Alternative Splicing, the Gene Concept, and Evolution

Stephen M. Downes

Department of Philosophy
University of Utah
260 S. Central Campus Drive, Room 341
Salt Lake City, UT 84112, USA

ABSTRACT - Alternative splicing allows for the production of many gene products from a single coding sequence. I introduce the concept of alternative splicing via some examples. I then discuss some current hypotheses about the explanatory role of alternative splicing, including the claim that splicing is a significant contributor to the difference in complexity between the human genome and proteosome. Hypotheses such as these bring into question our working concepts of the gene. I examine several gene concepts introduced to cope with processes such as alternative splicing. Next I introduce some hypotheses about the evolution of mechanisms alternative splicing in higher organisms. I conclude that attention to alternative splicing reveals that we adopt an attitude that developmental theorizing must inform evolutionary theorizing and vice versa.

KEYWORDS: alternative splicing; evolution; gene

1. Introduction

Alternative splicing is one of a number of post-transcriptional controls known to operate between the transcription of DNA and the ultimate production of proteins in the cell. Recent estimates (catalogued in Modrek and Lee 2002, 14) place the number of human genes that are alternatively spliced between 22% and 59% of our genetic complement. Work on alternative splicing, as well as work on other post-transcriptional controls, introduces new questions about both the gene concept itself and about evolution. I briefly address both these issues here. First, I introduce alternative splicing via a few illustrative examples. I then discuss relations between alternative splicing and organismal complexity. This relation is illustrated by comparing an organism's genomic complexity with the complexity of its proteosome. This discussion leads us to question the referent for the term 'gene'. In the following section I examine two recent definitions of the gene introduced to deal with related difficulties to the ones introduced here: Lenny Moss' (Moss 2001; Moss 2003) Gene-D and Eva Neuman-Held's (Neumann-Held 2001) Process Molecular Gene (PMG). I defend a modified version of Moss' Gene-D as a device to

help understand the referent of the term 'gene' in much contemporary molecular biology. Finally, I turn to the evolution of alternative splicing. How we approach the evolution of alternative splicing should be connected to our overall approach to understanding evolution. I argue for an approach to the evolution of alternative splicing that shares input from both developmental and evolutionary theorizing.

2. Alternative Splicing, Splice Variants and Some Examples

RNA splicing is known to occur in a huge range of organisms. The existence of the process is familiar to all molecular biologists and an outline of the process is presented in all introductory texts in molecular biology. Primary transcript RNA molecules in the nucleus of eukaryotes contain on average 6000 nucleotides while mature m-RNA molecules contain on average 500 nucleotides. The main process responsible for this reduction in nucleotide number is splicing, the removal of introns from the transcribed RNA (often called precursor-mRNA or pre-mRNA). What remains after this process is an mRNA strand that only contains the RNA version of the code from the exons in the original DNA sequence. (Figure 1) The outcome of this process can be modified by alternative splicing. Alternative splicing involves the production of a mature mRNA molecule that contains a selection

Illustration of Intron Splicing from Pre-mRNA

Intron removal/splicing.

Pre-mRNA

Mature mRNA

= Exon
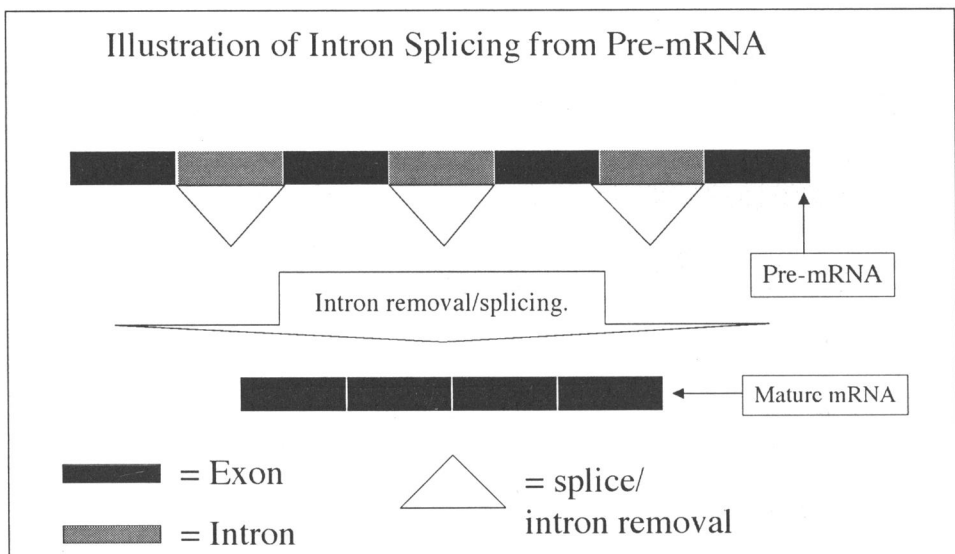= Intron
= splice/ intron removal

FIGURE 1

of the available exons present in the pre-mRNA molecule. Splice variants are the results of this process and are the different sequences of mature mRNA produced from an identical strand of pre-mRNA. (Figures 2 & 3) Important examples of splice variants include both
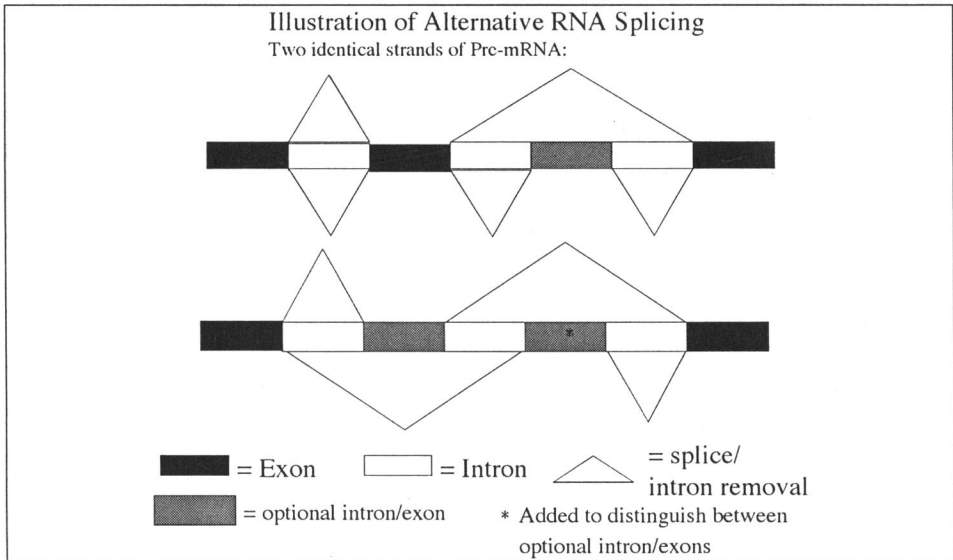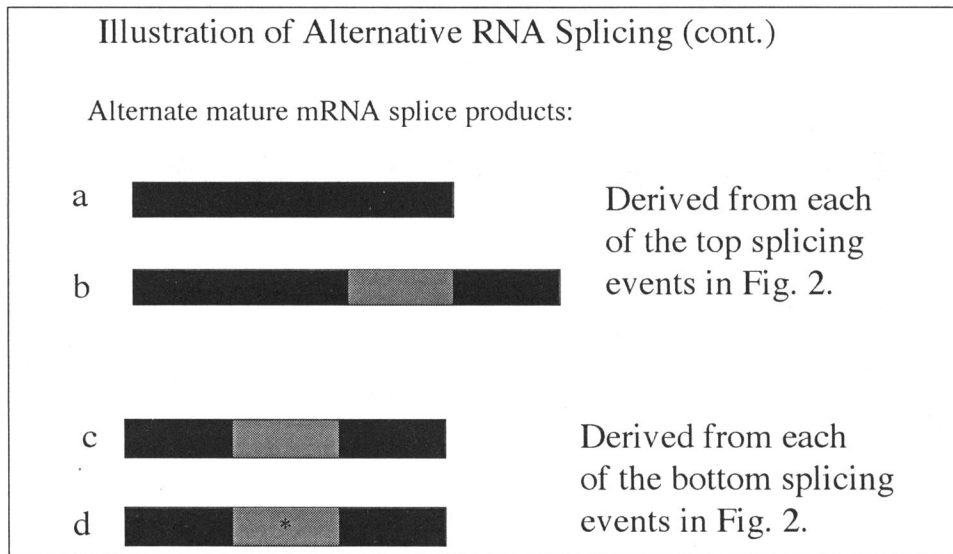

FIGURE 2

FIGURE 3

those that code for variants on a protein (protein isoforms) leading to a limited change in function and those that code for completely different proteins and hence different subsequent biological functions. (This latter form of post-transcriptional control is also called 'gene switching' or 'gene sharing' or explained in terms of 'overlapping genes' [Alberts *et al.* 2002, 438]).[1]

Alternative splicing is found in many organisms, including humans as I mentioned in the introduction. Here I mention just a few examples of alternative splicing.

Perhaps the most well known and most cited example of alternative splicing (especially in textbooks) (see e.g. Li and Graur 1991; Alberts *et al.* 2002) is the process that leads to sex determination in *Drosophila m.* (reviewed in Baker 1989) (Figure 4). Sex determination in *Drosophila m.* is primarily controlled by the ratio of X chromosomes relative to sets of
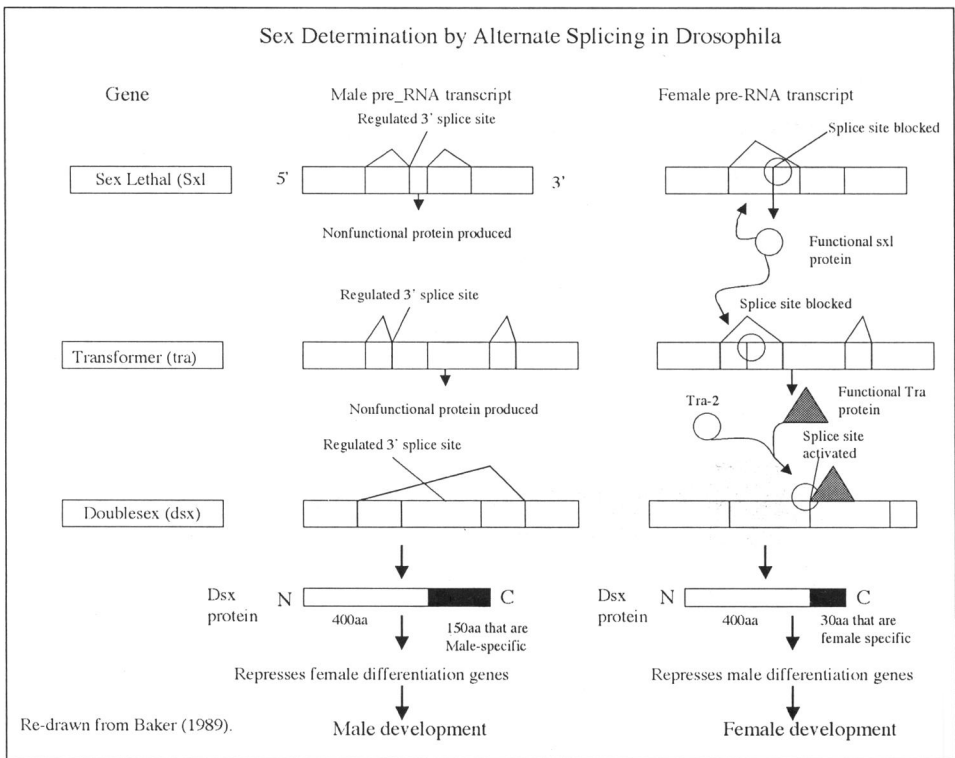

FIGURE 4

[1] Burian (this volume) discusses these cases among others.

autosomes, the X:A ratio (Baker 1989). Flies with a ratio of 1 are female, while those with a ratio of 0.5 are male. Whether the embryonic fly becomes a male or female is determined by a series of RNA splicing events initiated by these ratios. The male development pathway is the 'default' pathway. The X:A ratio of 1 triggers an alternate pathway leading to female sex characteristics. Let's focus on this pathway. The connection between the X:A ratio of 1 and the initiation of the pathway, the blocking of the splice site leading to the production of functional Sxl protein (seen at the top of the right hand column of Figure 4.) appears to be regulated by both maternal and zygotic gene products. Sxl is the first of two splicing regulatory proteins; one that blocks a splice and the other that activates a splice. The effect of these alternate splicings is the production of the female specific form of the protein Dsx. (The process for each sex is illustrated in Figure 4.) Sex determination in *Drosophila m.* is apparently not determined for by a specific DNA sequence but rather by alternative splicing regulators.

Work on various species of Drosophila has revealed another striking example of alternative splicing: the alternative splicing of the RNA transcripts of the Drosophila DSCAM gene. DSCAM proteins help direct growth of cells in the Drosophila nervous system. The pre-mRNA transcript of DSCAM contains 115 exons and each mature mRNA contains twenty four exons and four of these are each selected from four groups (of 12, 48, 33 and 2) of the original 115 (Figure 5).
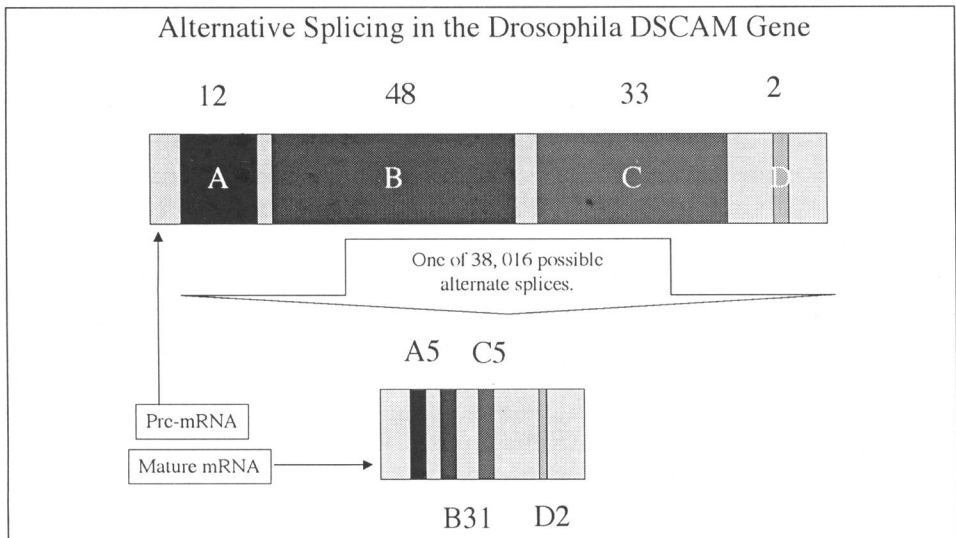

FIGURE 5

The remaining 20 exons in the pre-mRNA are always in the mature mRNA. The combinatorics here allow for 38,016 possible splice variants. Each of the variant DSCAM proteins has a similar structure, so there is not as much contrast between the outcomes of the alternative splicings as there is in the sex determination case but none of the proteins could be produced without alternative splicing occurring (adapted from Alberts *et al.* 2002).

Moss discusses a similar example presented by the human NCAM gene. The human NCAM gene has 19 exons but Moss points out 'there are no NCAM proteins that are composed of the protein domains encoded for by all 19 exons' (Moss 2003, 186). The alternate NCAM proteins are produced in a similar way to the alternate DSCAM proteins in Drosophila: each sequence of mature mRNA producing an NCAM protein is a splice variant.

There are many more examples of alternative splicing in the literature and doubtless many more will be appearing in the human molecular literature if the estimates from the bio-informatics work cited above are correct.

3. Alternative Splicing and Organismal Complexity

One of the more surprising findings to come out of the human genome project is the figure proposed for the number of human genes. While our chromosome complement contains a relatively large number of DNA base pairs, about 3 billion (nothing compared to a lot of plants and a few sharks), recent estimates put our number of genes at around 30,000. (For comparison, Drosophila appear to have in the region of 15,000 genes.) The reason that the proposed number of genes is surprising is that if each gene coded for only one protein, we would come in well under our protein complement, what is known as our proteosome. The number of proteins an organism can produce is a rough guide to the organism's overall complexity. (I leave out any further discussion of the concept of complexity in this paper.) Our proteosome is enormous and is roughly five times that of Drosophila.

Alternative splicing provides one explanation for disparity between our gene count and our protein count. As we have seen, alternative splicing can produce many proteins from one pre-mRNA transcript. Several additional explanations of this disparity have been proposed but most of these additional explanations should be understood as complementary and not competing. From now on I focus on

alternative splicing as an explanation for the gene to protein disparity but I am not proposing that this is a sufficient explanation. Molecular biology is beginning to provide us with numerous complementary explanations of this phenomenon.

Two classes of questions confront us when we propose alternative splicing as an explanation for the disparity between gene number and protein number. The first has to do with just exactly what the term 'gene' refers to and the second has to do with the evolution of alternative splicing machinery. I deal with each in turn.

4. Alternative Splicing and the Gene Concept

What Sterelny and Griffiths (1999) call the 'classic molecular gene concept', that a gene is 'a stretch of DNA that codes for a single polypeptide chain' (1999, 132) does not seem to help us make sense of the predicament outlined above. Molecular biologists talk about the number of genes in an organism and the number of proteins an organism can produce and point out that these numbers are different. Hence we need to be able to distinguish a gene from its RNA splice variants and their corresponding proteins. While many molecular biologists use a similar gene concept to Sterelny and Griffiths' classical molecular gene concept in some contexts, they explicitly invoke alternative gene concepts in others. Alberts *et al.* (2002), in the fourth edition of their text, tackle worries about the gene concept head on. They say that 'the discovery of split genes and introns in the late 1970's could be readily accommodated in the original definition of a gene, provided that a single polypeptide chain was specified by the RNA from any one transcribed DNA sequence. But it is now clear that many DNA sequences in higher eukaryotic cells can produce a set of distinct (but related) proteins by means of alternative RNA splicing' and go on to ask 'How then is a gene to be defined?' (2002, 438). Let's look at a few suggestions.

There are two distinct types of gene definition that have arisen in response to worries such as the one articulated here: gene concepts that locate the referent for the term gene in the DNA complement of the cell (DNA sequence gene concepts) and gene concepts that disperse the referent of the term gene over varying parts of the cellular machinery (inclusive or wide gene concepts). I present these two first and then briefly discuss definitions of another (intermediary) type (pragmatic gene concepts).

First type: DNA sequence gene concepts.

Alberts *et al.* (2002) answer their own question with a proposal to 'count as a gene any DNA sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains (protein isoforms)' (438). They go on to say that 'this definition of a gene also accommodates those DNA sequences that encode protein variants produced by post-transcriptional processes other than RNA splicing' (438). This definition allows us to talk about genes that lead to the production of a large number of proteins. A similar approach is adopted by Moss but his starting point is different.

Moss distinguishes two types of genes referred to in biological literature, gene-P and gene-D. Gene-P stands for phenotype or predictor gene and is 'defined in its relationship to a phenotype [...] when one speaks of a gene in the sense of Gene-P, one simply speaks *as if* it causes the phenotype' (Moss 2001, 87). This is the definition that best captures the usage of the term gene in medical genetics and some parts of population genetics but does not adequately capture the use of the term in molecular biology. Moss proposes gene-D, standing for developmental resource gene, for this purpose: 'Gene-D is defined by its molecular sequence [...] to be a Gene-D is to be a transcriptional unit (extending from start to stop codons) within which are contained molecular template resources' (Moss 2001, 88). My take on Moss' definition is that he intends genes-D to be DNA sequences and therefore his definition is consistent with Alberts *et al.*'s. This is supported by Moss' discussion of the NCAM genes where he says that the gene-D for NCAM contains 19 modular units or exons and that it is a 'resource for making a protein' (Moss 2003, 186). He also puts the discussion of genomic vs. proteosome complexity in these terms: 'The human genome has twice the number of Genes-D as that of the fly or worm' (Moss 2003, 187). Again, Moss' gene-D allows us to talk about genes and their multiple protein products. An advantage of Moss' definition over Alberts *et al.*'s is that it comes with a handy label: gene-D. But more importantly Moss' definition is more inclusive than Alberts' as gene-D applies to both DNA strands that are transcribed into polypeptides and DNA strands that produce r-RNA and t-RNA molecules and no polypeptides.

Second type: inclusive or wide gene concepts.

Neuman-Held (2001) proposes a definition of the gene that is also developed in response to problems for the gene concept generated by complexities of molecular development. She says that her goal is to establish a gene concept that applies to *developmental processes* on

those *molecular* levels of interactions, which have to do with *DNA* and end with the synthesis of *linear polypeptide* chains' (75; her italics). She proposes and defends the process molecular gene concept (PMG): 'PMG […] allows for inclusion of not only DNA, but also non-DNA located entities, thereby integrating into the gene concept those relevant entities that are necessary for the functional specification of the DNA sequences involved' (80). This is a wide gene concept. The concept tries to capture the idea that the gene produces the relevant protein. Given that numerous cellular processes are involved in producing protein from a DNA strand, Neumann-Held includes them in the referent of the term 'gene'. This approach solves some problems, for example there could be one PMG for each polypeptide chain, but still makes it hard for us to understand what molecular biologists are saying when they say that the human genome only contains 30,000 genes. The human genome (unless it also is redefined in PMG terms) contains no PMG's.

Intermediary types: Pragmatically derived gene concepts.

In a much cited passage discussing the gene concept Philip Kitcher says 'it is hard to see what would be lost by dropping talk of genes from molecular biology and simply discussing the properties of various interesting regions of nucleic acid' (Kitcher 1992, 130). The implied definition of a gene here is that a gene is any region of interesting nucleic acid. Interesting regions of nucleic acid include all the DNA strands accounted for by gene-D but also much of the machinery Neumann-Held wants to include in PMG as much of that machinery is RNA. Kitcher's motivation for proposing this move is that cataloguing the uses of the term 'gene' leads us to a far too hazy and ambiguous concept. The two alternative gene definitions considered above work by dividing the reference of the term 'gene' and introducing new terms for the partitioned referent. This approach seems more promising than abandoning the term 'gene' altogether. Aside from the fact that abandoning the term gene would require a super-Orwellian effort at re-writing molecular biology.

A related pragmatic approach to the term 'gene' is proposed by Sterelny and Griffiths (1999) when they say: 'molecular biologists do not seem to use the term *gene* as a name of a specific molecular structure. Rather, it is used as a floating label whose reference is fixed by the local contexts of use. Molecular biologists often seem to use *genes* to mean 'sequences of the sort(s) that are of interest in the process I am working on'. Their rich background of shared assumptions make this usage perfectly satisfactory' (1999, 133). (Their

proposal is close to Waters' gene concept (Waters 1994).[2] At this point their proposal is intermediary between gene-D and PMG. Later on in their discussion they seem to come closer to a gene-D concept for molecular biology: 'The concepts of classic genetics, most notably *gene* itself, continue to play a role in molecular biology, although perhaps as little more than shorthand for the various DNA sequences and collections of interacting DNA sequences used in molecular biological explanations of organisms and their traits' (148).

The gene concept that best accounts for the use of the term gene in molecular biology practice is going to have an element of stipulation to it. The issues generated by focusing on alternative splicing and other post-transcriptional controls lead me to proposing gene-D, restricted to DNA, as the best definition of the term 'gene' of those considered here. Closer examination of other molecular developmental processes, or consideration of alternative gene concepts, may require a revision of this position.

5. The Evolution of Alternative Splicing: Two Contrasting Perspectives

My discussion so far has made no reference to evolution. The background to the discussion of the definition of the term 'gene' presented here resides entirely in molecular developmental biology: the articulation of the processes involved in the production of proteins in cells. Many would argue that the relevant constraints on the gene concept come from articulating its explanatory role in evolutionary biology and not developmental biology. Maynard Smith, for example, argues that the evolutionary gene concept should be imported into developmental biology and that this would be an instructive and useful move for developmental biology (see e.g. Maynard Smith 1998). So why the emphasis on molecular developmental biology here?

. Here are a few brief general responses: First, alternative splicing has been proposed as an explanation of the existence of higher order diversity (see discussion in Brett *et al.* 2002). This presents discussion of alternative splicing in an evolutionary context. If we ask how higher orders of complexity arose in higher organisms, one answer could be by alternative splicing. Hence explaining how these organisms evolved

---

[2] Waters has developed a more complex and inclusive gene concept since his 1994 paper. The criticisms in my paper do not target an important component of Waters' newer gene concept: his technical definition of the molecular gene. My arguments are directed at pragmatic gene concepts, which owe a lot to Waters' earlier paper. Assessing whether my arguments apply to Waters' mature gene concept is a subject for a different paper.

requires invoking the cellular processes involved in alternative splicing. Second, a process like alternative splicing is important for evolutionary theorists to focus on because it is just one, of many, processes that lead to the production of proteins in cells. If any of the systems for controlling these processes are heritable in ways that parallel and accompany DNA transmission, then these systems have evolutionary significance. My view is that to confront these evolutionary questions we need to pay careful attention to theoretical developments in molecular developmental biology and work towards an account of evolution that is consistent with these findings. So I resist Maynard Smith's proposal, not by reversing the direction of his proposal and suggesting that evolutionary biology must import concepts from developmental biology, but by recommending theoretical influence in both directions. This suggestion is consistent with the goal of evolutionary developmental biology as defended by Hall who suggests that evolutionary developmental biology is 'a synthesis of evolution and development with emergent properties not found from analysis of development or evolution alone' (Hall 2000, 177-178).

I now look at some specific suggestions about the evolution of alternative splicing. There are two positions in the discussion of the evolution of alternative splicing or evolution resulting from new alternative splicing events. The first emphasizes change as a result of mutations in DNA sequences and the second emphasizes change as a result of changes in RNA and other splice controlling mechanisms. Proponents of both perspectives agree that the production of new splice variants leads to greater diversity of phenotypes.

Li and Gruar (1991) represent the first perspective. They argue that 'the evolution of alternative splicing requires that an alternative splice junction be created de novo. Since splicing signals are usually 5-10 nucleotides long, it is possible that such splice sites are created with an appreciable frequency by mutation' (160). They discuss one example of this process, the b+-Thalassemia gene. Unfortunately, this is not the best illustration of their point as possession of the mutation is lethal. The general principle of their idea is grounded in the distinction between weak and strong splice sites. The b+-Thalassemia mutation creates a strong splice site that leads the cell to always produce the deleterious protein. The production of a weak splice site will provide an opportunity for the cell to produce both the original protein and the new one, hence giving the cell the potential to produce a new protein with perhaps a new function. Alberts *et al.*

(2002) add that there is an interplay between weak and strong splice sites in pre-mRNA. If a strong splice site is blocked (as in the Drosophila sex determination example above) a weak splice site may be exposed to produce a different splicing pattern. But this added explanation exposes a weakness in Li and Gruar's perspective: whatever novelty is produced is not a result simply of a mutation in the DNA, the gene-D, but also the result of the differential effects of regulatory proteins and RNA machinery.

Alberts *et al.* (2002) defend an alternative perspective. They argue that the 'RNA-splicing cascade is an ancient control device, left over from a stage of evolution where RNA was the predominant biological molecule and controls of gene expression had to be based almost entirely on RNA-RNA interactions' (Alberts 1994, 456; Alberts *et al.* 2002, 439). As a result they emphasize an examination of the processes that lead to changes in these regulatory structures. Now of course this could amount to the suggestion that we look back to the DNA but if various regulatory structures involved in RNA splicing are inherited independently from DNA transmission, then their proposal is different than and supplementary to Li and Gruar's.

Moss (2001; 2003), Gerhardt and Kirshner (1997) and several proponents of developmental systems theory and its variants (see e.g. Jablonka 2001) hold out for this latter perspective. Moss' approach is illustrative, his view is that evolution is not achieved by the elaboration of a master code script in DNA (e.g. simply by mutation and selection) but rather 'by the fragmentation of the functional resources of the cell into many modular units whose linkages to one another have become contingent' (Moss 2003, 188-189). Exploitation of various combinations of these modular units in varying ways leads to the production of novel proteins and structures. Moss supports the emphasis of Alberts *et al.* in approaching the evolutionary problem via looking at the inheritance of mechanisms other than DNA sequences that guide splicing and other post-transcriptional processes.

My sense is that neither perspective on the evolution of alternative splicing should overwhelm the other. If we adopt the gene-D account, then it seems consistent to say that mutations in genes-D provide new opportunities for developmental processes. We can say this without ruling out investigation into the inheritance and variation over time in developmental regulatory systems outside the DNA.

6. Conclusion

Looking at the process of alternative splicing provides an opportunity to examine both the gene concept and our views about what perspective to emphasize when explaining the evolution of cellular processes. I have argued that a slightly modified version of Moss' gene-D best fits the concept of gene invoked in discussions of alternative splicing. I have also argued that explaining the evolution of cellular processes requires adopting (at least) two perspectives on evolution. This move requires adopting an attitude that developmental theorizing must inform evolutionary theorizing and vice versa.

Acknowledgements

References

Alberts B., 1994, *The Molecular Biology of the Cell*, New York: Garland.

Alberts B.A., Johnson *et al.*, 2002, *The Molecular Biology of the Cell*, New York: Garland.

Baker B.S., 1989, 'Sex in Flies: The Splice of Life', *Nature,* 340: 521-524.

Brett D., H. Pospisil H. *et al.*, 2002, 'Alternative Splicing and Genome Complexity', *Nature Genetics,* 30: 29-30.

Gerhart J., Kirshner M., 1997, *Cells, Embryos, and Evolution*, Malden, MA: Blackwell Science.

Hall B.K., 2000, *Evolutionary Developmental Biology*, New York: Chapman and Hall.

Jablonka E., 2001, 'The Systems of Inheritance'. In: Oyama S., Griffiths P., Gray R.D. (eds), *Cycles of Contingency*, Cambridge: MIT Press.

Kitcher P., 1992, 'Gene: Current Usages'. In: Fox Keller E. and Lloyd E.A. (eds), *Keywords in Evolutionary Biology*, Cambridge: Harvard University Press, 128-131.

Li W.-H., Graur D., 1991, *Fundamentals of Molecular Evolution*, Sunderland: Sinauer Associates.

Maynard Smith J., 1998, *Shaping Life*, New Haven: Yale University Press.

Modrek B., Lee C., 2002, 'A Genomic View of Alternative Splicing', *Nature Genetics,* 30: 13-19.

Moss L., 2001, 'Deconstructing the Gene and Reconstructing Molecular Developmental Systems'. In: Oyama S., Griffiths P.E., Gray R.D. (eds), *Cycles of Contingency*, Cambridge: MIT Press.

Moss L., 2003, *What Genes Can't Do*, Cambridge: MIT Press.

Neumann-Held E.M., 2001, 'Let's Talk About Genes: The Process Molecular Gene Concept and Its Context'. In: Oyama S., Griffiths P.E., Gray R.D. (eds), *Cycles of Contingency*, Cambridge: MIT Press.

Sterelny K., Griffiths P.E., 1999, *Sex and Death*, Chicago: University of Chicago Press.

Waters K., 1994, 'Genes Made Molecular', *Philosophy of Science,* 61: 163-185.