International Journal of Biological Sciences
2009; 5(1):13-19

Research Paper

Using Profiles Based on Nucleotide Hydrophobicity to Define Essential Regions for Splicing

Galina Boldina[1] ✉, Anatoly Ivashchenko[1], Mireille Régnier[2]

1. The Kazakh National University named after al-Farabi, av. al-Farabi, build. 71, 050038 Almaty, Kazakhstan
2. INRIA (Institut National de Recherche en Informatique et en Automatique), BP 105, Le Chesnay, France

✉ Correspondence to: Galina Boldina, 2, rue Robert Escarpit, 33607 PESSAC Cedex Institut Européen de Chimie et Biologie. Phone +33622259648; e-mail: g.boldina@iecb.u-bordeaux.fr.

Received: 2008.03.17; Accepted: 2008.12.01; Published: 2008.12.03

The splice-site sequences of U2-type introns are highly degenerate, so many different sequences can function as U2-type splice sites. Using our new profiles based on hydrophobicity properties we pointed out specific properties for regions surrounding splice sites. We built a set *T* of flanking regions of genes with 1-3 introns from 21st and 22nd chromosomes extracted from *GenBank* to define positions having conserved properties, namely hydrophobicity, that are potentially essential for recognition by spliceosome.

GT–AG introns exist in U2 and U12-types. Therefore, intron type cannot be simply determined by the dinucleotide termini. We attempted to distinguish U2 and U12-types introns with help of hydrophobicity profiles on sets of spice sites for U2 or U12-type introns extracted from *SpliceRack* database. The positions given by our method, which may be important for recognition by spliceosome, were compared to the nucleotide consensus provided by a classical method, *Pictogram*. We showed that there is a similarity of hydrophobicity profiles inside intron types. On one hand, GT–AG and GC–AG introns belonging to U2-type have resembling hydrophobicity profiles as well as AT–AC and GT–AG introns belonging to U12-type. On the other hand, hydrophobicity profiles of U2 and U12-types GT–AG introns are completely different. We suggest that hydrophobicity profiles facilitate definition of intron type, distinguishing U2 and U12 intron types and can be used to build programs to search splice site and to evaluate their strength.

Therefore, our study proves that hydrophobicity profiles are informative method providing insights into mechanisms of splice sites recognition.

Key words: hydrophobicity, splice sites, U2-type introns, U12-type introns, Pictograms, *P*-value, consensus sequences, splice sites recognition.

Introduction

Pre-mRNA splicing is a nuclear process that is conserved across eukaryotes [1]. The spliceosome recognizes conserved sequences at the exon–intron boundaries, namely the 5' splice site (5'ss) and the 3' splice site (3'ss). In addition, a third conserved intronic sequence that is known to be functionally important in splicing is the so-called branch point site (BPS) which is usually located very close to the end of the intron, at most 40 nucleotides before the terminal dinucleotide. The splicing mechanism involves the following steps: cleavage at 5′ splice site, nucleolytic attack of the splice donor site at the invariant A of the branch site to form a lariat-shaped structure, cleavage at 3′ splice site, leading to a release of the intronic RNA as a lariat, and ligation of the exons. The above reactions are mediated by a large RNA-protein complex, the spliceosome, which consists of five types of snRNA (small nuclear RNA) and more than 200 proteins [2].

There are at least two classes of pre-mRNA introns, based on the splicing machineries that catalyze the reaction. U2 snRNP-dependent introns make up the majority of all introns and are excised by spliceosomes containing the U1, U2, U4, U5 and U6 snRNPs. These introns consist of three subtypes, according to their terminal dinucleotides: GT–AG,

GC–AG and AT–AC introns. U12 snRNP-dependent introns are the minor class of introns and are excised by spliceosomes containing U11, U12, U4atac, U6atac and U5 snRNPs. The overall similarity in the predicted secondary structure between analogous U2 and U12-type snRNAs suggests that the spliceosome rearrangements during catalysis are conserved between the two spliceosomes [3-4]. U12 introns mainly consist of two subtypes, as defined by their terminal dinucleotides: AT–AC and GT–AG introns. In addition, a small fraction of the U12-type introns exhibit other terminal dinucleotides [5-7]. Whereas U2-type introns have been found in virtually all eukaryotes [1] and comprise the vast majority of the splice sites found in any organism, U12-type introns have only been identified in vertebrates, insects, jellyfish and plants [8].

Moreover, the U12-type of introns is characterized by highly conserved consensus sequences at the donor and branch sites [9]. Therefore, Burge et al. [8] designed a computer program, named *U12Scan*, to identify U12-type introns based on search of conserved motives.

Later, Levine and Durbin applied resembling methods to recognize human U12-type introns [6]. The U12-type introns in the human genome were predicted first, and then confirmed by expressed sequence data from dbEST downloaded from the NCBI (available from http://www.ncbi.nlm.nih.gov/Genbank/). However, both methods have some limitations: any U12-type introns having DistBAs shorter than *8nt* or longer than *21nt* would not be located. In addition, the later approach would not identify any U12-type introns if they are flanked by exons shorter than *32bp*. Therefore, a universal approach to recognized a splice site and distinguish its type has not been built yet. Partly, it can be explained by the fact that the U2-type splicing signals have highly degenerate sequence motifs; many different sequences can function as U2-type splice sites. It is not clear how degenerate sequences at the splice sites of U2 intron types are recognized by spliceosomal complex. Clearly, if terminal dinucleotides were sufficient to be recognized, then GT-AG introns of U2 and U12-type introns would be excised by both U2 and U12 spliceosomes, however nothing of this kind has been described. Findings of Shellenberg et al. in 2006 [10] indicate an intimate association between protein p14 and RNA at the heart of mammalian spliceosome during the course of splicing. Subsequent experiments showed that this protein crosselinked directly to the BPS. Kauzman [11] showed that the nature of protein–protein and RNA-protein cooperation is based on hydrophobic-hydrophilic interactions.

Research Purpose

We attempt to use our approach based on hydrophobicity evaluation to find out regions with conserved properties, namely hydrophobicity, which must be functionally important to be recognized by spliceosomal machinery. Hydrophobic-hydrophilic properties of amino acids and nucleotides may be predicted on the base of half-empirical coefficients of hydrophobicity. [12]. We expect profiles based on hydrophobicity properties pave the road towards general understanding of recognition mechanisms of exon-intron junctions by spliceosome. Also we expect to use hydrophobicity profiles to distinguish U2 and U12-types of introns.

Methods and Programs

In order to distinguish regions having conserved properties, namely hydrophobicity, we defined a general hydrophobicity profile. Regions whose hydrophobicity differs from an average value are expected to be essential for recognition by spliceosome. We first tested this hypothesis on one set *T* then compared out method to a classical one.

Hydrophobicity profile

First, genes of 21st and 22nd chromosomes containing 1 to 3 introns were extracted from *GenBank* (http://www.ncbi.nlm.nih.gov). This yielded 313 introns of length greater than *200 nt* and ones of 385 exons of length greater than *60 nt*. In order to compute an average hydrophobicity value, we determined the background frequency of the bases in exons and introns separately. We counted the bases distributions along the full length of exons and introns. There is no statistically confirmed difference between the two ones. Base frequencies were found to be $q(A) = 0.237$, $q(C) = 0.283$, $q(G) = 0.276$ and $q(U) = 0.204$. For introns of lengths greater than 200, we counted the base distribution at the 5′ boundaries only. Indeed, it is well known that a polypyrimidine track exists near the 3′ boundaries of introns that does not allow taking them into account. Base frequencies were found to be $q(A) = 0.210$, $q(C) = 0.249$, $q(G) = 0.299$ and $q(U) = 0.241$.

To build hydrophobicity profile we built the set T of flanking sequences (*30 nt* within the exon and *30 nt* within the intron) extracted at both exon–intron junctions, at 5'ss and 3'ss boundaries. Given a set of splice sites, one defines a hydrophobicity profile as follows. *A* hydrophobicity coefficient $hc(i)$ may be associated to each base i . DNA basic hydrophobicity coefficients given in [12] have been used to compute

an average hydrophobicity value for each splice site position. DNA basic coefficients were suggested appropriate to be extrapolated to ribonucleotides since nucleotide hydrophobicity is mostly determined by the "bases" moieties varied in size, shape and polarity not by the hydrocarbon component. Values of coefficients are *hc(A)=-1.07, hc(C)=-0.76, hc(G)=-1.36* and *hc(U)=-0.76*. As hydrophobicity coefficient for uracil was not determined experimentally in [12], we used the hydrophobicity value for cytosine (-0.76) for uracil because both ribonucleotides belong to pyrimidines and correspondingly have resembling size and shape correlating quite well with physical properties [13]. However, there is NH_2 group at the fourth position of cytosine basis and oxygen at the same position of uracil. This difference is graded by close hydrophobicity values of these groups [13].

For each position *j* and each base *i*, let $n_i(j)$ be the number of occurrences of base *i* at position *j* in this set. The average hydrophobicity value at position *j* is

$$h(j) = \text{sum}\_\{i \text{ in } \{A,C,G,U\}\} \; n_i(j) hc(i).$$

For our set *T* we aligned all flanking sequences of the 5′ boundaries, that are *60 nt* long, using terminal dinucleotides as an anchor. The hydrophobicity values were computed for each position of splice site set. We proceeded similarly for 3' boundaries of the set *T* as well.

Then, we compared this result to the background. Hydrophobicity value at a given position can be interpreted as the sum of *k* random variables. We compared it to the sum of *k* random variables with a Bernoulli distribution given by the background frequencies. The associated distribution is approximately normal. We computed the average background value, *E*= sum _{i in {A,C,G,U}} q(i)*hc(i)* that is *-0.999* for exons and *-1.003* for introns. The corresponding variances, e.g sum _{i in {A,C,G,U}} q(i)*hc(i)*^2 -*E*^2 were found to be $V_E = 0.0649$ and $V_I = 0.0689$. We computed the limits of 0.9995 of confidence interval, that we denoted

$$hp = E + Z(0.9995)\, sqrt(V/k), \text{ with } Z(0.9995) = 3.2905.$$

Tables of normal distribution give "*Z*(1-*alpha*/2)" for a confidence level alpha (=0.001 here):

$$P(Z>=Z(1-alpha/2))=alpha/2.$$

When *h(i)* was out of this range, we used the large deviation formula [14] to compute the *P*-value of *h(i)*.

Consensus sequences and hydrophobicity profiles for two types of introns

We will compare two methods that identify common and distinguishing features in each splice-site type. We built four sets of 200 introns with two confirmed splice sites that were extracted from a user-friendly resource, *SpliceRack* (http://katahdin.cshl.edu:9331/SpliceRack/). Two sets are associated to human U2-type introns, with GT–AG termini for one set and GC–AG termini for the other. Two sets are associated to human U12-type introns with GT–AG termini for one set and AT–AC termini for the other. On the set *T* it was shown that in exons only the nearest positions to exon–intron junction deviate from an average hydrophobicity value, since that for each site one extracts a *38 nt* long region around each exon–intron junction at the 5'ss and at the 3'ss. More precisely, we extracted *8 nt* within the exon and *30 nt* within the intron. These new sets are spread out on several chromosomes; therefore we use a different background. The background frequencies of nucleotides were evaluated from a set of 1228 human genes extracted from *GenBank (*http://www.ncbi.nlm.nih.gov*)* and were found to be *q(A) = 0.219, q(C) = 0.269, q(G) = 0.280* and *q(U) = 0.233*.

An average hydrophobicity value and a variance were *E = 0.996* and *V = 0.0652* correspondingly**.** Here we did not compute the average hydrophobicity values separately for exons and introns since they were not significantly different as it was shown in calculations on the set *T*.

The splice sites of each set were separately aligned using *Pictogram* program designed by Chris Burge and Frank White (http://genes.mit.edu/pictogram.html*)*. *Pictogram* is a handy tool to visualize consensus sequences. Its input is an array of sequences of equal length. For each position *j*, it computes for each base *i* its relative frequency, e.g. the ratio $p_i(j)/q_i$ of the frequency of each nucleotide $p_i(j)$ to the background frequency q_i. It also computes the information content that is defined as $\sum p_i(j) \; log \; p_i(j)/q_i$. Information content is commonly used to study the variability (or level of conservation) at each position, varying from 2 (one event is certain) to 0 (all observed frequencies are equal to background probabilities. *Pictogram* output is a diagram of letters. It must be said that *Pictogram* program replaces U to T to plot output diagrams (Figure 2 and 3) thereupon to be consistent with pictograms we used GT-AG and AT-AC designations for intron terminal dinucleotides. For each position *j*, the height of *i*-th letter is proportional to the ratio $p_i(j)/q_i$. Additionally, information content is written below.

Our method based on hydrophobicity evaluation allowed pointing out regions which may be essential for recognition by spliceosome since they have conserved properties, namely hydrophobicity. We

built hydrophobicity profiles for four sets of splice sites of both U2 and U12-type introns and compared obtained results to the results derived by *Pictogram* method.

Results and Discussion

Distinguishing positions with conserved properties, namely hydrophobicity

Sliding along all regions of our set *T*, we computed an average hydrophobicity value for each position. Hydrophobicity profile is illustrated in Figure 1. The terminal dinucleotides of the introns are marked in red. *P*-values were computed for all positions of exons and introns using our large deviation formula (Supplementary material).

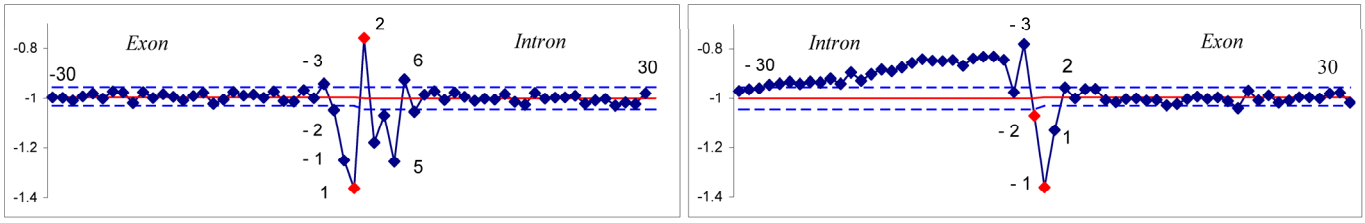
The hydrophobicity profiles of the donor (left) and the acceptor (right) sites are shown.

The positions of nucleotides are marked on the *x*-axis and hydrophobicity values are indicated by the scale on *y*-axis. Average hydrophobicity values are marked by red line. Limits of 0.9995 confidence intervals are given by blue dotted lines.

An approximately normal distribution of hydrophobicity values is observed at positions -30 to -3 and +8 to +30 at the 5′ss as well as at positions +2 to +30 at the 3′ss.

Regions at the positions -2 to +6 at the 5′ss and -26 to +1 at the 3′ss deviate from the background hydrophobicity with significant *P*-values (Supplementary material). Slow decay at positions -26 to -5 due to pyrimidine abundance correspond to polypyrimidine track. General hydrophobicity profile of splice sites of genes with 1- 3 introns from 21st and 22nd chromosomes resembles to the U2-type introns hydrophobicity profile (Figure 2 A, E and 2 B, F), because of the low proportion of U12-type introns that does not exceed 0,34% [6].

Figure 1. The hydrophobicity profiles of the 5′and 3′ splice sites from the set *T*.



Consensus sequences and hydrophobicity profiles for two types of introns

In this section, we report results concerning a comparison of two methods identifying the variability of each splice site position by estimation of nucleotide consensus (method *Pictograms* by Chris Burge and Frank White) and conservation of properties, namely hydrophobicity in order to find out potential binding sites of spliceosomal particles. In Figures 2 and 3, the hydrophobicity profiles of U2 and U12- dependent introns with different termini and corresponding pictograms are depicted and may be compared. For all pictures of hydrophobicity profiles, the numbers of nucleotides are marked on the *x*-axis and hydrophobicity values are indicated by the scale on *y*-axis. The terminal dinucleotides of the introns are marked in red. Average hydrophobicity values are marked by red line. Limits of 0.9995 confidence intervals are given by blue dotted lines. In pictograms the nucleotide consensus is displayed by the letter size that indicates the base frequency distribution at each position of the splice sites.

U2-type introns

200 splice sites with GT-AG and GC-AG terminal dinucleotides extracted from *SpliceRack* were used to construct each pictogram. GT–AG and GC–AG subtypes of U2-type introns are considered separately in order to compare our results to the ones obtained by classical methods.

The hydrophobicity profiles and corresponding pictograms of the donor (left) and the acceptor (right) sites of two subtypes U2-type introns with GT-AG (Figures 2a-d) and GC-AG (Figures 2e-h) terminal dinucleotides are shown for pairwise comparison.

Although the pictograms of GT–AG and GC–AG subtypes are significantly different (Figure 2c, d, g, h), the hydrophobicity profiles are quite similar. Indeed, nucleotide consensus at the 5′ss of U2-type introns mainly contain quite hydrophilic purines when termini are either GT-AG (Figure 2a) or and GC-AG (Figure 2e).

The consensus sequence for 5'ss motif of U2-type GT–AG introns corresponds to a perfect base pairing to the U1 snRNA 5' end, in spite of more conservation in the exonic nucleotides for 5'ss of U2 GC–AG introns, in comparison with the 5'ss of U2 GC–AG in-

trons. In GC–AG introns, the substitution at position +2 of the 5'ss introduces a mismatch in the U1: 5'ss helix. The identity of hydrophobicity coefficients of T and C that are both hydrophobic pyrimidines apparently compensates the mismatch at position +2 of U2 GC-AG introns. As it is seen from the pictograms 2 C nucleotide content is highly degenerate at the 5′ss of U2 GT-AG introns. Indeed, both G and A nucleotides can be found at its third intronic position. We suggest that they are competent for splicing since both belong to purines having small values of hydrophobicity coefficients.

Pictograms method, that attempts to point out a nucleotide consensus, does not show high nucleotide conservation at the 3′ss among U2 GT–AG (Figure 2d) and GC–AG (Figure 2h). Moreover, the polypyrimidine track (PPT) of U2-type GT–AG is C-rich while one of U2-type GC-AG is T- rich.

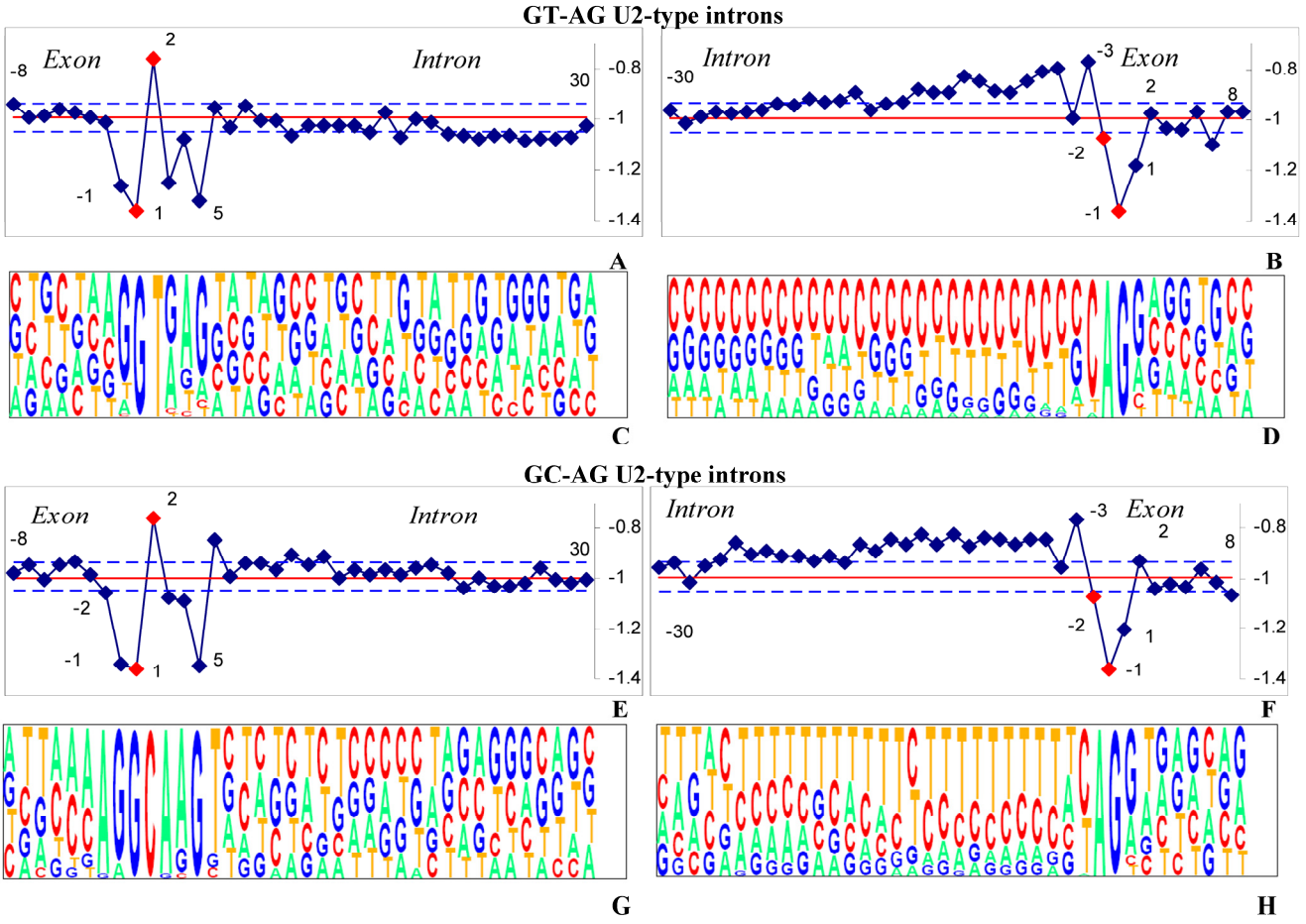


Figure 2. The hydrophobicity profiles and corresponding pictograms for pairwise comparison of the U2-type introns.

We are particularly interested to reveal mechanisms of the competence of both U2-subtypes with inconsistent 3′ss for splicing machinery. We attempt to point out a consensus of properties, namely hydrophobicity. We show that the 3′ss of both subtypes of U2-type introns have resembling hydrophobicity profiles (Figure 2b, f). Intronic nucleotides of both subtypes of U2-type introns are enriched in pyrimidines and as a result hydrophobic. Therefore, our results indicate that the mechanism of functional regions definition is probably based on recognition of conserved features, but not only on nucleotide base pairing.

According to pictograms methods the PPT of both U2-type GC-AG and U2-type GT-AG start at the position -30. Using *P*-values we reveal that the large *P*-values corresponding to poly-pyrimidine track only start at the position -21 in intones of U2-type GT–AG and at the position -26 in introns of U2-type GC-AG (Supplementary material).

U12-type introns

The hydrophobicity profiles and corresponding pictograms of the donor (left) and acceptor (right) sites of two subtypes U12-type introns with AT-AC (Figures 3a-d) and GT-AG (Figures 3e-h) termini are shown for pairwise comparison.

The pictograms of the 5'ss of U12-type introns (Figure 3c, g) show a high degree of conservation and the large *P*-values (Supplementary material) at intronic positions +1 to +9. This sequence consensus is enriched in hydrophobic pyrimidines that are a perfect splicing signal intensified by the region, lying to the right of position +10, with hydrophobicity values close to the average ones.

Quite high degree of conservation of hydrophobicity was found by our method for the BPS of U12-type introns joined to the 3′ of introns. Indeed, the BPS is rich of pirimidines for both U12-type GT-AG and AT-AC. However, we reveal inconstancy of nucleotide content at positions -14 to -3 at the 3′ss for U12-type AT-AC and GT-AG as it shown in pictograms (Figure 3d and h).

In order to explain recognition mechanisms of the U2 and the U12 – dependent introns with the same terminal dinucleotides we compared the hydrophobicity profiles of the U2 and the U12 – dependent introns (Figures 2a, 3e and 2b, 3f). The 5'ss profile for U12-type GT-AG is different from the U2-type GT-AG. Indeed, exonic and intronic nucleotides of the 5′ss of U2 – dependent introns have hydrophilic purine rich consequence, while 5′ss of U12-dependent introns match hydrophobic pyrimidine rich canonical ATCCTTT consensus (plateau in Figures 3a and e) that succeeds terminal dinucleotides.

The 3'ss profile for U12-type GT-AG is different from the U2-type GT-AG. U12-type introns lack obvious PPT at the 3′ss and the BPS lies close to the 3′ end of introns (Figures 2b, 3f and 2d, 3h).

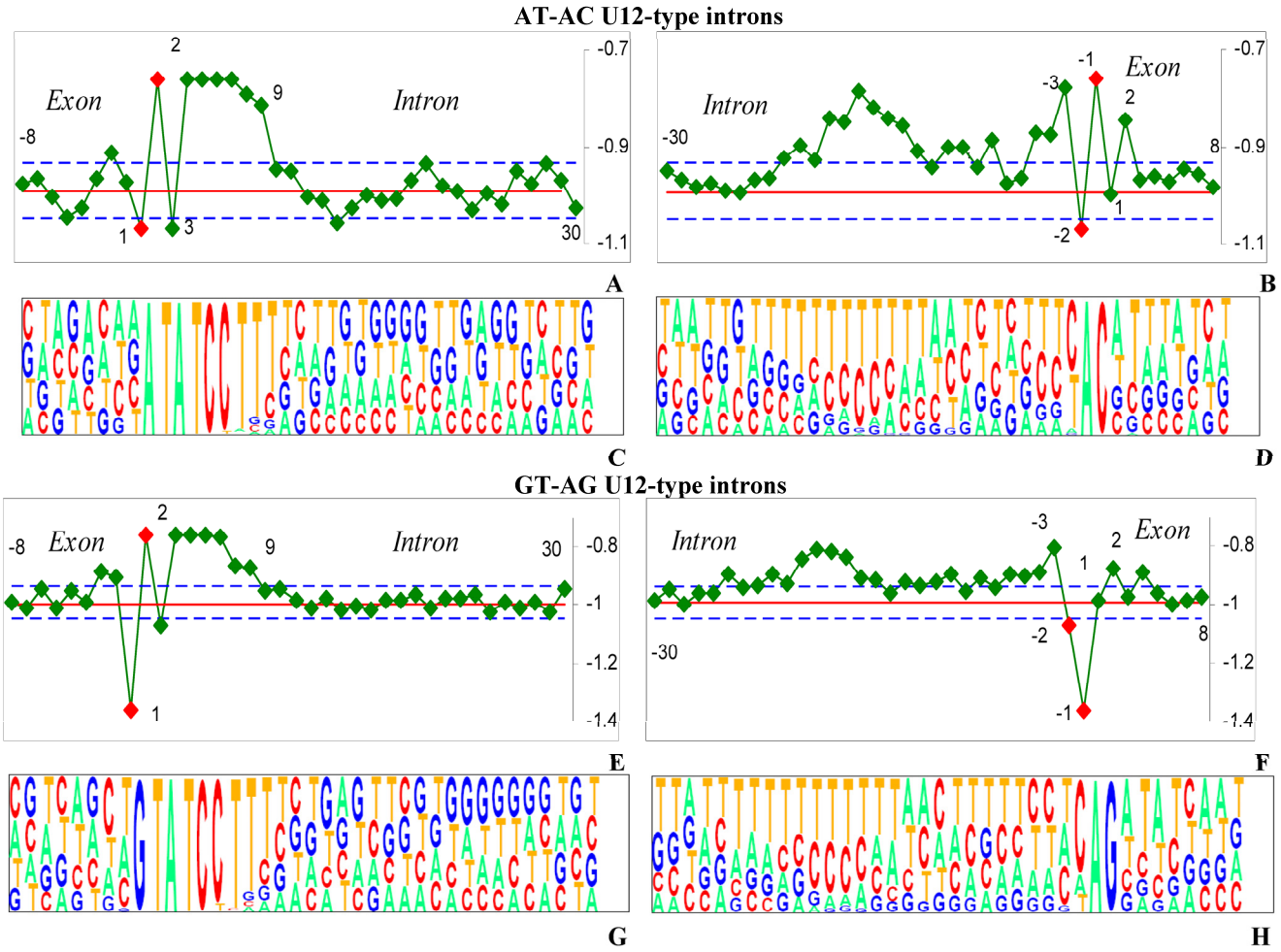


Figure 3. The hydrophobicity profiles and corresponding pictograms for pairwise comparison of the U12-type introns.

We suggest our findings partly explain accurate distinguishing of U2 and U12– types GT-AG by U2 and U12 spliceosomes.

Limitation of hydrophobicity profiles method

The method of hydrophobicity profiles was suggested to visualize the difference between two intron types and to show similarity of properties, namely hydrophobicity inside intron types.

Consequently, for estimation of nucleotide consensus *Pictogram* method is more suitable.

Indeed, when a strong consensus with value A is found (see Figures 3a, b and f) our large deviation formula does not give large P-values because hydrophobicity value for A is close to the average hydrophobicity value. Nevertheless, if we increase the number (n) of sequences (n), the hydrophobicity values slowly depart from the average. We plan to improve our P-value approach to deal with this case. Therefore, the methods of hydrophobicity profiles and *Pictograms* are complementary.

Conclusion

We show that there is a similarity of hydrophobicity profiles inside intron types. On the one hand, GT–AG and GC–AG introns belonging to U2-type have resembling hydrophobicity profiles as well as AT–AC and GT–AG introns belonging to U12-type. On the other hand, hydrophobicity profiles of U2 and U12-types GT–AG introns are completely different. Our analysis should be a step forward for a general understanding of recognition of regions, which are essential for splicing, by spliceosome and for a distinction of U2 and U12-types of introns. We suggest that hydrophobicity profiles facilitate definition of intron type, distinguishing U2 and U12 intron types and can be used to build programs to search splice site and to evaluate their strength.

Therefore, our study proves that hydrophobicity profiles are informative method providing insights into mechanisms of splice sites recognition.

Supplementary Material

Supplementary Material [http://www.biolsci.org/v05p0013s1.pdf]

Conflict of Interest

The authors have declared that no conflict of interest exists.

References

1. Collins L and Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 2005; 22: 1053–1066.
2. Staley JP, Guthrie C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell. 1998; 92(3): 315-26.
3. Tarn WY and Steitz JA. Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. Trends Biochem Sci. 1997; 22: 132–137.
4. Frilander MJ and Steitz JA. Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome. Mol Cell. 2001; 7: 217–226.
5. Wu Q and Krainer AR. Splicing of a divergent subclass of AT-AC introns require the major spliceosomal snRNAs. RNA. 1997; 3: 586–601.
6. Levine A and Durbin RA. Computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. 2001; 29: 4006–4013.
7. Dietrich RC, Fuller JD, Padgett RA. A mutational analysis of U12-dependent splice site dinucleotides. RNA. 2005; 11: 1430–1440.
8. Burge CB, Padgett RA, Sharp PA. Evolutionary fates and origins of U12-type introns. Mol Cell 1998; 2: 773–785.
9. Zhu W and Brendel V. Identification, characterization and molecular phylogeny of U12-type introns in the Arabidopsis thaliana genome. Nucleic Acids Res. 2003; 1: 4561-4572.
10. Shellenberg M, Edwards RA, Ritchie D, , et al. Crystal structure of a core spliceosomal protein interface. PNAS. 2006; 103: 1266-1271.
11. Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem. 1959; 15: 1-63.
12. Guckian KM, Schweitzer BA, Ren R, Sheils CJ, Tahmassebi DC, Kool ET. Factors contributing to aromatic stacking in water: evaluation in context of DNA. J Am Chem Soc. 2000; 122: 2213-2222.
13. Leo P, Hansch C, Elkins A. Partition coefficients and their uses. Chem Reviews. 1971, 71(6): 526 – 616.
14. Régnier M, Vandenbogaert M. Comparison of statistical significance criteria. Journal of Bioinformatics and Computational Biology. 2006; 4: 537–551.