**PRE-MRNA SPLICING IN EUKARYOTES - INTRON STRUCTURE, INTRON DETECTION, ALGORITHMS AND DATA STRUCTURES.**

CHEKMENEV D.S.

GNII genetika, 1st Dorozhny proezd, Moscow, 113545, Russia;
e-mail: chicha@mail.cir.ru;

Keywords: pre-mRNA, splicing, gene expression, intron structure, intron detection, RNA secondary structure

**Introduction**

More than 20 years have passed since the discovery of the pre-mRNA splicing. Many components of spliceosome and snRNP moieties interactions were identified, but intron/exon detection algorithms have not reached accuracy of splicing machinery. In this abstract I sugest the way to improve intron detection methods with respect to identified pathways of splicing process.

**1. PRE-mRNA splicing**

***1.1. Role of splicing in gene expression***

Genes in eukaryotes are often interrupted by intervening sequences (IVSs or introns) that must be removed during gene expression. RNA splicing is the process by which these intrervening sequences are precisely removed and the flanking, functional sequences (exons) are joined together [1, 2]. So RNA splicing as significant part of pre-mRNA processing is one of the major steps in the control of gene expression in eukaryotes.

Regulated mechanism of alternative splicing allow multiple different proteins to be translated from the single RNA transcript. By alternative splicing, a single sequence has been found to be able to code for dozen different proteins, depending on how its exons are assembled. Alternative splicing is regulated in a developmental or tissue specific manner. For example, a gene in thyroid tissue produces calcitonin in rats. The same gene in brain tissue produces a neuropeptide by using a different exon combination.

Mutations can affect splicing of certain introns, leading to abnormal conditions. For example a form of thalassemia, a blood disorder, is due to a mutation causing splicing failure of an intron in a globin transcript, which then becomes untranslatable. Abnormal beta-amyloid in Alzheimer disease is result of intron mutation that impair splicing.

So elucidation of splicing mechanism will help us to find new ways in genetic diseases treatment and better understanding of genetic information organisation and gene expression.

***1.2. Splicing mechanism***

Splicing of nuclear introns occurrs by a two step pathway. In the first step, the phosphodiester bond at the 5’ splice site is attacked by the 2’-OH of an adenosine residue in the intron, the branch point. This reaction produces a free upstream exon and a lariat intermediate molecule containing both the downstream exon and the intron with its 5’ end covalently linked to the branch nucleotide. During the second cleavage-ligation step, the 3’ hydroxyl of the 5’ exon attacks the phosphate at the 3’ splice site. This results in the ligation of the two exons and the release of the intron in lariat form [1, 2, 3].

Removal of introns is catalysed by a large ribonucleoprotein complex called the spliceosome, which consists of four small nuclear ribonucleoprotein particles (U1, U2, U5, and U4/U6 snRNPs) and auxiliary protein factors [3, 4]. A minor type of AT-AC introns require U11, U12, U5, U4atac and U6atac snRNAs [5, 6].

Before the two steps of splicing, the pre-mRNA has to be assembled into a highly complex ribonucleoprotein structure, the spliceosome. RNA interactions are thought to be central to the splicing process and may play an important role in the catalytic core of the active spliceosome [3, 4]. U1 snRNP interacts with the 5’ splice site [10, 11] and U2 snRNP with the branch site of pre-mRNA [12, 13] both of this interactions involve Watson-Crick base pairing. U1 binds at an early step in spliceosome assembly and commits the pre-mRNA to the splicing pathway [14, 15, 16]. Genetic and biochemical data place U5 in close proximity to the 5’ and 3’ exon sequences [17, 18, 19]. Crosslinking experiments in mammalian [18] and yeast [20] extracts revealed 5’ splice site - conserved domain of U6 (ACAGAG) interaction [21, 22]. The conserved domain of U6 is immediately upstream of a helix formed by base-pairing interactions between U6 and U2 [23, 24]. This helix juxtapose 5’ splice site with the branch point interaction domain of U2 [25]. So active site of spliceosome are formed by Watson-Crick RNA-RNA interaction.

**2. Role of RNA - RNA interactions**

***2.1. snRNA - intron interactions (intron primary structure)***

The hypothesis that, splicing is RNA-catalyzed process mediated by the spliceosomal snRNAs, was galvanized by the observation that Group II self-splicing introns are removed by a two-step chemical pathway that is highly similar if not identical to that which accomplishes nuclear pre-mRNA splicing [7, 8, 9, 26, 27]. Most actual for intron detection is interaction of snRNAs and pre-mRNA (intron). The early U1 snRNA interaction with the 5’ splice site is important to recruit RNA sequences into commitment complexes and pre-spliceosomes [15]. The presence of U1 at 5’ splice site is necessary for binding U2 snRNA to branch point of intron. All these initial intron - snRNA interaction require Watson-Crick basepairing.

As might be expected from the fact that exons must encode diverse sequences, conserved information at the 5’ and 3’ splice sites residues almost completely in the intron [3, 28]. For yeast introns it are /GUauGu for 5’ and YAG/ for 3’ (/ - splice site; upper case - most conserved, lower case - less conserved). Branch point has UACUAACA (branch point adenosine is underlined). Mammalian introns are less conservative especially in branch point [3].

***2.2. Methods applied in splice sites detection***

A common approach to locating sites of all kinds is to search for similarities to ‘consensus sequences’. The method suffers from the fact that individual sites are not usually identical to the consensus, and different positions vary in their importance within the consensus. For example, realy conserved are dinucleotides at 3’ and 5’ splice sites, G at position 5 of intron and branch point adenosine. Other nucleotides are significantly lower conserved. A superior method is to search using a matrix. The matrix contains an element for each posible base at every position within a site. The evaluation of each potential site involves summing the elements that correspond to the sequence at that site. Such matrix can be used to find all sites within some range of similarity (Tables 1 and 2. Method from [Mount S.M. Nucleic Acids Res. **10**, 459 (1982)] was used for calculation).

Table 1. Matrix to find 5’ splice sites. Intron begins at position 0.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pos: | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
| A | 5 | 9 | -11 | -35 | -35 | 9 | 10 | -11 | -4 |
| C | 5 | -8 | -15 | -35 | -35 | -24 | -10 | -35 | -7 |
| G | -10 | -8 | 11 | 14 | -35 | 2 | -8 | 12 | -11 |
| T | -12 | -7 | -7 | -35 | 14 | -14 | -8 | -16 | 9 |

Table 2. Matrix to find 3’ splice site Exon begins at position 0.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pos: | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 0 | 1 |
| A | -14 | -5 | -8 | -3 | -7 | -21 | -9 | 0 | -19 | 14 | -35 | -1 | -4 |
| C | 0 | 2 | 3 | 1 | 4 | 4 | 1 | -1 | 9 | -35 | -35 | -3 | -1 |
| G | -9 | -15 | -13 | -11 | -13 | -17 | -17 | 0 | -35 | -35 | 14 | 7 | 0 |
| T | 9 | 7 | 7 | 6 | 6 | 8 | 8 | 2 | 2 | -35 | -35 | -11 | 4 |

Unfortunately there is quite a lot of overlap between the values of real sites and unused sites. There are too many posible sites found by the matrices.

Real algorithms using for finding splice sites are using other information than intron conserved sequences. It is coding potential of exons. Base/position preferences and codon bias also can be taken in consideration for detrmining proper reading frame in exon. But this features hardly reflect real processes that take place in spliceosome operation. Exon mutations do not impair splicing [19].

Probably other information besides the primary sequence is used, such as the secondary structure of the RNA [32].

***2.3. Role of intron secondary structure in splicing***

Conserved secondary structure motifs of intron can play importanat role in intron recognition by spliceosome. This can be supported by possible evolutionary origin of pre-mRNA splicing from self-excised introns of group II. Self-splicing of group II introns mediated only by it secondary structure. A number of secondary structure motifs from group II introns were found in snRNAs. U2-U6 snRNA helix is similar to domain 5 of group II intron [25]. U6 - 5’ splice site helix functioning analogously to epsilon, a sequence that pairs with intron nucleotides near the 5’ splice site of group II self-splicing [29]. The U5 conserved loop can be viewed as the spliceosomal counterpart of the exon binding site (EBS1) of group II introns [17, 30]. In Group II introns, the branchpoint adenosine is found bulged out of a duplex, termed domain 6. Similarly, in nuclear pre-mRNAs, the branchpoint is identified in part through a basepairing interaction with U2 snRNA in which the adenosine nucleophile is bulged out of the U2 - pre-mRNA duplex [31]. I suppose that in pre-mRNA intron can be found secondary structure motifs similar to those in group II introns that will help us significantly improve algorithms of intron detection.

**3. Conserved structures search method requirements**

***3.1. Requirements for conserved structures search methods***

For finding conserved secondary structure motifs some enhancements to the search matrices can be applied.

Conserved RNA secondary structures motives can be treated as RNA double helices hairpins in conserved positions of intron. The existence of such helices can’t be found by analysing of nucleotide positions in RNA sequence.

To identify conserved secondary structure we must add probabilities for nucleotide complementarity to the search matrices. So every element of search matrix must contain not only four probabilities for nucleotide appearance, but also vector of complementarity of current position to other sites of intron. Calculation of such probability vector can take great computational resources, because intron length is high enough. (up to several thousand of nucleotides).

Some restriction can be taken in consideration to reduce computational resources. First, length of hairpin loops can be reduce to dozen of nucleotides. Hairpins with very long internal loops hardly can be formed during pre-mRNA splicing, because it is relatively fast running, almost co-transcriptional event [34]. Second, such conserved structures can be awayted strictly in particular regions of intron. It must be found near splice sites and especially near branch point or in the polypyrimidine tract between branch point and 3’ splice site. Third, search of conserved structures can be fulfiled on the relatively short introns (up to one hundred nucleotides).

***3.2 Requirements for splice site detection method***

The idea of high role of RNA secondary structure contribution to particular RNA/DNA site detection is not only restricted by splice site. RNA secondary structures can also contribute to determination of translation initiation sites and even to transcription initiation and promoter regions [33]. So information on conserved structural motifs must be integrated in methods of particular regions detection and posible in genetic data banks.

**References**

1. M.R. Green, "Biochemical mechanisms of constitutive and regulated pre-mRNA splicing" Annu. Rev. Cell Biol. **7**, 559-599 (1991)
2. M.J. Moore, C.C. Query, P.A. Sharp, "Splicing of precursors to mRNA by the spliceosome" p.p. 303-358 in R.F. Gesteland and J.F. Atkins (ed.) "The RNA World" (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., (1993)
3. H.D. Madhani, C. Guthrie, "Dynamic RNA - RNA interactions in the spliceosome" Annu. Rev. Genet. **28**, 1-26 (1994)
4. J.A. Steitz, D.L. Black, V. Gerke, K.A. Parker, A. Kramer, "Functions of the abundant U-snRNPs" in "Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles", 115-154 (1988)
5. W.-Y. Tarn, J.A. Steitz, "Highly Diverged U4 and U6 Small Nuclear RNAs Required for Splicing Rare AT-AC Introns", Science **273**, 1824-1833 (1996)
6. T.W. Nilsen, "A Parallel Spliceosome", Science **273**, 1813 (1996)
7. P.A. Sharp, "On the Origin of RNA Splicing and Introns", Cell **42**, 397-400 (1985)
8. A.M. Weiner "mRNA Splicing and Autocatalytic Introns: Distant Cousins or the Products of Chemical Determinism?", Cell **72**, 161-164 (1993)
9. M.Belfort, M.E. Reaban, T. Coetzee, J.Z. Dalgaard, "Prokaryotic Introns and Inteins: a Panoply of Form and Function", J. of Bacteriology **177**, 3897-3903 (1995)
10. B. Seraphin, L. Kretzner, M. Rosbash, "A U1 snRNA: pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5’ cleavage site", EMBO J. **7**, 2533-2538 (1988)
11. Y. Zhuang, A.M. Weiner, "A compensatory base change in U1 snRNA suppresses a 5’ splice site mutation", Cell **46**, 827-835 (1986)
12. R. Parker, P.G. Siliciano, C. Guthrie, "Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA", Cell **49**, 229-239 (1987)
13. J.A. Wu, J.L. Manley, "Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing", Genes Dev. **3**, 1553-1561 (1989)
14. P. Legrain, B. Seraphin, M. Rosbash, "Early commitment of yeast pre-mRNA to the spliceosome pathway", Mol. Cell Biol. **8**, 3755-3760 (1988)
15. S.W. Ruby, J.N. Abelson, "An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly", Science **242**, 1028-1035 (1988)
16. B. Seraphin, M. Rosbash, "Identification of functional U1 snRNA - pre-mRNA complexes committed to spliceosome assembly and splicing", Cell **59**, 349-358 (1989)
17. A.J.Newman, C. Norman, "U5 snRNA interacts with exon sequences at 5’ and 3’ splice sites", Cell **68**, 743-754 (1992)
18. D.A. Wassarman, J.A. Steitz, "Interactions of small nuclear RNAs with precursor messenger RNA during in vitro splicing", Science **257**, 1918-1925 (1992)
19. J.R. Wyatt, E.J. Sontheimer, J.A. Steitz, "Site-specific cross-linking of mammalian U5 snRNP to the 5’ splice site before the first step of pre-mRNA splicing", Genes Dev. **6**, 2542-2553 (1992)
20. H. Sawa, J.N. Abelson, "Evidence for a base-pairing interaction between U6 small nuclear RNA and 5’ splice site during the splicing reaction in yeast", Proc. Natl. Acaad. Sci. USA **89**, 11269-11273 (1992)
21. S. Kandels-Lewis, B. Seraphin, "Role of U6 snRNA in 5’ splice site selection", Science **262**, 2035-2039 (1993)
22. C.F. Lesser, C. Guthrie, "Mutations in U6 snRNA that alter splice site specificity: implications for the active site", Science **262**, 1982-1988 (1993)
23. J.A. Wu, J.L. Manley, "Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA", Nature **352**, 818-821 (1991)
24. B. Datta, A.M. Weiner, "Genetic evidence for base pairing between U2 and U6 anRNA in mammalian mRNA splicing", Nature **352**, 821-824, (1991)
25. H.D. Madhani, C. Guthrie, "A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome", Cell **71**, 803-817 (1992)
26. C.L. Peebles, P.S. Perlman, K.L. Mecklenburg, M.L. Petrillo, J.H. Tabor, "A self-splicing RNA excises an intron lariat", Cell**44**, 213-223 (1986)
27. R. van der Veen, A.C. Arnbegr, G. van der Horst, L. Bonen, H.F. Tabak, L.A. Grivell, "Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro", Cell **44**, 225-234 (1986)
28. F.E. Penotti, "Human pre-mRNA splicing signals" J. Theor. Biol. **150**, 385-420 (1991)
29. E.J. Sontheimer, J.A. Steitz "The U5 and U6 small nuclear RNAs as active site components of the spliceosome", Science **262**, 1989-1996 (1993)
30. A. Jacquier, N. Jacquesson-Breuleux "Splice site selection and role of the lariat in a group II intron", J. Mol. Biol. **219**, 415-428 (1991)
31. C.C. Query, M.J. Moore, P.A. Sharp "Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model", Genes Dev. **8**, 587-597 (1994)
32. G.D. Stormo "Identifying coding sequences" in "Nucleic acid and protein sequence analysis, a practical approach" (IRL Press, Oxford Washington DC, ed. M.J. Bishop, C.J. Rawlings), 231-258 (1987)
33. M. Gouy "Secondary structure prediction of RNA" in "Nucleic acid and protein sequence analysis, a practical approach" (IRL Press, Oxford Washington DC, ed. M.J. Bishop, C.J. Rawlings), 259-284 (1987)
34. G. Zhang, K.L. Taneja, R.H. Singer, M.R. Green "Localization of pre-mRNA splicing in mammalian nuclei", Nature **372**, 809-812 (1994)