The Effects of Alternative Splicing on Transmembrane Proteins in the Mouse Genome

M.S. Cline, R. Shigeta, R.L. Wheeler, M.A. Siani-Rose, D. Kulp, and A.E. Loraine

Pacific Symposium on Biocomputing 9:17-28(2004)

THE EFFECTS OF ALTERNATIVE SPLICING ON TRANSMEMBRANE PROTEINS IN THE MOUSE GENOME

M. S. CLINE, R. SHIGETA, R. L. WHEELER, M. A. SIANI-ROSE, D. KULP, A. E. LORAINE
Affymetrix Inc., 6550 Vallejo Street, Suite 100
Emeryville, CA, 94608, USA

Alternative splicing is a major source of variety in mammalian mRNAs, yet many questions remain on its downstream effects on protein function. To this end, we assessed the impact of gene structure and splice variation on signal peptide and transmembrane regions in proteins. Transmembrane proteins perform several key functions in cell signaling and transport, with their function tied closely to their transmembrane architecture. Signal peptides and transmembrane regions both provide key information on protein localization. Thus, any modification to such regions will likely alter protein destination and function. We applied TMHMM and SignalP to a nonredundant set of proteins, and assessed the effects of gene structure and alternative splicing on predicted transmembrane and signal peptide regions. These regions were altered by alternative splicing in roughly half of the cases studied. Transmembrane regions are divided by introns slightly less often than expected given gene structure and transmembrane region size. However, the transmembrane regions in single-pass transmembranes are divided substantially less often than expected. This suggests that intron placement might be subject to some evolutionary pressure to preserve function in these signaling proteins. The data described in this paper is available online at http://www.affymetrix.com/community/publications/affymetrix/tmsplice/.

1 Introduction

Attention on alternative splicing has increased. Numerous groups have published analyses estimating alternative splicing frequency [1, 2], and the degree of conservation of splicing patterns [3, 4]. Consequently, alternative splicing is now recognized as a major source of protein diversity in mammals. Yet questions remain on its functional significance [5]. A relation has been observed between intron positions and compact units of protein tertiary structure [6], and we previously observed that alternative splicing altered the pattern of domains and motifs in roughly one third of the genes studied [7]. Here, we focus on protein motifs of distinct structural and functional relevance: signal peptides and transmembrane helices. Thus, we explored the effects of gene structure and splice variation on predictions by TMHMM [8] and SignalP [9].

TMHMM is the prevalent method for identifying putative transmembrane helices in membrane-spanning proteins [10]. These include transporters, channels, and signaling proteins. SignalP is the prevalent method for predicting signal sequences [11]. Signal sequences help to guide secreted proteins into the endoplasmic reticulum, and are frequently present in transmembrane proteins. Because signal sequences and transmembrane regions are easily confused, transmembrane and signal peptide predictors are best used together, with the signal peptide predictor acting as a screen for the transmembrane predictor [10].

By analysis of genomic alignments, we identified the genomic coordinates of a number of proteins, associating a gene structure with the protein sequence. To focus our analysis on splice variation rather than genetic variation, we derived putative protein translations from the genomic sequence. We then applied SignalP and TMHMM to each translated protein, and determined the genomic coordinates of each predicted signal and transmembrane region We compared these genomic coordinates to the gene structures to determine how often intron boundaries avoid transmembrane regions. For perspective, we estimated how often intron boundaries might divide equivalently-sized segments of the same protein, selected at random. Finally, we assessed how often splice variation deletes or alters a signal peptide or transmembrane region of a protein. Because of the significance of these regions, any such alterations will have major consequences in protein localization and function.

2 Methods

2.1 . Gene structures and cDNA organization

We chose the mouse genome for this investigation to build upon and support other investigations underway at our organization. We aligned all of the mouse cDNA sequences from GenBank (release 128) to the mouse genome (Whitehead Institute Center for Genome Research, April 2002) using blat [12]. Of the 55997 sequences that aligned, we explored 13864 that aligned with coverage of at least 90% and a sequence identity of at least 95%; contained CDS annotations; and had no cDNA inserts in alignment of the CDS regions to the genome.

Exon structures and transcript orientation were derived from the alignments as follows. Successive segments of matching sequence were joined if they were 20 bases or closer; otherwise, they were considered introns. MRNA orientation was determined by a weighted calculation on the directions inferred by the labeled GenBank direction, the polyA site and signal evidence on the mRNA, and the dinucleotide splice pairs derived from the genomic alignment.

We dynamically grouped transcripts together by gene according to their exon structure. We considered two transcripts to be from the same gene if they had overlapping genomic coordinates, and shared at least one intron junction. We grouped these transcripts by splice variation as follows: if an intron in one transcript alignment overlapped an exon in another, or if the two transcripts had start or stop codons at different locations, then the transcripts are considered products of different splice variants. Note that this scheme is not perfect: it might miss cases where one transcript is a genuine longer form of another, with additional exons outside the coding region. However, due to limitations in sequencing technology, a cDNA sequence annotated as “full length” might not necessarily represent the full length of the sequence. Consequently, we chose the conservative route, and consider two sequences to be examples of the same splice variant unless there is strong evidence that they are not.

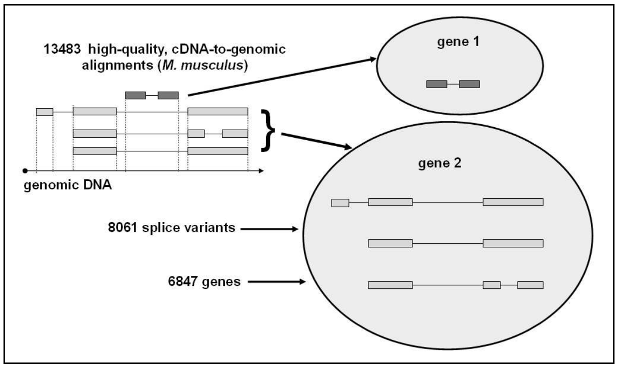
13483 high-quality, cDNA-to-genomic alignments (*M. musculus*)
gene 1
gene 2
genomic DNA
8061 splice variants
6847 genes

Figure 1: Illustration of grouping transcripts by genes and splice variants.

Next, we pruned the gene set to ensure that no UniGene cluster was associated with more than one gene. This step provided a safeguard against bias due to a large population of paralogs. This generated a set of 13483 transcripts of 6847 genes and 8061 splice variants. From each splice variant, we arbitrarily selected one protein for subsequent analysis.

Only 904 genes had multiple variants at the protein level. This should not be regarded as an indication of alternative splicing frequency, as protein-level evidence represents a high evidence standard. A greater degree of alternative splicing can be observed by compiling putative transcripts from cDNA and EST evidence [13], but such transcripts often have no clear protein translation.

2.2 Protein Sequence Analysis

For each cDNA sequence, we derived a protein sequence by assembling an mRNA from the genomic sequence, and inferring a protein translation from its CDS annotation. Note that this protein sequence might differ from the sequence associated with the cDNA, as this scheme does not account for genetic variation. This was deliberate. We chose to focus on splice variation. Other forms of variation, including genetic variation, are outside the scope of this work.

Next, we applied TMHMM [8]and SignalP [9]to the translated proteins, using default parameters for both. From the TMHMM output, we discarded transmembrane regions with a score of 0.3 or less, or those that overlapped with regions predicted as signal peptides. These methods allow identification of three classes of proteins routed through the endoplasmic reticulum (ER). Proteins which have a predicted signal peptide but no predicted cleavage may be routed to the cell surface, but will remain anchored there; these are called Anchor proteins. Those predicted signal peptides with a predicted peptide cleavage may be released into the extracellular environment, and are denoted as secreted proteins. Finally, transmembrane proteins bridge the cell membrane, but are not released into the extracellular environment.

2.3 Genome-level analysis of protein transmembrane regions

Each transmembrane protein region was mapped to genomic coordinates according to the CDS annotations of the associated cDNA and the protein coordinates of the transmembrane region. Each transmembrane region was divided into one more genomic spans, where a genomic span represents the ungapped alignment of a protein segment onto the genomic sequence. If the entire transmembrane region mapped onto one exon, then it had one genomic span; if it was divided by an intron, then it had two genomic spans. For each genomic span, we recorded its start and stop coordinates in the genomic sequence and the protein sequence, and inferred the translation frame from the corresponding CDS region.

Next, we divided the transmembrane regions into two sets: those appearing in all transcripts of a gene, and those not. A region was placed into the first set only if all transcripts contained a region of the same type (signal or transmembrane), with the same genomic coordinates and translation frame.

ProtAnnot, a program designed to allow visualization of protein motifs in the context of genomic sequence, was used to view protein sequence annotations in the context of gene structures [7]. The software is freely available from Affymetrix at http://www.affymetrix.com/analysis/biotools/protannot/index.affx.

3 Results

We applied SignalP and TMHMM to a nonredundant set of 8061 genome-derived protein translations. 1156 proteins contained putative signal peptides, and 1714 contained putative transmembrane segments. Altogether, 2039 of the 8061 proteins contained a transmembrane region of some form.

3.1 Relation between exon boundaries and transmembrane protein regions

Prior evidence suggests some correspondence between modules, compact subunits of protein domains, and intron boundaries [6]. Along those lines, we would expect intron boundaries to typically avoid transmembrane regions. Thus, we assessed how often this is the case. Overall, intron boundaries did not split 695 of 1116 signal peptides (62.3%), 28 of 40 anchor peptides (70.0%), and 3628 of 5895 individual transmembrane regions (61.2%). The transmembrane regions in single-pass transmembranes were divided by introns the least: 687 of 812 (84.6%) were not divided by introns. For seven-transmembrane proteins, 793 of 980 (81.2%) individual transmembrane regions in 120 proteins were not split by introns. This follows the observation that genes encoding GPCRs, in particular, consist of a small number of large exons [14].

To put this into perspective, we estimated the background likelihood of a 22-residue segment being divided by an intron, given observed gene structures and transmembrane topologies. Note that 22 residues is the average length of a region predicted by TMHMM. The likelihood estimation was as follows. For each protein of *n* transmembrane regions, we identified the all positions in the protein corresponding to a splice junction. Then, we selected *n* 22-residue segments at random. If these *n* random 22-mers did not overlap, and were separated by at least five residues (representing a minimal distance for turns between adjacent transmembrane segments), then we noted the number of segments placed *n* and the number *m* of segments that did not span any splice junctions. This process was repeated 100,000 times to sample the protein's conformational space, yielding a total of *N* total segments placed, and *M* not divided by introns. The likelihood *l* of a 22-mer segment being divided by an intron, given the gene structure, was estimated as *M/N*. Finally, the overall likelihood *L* of any 22-residue segment being divided in any *K*-pass transmembrane was estimated as the average likelihood *l* for all *K*-TM proteins analyzed. This data is shown in Figure 2.

In general, the likelihood that transmembrane regions are kept intact is only slightly greater than background. Even the transmembrane regions in 7-TM proteins are kept intact at a rate only slightly higher than expected, even though they are kept intact at a high rate of 81.2%. 7-TM proteins tend to be encoded by genes of few exons. This data indicates that transmembrane regions in 7-TM proteins span introns infrequently because they have few introns, not because introns are placed elsewhere in the gene. For contrast, the single transmembrane region in 1-TM proteins is kept intact at a rate of 84.6%, versus a background expectation of 58.5%. Thus, if there is some selective pressure to keep the transmembrane regions intact in the genomic sequence, this is evidenced to the greatest extent by single-pass transmembranes.

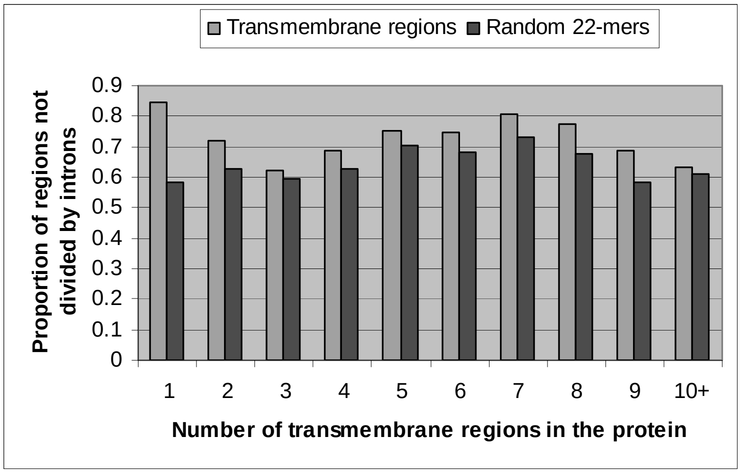

Figure 2: Shown by topology is the proportion of transmembrane regions not divided by an intron. This is compared to the likelihood of a random 22-mer amino acid sequence not being divided by an intron, as estimated by placing the equivalent number of 22-length segments on the protein sequence at random in 100,000 trials per protein. The trend towards increased intact, single exon TM sections for 5, 6, and 7TM proteins correlates well with the prevalence and importance of TM-bound receptors, particularly for the large class of important GPCRs which contain 7TM segments.

3.2 Effects of alternative splicing on transmembrane protein regions

Previously, we analyzed all proteins with a plausible genomic alignment. Here, we analyze only those proteins from 904 genes with protein-level evidence of splice variation. Of the 904 genes, 240 contained some form of transmembrane annotation. These genes yielded a total of 790 annotations in 553 distinct proteins, each representing a distinct splice variant. We divided the annotations into two sets: those common to all observed splice variants, and those not. Annotations were considered common to all splice variants only if all variants contained a region of the same type (signal or transmembrane), produced from the same genomic coordinates and in the same translation frame. Additionally, for an annotation to be common, we required the same class of annotation: the same number of transmembrane spans for a TMHMM prediction, and the same Anchor or Signal classification for SignalP predictions. As shown in Table 1, alternative splicing was associated with changes in transmembrane topology for about half of the genes studied, and about half of the annotations in each class.

Overall, 7-TM regions were altered by alternative splicing at a lower rate than others, although the sample size is too small to suggest a significant trend. We did not observe any general trends, such as whether the variants of a gene tended differ in their their transmembrane span count by multiples of two, a trend which would suggest that the terminal domains of the protein stayed in the same cellular region even if the number of transmembrane spans varied.

Table 1: For each transmembrane architecture, listed are the total examples observed, and the number that differ in some other variant of the same gene. Overall, half of the genes contained splice variants with differing transmembrane architectures.

Topology	***Total***	***Changed***	***Topology***	***Total***	***Changed***
Signal Peptide	145	79	Anchor Peptide	7	4
1-pass TM	128	65	6-pass TM	24	16
2-pass TM	17	15	7-pass TM	16	6
3-pass TM	17	9	8-pass TM	5	5
4-pass TM	15	13	9-pass TM	9	5
5-pass TM	14	10	10+ pass TM	12	7

For all transmembrane proteins, the function of the protein is intrinsically related to the number of transmembrane spans. Yet the effect is most vivid for single-pass transmembranes. There are numerous documented cases of genes with a single-pass transmembrane variant and a secreted variant; both variants contain the same extracellular domain, and the secreted variant inhibits the activity of the transmembrane variant. Two examples include the fibroblast growth factor receptor 1 (FGF-R1) [15] and the neuropilins [16]. Roughly half of the single-pass transmembranes we analyzed contained a variant with no transmembrane region. This data suggests that these cases might not be examples of isolated phenomena, but part of a general trend.

In most cases when the transmembrane architecture was modified, one or more transmembrane region was deleted. Yet in a small number of cases, verified by hand, the genomic coordinates of one transmembrane region were moved in one variant relative to another. Thus, the gene contained transmembrane-coding regions in the exons not constitutively expressed; by selective use of these regions, the splice variants contained the same transmembrane composition. One example of this is MDR/TAP, the multi drug-resistant ATP binding cassette, subfamily B. The splice variants of this gene map to different 5’ exons, suggesting alternative promoters. Yet all variants encode a signal peptide in the 5’ exons. So curiously, the presence of the signal peptide is preserved in splice variation, even at the expense of maintaining two different sets of genomic coordinates. Other genes showing similar behavior include the interferon gamma receptor IFNGR, the poliovirus-receptor-related gene PVR13, and the tyrosine kinase TYR03.

3.3 Case Study1: Alternative splicing of GPCRs

GPCRs typically feature a simple gene structure, comprised of a small number of large exons. Yet even so, they exhibit splice variation. Figure 4 shows the kappa-3 opiate receptor (KOR3) gene explored by Pan et al. [17] In this gene, individual differences in splice variation are believed to have distinct phenotypic consequences. Incomplete cross-tolerance, where patients are highly tolerant of one opiate yet react to a second at surprisingly low doses, is believed to stem from differences in splice variation.

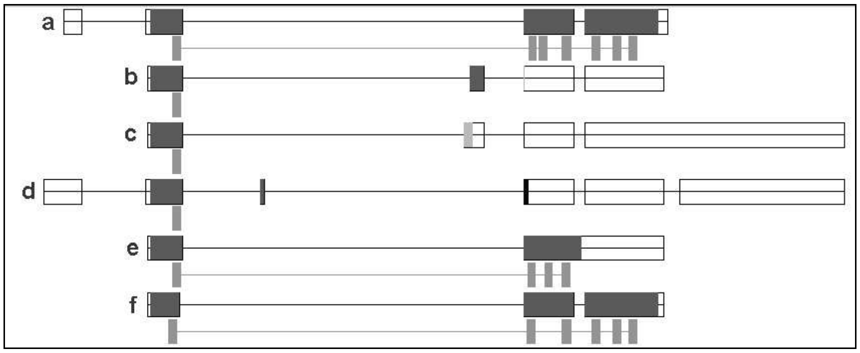

Figure 3: Alternative splicing of the mu opiate GPCR. In this image generated by ProtAnnot, the six splice variants for this gene are labeled with the letters a-f. Empty rectangles represent non-coding exons. Filled rectangles represent translated exons, with the translation frame indicated by the shade of grey. The small rectangles below each transcript indicate the locations of the transmembrane regions.

This gene has several documented splice variants: ordinary 7-TM GPCRs (a); N-terminal anchored 1-TMs (b-d), and 4-TM variants with extracellular C-terminal domains (e) [17]. We observed a 6-TM variant in addition (f). Given the complex interactions between membrane-bound receptors [18], the non-7-TM receptors are not necessarily dead variants, but may be part of the complex interplay between receptors in regulating response to outside influences and neuronal states.

3.4 Case Study 2: Alternative splicing and nonsense-mediated decay

In 30 randomly-selected genes, we found five examples in which alternative splicing caused shifts in the translation frame and introduced premature termination codons (PTCs). Although such events can stem from artifacts in the cDNA library, we emphasize that all five sequences were documented as full-length, with protein translations. The changes in translation frame stemmed from shifts in the exon boundaries, and conditional inclusion or exclusion of cassette exons. Two examples are shown in Figure 4.

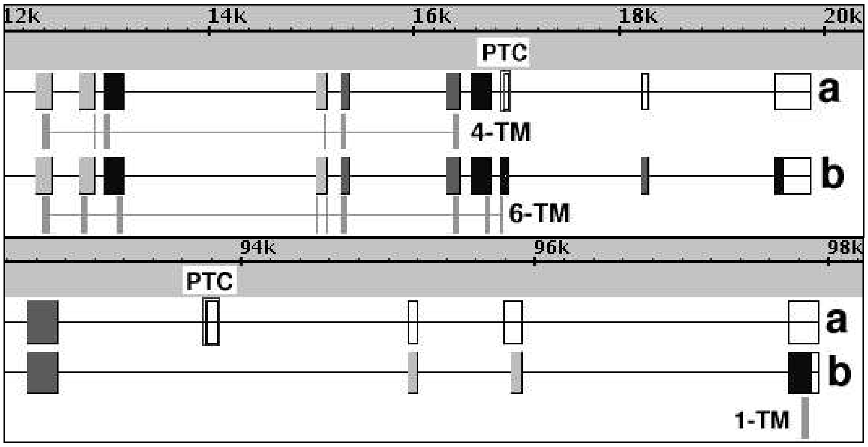

Figure 4. Alternative splicing introduces premature termination codons (PTCs) via two different mechanisms: variable splice site selection and optional inclusion of an alternative exon in. In both examples, the termination codon in one of the transcripts is more than 50 bases upstream of a splice junction, thus exposing it them to regulation by nonsense-mediated decay pathway. (Top) TMC6 encodes putative 4-pass (a) and 6-pass (b) transmembrane membrane-bound proteins are shown. The exon beneath the PTC contains a shorter 5' leg in (a) than in (b), indicating variation in the 5' boundary of the affected exon in the two transcripts. (Bottom) Shown is Chnn (calmin), a putative actin-binding protein**.** Inclusion of an optional exon in (a) introduces a PTC which deletes a downstream single-pass transmembrane region present in (b).

Curiously, many of these PTC-containing variants contained splice junctions downstream of the termination codon. According to current theories, this should target these proteins for nonsense-mediated decay (NMD). After splicing, components of the splicing machinery are thought to remain attached to the mRNA near former splice junctions, marking the positions of former introns [19]. They are usually displaced during translation, but might not detach if the mRNA contains splice boundaries 50 bases or more downstream from the termination codon [20]. Their presence is believed to activate the nonsense-mediated decay pathway, resulting in degradation of the affected molecule.

The effects of NMD vary from gene to gene [21]. Recently, it was proposed as a genome-wide mechanism by which cells ensure splicing fidelity and avoid the production of potentially toxic, nonfunctional proteins [22]. Yet give our results, are all classes of protein-coding transcripts equally susceptible to NMD? We observed 3 examples of NMD-susceptible transmembrane protein encoding transcripts (Tmc6, Clmm, Il17rb) in 30 genes examined. Perhaps mRNAs encoding membrane-spanning proteins, which are co-translationally inserted into the ER, might be subject to NMD to a lesser degree than other proteins.

4 Conclusions

Transmembrane proteins perform a number of key roles, including inter-cellular signaling and transport. Their function is tied closely to their organization of transmembrane spans. Alternative splicing modified this organization in about half of the genes studied, almost certainly altering the functions of the proteins produced. Thus, the process of alternative splicing could have a substantial impact on any cellular processes in which these proteins are involved.

One cannot consider splicing without of gene structure. Associations have been observed between exons and units of protein structure [6]. Given the functional importance of transmembrane regions, plus their short length, we might expect them to be divided by introns rarely. On the surface, this seems true. However, when compared to the likelihood of an intron dividing an equivalently-sized protein segment, we observed that most transmembrane regions were kept intact at a rate barely higher than expected. The exception is the single pass in 1-TM proteins, which are kept intact far more frequently than expected. Few protein regions have such clear functional interpretation as these. There are numerous documented cases of 1-TMs with a secreted splice variant, where the two variants contain the same extracellular domain and the secreted variant inhibits the function of the transmembrane variant. These facts together support the idea of an evolutionary mechanism that avoids fragmentation of critical portions of the protein.

While this work represents a starting point. Here, our interpretation of the results is limited by small data set sizes, resulting from the small amount of cDNA data for the mouse. In future work, we are considering repeating this analysis on other genomes where the cDNA data is more abundant.

Any analysis based on genomic data tells only half of the story. Any cDNA sequence represents a splicing event that has been documented at least once. The trends we reported here based on in-silico observations, but cannot describe the conditions under which such trends arise. Questions remain, such as when alternative splicing events are regulated, and when they represent random consequences of a noisy process. Addressing such questions would require the genomic data to be coupled with the proper measurement technology. In related future work, we hope to shed more light on some of the events described here, and the circumstances under which they occur.

Acknowledgments

We wish to thank several colleagues for insightful discussions on the science of splicing: John Blume, Jing-Shan Hu, Gang Lu, Tyson Clark, Gangwu Mei, Manny Ares, Chuck Sugnet, Bruce Conklin, Nathan Salomonis, and especially Hui Wang. The analysis reported here would have been nearly impossible without the elegant data analysis pipelines developed by Alan Williams, Brant Wong, and Harley Gorrell. Additionally, we wish to thank Harley for his generous assistance with the compute cluster and postgresql databases.

References

1. Modrek, B., et al., *Genome-wide detection of alternative splicing in expressed sequences of human genes.* Nucleic Acids Res, 2001. **29**(13): p. 2850-9.
2. Mironov, A.A., J.W. Fickett, and M.S. Gelfand, *Frequent alternative splicing of human genes.* Genome Res, 1999. **9**(12): p. 1288-93.
3. Nurtdinov, R.N., et al., *Low conservation of alternative splicing patterns in the human and mouse genomes.* Human Molecular Genetics, 2003. **12**(11): p. 1313-20.
4. Thanaraj, T.A., F. Clark, and J. Muilu, *Conservation of human alternative splice events in mouse.* Nucleic Acids Res, 2003. **31**(10): p. 2544-52.
5. Modrek, B. and C. Lee, *A genomic view of alternative splicing.* Nat Genet, 2002. **30**(1): p. 13-9.
6. Fedorov, A., et al., *Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(23): p. 13177-82.
7. Loraine, A., et al. *Protein-based analysis of alternative splicing in the human genome*. in *IEEE Computer Society Bioinformatics Conference*. 2002. Stanford University.
8. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
9. Nielsen, H., S. Brunak, and G. von Heijne, *Machine learning approaches for the prediction of signal peptides and other protein sorting signals.* Protein Engineering, 1999. **12**(1): p. 3-9.
10. Moller, S., M.D. Croning, and R. Apweiler, *Evaluation of methods for the prediction of membrane spanning regions.* Bioinformatics, 2001. **17**(7): p. 646-53.

11. Menne, K.M., H. Hermjakob, and R. Apweiler, *A comparison of signal sequence prediction methods using a test set of signal peptides.* Bioinformatics, 2000. **16**(8): p. 741-2.
12. Kent, W.J., *BLAT-the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
13. Reese, M.G., et al., *Improved splice site detection in Genie.* J Comput Biol, 1997. **4**(3): p. 311-23.
14. Kilpatrick, G.J., et al., *7TM receptors: the splicing on the cake.* Trends in Pharmacological Sciences, 1999. **20**(7): p. 294-301.
15. Kornmann, M., et al., *Expression of the IIIc Variant of FGF Receptor-1 Confers Mitogenic Responsiveness to Heparin and FGF-5 in TAKA-1 Pancreatic Ductal Cells.* International Journal of Gastrointenstinal Cancer, 2001. **29**(2): p. 85-92.
16. Nakamura, F. and Y. Goshima, *Structural and functional relation of neuropilins.* Advances in Experimental Medicine and Biology, 2002. **515**: p. 55-69.
17. Pan, Y.X., *Identification of alternatively spliced variants from opioid receptor genes.* Methods in Molecular Enzymology, 2003. **84**: p. 65-75.
18. Abbadie, C., et al., *Anatomical and functional correlation of the endomorphins with mu opioid receptor splice variants.* European Journal of Neuroscience, 2002. **16**(6): p. 1075-82.
19. Le Hir, H., et al., *The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay.* Embo J, 2001. **20**(17): p. 4987-97.
20. Nagy, E. and L.E. Maquat, *A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.* Trends Biochem Sci, 1998. **23**(6): p. 198-9.
21. Gudikote, J.P. and M.F. Wilkinson, *T-cell receptor sequences that elicit strong down-regulation of premature termination codon-bearing transcripts.* EMBO Journal, 2002. **21**(1-2): p. 125-34.
22. Lewis, B.P., R.E. Green, and S.E. Brenner, *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A, 2003. **100**(1): p. 189-92.