
[1] These authors contributed equally to this work.
[2] Tel.: +55 21 3207 6546.
[3] Tel.: +55 19 3512 1010.
[4] Tel.: +55 11 2189 5000x2955.

1476-9271/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2012.01.002

exonic and intronic sequences flanking the splice sites of a given intron (Zhuo et al., 2007). This study has revealed that pairs of these sequences with perfect matches were overrepresented in the human genome by comparing the number of identical nucleotides in the splice sites of introns to a control set of randomly selected splice sites. However, how the nucleotide bias in the vicinity of splice sites influenced the levels of sequence identity was not studied as thoroughly. In addition, it was left to verify whether selecting the splice sites from the same gene rather than selecting them from different genes would have any effect on the levels of sequence identity for the control set. Our study aims at measuring the average identity length between the last 15 nucleotides of an exon and the last 15 nucleotides of introns using a set of 471 protein-coding genes for which no intron retention was detected. We show that the end of exons/introns have higher levels of sequence identity than random exonic/intronic sequences and than these identity levels increase if the pairs are taken from the same gene. We also show that selecting these sequences from consecutive pairs of exons and introns instead of choosing random non-consecutive pairs within the same gene does not seem to affect the average sequence identity length.

2. Materials and methods

Briefly, we generalize our analysis by using randomized sets of protein-coding genes and compare the average identity length to the one calculated on non-consecutive pairs of exons and introns from the same gene. We then compare this average length with the one obtained for pairs of exons and introns that were from different genes and also for pairs of sequences not necessarily stemming from the end of exons and introns.

2.1. Detection of repeated sequences in 3′ of exons and introns

The first step was to select a set of protein-coding genes without intron retention since we cannot state for sure that exons that appear to be consecutive are in fact consecutive *in vivo*. Picking the end of an intron that would be in fact an exon would interfere with our analysis. To this end, we discarded any gene having either a RefSeq transcript or an EST with an exon that would span an entire intron from another transcript or EST. Most human genes had at least one transcribed sequence, usually an EST, presenting this feature. This could indicate that either intron retention is a common event or, and more likely, that the contamination of the database with ESTs derived from immature (partially spliced) mRNAs is rife. To detect this, we used a methodology called ternary matrices (manuscript in preparation). We used the RefSeq dataset (version 28) as well as Unigene (version 207) from NCBI, aligned to the human genome (UCSC version hg18) using SIBsim4 (http://sibsim4.sourceforge.net/). The scripts used to perform these analyses were written in Perl. Thus, from the original set of 20,311 human protein-coding genes, we were left with 471 genes without any transcript or EST that presented the possibility of intron retention.

The second step was to define an identity hit between end of exon sequences and the end of intron sequences. We picked 15 nucleotides upstream of the splice site in exons and, from the same gene, 15 nucleotides upstream of the splice site in the consecutive intron. The length of 15 nucleotides was picked arbitrarily because it is unlikely that an identity hit with more than 15 base pairs would be expected. We compared both sequences by counting the number of identical base pairs starting from the 3′ end until we met a mismatch and defined an identity hit if both sequences shared at least 2 identical nucleotides in the last 2 positions.

2.2. Analysis of nucleotide prevalence in 3′ end of exons and introns

The tool Weblogo was used to determine the nucleotide prevalence at the 3′ end of exons and introns according to the methodology described by Crooks et al. (2004).

2.3. Statistical validation

The March 2006 version of the Human Genome and the RefSeq track, both from the UCSC Genome Browser, were used. A Perl script was developed for this analysis to cluster RefSeq genes and their associated transcripts. Transcripts were clustered given their Locus Link ID and transcripts mapping to more than a single location in the genome were discarded. The rand() subroutine from Perl version 5.8.8 was used for random number generation. The coordinates of the exons and introns based on the mapping information from the UCSC Genome Browser were used to extract the genomic sequence corresponding to the sequence preceding the splice sites. Identity was computed using a Perl subroutine by counting how many base pairs were identical starting from the 3′ end. Genes that did not have at least one identity hit between a pair of consecutive exon/intron were discarded. The selection of 471 protein-coding genes has been repeated 60,000 times. For each set, the number of exon/intron sequence pairs for each pattern length for each random category was measured. To compute the averages presented in Table 3, we determined that each set of 471 used on average 2922 pairs of exons/introns. This number was used to normalize the raw counts allowing us to compute the overall averages.

The pairwise *t*-test was performed using the pairwise.t.test() from R version 2.9.0. We visually evaluated whether our data fitted a normal distribution using a quantile-normal graph using the qqnorm() function and qqline() (Fig. S1).

3. Results

3.1. Search for patterns of repetition in the 3′ ends of consecutive exons and introns

Our first approach was to search for identical sequence patterns in the 3′ end of consecutive exons and introns from the same gene, by counting the number of identical nucleotides (starting from the 3′ end and going to the 5′ end), until a mismatch was found (Fig. 1A). We used a dataset consisting of 20,311 human protein-coding genes having at least one RefSeq RNA sequence mapped onto the human genome. To avoid introns that could be expressed as exons in certain splicing isoforms, we removed any gene that displayed intron retention in any of its RefSeq transcripts or ESTs which left us with a list of 471 genes to run our search on. For this set of 471 genes, we found that 58.4% (standard deviation of 20.9%) of them had at least one exon/intron pair that shared at least 2 nucleotides and in 99.9% of cases those nucleotides were AG. Hence, we set the minimum length for flagging 2 sequences as displaying an identity pattern at 2 identical nucleotides.

The maximum length that was found for all the sequences flagged as having a pairwise sequence identity pattern was of 7 nucleotides and found in 4 distinct genes (Table 1). As expected, the number of genes exhibiting the sequence identity pattern decreased as the length of the identity pattern increased. The average length of all exon/intron pairs having this identity pattern in a given gene was also calculated and an average of 2.47 identical nucleotides (standard deviation of 0.53) was found for any given gene in our set.

The human gene which presented the highest frequency of splice sites having the sequence pattern focus of this article was

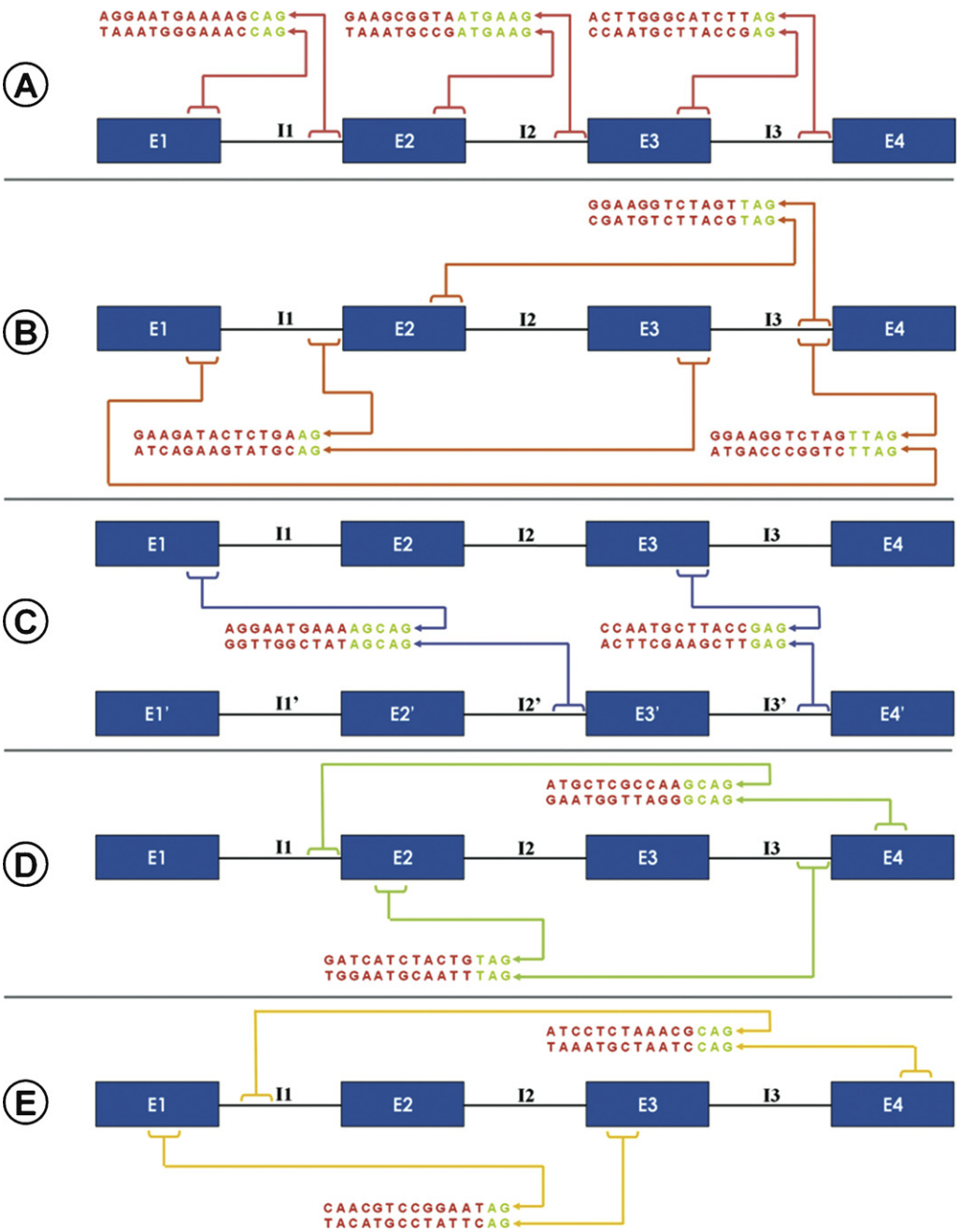

Fig. 1. Overview of the methodology for the 5 groups of sequences that were measured for the randomizations. The top part (A) represents our approach for the 3′ end sequence of consecutive exons and introns. The second part (B) represents the approach for the 3′ end sequence of non-consecutive exons and introns. We sought to compare the level of identity of our data against randomized sequences to determine if our data displayed higher levels of identity than expected. We designed three approaches to stochastic generation. The first one (C) involved comparing sequences from the end of exons and the end of introns from random genes (middle part). The second (D) involved comparing sequences from the same gene where one was taken from the end of an intron and the other was taken from an exon regardless of its position as long as it ended with AG while the third one (E) involved comparing sequences from exons and introns, regardless of their position, that both ended with AG and came from the same gene.

Table 1
Number of genes with a confirmed identity sequence pattern between consecutive exons and introns.

Length of identity pattern (nt)	Genes with pattern	Genes with no pattern	Non-cumulative
2	469	2	133
3	336	135	211
4	125	346	85
5	40	431	21
6	19	452	15
7	4	467	4

LIN54, a subunit of the DREAM/LINC complex, involved in the regulation of cell cycles genes (Schmit et al., 2009). As can be depicted in Fig. 2, *LIN54* presented multiple repeat sequence patterns and the longest is TTGCAG, presented both in the end of exon 3 and the end of intron 3.

3.2. Analysis of nucleotide prevalence in 3′ regions of consecutive exons and introns

Using sequence logos, we have noticed some features that were present in exons and introns separately and in the analysis performed on superimposed exons/intros sequences as well. We

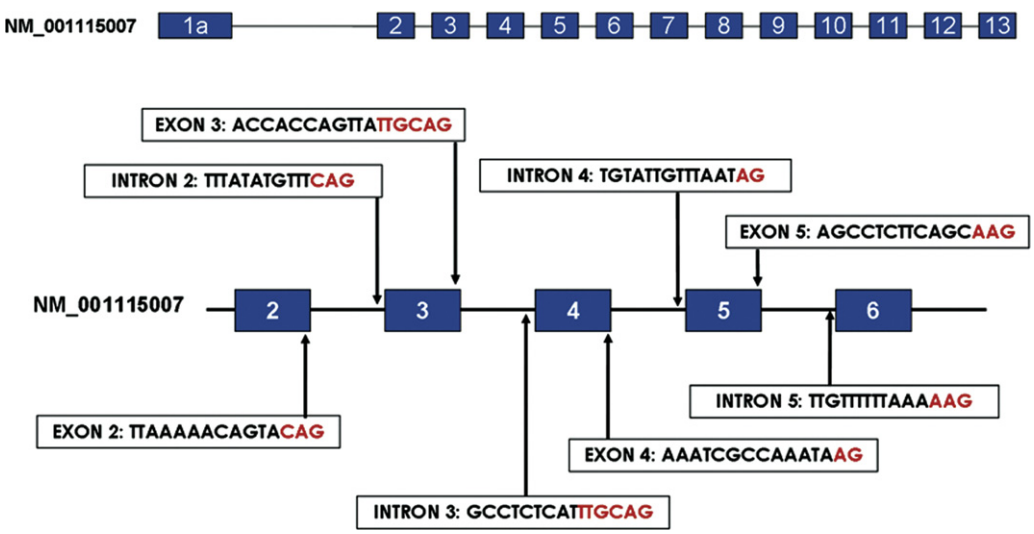

Fig. 2. Example of a gene with various identity patterns. A manual inspection of our results revealed a gene with identical sequences between four pairs of consecutive exons and introns. The gene *LIN54* has the pattern CAG in the second exon, TTGCAG in exon three, AG in exon four and the pattern AAG in the fifth exon.

observed the prevalence of the nucleotides adenine (A) and guanine (G) at positions −2 and −1, respectively, and a preference for cytosine (C) at position −3 when a cutoff length of 3 identical nucleotides was applied. Another interesting characteristic was the equal frequency for all nucleotides (∼25%) at position −4. Unsurprisingly, a slight preference for pyrimidines in the remaining positions −15 to −5 was observed in introns only. For the same positions, exons did not exhibit this property when analyzed separately and both sequences seem to lose this stretch of pyrimidines once superimposed (Fig. 3).

3.3. Assessing statistical validity

To assess the statistical validity of our results and to verify whether we would see the same identity pattern length in a different set of genes, we selected a random set of 471 human genes and performed 5 distinct analyses for every gene that was picked, we searched for our identity pattern among (Fig. 1):

(A) The exonic sequences preceding the splice site and the intronic sequence preceding the splice site from the intron immediately downstream in the same gene (Fig. 1A). Let **N** be the number of exon/intron pairs that share at least 2 identical base pairs at their 3′ end. These sequences correspond to the same approach that was detailed in the first part of the results section and whose average pattern length is left to statistically validate.
(B) **N** pairs of exon/intron sequences from the same gene that was selected from a) but from non-consecutive exons and introns (Fig. 1b). We required that the pairs shared at least 2 identical base pairs at their 3′ end and selecting the same exon or intron more than once was allowed.
(C) **N** pairs of exon/intron sequences, not necessarily from the same gene, from a pre-compiled genome-wide list of exonic sequences preceding the splice site and intronic sequences preceding the splice site (Fig. 1C). This meant that these exon/intron pairs were not necessarily evaluated in a consecutive order as they were for A.
(D) **N** pairs of exon/intron sequences, one taken from an exon regardless of its position with respect to the splice site and ending with AG (see explanation below) and the other taken from the 3′ end of an intron, both from the same gene that was selected for A (Fig. 1D).
(E) **N** pairs of exon/intron sequences, both ending in AG, one taken from an exon, the other taken from an intron, both from the same gene that was selected for A and not necessarily preceding the splice site (Fig. 1E).

As mentioned above, the requirement was that sequences picked from exons and introns that were not necessarily from the 3′ end ended with AG. This was done since we evaluated that, for any given human gene, on average 58.9% (standard deviation of 20.7%) of its consecutive exonic/intronic sequence preceding the splice site shared at least 2 base pairs and that, in the overwhelming majority of cases (99.9%) that sequence was AG. As previously presented, for our 471 initial genes, we had 58.4% (standard deviation of 20.9%) of exon/intron pairs that shared at least 2 nucleotides and, in 99.9% of cases, those nucleotides were AG. Hence, our initial dataset may be considered a representative set of the entire human genome and

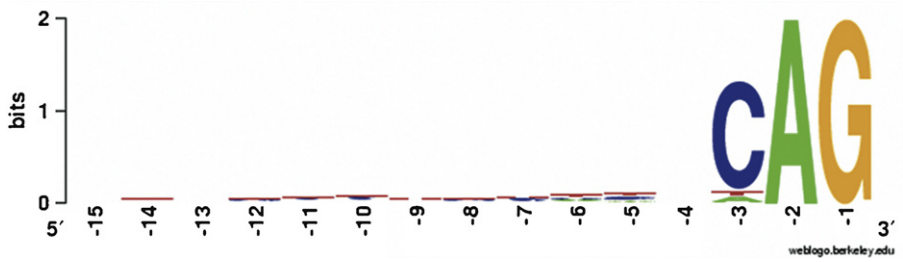

Fig. 3. Sequence logo for exon/intron overlap at a cutoff of 3 identical nucleotides. The sequence logo shows that the dinucleotides AG are prevalent at positions −2 and −1 and that cytosine is overrepresented at position −3.

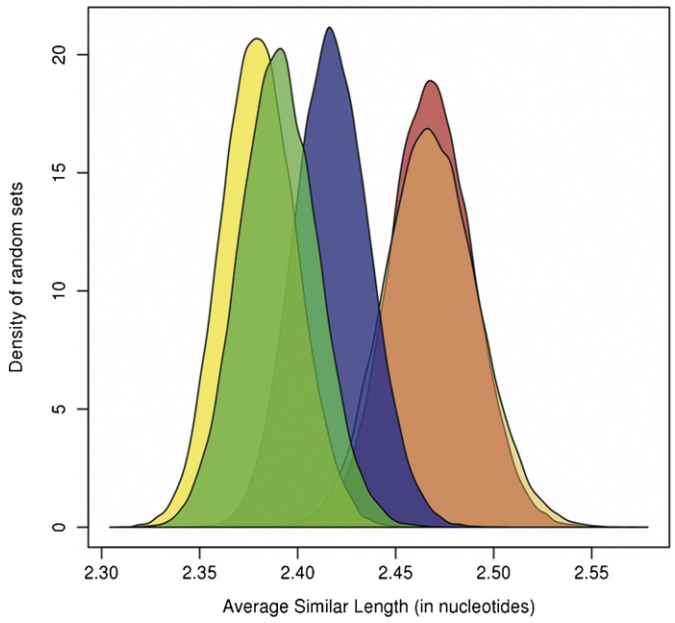

Fig. 4. Distribution of the average identity length for all 5 groups of sequences that were extracted from the random genes sets. All 5 curves represent the distribution of the length average for identity hits in a given gene set. The first group (A) for consecutive exons and introns next to the splice site is represented is in red, the second group (B) for non-consecutive exons and introns next to the splice site is in orange, the third group (C) for exons and introns next to the splice site from random genes is in blue, the fourth group (D) for introns next to the splice site and exonic sequences ending with AG not necessarily next to the splice site is in green and random exonic and intronic sequences (E) both ending with AG not necessarily next to the splice site in yellow. We visually verified that our curves corresponded to a normal distribution using a quantile-normal graph (Fig. S1).

anchoring our random exonic/intronic sequences to an AG would provide an adequate stochastic benchmark.

As depicted in Fig. 4, after 60,000 random selections of sets of 471 protein-coding genes among all possible human protein-coding genes, we observed 5 distinct normal distributions for the analyses that were performed. According to our results, a pairwise comparison of the distributions revealed that they all had different means (Table 2) with high statistical significance (p-value < 2e−16) except for the comparison between groups A and B (p-value 0.094). Although the overall averages were very close to each other, a greater difference could be observed in the averages of exon/intron sequence pairs for each pattern length cutoff (Table 3). The distributions of the average number of exon/intron sequence pairs for a given data group were plotted separately according to the length of the pattern that was found (Fig. 5).

Table 2
Mean and average length for the 5 groups of sequences measured during the stochastic evaluation.

Group	Mean	Standard deviation
A	2.468	0.021
B	2.468	0.024
C	2.418	0.019
D	2.391	0.019
E	2.380	0.020

Table 3
Average number of exon/intron pairs per random group for each pattern length.

Group	2	3	4	5	6	7–15
A	1984.6	683.5	178.2	52.9	16.0	6.9
B	1990.3	682.7	175.8	52.6	15.2	5.8
C	2017.6	677.1	163.4	46.7	12.8	4.7
D	2105.8	600.4	155.5	43.3	12.7	4.8
E	2116.4	589.7	157.1	43.1	11.7	4.4

4. Discussion

Our initial goal was to study the patterns of identity between the exonic sequences and intronic sequences close to their respective 3′ splice site. To avoid the problem of having false positives contaminating our results due to intron retention, a set of 471 genes that did not show any signs of intron retention was built and an analysis of the identity patterns between the end of exons and the end of their consecutive intron was performed. To verify whether or not the results held on a genome-wide scale and if they were statistically significant, various measures were performed on a randomized of the set of genes. In our initial set of 471 genes, a maximum length of 7 identical nucleotides between the ends of pairs of exons and introns was observed. The vast majority of sequences exhibited a pattern of at least 2 identical nucleotides since only 2 out of 471 genes did not meet this criterion. Moreover, the majority, 336 out of 471 (71.3%), of genes exhibited a pattern of at least 3 identical nucleotides in at least one of their exon/intron pairs.

The high prevalence of dinucleotide AG in positions −2 and −1, the relatively high frequency of cytosine in position −3 both in exons and introns and the absence of preference for any nucleotide at position −4 were consistent with previous studies that investigated the composition of nucleotides in exon/intron junctions for 5′ and 3′ splice sites (Padgett et al., 1986; Shapiro and Senapathy, 1987). Probably this sequence composition could be related with snRNA U1 which is responsible for identifying the 5′ splice site, since there is base complementarity in this region between pre-mRNA and this component of the spliceosome (Lund and Kjems, 2002). Finally, there was a slight preference for pyrimidines for positions −15 to −5, respecting the polypyrimidine tract previously described in the literature (Baralle and Baralle, 2005). The frequency of C observed at position −3 in introns could also be related with the pattern NAGNAG and, according to various studies, the N within the NAG exhibited a nucleotide preference in the following order C > T > A > G. It would also seem that this nucleotide order influences the strength of 3′ splice site (Akerman and Mandel-Gutfreund, 2006; Sinha et al., 2009).

To what point the selection of our gene list influenced our identity pattern length was left to determine. Furthermore, the influence of selecting sequences close to the splice site and from consecutive exon/intron pairs from the same gene needed to be evaluated. We noticed that the average identity length of 2.47 (standard deviation of 0.53) that was measured for the initial 471 genes (versus 2.48 genome-wide with a standard deviation of 0.50) fell perfectly within the mean of the A group for random sets. This indicated that our initial group of gene was not biased towards any particular selection criteria.

The presence of the shadow sequence in exons of the AG dinucleotide that normally occurs at the end of introns and the nucleotide distribution for the positions preceding the AG in both the exons and introns has been known for years. However, whether the average identity length between consecutive exons and introns naturally follows from this nucleotide distribution was left to determine. First, the comparison between group D and group E (see Fig. 1) indicates that taking the intronic sequence from the 3′ end rather than from any part of the intron slightly increases the levels of identity. Comparing the average identity length of the 2 previous groups to group C indicates that taking the exonic sequence from the end rather than from a random position significantly increases the levels of identity despite being from a different gene. However, stating that the higher levels of identity is solely due to the proximity of the sequence to the splice site or whether it is simply due to

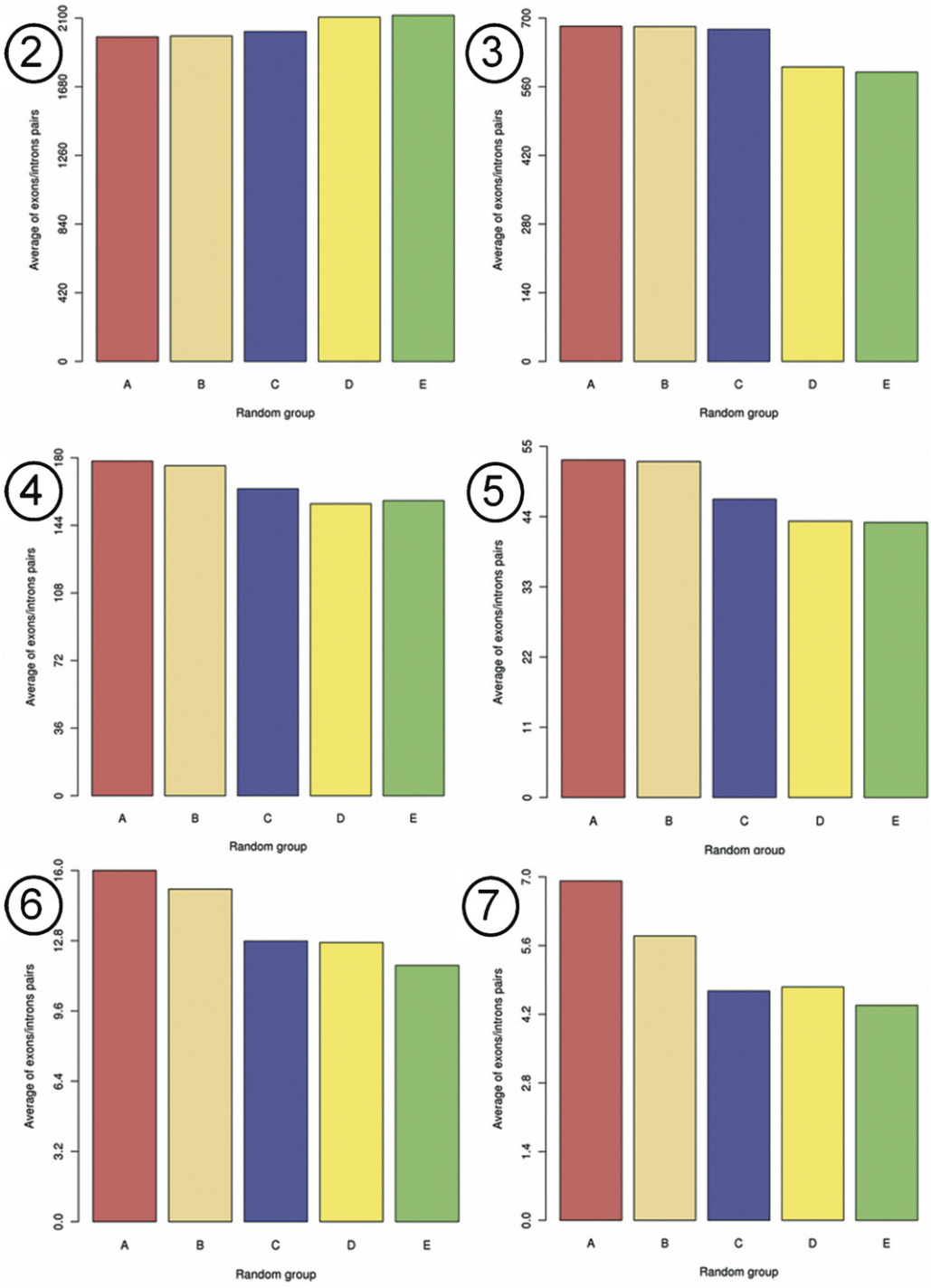

Fig. 5. Distributions of the averages for various identity pattern lengths for all 60,000 random sets. For pattern lengths greater than 3 (3), we can see that groups A and B are approximately equal. However, for pattern lengths greater than 7 (7), we notice a slight increase for group A with respect to group B. A clear conclusion is hard to make since the scale is very small since finding patterns of length greater than 7 is an unlikely event. Groups D and E generally follow the same trend while group C seems to be in between the A and B groups and the D and E groups.

the nucleotide bias due to the presence of the "AG" is harder to state unequivocally. This is because the averages measured on groups D and E provide a benchmark for comparison that is slightly inaccurate since the nucleotide distribution is likely to be different in the vicinity of the splice site than in the rest of the exon (Korzinov et al., 2008).

The comparison between the average lengths for group C and the A and B groups yields a clearer conclusion since both groups use sequences at the end of exons/introns and the only variable is the choice of the same gene. This allows us to state that the ends of exons and introns have a higher degree of identity if both sequences are taken from the same gene. Furthermore, the comparison between group A and group B seems to indicate that taking exonic and intronic sequences on each side of the same intron or from different introns does not affect the average identity length. When using the average number of identical nucleotide for comparison, we notice that the differences between the 5 different data sets are very small. However, when looking at Table 3, we notice

that the even though the differences are significant within each pattern length cutoff, they will have little influence on the overall average for identity length due to the abundance of exon/intron sequence pairs with a length of 2.

A closer look at Fig. 5 seems to indicate that group A, group B and group C do not exhibit a considerable difference for patterns of length 3. However, for identity pattern lengths from 4 to 6, we can recognize the differences between the various groups for the averages of identity pattern length that we previously elaborated on.

From the perspective of the average identity length, selecting the sequences from the same gene rather than from the boundaries of the same intron seems to be the determining factor. We can refer to the study by Zhuo et al. (2007) which reported an overrepresentation of introns with high levels of identity between the sequences at each splice site which would be, according to the authors, a consequence of the duplication of small exonic sequences as a source of boundaries for new introns, sometimes referred to in the literature as tandem genomic duplications (Roy and Gilbert, 2006). This could explain the difference in sequence identity length between groups A and C but fails to offer an explanation for the difference between groups B and C and the apparent similarity between groups A and B. Although the exact biological cause remains elusive, this would indicate a preference for similar splice sites within a given gene.

5. Conclusion

In the present study, we extensively analyzed the ends of exons and introns of human protein-coding genes and discovered identical sequence patterns with a tendency of longer sequences if they are found within the same gene. Our data gives an additional sequence feature not previously described that may rule the influence of pre-mRNAs splicing and may be used to improve mathematical models of alternative splicing predictors.

Acknowledgements

GR is supported by CNPq (382791/2009-6). FP and the Bioinformatics Unit at the Clinical Research Coordination at INCA acknowledge the support of CNPq, MCT/CT-Saúde and DECIT/SCTIE/MS (#577593/2008-0 and #312733/2009-7), Swiss Bridge Foundation and Fundação do Câncer. RT was supported by INCA/MS. EDN and the LIM27 acknowledge the support of Associação Beneficente Alzira Denise Hertzog Silva (ABADHS).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2012.01.002.

References

Akerman, M., Mandel-Gutfreund, Y., 2006. Alternative splicing regulation at tandem 3′ splice sites. Nucleic Acids Res. 34, 23–31.
Baralle, D., Baralle, M., 2005. Splicing in action: assessing disease causing sequence changes. J. Med. Genet. 42, 737–748.
Bernard, E., Michel, J., 2009. Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs): an application to splice sites of human genome introns. Comput. Biol. Chem. 33, 245–252.
Breathnach, R., Chambon, P., 1981. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. 50, 349–393.
Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28, 4364–4375.
Crooks, G.E., Hon, G., Chandonia, J., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.
Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., Platzer, M., 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat. Genet. 36, 1255–1257.
Hiller, M., Huse, K., Szafranski, K., Rosenstiel, P., Schreiber, S., Backofen, R., et al., 2006. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. Genome Biol. 7, 65, 1-65.12.
Korzinov, O.M., Astakhova, T.V., Vlasov, P.K., Roytberg, M.A., 2008. Statistical analysis of DNA sequences in the neighborhood of splice sites. Mol. Biol. 42, 133–145.
Lund, M., Kjems, J., 2002. Defining a 5′ splice site by functional selection in the presence and absence of U1 snRNA 5′ end. RNA 8, 166–179.
Modrek, B., Lee, C., 2002. A genomic view of alternative splicing. Nat. Genet. 30, 13–19.
Mount, S.M., 1982. A catalogue of splice junction sequences. Nucleic Acids Res. 10, 459–472.
Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., Fields, C., 1992. Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res. 20, 4255–4262.
Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S., Sharp, P.A., 1986. Splicing of messenger RNA precursors. Annu. Rev. Biochem. 55, 1119–1150.
Qiu, W.G., Schisler, N., Stoltzfus, A., 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol. Biol. Evol. 21, 1252–1263.
Roy, S.W., Gilbert, W., 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat. Rev. Genet. 7, 211–221.
Schmit, F., Cremer, S., Gaubatz, S., 2009. LIN54 is an essential core subunit of the DREAM/LINC complex that binds to the cdc2 promoter in a sequence-specific manner. FEBS J. 276, 5703–5716.
Shapiro, M.B., Senapathy, P., 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 15, 7155–7174.
Sinha, R., Nikolajewa, S., Szafranski, K., Hiller, M., Jahn, N., Huse, K., Platzer, M., Backofen, R., 2009. Accurate prediction of NAGNAG alternative splicing. Nucleic Acids Res. 37, 3569–3579.
Zhang, M.Q., 1998. Statistical features of human exons and their flanking regions. Hum. Mol. Genet. 7, 919–932.
Zhuo, D., Madden, R., Elela, S.A., Chabot, B., 2007. Modern origin of numerous alternatively spliced human introns from tandem arrays. Proc. Natl. Acad. Sci. U.S.A. 104, 882–886.