Christopher Lee
is an Associate Professor of Chemistry and Biochemistry at the University of California, Los Angeles.

Qi Wang
is a graduate student in the Interdepartmental PhD programme of the Molecular Biology Institute, at the University of California, Los Angeles.

Bioinformatics analysis of alternative splicing

Christopher Lee and Qi Wang

Date received (in revised form): 6th January 2005

Abstract

Over the past few years, the analysis of alternative splicing using bioinformatics has emerged as an important new field, and has significantly changed our view of genome function. One exciting front has been the analysis of microarray data to measure alternative splicing genome-wide. Pioneering studies of both human and mouse data have produced algorithms for discerning evidence of alternative splicing and clustering genes and samples by their alternative splicing patterns. Moreover, these data indicate the presence of alternative splice forms in up to 80 per cent of human genes. Comparative genomics studies in both mammals and insects have demonstrated that alternative splicing can in some cases be predicted directly from comparisons of genome sequences, based on heightened sequence conservation and exon length. Such studies have also provided new insights into the connection between alternative splicing and a variety of evolutionary processes such as Alu-based exonisation, exon creation and loss. A number of groups have used a combination of bioinformatics, comparative genomics and experimental validation to identify new motifs for splice regulatory factors, analyse the balance of factors that regulate alternative splicing, and propose a new mechanism for regulation based on the interaction of alternative splicing and nonsense-mediated decay. Bioinformatics studies of the functional impact of alternative splicing have revealed a wide range of regulatory mechanisms, from NAGNAG sites that add a single amino acid; to short peptide segments that can play surprisingly complex roles in switching protein conformation and function (as in the Piccolo C2A domain); to events that entirely remove a specific protein interaction domain or membrane anchoring domain. Common to many bioinformatics studies is a new emphasis on graph representations of alternative splicing structures, which have many advantages for analysis.

Keywords: *alternative splicing, microarrays, comparative genomics, graph algorithms, regulation*

Christopher Lee,
Molecular Biology Institute,
Center for Genomics and Proteomics,
Dept. of Chemistry & Biochemistry,
University of California,
Los Angeles, Los Angeles,
CA 90095–1570, USA

Tel: +1 310 825 7374
Fax: +1 310 206 7286
E-mail: leec@mbi.ucla.edu

INTRODUCTION

Over the past few years, the analysis of alternative splicing using bioinformatics has emerged as an important new field, and has significantly changed our view of genome function. Instead of focusing mainly on an organism's total number of genes to explain its functional and behavioural complexity, researchers are now interested in the many ways each gene can be 'reused' to create multiple functions and new modes of regulation. Bioinformatics studies were the first to show that alternative splicing, long considered to be a relatively unusual form of regulation, is actually ubiquitous throughout the human genome and other genomes. Now interest and activity in this field are exploding, through new technologies (eg microarray detection of alternative splicing), new data and perspectives (eg use of multiple genome data to study alternative splicing's evolution) and new computer science (eg analysis based on graph representations). Growth in new publications in the field has been rapid, and new international meetings have emerged focusing specifically on the interface between bioinformatics and the biology of alternative splicing, such as the 2004 Workshop on Alternative Splicing at the Pacific Symposium on Biocomputing (January 2004), the Alternative Transcript Diversity meeting at EBI-Hinxton (November 2004) and the forthcoming Alternative Splicing SIG at the 2005 ISMB meeting (June 2005).

While it would be impossible to survey all the latest work in this field, in this review we will try to summarise new findings from several areas of the field: *detection* using new approaches such as microarrays and comparative genomics; *representation* and analysis of complex alternative splicing patterns; *regulation* of alternative splicing; analysis of *functional impact*, particularly on the proteome; and comparative genomics studies of alternative splicing *evolution*.

MICROARRAY STUDIES OF ALTERNATIVE SPLICING

Analysis of expressed sequence tags (ESTs) was the first step in bringing the power of genomics to the study of alternative splicing, and has played a leading role in the development of the field so far. However, ESTs have many problems. Owing to many possible experimental artefacts and biases, great care in interpreting EST data is required (see Modrek and Lee[1] for a review).

Estimating the extent of alternative splicing using microarrays

Thus, one exciting recent development has been the transition of microarray studies of alternative splicing from the prototype-demonstration stage (in which the sole purpose was to show that microarrays *can* detect alternative splicing), to becoming a tool that researchers use to make discoveries. The work of Johnson *et al.* has been a major step forward, reporting a very large study of exon skipping (custom-designed microarrays based on the Agilent array platform, covering 10,000 human genes and surveying over 50 diverse human tissue samples).[2] Exon skipping events were detected by statistical analysis of the microarray data, and approximately half of these events were validated by independent reverse transcriptase polymerase chain reaction (RT-PCR) experiments. This microarray study was able not only to discover many new alternative splicing events (approximately 800 not previously detected using ESTs), but also to show that microarrays can be free of some of the limitations of EST data, such as their strong bias for the 3′ end of the gene, and inability to detect alternative splicing in regions of poor EST coverage. This paper appeared to give strong confirmation to the high levels of alternative splicing reported by EST studies, reporting that at least 74 per cent of human genes are alternatively spliced. However, this number combined results from ESTs and from the microarray data, with the EST data providing the larger share. Moreover, for genes in which no alternative splicing was detected by ESTs, the microarray analysis indicated exon skipping in only about 20 per cent of the genes. While this is hardly surprising, it leaves a significant range of uncertainty for the true average frequency of alternative splicing.

Another microarray study using Affymetrix arrays has filled this gap, by tiling probes every 35 bp along chromosomes 21 and 22, and profiling RNA samples from 11 different human cell lines.[3] In the absence of alternative splicing, the hybridisation profile for probes within a gene should have a similar shape in different samples. But Kampa *et al.* found strong differences between profiles from different cell lines, indicating the presence of multiple isoforms, in 79–88 per cent of human genes on chromosomes 21 and 22. This provides an independent confirmation of the high level of alternative splicing reported by EST studies and by Johnson *et al.*

One appealing feature of alternative splicing is that it produces a *qualitative* change in the gene product; that is, it is not just a change in the amount of a transcript, but a change in its sequence content. This suggests a general bioinformatics approach for detecting alternative splicing in microarray data, as a qualitative change in hybridisation profiles that can easily be distinguished from changes in gene expression. In simple cases such as a single exon skip, the adjacent constitutive exons provide an easy control for separating changes in gene expression from genuine alternative splicing; this simplification has been critical for the pioneering studies, which

Algorithms for distinguishing alternative splicing from simultaneous changes in gene expression

have focused on exon skipping, eg Johnson *et al.*[2] and Clark *et al.*[4] But how to solve the general case, in which a wide variety of complex alternative splicing events may be combined in a single gene, and it is not necessarily known which exons are truly constitutive? In this case, being able to deconvolute simultaneous changes in gene expression from changes in splicing becomes both harder and more important for reliable detection of alternative splicing by microarrays.

Le *et al.* have described a simple solution to this problem, based on comparing the hybridisation profiles of a gene in two different tissues, each normalised versus a pool combining both tissues.[5] In the absence of alternative splicing, an *XY* scatterplot of the individual probes will show an uncorrelated random scatter (whose radius reflects the level of noise in the array experiments), whose centroid position measures the change in gene expression (specifically its distance from the $y = x$ diagonal of the scatterplot, which represents zero change in gene expression). On the other hand, if the two tissues have different alternative splicing, the scatterplot will display a strong negative correlation. Since the shape of the scatterplot is evaluated independently of the position of its centroid, this cleanly separates detection of alternative splicing as a qualitative phenomenon from any changes in gene expression that may also occur. This procedure was tested using Agilent microarray data for five human tissues, and gave reliable detection of many types of alternative isoforms, including alternative-5′ and alternative-3′ splicing, alternative initiation, alternative termination and exon skipping. It also provides a very simple way for clustering microarray data by similarity of alternative splicing patterns, which was shown to yield distinct (and often simpler) patterns compared with gene expression clustering of the same microarray data.

Clustering genes and samples by their patterns of alternative splicing reveals new groupings

As a demonstration of reliable *quantitative* analysis of alternative splicing using microarrays, Pan *et al.* have reported studies of over 3,000 exon skipping events from mouse, using Agilent microarrays and a panel of ten mouse tissue samples.[6] Employing a Bayesian network approach, they developed a reliable method for calculating the exon inclusion level (percentage of transcripts of a gene that include a specific exon) for each alternatively spliced exon in each sample. They were able to show that this measure was highly reproducible, and closely matched results from RT-PCR (Pearson correlation >0.8). Moreover, the exon inclusion level can also be used to cluster genes and samples according to their similarity of alternative splicing. While this clustering shared many features with a clustering based on gene expression, it also revealed new functional groupings of genes.

Together, these data indicate that two major microarray platforms (Affymetrix and Agilent) can reliably detect alternative splicing, that these approaches can be scaled up to genome-wide studies, and that the resulting data yield a new view of gene functions that is not identical to simple gene-expression based analysis (even when that analysis is applied to the exact same microarray data).

PREDICTING ALTERNATIVELY SPLICED EXONS USING COMPARATIVE GENOMICS

The availability of multiple genome sequences opens new paths for discovering alternative splicing. Kan *et al.* showed that EST and genome data for related organisms such as human and mouse can be combined to obtain predictions of alternative splicing.[7] In such similar organisms, gene sequences and structures are similar enough that human ESTs can be aligned to the mouse genome, and will successfully identify most constitutive exons in mouse, and about half of the known alternatively spliced exons.

Sorek *et al.* have taken this idea much further, by simply using the comparison of human and mouse genomic sequences

Comparative genomics enables direct prediction of alternatively spliced exons

to predict alternatively spliced exons, without any consideration of EST data.[8] They exploited three distinguishing characteristics of alternatively spliced exons: higher sequence identity (between the corresponding human and mouse genomic sequence regions; they found >95 per cent identity to be the best threshold); a stronger tendency for the exon length to be an exact multiple of 3 nt (preserving the protein reading frame); and higher sequence conservation in the flanking introns. Their analysis suggests that it should be possible to detect up to half of true alternatively spliced exons, with an error (false positive) rate of only 25 per cent. RT-PCR testing of a sample of 15 of their predicted exons found that 40–60 per cent were visibly alternatively spliced in a survey of 14 tissue samples. This very exciting prospect has been studied in *Drosophila*, where the availability of genome sequence for *D. melanogaster* and *D. pseudoobscura*, separated by 30 million years, enables a similar prediction scheme.

Philipps *et al.* searched the *D. melanogaster* genome for annotated exons with >95 per cent identity to *D. pseudoobscura* (relative to the average level of identity for annotated exons, which is 79 per cent), with at least 10 bp of flanking intron sequence of at least 75 per cent identity.[9] Of the 162 exons meeting this criteria, 45 were found to be alternatively spliced in available transcript data, and 23 more were visibly alternatively spliced by RT-PCR of mixed *D. melanogaster* RNA. The overall success rate of their predictions (42 per cent) was thus similar to that of Sorek *et al.* It appears that such approaches could open up the study of alternative splicing in many genomes where sufficient EST data have been lacking.

SPLICING GRAPHS, ISOFORMS AND SPLICING PATTERNS

Representing alternative splicing as directed acyclic graphs solves many analysis problems

It has been customary to represent sequences as strings, and to analyse them using string algorithms. When this approach is applied to gene structures, however, it brings a hidden assumption that the alignment of the genomic and transcript sequence data also forms a linear structure. This assumption affects many basic things. For example, it is customary to represent 'the open reading frame' of a gene as an interval composed of a start position (ATG) and end position (stop codon). On this basis it is extremely simple to check whether a given alternative splicing event is 'inside' the protein coding region, by comparing its location to these two positions. However, alternative splicing often changes the end-points of the open reading frame, rendering this assumption incorrect. Alternative splicing can be thought of as introducing *branching* to the alignment – places where there is more than one possible way to splice the gene. This branching structure is not represented well by linear strings. To properly represent the set of possible open reading frames on top of the set of possible alternative splicing events and isoforms, it is best to use a graph representation in which exons are represented by nodes, and splices by directed edges that connect them. In this framework, nearly all questions about the splicing of a gene are answered by simple graph algorithms, typically *graph traversal* (eg 'find the best path from *A* to *B*. . .'). This theoretical foundation for alternative splicing analysis has recently been highlighted by many groups, whose work illustrates the power of this approach.[10–17] Here we will try to compare the work of these groups, to draw out their similarities and differences.

First, these methods share a basic model of how to represent alternative splicing data, as *directed acyclic graphs*. Full-length transcripts correspond to paths through the graph, which begin at a 'start' node (node with no incoming edges) and terminate at an end node (no outgoing edges). All of the above reports use this model except Eyras *et al.*, which we will discuss in a moment. Methods differ in how they construct this basic graph from a large set of EST observations. Clearly, exon observations (from different ESTs) that share the same start site and end site may be merged as a single node representing that exon. Fundamentally, this assumes an alignment of all the ESTs to each other (and to genomic sequence, if available). In all of our work on alternative splicing (beginning with Modrek *et al.*[18]), we have constructed this alignment using partial order alignment, which differs from traditional alignment methods in that it models alignment itself as a directed acyclic graph, to which algorithms such as dynamic programming can be applied very efficiently.[11] Thus the basic partial order alignment of ESTs and genomic sequence is itself a splice graph,[11] generated via dynamic programming alignment of ESTs to the graph itself. By contrast, the splice graphs of Heber *et al.* are constructed as *de Bruijn graphs*, by joining matching 20-mers from different ESTs (with appropriate guards against sequencing error), so that the individual EST strings fuse to form a directed graph like that of partial order alignment. Eyras *et al.* used the program Exonerate to generate the initial alignment,[13] while Lee *et al.* simply used supplied genome annotation as the initial alignment.[15]

'Liberal' *v.* 'conservative' transcript assembly algorithms

One obvious application of splicing graphs is the construction of full-length transcript isoforms from EST fragments, ie EST assembly that explicitly takes into account alternative splicing. Three distinct approaches have emerged. Heber and coworkers generate all possible paths through the splicing graph.[10,16] This ensures that no possible solution is left out but, owing to the explosive growth of combinatorix, enormous numbers of isoforms can be generated (eg more than 5,000 isoforms for a single gene). By contrast, the principle of maximum likelihood can be used to generate a 'minimal' set of isoforms sufficient to explain the actual experimental observations (ESTs). This is implemented via a trivially simple dynamic programming algorithm that traverses the graph by simply following the 'heaviest' edge (edge with the most observational evidence) from each node.[12] This 'heaviest bundling' algorithm can be used both with and without aligned genomic sequence, and has been shown to produce good quality predicted protein isoform sequences.[17]

Eyras *et al.* have adopted a different representation, in which each EST is a node, and the total 'alignment relationship' between each pair of ESTs can be shown as an edge. Specifically, edges are assigned for the two types of sequence relationship that are strictly linear (inclusion and extension), and graph traversals (following the 'extension' directed edges) are sought.[13] These graph traversals again represent full-length isoforms. Like heaviest bundling, Eyras *et al.* generate a 'minimal isoform set', ie the minimum number of traversals that include all of the observations. This minimal set should reflect patterns of coupling that are observed within the EST data (ie if two different alternative splices occur together only in the observations, they should also be joined together in a single generated isoform). This again is like heaviest bundling, but contrary to Heber *et al.* Overall, these methods can be viewed as opposite ends of the traditional ROC (receiver operating characteristic) curve: minimising *false negatives* (by generating all possibilities, in the 'liberal' method of Heber *et al.*), versus minimising *false positives* (by generating only isoforms that are necessary to explain the actual observations, and choosing the candidates according to maximum likelihood).

A very nice use of splice graphs is illustrated by Ranganathan and coworkers, who categorise distinct patterns of alternative splicing (eg exon skipping *v.* alternative-3′ splicing) according to their distinct graph structures.[15] Tools that make such patterns easy to search for in databases, and easy to look at in real data, will make splice graphs a natural way for biologists to think about alternative splicing. A number of such visualisation tools have already been developed,[10,15,16,19] but more work is needed.

REGULATION OF ALTERNATIVE SPLICING

Bioinformatics and experimental validation identify new splicing regulatory motifs

Making the transition from identifying alternative splice forms to fully understanding how each one is regulated is a very difficult project, which for most genes has only just begun. Several stages of this process are evident in recent research. First, many studies have sought to assess the specificity of alternative splice forms across tissues and cellular development, but we will not attempt to review this field here. Second, researchers have used statistical analysis and experimental validation to seek understanding of how different regulation mechanisms are used, and to identify specific motifs and factors that are likely to be involved. Finally, bioinformatics analysis has also been able to suggest genuinely new hypotheses for how gene function can be regulated through alternative splicing.

Alternative exons have weaker splice sites

Recently, many studies have focused on the relative contributions of splice site sequences – the essential recognition sites for the spliceosome – versus splicing enhancer and silencer sites, the binding sites for splicing regulatory factors such as SR proteins. Several groups have studied whether alternative splice sites can be distinguished from constitutive splice sites on the basis of splice site sequence. Roca *et al.*[20] compared the strength of authentic 5′ splice sites with that of cryptic 5′ splice sites that were used only when the authentic splice site was disrupted by mutation, by different scoring methods. All methods gave highest ranking to the authentic sites and lowest to the mutant sites, and intermediate scores for the cryptic sites. However, the cryptic sites observed to be used were not necessarily the best scoring alternative sites in the vicinity of the authentic site, suggesting that other factors besides splice site strength contribute to splice site selection. Itoh *et al.* compared the splice site strength of alternative exons versus that of the constitutive exons and showed that alternative exons generally have weaker splice sites than constitutive exons.[21] Moreover, they also noted that alternative exons displayed stronger conservation (eg between human *v.* mouse orthologous sequences) than did constitutive exons. Based on these data, they suggest that alternative exons have weaker splice sites and a greater requirement for specific splicing enhancer motifs.

On this basis, the search for splicing enhancer and silencer motifs (and their corresponding regulatory factors) is a very important direction for deciphering the regulation of alternative splicing. Burge and colleagues have studied exonic splicing regulatory elements in a combination of computational and experimental approaches. Fairbrother *et al.*[22] predicted exonic splicing enhancers sequences by statistical analysis. They looked for hexameric sequences that were enriched in exons relative to introns and that were enriched in weak splice site flanking regions relative to that of strong splice sites. Then they successfully verified the enhancers' activity experimentally for nine out of ten enhancers they predicted. Use of single nucleotide polymorphism data and comparison with the chimpanzee genome enabled them to confirm that exonic splicing enhancers predicted by their method have been under evolutionary selection pressure in human exons, particularly when they are adjacent to splice sites.[23] They have also developed an *in vivo* splicing reporter system to screen a random decanucleotide library for exonic splicing silencers. The resulting silencer decamers were then clustered to yield seven putative motifs, some of which were suggested to play important role in alternative splicing.[24] Itoh *et al.* were also able to identify heptamer motifs that were enriched in alternatively spliced exons, one of which closely resembled the known ASF/SF2 binding site.[21] Sakai and Maruyama have suggested that specific types of alternative splicing events may have distinct patterns of regulatory motifs, and have extracted sequence motifs associated with different types of alternative splicing.[25] Mutations in splice regulatory motifs may play a significant role in human disease.[26,27]

Alternative splicing and NMD can work together to regulate gene function

Lewis *et al.*[28] have proposed that gene expression could be regulated through a novel mechanism based on a combination of alternative splicing and nonsense-mediated decay (NMD), which they have dubbed regulated unproductive splicing and translation (RUST). They found that one-third of the alternative transcripts examined were likely candidates for nonsense-mediated mRNA decay. In this case, the alternative splice form may simply result in degradation of the transcript, and thus down-regulation of the protein product. They later found that 144 of human alternative spliced isoforms in the human-curated Swiss-Prot database are NMD candidates, which in a number of cases appeared to be consistent with available data about these isoforms.[29] For further information on the role of RUST in regulating alternative splicing, see the extensive review of Brenner and coworkers.[30]

FUNCTIONAL IMPACT OF ALTERNATIVE SPLICING

Alternative splicing can modify protein domain structure and interactions

Bioinformatics can tell us how alternative splicing changes a transcript's nucleotide sequence, but predicting the functional impact requires knowing how it will change the translated protein's structure and interactions. This fundamental difficulty has been an unavoidable theme for many recent papers. We will review recent progress in several areas. First, researchers have sought to assess whether alternative splicing obeys a clear 'structural logic' in how it alters proteins. Second, is it possible to identify particular kinds of structural changes that have clear functional implications? Finally, experimental studies of the structure of one protein (Piccolo) provide a glimpse of the future of this field, showing both how interesting alternative splicing's effects on protein structure and function can be, and how complex.

How do alternative splicing events correlate with basic units of protein structure such as globular domains? Loraine *et al.*[31] found that 30 per cent of the multi-variant genes in their test set had alternatively spliced regions that coincided with conserved motifs. Kriventseva *et al.*[32] found that alternative splicing tended to insert or delete complete protein domains more frequently than expected by chance, whereas the disruption of domains and other structure modules was less frequent. Taneri *et al.*[33] provided evidence of the same pattern for transcription factors. Homma *et al.*[34] observed a slight tendency of alternative splice junctions to avoid the interior of SCOP domains and a strong tendency of alternative regions to coincide with SCOP domain boundaries when they studied the structures of the proteins encoded by the alternative splice variant of human brain cDNAs. Researchers also observed this pattern when they focused on specific domain types. Cline *et al.*[35] found that single-pass transmembrane (TM) regions in mouse were divided by introns substantially less often than expected by random chance. Overall, these results suggest that alternative splicing has been under some selection pressure to avoid structurally unsound changes that are inconsistent with protein domain structure. However, these patterns are not necessarily simple. For example, Offman *et al.* found no significant correlation between alternatively spliced regions and protein interaction sites in a test set of 21 alternatively spliced genes,[36] but as they point out, this does not necessarily mean that alternative splicing is not affecting these interaction sites.

One category of structural alteration that can have clear functional implications is *domain removal*, in which alternative splicing removes an entire domain of a protein. Some types of protein domains appear to be removed by alternative splicing much more frequently than would be expected by random chance, such as KRAB (Kruppel-associated box) domains and ankyrin repeats.[37] A common theme is the removal of domains that mediate protein–protein interactions, resulting in an alternative splice form that alters the default protein interaction network and thus can redirect a pathway (by changing specific linkages within that pathway). Another striking pattern is the removal of the transmembrane domain anchoring a single-pass membrane protein, resulting in an alternative splice form that produces a secreted protein.[38] This is equivalent to well-known regulatory mechanisms such as ectodomain shedding (in which an extracellular domain is released from its transmembrane anchor by proteolysis, resulting in both agonistic and antagonistic signalling effects), and has been identified in over 180 human genes.

Alternative splicing of very short peptide segments can change protein conformation in surprising ways

At the opposite end of the spectrum, alternative splicing can also modify protein structure in small, subtle ways, such as the insertion of just a few amino acids. For example, Hiller *et al.* report that NAGNAG motifs can produce alternative splice forms that insert a single amino acid, and that at least 5 per cent of human genes appear to contain such alternative splicing events.[39] In a similar vein, Wen *et al.* analysed the effects of very short alternatively spliced regions (insertions or removals of less than 50 nt, ie less than 17 aa), and suggest that such alternative splicing events make a large contribution to proteomic diversity.[40] One might expect that such 'minimal' alterations would be the easiest for predicting their detailed effect on protein structure and function, by using standard homology modelling methods to predict how these very slight sequence changes would be incorporated into the surrounding protein structure.

However, a recent structural study of the Piccolo protein shows how interesting (and surprising) even small sequence changes can be.[41] In this protein, alternative splicing inserts nine amino acids into a domain of previously known structure, the C_2A domain. Modelling of the short *v.* long forms of the Piccolo protein predicted that the extra nine amino acids are inserted in a surface loop of the protein (elongating this loop), distant from the protein's Ca^{2+} binding site. However, experimental determination of the structure of the Piccolo long isoform by NMR yielded a very surprising result. In reality, the 9 aa insert occupied a beta-strand in the protein core, displacing the sequence segment that normally forms this strand, which instead takes on a helical conformation, dramatically altering the Ca^{2+} binding site. Thus alternative splicing appears to cause a large conformational change in the protein structure, affecting both its Ca^{2+} affinity and protein interactions (the long form undergoes Ca^{2+}-dependent dimerization, while the short form does not). On the one hand, this result is exciting, because it shows in detail how alternative splicing can produce specific conformational changes that regulate function. On the other hand, it shows how structurally complex even such 'small' sequence changes can be. The standard assumption of modelling – that sequence regions that are unchanged will remain similar structurally – simply fails in this case, suggesting that modelling the protein structure effects of alternative splicing will be very challenging. Clearly, experimental studies of many more alternatively spliced protein structures are needed, and based on this result, are likely to produce very interesting insights into functional regulation and conformational change.

ALTERNATIVE SPLICING AND EVOLUTION

As additional genome sequences become available, comparative genomics is providing many new insights into alternative splicing and its role in genome evolution. Two related directions of research have recently emerged: first, the role of retronuons (and specifically Alu sequences in primates) in gene evolution (for a recent review see Kreahling and Graveley[42]); and second the contribution of alternative splicing to extending proteomic diversity during evolution (for reviews see Lareau *et al.*[30] and Boue *et al.*[43]). Here we will briefly mention some recent developments.

Alu exonisation may be a major contributor to human disease

Sorek and coworkers have studied in detail the mutations necessary to convert an intronic Alu into an exon that is constitutively or alternatively spliced.[44,45] Their data indicate that Alu 'exonisation' is a ubiquitous phenomenon that should have an important effect on genome evolution and human disease. For example, they estimate that more than 75,000 intronic Alu elements in the human genome are only a single mutation away from being converted into constitutive exons – in most cases, disrupting the product of the gene they are located in. Over 7,800 Alu elements in the human genome meet their criteria for exonisation, and may actually be exonic. It seems likely that mutations triggering Alu exonisation play a role in many human diseases; this has already been demonstrated in a number of cases.

Alternative splicing can modulate evolutionary selection pressures and 'accelerate' evolution

One final theme that has emerged from many studies is the measurement of different kinds of selection pressures on alternative splicing, to gain insight into its functional evolution. For example, Sorek *et al.* examined alternatively spliced exons for the occurrence of repeat elements (such as Alu) and frameshifts.[46] They found that human alternatively spliced exons that were conserved in mouse displayed much stronger selection pressure against both of these phenomena than was observed in alternatively spliced exons that were not conserved in mouse. This can be interpreted as evidence of reduced selection pressure on the latter category of exons. Similarly, both alternative splicing and diploidy were observed to reduce selection pressure against premature protein truncation codons.[47] Homma *et al.* used structural criteria to identify alternative splice events that were likely to produce an unstable protein product, and found via RT-PCR experiments that such splice forms were expressed at much lower levels than other variants.[34] Resch *et al.* examined selection pressure for reading-frame preservation (specifically, whether an exon is an exact multiple of 3 nt in length), and found that exons that were alternatively spliced in multiple genomes (eg human and mouse) displayed much higher selection pressure for frame-preservation.[48] This was particularly true for 'minor-form' exons with low exon inclusion levels. Philipps *et al.* discovered a very similar pattern in their studies of *Drosophila melanogaster v. D. pseudoobscura*.[9] Finally, analysis of microarray data has confirmed that alternatively spliced exons with low exon inclusion levels (in this case, measured in mouse tissue samples) were much less likely to be conserved in the human genome[6] (ie they are recent exon creation or loss events[49]), and also raised the interesting question of 'species-specific' splice forms.[6]

Overall, these data provide a rapidly converging picture of alternative splicing in genome evolution, in which alternative splicing both accelerates many kinds of evolutionary change (eg 'exonisation' and exon creation) by reducing negative selection pressure against such changes, but also undergoes special patterns of increased selection pressure (eg for exon lengths that are exact multiples of the 3 nt, so that alternative splicing of the exon will not alter the reading frame). Stay tuned – this story will grow as we dig deeper into the complexities of genome evolution.

References

* Papers of particular interest published within the period of this review.

** Papers of extreme interest published within the period of this review.

1. Modrek, B. and Lee, C. (2002), 'A genomic view of alternative splicing', *Nature Genet.*, Vol. 30(1), pp. 13–19.
2. **Johnson, J. M., Castle, J., Garrett-Engele, P. *et al.* (2003), 'Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays', *Science*, Vol. 302(5653), pp. 2141–2144.
3. Kampa, D., Cheng, J., Kapranov, P. *et al.* (2004), 'Novel RNAs identified from an in-depth analysis of the transcriptome of human

chromosomes 21 and 22', *Genome Res.*, Vol. 14(3), pp. 331–342.

4. Clark, T. A., Sugnet, C. W. and Ares, M. J. (2002), 'Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays', *Science*, Vol. 296(5569), pp. 907–910.

5. Le, K. Mitsouras, K., Roy, M. *et al.* (2004), 'Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data', *Nucleic Acids Res.*, Vol. 32(22), p. e180.

6. ⋆Pan, Q., Shai, O., Misquitta, C. *et al.* (2004), 'Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform', *Mol. Cell*, Vol. 16(6), pp. 929–941.

7. Kan, Z., Castle, J., Johnson, J. M. and Tsinoremas, N. F. (2004), 'Detection of novel splice forms in human and mouse using cross-species approach', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–10th January, Hawaii, pp. 42–53.

8. ⋆⋆Sorek, R., Shemesh, R., Cohen, Y. *et al.* (2004), 'A non-EST-based method for exon-skipping prediction', *Genome Res.*, Vol. 14(8), pp. 1617–1623.

9. ⋆Philipps, D. L., Park, J. W. and Graveley, B. R. (2004), 'A computational and experimental approach toward a priori identification of alternatively spliced exons', *RNA*, Vol. 10(12), pp. 1838–1844.

10. ⋆Heber, S., Alekseyev, M., Sze, S. H. *et al.* (2002), 'Splicing graphs and EST assembly problem', *Bioinformatics*, Vol. 18 Suppl. 1, pp. S181–188.

11. ⋆Lee, C., Grasso, C. and Sharlow, M. (2002), 'Multiple sequence alignment using partial order graphs', *Bioinformatics*, Vol. 18(3), pp. 452–464.

12. Lee, C. (2003), 'Generating consensus sequences from partial order multiple sequence alignment graphs', *Bioinformatics*, Vol. 19(8), pp. 999–1008.

13. Eyras, E., Caccamo, M., Curwen, V. and Clamp, M. (2004), 'ESTGenes: Alternative splicing from ESTs in Ensembl', *Genome Res.*, Vol. 14(5), pp. 976–987.

14. Grasso, C., Modrek, B., Xing, Y. and Lee, C. (2004), 'Genome-wide detection of alternative-splicing in expressed sequences using partial order multiple sequence alignment graphs', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–10th January, Hawaii, pp. 29–41.

15. ⋆Lee, B. T., Tan, T. W. and Ranganathan, S. (2004), 'DEDB: A database of *Drosophila melanogaster* exons in splicing graph form', *BMC Bioinformatics*, Vol. 5(1), p. 189.

16. ⋆Leipzig, J., Pevzner, P. and Heber, S. (2004), 'The Alternative Splicing Gallery (ASG): Bridging the gap between genome and transcriptome', *Nucleic Acids Res.*, Vol. 32(13), pp. 3977–3983.

17. ⋆Xing, Y., Resch, A. and Lee, C. (2004), 'The Multiassembly Problem: Reconstructing multiple transcript isoforms from EST fragment mixtures.' *Genome Res.*, Vol. 14(3), pp. 426–441.

18. Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001), 'Genome-wide detection of alternative splicing in expressed sequences of human genes', *Nucleic Acids Res.*, Vol. 29(13), pp. 2850–2859.

19. Grasso, C., Quist, M., Ke, K. and Lee, C. (2003), 'POAVIZ: A partial order multiple sequence alignment visualizer', *Bioinformatics*, Vol. 19(11), pp. 1446–1448.

20. Roca, X., Sachidanandam, R. and Krainer, A. R. (2003), 'Intrinsic differences between authentic and cryptic 5′ splice sites', *Nucleic Acids Res.* , Vol. 31(21), pp. 6321–6333.

21. ⋆Itoh, H., Washio, T. and Tomita, M. (2004), 'Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes', *RNA*, Vol. 10(7), pp. 1005–1018.

22. ⋆⋆Fairbrother, W. G., Yeh, R. F., Sharp, P. A. and Burge, C. B. (2002), 'Predictive identification of exonic splicing enhancers in human genes', *Science*, Vol. 297(5583), pp. 1007–1013.

23. Fairbrother, W. G., Holste, D., Burge, C. B. and Sharp, P. A. (2004), 'Single nucleotide polymorphism-based validation of exonic splicing enhancers', *PLoS Biol.*, Vol. 2(9), p. E268.

24. ⋆Wang, Z., Rolish, M. E., Yeo, G. *et al.* (2004), 'Systematic identification and analysis of exonic splicing silencers', *Cell*, Vol. 119(6), pp. 831–845.

25. Sakai, H. and Maruyama, O. (2004), 'Extensive search for discriminative features of alternative splicing', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–10th January, Hawaii, pp. 54–65.

26. ⋆Cartegni, L., Chew, S. L. and Krainer, A. R. (2002), 'Listening to silence and understanding nonsense: exonic mutations that affect splicing', *Nat. Rev. Genet.*, Vol. 3(4), pp. 285–298.

27. Pagani, F. and Baralle, F. E. (2004), 'Genomic variants in exons and introns: Identifying the splicing spoilers', *Nat. Rev. Genet.*, Vol. 5(5), pp. 389–396.

28. ⋆⋆Lewis, B. P., Green, R. E. and Brenner, S. E. (2003), 'Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans', *Proc. Natl Acad. Sci. USA*, Vol. 100(1), pp. 189–192.

29. Hillman, R. T., Green, R. E. and Brenner, S. E. (2004), 'An unappreciated role for RNA surveillance', *Genome Biol.*, Vol. 5(2), p. R8.

30. ⋆Lareau, L. F., Green, R. E., Bhatnagar, R. S. and Brenner, S. E. (2004), 'The evolving roles of alternative splicing', *Curr. Opin. Struct. Biol.*, Vol. 14(3), pp. 273–282.

31. Loraine, A. E., Helt, G. A., Cline, M. S. and Siani-Rose, M. A. (2003), 'Exploring alternative transcript structure in the human genome using blocks and InterPro', *J. Bioinform. Comput. Biol.*, Vol. 1(2), pp. 289–306.

32. ⋆Kriventseva, E. V., Koch, I., Apweiler, R. *et al.* (2003), 'Increase of functional diversity by alternative splicing', *Trends Genet.*, Vol. 19(3), pp. 124–128.

33. Taneri, B., Snyder, B., Novoradovsky, A. and Gaasterland, T. (2004), 'Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific', *Genome Biol.*, Vol. 5(10), p. R75.

34. Homma, K., Kikuno, R. F., Nagase, T. *et al.*(2004), 'Alternative splice variants encoding unstable protein domains exist in the human brain', *J. Mol. Biol.*, Vol. 343(5), pp. 1207–1220.

35. Cline, M. S., Shigeta, R., Wheeler, R. L. *et al.* (2004), 'The effects of alternative splicing on transmembrane proteins in the mouse genome', in 'Proceedings of the 9th Pacific Symposium on Biocomputing', 6th–9th January, Hawaii, pp. 17–28.

36. Offman, M. N., Nurtdinov, R. N., Gelfand, M. S. and Frishman, D. (2004), 'No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions', *BMC Bioinformatics*, Vol. 5(1), p. 41.

37. Resch, A., Xing, Y., Modrek, B. *et al.* (2004), 'Assessing the impact of alternative splicing on domain interactions in the human proteome', *J. Proteome Res.*, Vol. 3(1), pp. 76–83.

38. ⋆Xing, Y., Xu, Q. and Lee, C. (2003), 'Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains.' *FEBS Lett.*, Vol. 555(3), pp. 572–578.

39. ⋆Hiller, M., Huse, K., Szafranski, K. *et al.* (2004), 'Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity', *Nat. Genet.* Vol. 36(12), pp. 1255–1257.

40. Wen, F., Li, F., Xia, H. *et al.* (2004), 'The impact of very short alternative splicing on protein structures and functions in the human genome', *Trends Genet.*, Vol. 20(5), pp. 232–236.

41. ⋆⋆Garcia, J., Gerber, S. H., Sugita, S. *et al.* (2004), 'A conformational switch in the Piccolo C2A domain regulated by alternative splicing', *Nat. Struct. Mol. Biol.*, Vol. 11(1), pp. 45–53.

42. ⋆Kreahling, J. and Graveley, B. R. (2004), 'The origins and implications of Aluternative splicing', *Trends Genet.*, Vol. 20(1), pp. 1–4.

43. Boue, S., Letunic, I. and Bork, P. (2003), 'Alternative splicing and evolution', *Bioessays*, Vol. 25(11), pp. 1031–1034.

44. ⋆Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003), 'The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons', *Science*, Vol. 300(5623), pp. 1288–1291.

45. Sorek, R., Lev-Maor, G., Reznick, M. *et al.* (2004), 'Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons', *Mol. Cell*, Vol. 14(2), pp. 221–231.

46. Sorek, R., Shamir, R. and Ast, G. (2004), 'How prevalent is functional alternative splicing in the human genome?', *Trends Genet.*, Vol. 20(2), pp. 68–71.

47. Xing, Y. and Lee, C. (2004), 'Negative selection pressure against premature protein truncation is reduced by both alternative splicing and diploidy', *Trends Genet.*, Vol. 20(10), pp. 472–475.

48. Resch, A., Xing, Y., Alekseyenko, A. *et al.* (2004), 'Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation', *Nucleic Acids Res.*, Vol. 32(4), pp. 1261–1269.

49. Modrek, B. and Lee, C. (2003), 'Alternative splicing in the human, mouse and rat genomes is associated with an increased rate of exon creation/loss', *Nature Genet.*, Vol. 34(2), pp. 177–180.