Thomas Blumenthal
is Professor and Chairman in the Department of Biochemistry and Molecular Genetics at the University of Colorado School of Medicine.

Operons in eukaryotes

Thomas Blumenthal

Date received (in revised form): 1st July 2004

Abstract

It was thought that polycistronic transcription is a characteristic of bacteria and archaea, where many of the genes are clustered in operons composed of two to more than ten genes. By contrast, the genes of eukaryotes are generally considered to be monocistronic, each with its own promoter at the 5′ end and a transcription terminator at the 3′ end; however, it has recently become clear that not all eukaryotic genes are transcribed monocistronically. Numerous instances of polycistronic transcription in eukaryotes, from protists to chordates, have been reported. These can be divided into two broad types. Dicistronic transcription units specify a messenger RNA (mRNA) encoding two separate genes that is transported to the cytoplasm and translated in that form. Presumably, internal ribosome entry sites (IRES), or some form of translational re-initiation following the stop codon, are responsible for allowing translation of the downstream gene. In the other type, the initial transcript is processed by 3′ end cleavage and *trans*-splicing to create monocistronic mRNAs that are transported to the cytoplasm and translated. Like bacterial operons, eukaryotic operons often result in co-expression of functionally related proteins.

Keywords: *C. elegans, Drosophila, trans-splicing, gene organisation, genome organisation*

INTRODUCTION

Operons were described by Jacob *et al.*[1] as a cluster of genes under the control of a single regulatory signal or promoter. In the case of bacterial operons (the only ones known until recently), it was subsequently shown that the co-regulated genes of an operon are expressed from a single polycistronic messenger RNA (mRNA) that is translated in that form; however, production of polycistronic mRNA is not a part of the definition of an operon and is not a property of all kinds of operon. This paper will discuss two different types of eukaryotic operon. Polycistronic transcription in eukaryotes was first found in 1988 in trypanosomes, although these polycistronically-transcribed genes do not represent operons in the sense of co-regulation (reviewed in ref. 2). Widespread operons in an animal were first discovered in the nematode, *Caenorhabditis elegans*, in 1993.[3] Nematode operons are transcribed to produce polycistronic initial transcripts that are co-transcriptionally processed to make monocistronic mRNAs (Figure 1).[3,9] This is also true for the operons discovered more recently in flatworms and primitive chordates.[10,11] The other type of operon consists of dicistronic gene clusters first found by Lee in 1991.[12] Similar gene arrangements were found in flies and plants (Figure 1). These are more like bacterial operons — they make polycistronic mRNAs that are translated in that form. Several recent reviews have discussed various aspects of eukaryotic operons.[13–16]

RECOGNISING POLYCISTRONIC TRANSCRIPTION

Three kinds of observation have led to the discovery of operons: (1) the presence of stable dicistronic mRNAs on Northern blots; (2) a special kind of *trans*-splicing restricted to use within operons; and (3) the presence of *trans*-spliced genes in such closely spaced clusters that they almost must be co-transcribed. Stable dicistronic mRNAs were first discovered in vertebrates by Lee.[12] Two related genes, *GDF-1* and *UOG-1*, are translated from a single mRNA, and this arrangement is conserved between humans and mice. In *Drosophila*, Brogna and Ashburner[5] discovered that *Adh* and *Adhr*, two similar genes, are translated from a dicistronic mRNA, and Andrews *et al.*[6] discovered

Thomas Blumenthal,
Department of Biochemistry and Molecular Genetics,
University of Colorado School of Medicine at Fitzsimons,
Mail Stop 8101,
PO Box 6511,
Aurora, CO 80045, USA

Tel: +1 303 724 3203
Fax: +1 303 724 3215
E-mail: tom.blumenthal@uchsc.edu

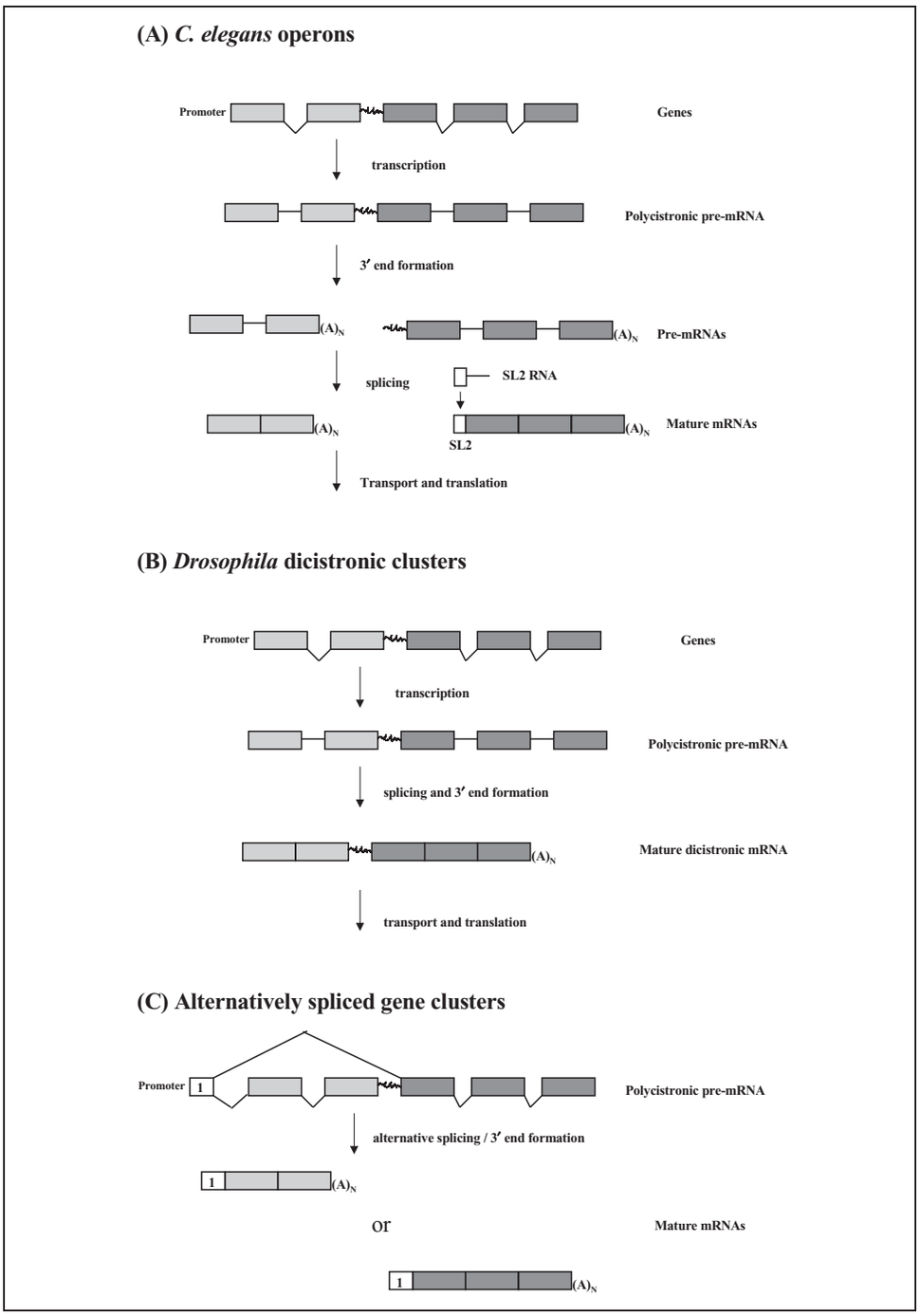

Figure 1: Three types of gene clusters give rise to different products. (A) The *C. elegans* operons are co-transcriptionally processed to give a polycistronic pre-mRNA that is processed between the genes to give monocistronic mature mRNAs, both of which are polyadenylated by conventional mechanisms. The downstream product is *trans*-spliced to the 22 nucleotide leader, SL2. This trims off the inter-cistronic sequence that is normally ~100 nucleotides long.[4] (B) The *Drosophila* dicistronic clusters are transcribed to make dicistronic mature mRNAs that are transported and translated in that form (eg refs. 5 and 6). (C) Alternatively spliced gene clusters splice a common first exon to different gene bodies so that for each pre-mRNA either the first gene or second gene mRNA, but not both, can be produced (eg refs. 7 and 8). Notes: Boxes, exons, with different fills representing different genes; lines, introns; wavy lines, inter-cistronic sequences; (A_N), poly(A).

that the two genes that comprise the stoned locus are made from a dicistronic mRNA. About 40 additional examples were uncovered following the sequencing of the entire fly genome.[17]

In some primitive protists, notably the trypanosomes, all transcription is polycistronic and the pre-mRNAs are all processed by *trans*-splicing coupled tightly with 3′ end formation (reviewed in ref. 2). It has recently been demonstrated that one chromosome in the trypanosome, *Leishmania major*, contains two long transcription units, each beginning near a telomere and meeting at a central transfer RNA (tRNA) gene at which transcription termination occurs.[18,19] In general, the genes in these polycistronic transcription units lack introns, so *trans*-splicing is the only splicing they undergo. While it is not clear if most co-transcribed clusters in the trypanosome genome are co-regulated in any way, there are examples of functionally related genes being co-transcribed (eg five genes encoding the entire *de novo* pyrimidine biosynthesis pathway).[20]

Proving the existence of polycistronic transcription has been difficult

The discovery of operons in disparate lineages has occurred in quite different ways

In nematodes, operons were discovered as a consequence of the strong correlation between genes with the same 5′ to 3′ orientation existing in unusually tight clusters, and the observation that the downstream genes were *trans*-spliced with the relatively rare spliced leader, SL2.[3] This correlation resulted in the hypothesis that genes were co-transcribed and that SL2 was a form of spliced leader specialised for separating mRNAs from polycistronic pre-mRNA precursors. Subsequent work demonstrated that this hypothesis was correct and that ~15 per cent of all *C. elegans* genes are transcribed in polycistronic clusters ranging from two to eight genes in length.[21,22] The vast majority of genes in *C. elegans* operons contain multiple introns, so these pre-mRNAs undergo both *cis*- and *trans*-splicing, in contrast to the situation in trypanosomes.

Finally, operons have been discovered in several organisms by virtue of the tight clustering of genes. Examples include the tomPRO1 locus of tomato,[23] several operons of the eukaryotic endosymbiont nucleomorph of algae,[24] operons in the flatworm *Schistosoma mansoni*[10] and, very recently, operons in a primitive chordate — the appendicularian tunicate *Oikopleura dioica*.[11]

DEMONSTRATING POLYCISTRONIC TRANSCRIPTION

In most cases, it has been difficult to demonstrate that genes are co-transcribed except in cases where they produce a stable polycistronic mRNA, as is the case in the fly dicistronic mRNAs and the isolated examples found in mammals. When the polycistronic pre-mRNA is processed by *trans*-splicing to produce monocistronic mRNAs as in worms, primitive chordates and trypanosomes, the 3′ end formation and *trans*-splicing processes are so efficient that polycistronic precursor does not accumulate. There are examples where the precursor is stable enough to allow its detection (eg in *C. elegans*, *S. mansoni* and *O. dioica*), but even in most of these cases it has not proven possible to demonstrate unequivocally that the polycistronic RNAs are actually the progenitors of the mature mRNAs, since they could be dead-end products. Because it is not possible to eliminate the possibility of promoters sandwiched between the genes in these gene pairs, the conclusion that these clusters represent true operons, while very likely correct, must be considered tentative. The one exception is the nematode operon, where the strong correlation between gene clustering and SL2 *trans*-splicing is sufficiently strong to demonstrate that about 1,000 identified clusters where the downstream genes are known to be SL2 *trans*-spliced represent true operons.[22]

SL-type *trans*-splicing in animals

Normally, mRNA 3′ end formation by cleavage and polyadenylation leaves a phosphate on the 5′ end of the downstream RNA, and this is the signal

for rapid 5′ to 3′ exonuclease activity to degrade the RNA (reviewed in ref. 25). This then results in transcription termination.[26] How can multiple genes be expressed from a single promoter, as in operons, without the upstream 3′ end formation leading to degradation of downstream mRNA and transcription termination? In most organisms that express genes in operons, it is *trans*-splicing that makes this sort of gene arrangement possible. *Trans*-splicing involves transferring a small leader, called the spliced leader, or SL, from a short RNA donor called the SL RNA to the 5′ ends of some or all of the mRNAs in the organism (reviewed in ref. 27). In most *trans*-splicing events, the SL is spliced onto the first 3′ splice site in the pre-mRNA, replacing the RNA at the 5′ end of the pre-mRNA between the promoter and the *trans*-splice site, called the outron. In addition, *trans*-splicing can trim the inter-gene region in an operon pre-mRNA, putting a cap on the downstream transcript, thus preventing further exonucleolytic attack.

Trans-splicing provides a cap for downstream operon genes, ensuring their stability

The SL ranges from 16 nucleotides long in the primitive chordate *Ciona intestinalis*,[28] to 51 nucleotides long in the flatworm *Stylochus zebra*.[29] The SL RNA donors are also quite small, between 81 and 110 nucleotides long. *Trans*-splicing is a process that is very closely related to *cis*-splicing, the process by which introns are removed from pre-mRNA. Both processes are catalysed by spliceosomes made up mostly of the same small nuclear RNAs (snRNAs) and proteins. The one exception is the small nuclear ribonucleoprotein particle (U1 snRNP), which is required for recognition of the 5′ splice site in *cis*-splicing.[30] By contrast, in *trans*-splicing, the 5′ splice site is contained on a special snRNP, the SL snRNP. This snRNP is consumed in the process of splicing: its 5′ end forms the 5′ ends of the mature mRNAs. In nematodes, this snRNP has been found to contain two proteins that are not found on any other snRNP and, in fact, have not been found outside the nematode phylum.[31] The protein components of the SL snRNPs from other phyla have not yet been investigated.

It is likely that trans-splicing and operons have arisen multiple times during evolution

EVOLUTION OF *TRANS*-SPLICING AND OPERONS

The close similarity between *cis*- and *trans*-splicing begs the question of which came first. *Cis*-splicing is found throughout the eukaryotes, but *trans*-splicing is found in only a few protists, several lower animals (nematodes, flatworms and cnidarians)[32] and primitive chordates. It is clearly missing from arthropods, plants, most protists, fungi and vertebrates. Since *trans*-splicing is at least superficially quite similar in all of the diverse organisms in which it is found, how likely is it that it has arisen multiple times versus the alternative hypothesis that *trans*-splicing arose early and was maintained only in certain species? Based on phylogeny alone, either hypothesis is viable; however, there is a compelling reason to argue that *trans*-splicing has arisen many times from *cis*-splicing, perhaps even in each lineage in which it is found today. The argument is that once *trans*-splicing exists, it is very difficult or impossible for an organism to lose it — it is essentially a one-way street. This is because the region of the RNA between the promoter and the *trans*-splice site is spliced out before the RNA is translated, so it can and does accumulate out-of-frame AUGs. These cause no harm as long as *trans*-splicing removes them, but, if trans-splicing were lost, they would prevent the correct reading frame from being translated. This occurs on virtually every mRNA that is *trans*-spliced, so if the *trans*-splicing process itself were lost, an unacceptable number of genes would not be expressed. This argues that every lineage that ever had *trans*-splicing still does. For example, although primitive chordates do have *trans*-splicing, they must have evolved it after they separated from the lineage that gave rise to vertebrates.

A very similar argument can be made concerning operons. As explained above,

the existence of *trans*-splicing made operons possible; it allowed the pre-mRNA from downstream genes to escape exonucleolytic degradation. Operons are also an evolutionary one-way street; once they have evolved, it is very difficult to lose them because several genes share a promoter. Most breakpoints involving operons would be expected to leave downstream genes without a promoter, and hence no way to be expressed. Indeed, it has been found that operons have been extraordinarily highly conserved between two species — *C. elegans* and *C. briggsae* — that are estimated to have diverged more than 100 million years ago.[33] Operons are also present in more distant nematodes[34,35] and, indeed, may be found throughout the phylum.

Genes encoding the basic machinery of gene expression and energy generations are overrepresented in *C. elegans* operons

WHAT GENES ARE TRANSCRIBED IN *C. ELEGANS* OPERONS?

C. elegans operons may contain mostly genes that are primarily regulated at the level of mRNA stability or translation

There are two central issues in considering whether operons serve to perform some sort of regulatory function or whether they are random assemblages of genes. The first is whether operons preferentially contain some gene classes, and the second is whether functionally related genes are in operons together. An analysis of ~2,600 genes shown to be contained in the *C. elegans* operons indicated that some classes of genes have a very strong tendency to be in operons whereas other classes of genes are never, or almost never, transcribed in operons.[14] Genes whose function is to specify mitochondrial proteins and those that encode the basic machinery for gene expression, transcription, splicing and translation have a very strong tendency to be transcribed in operons (Table 1). Even though only about 15 per cent of all genes are identified as operon genes in *C. elegans*, nearly 50 per cent of genes in these classes are in operons. The class of genes with the strongest tendency to be in operons is that which encodes proteins involved in RNA decay — 80 per cent of these genes are in operons. By contrast, tissue-specific genes, such as transcription factors, collagens and cytochrome P450 proteins, tend to be excluded from operons (Table 1). This strong asymmetry in types of genes in operons could reflect the fact that genes that are regulated at the level of transcription are excluded from operons, since multiple genes would end up being co-regulated identically from the same promoter. There are two broad ideas about what selects for genes to be in operons. In the first, the genes in operons are more efficiently co-regulated than genes that do not share a single promoter and regulatory site. It is possible that the genes contained in operons need to be able to respond to global signals so they can be efficiently repressed or activated as a group. In the second broad class of explanation, the selection is simply for a compact genome: operons dramatically reduce both the DNA between genes and the amount of DNA expended on regulatory sites. Operons may mostly serve to transcribe genes that do not need to be regulated at all, but which just need to be turned on in all tissues at all times. Another reasonable idea is that genes that are regulated at some level other than transcription have accumulated in operons because they can be expressed from the same promoter, but then regulated

Table 1: Some types of genes are transcribed in operons far more than other types

Gene class (number in class)	Percentage in operons
Peroxisome (21)	0
Cuticle (46)	0
Secretory vesicle (87)	2
Homeobox (88)	3
Protein phosphatase (106)	8
Cytoskeletal (84)	11
Motor proteins (50)	12
DNA replication (19)	26
Ribosomal protein (115)	39
Mitochondrial (212)	42
Spliceosome (55)	45
Cyclophilin (18)	50
RNA pol subunit (15)	52
tRNA synthetase (31)	52
RNA decay (15)	80

Data is from Blumenthal and Gleason.[14]

differentially at the level of mRNA stability or translation. These explanations are not mutually exclusive; different modes of selection may have resulted in the creation of different operons.

Very recent results have demonstrated that the primary driving force for genes to be included in operons may be the need for expression in the oogenic germ line. Reinke[36] has shown that almost all operons are expressed in the female germ line and that germ line expression does not depend on the function of the encoded proteins. One possible explanation is that the sustained high level of transcription needed during oogenesis might be facilitated by the sharing of promoters by many genes. Clearly, the basic machinery of energy generation and gene expression would be required during oogenesis, so that may explain why these classes of genes have such a strong tendency to be contained in operons.

C. elegans operons are expressed in the female germ line

ARE FUNCTIONALLY RELATED GENES TRANSCRIBED IN *C. ELEGANS* OPERONS TOGETHER?

At least some operons are used to co-regulate genes that function together. Examples include: an operon that expresses two subunits of the acetylcholine receptor;[37] an operon that encodes two proteins needed for modifying collagen, expressed only in collagen-producing cells;[38] an operon that co-expresses an ion channel protein with a protein that modifies the activity of that channel;[39] and an operon that expresses two of three proteins responsible for 3′ splice site recognition as well as a cyclophilin/RRM protein that has recently been identified as a spliceosome component.[40–42]

Many _C. elegans_ operons contain genes that are functionally related

In addition, there are many intriguing relationships between genes in operons that strongly suggest that related genes are found clustered in operons.[14] For example, there are operons that contain RNA polymerase I or III subunits and transcription factors that activate genes transcribed by that type of polymerase. There are numerous operons that co-express mitochondrial proteins: there is an operon containing two subunits of the exosome; there is an operon that encodes survivin, an inhibitor of apoptosis, along with SKIP — a transcription coactivator;[43] there is another that encodes the snRNP proteins U1A and U2B′′; and many more. Furthermore, operons containing related genes occur more frequently than expected by chance.[22] Nonetheless, it is premature to conclude that the presence of genes in an operon together implies a functional relationship. There are many operons where it is impossible to discern any relationship among the genes. It may be that, for these operons, the genes are expressed from the same promoter solely because they need to be expressed in the oogenic germ line.[36] It seems likely that operons form randomly by genomic rearrangements, and that most are selected against because they create unfavourable co-regulatory circuitry. Some are selected for because they allow sharing of promoters when there is no cost, however, and others are more strongly selected for because they result in favourable co-regulatory circuitry. It is in this latter group where the functional relationships among the genes in the operon can be discerned. In fact, there are now several instances in which functional relationships between genes have been found by virtue of their presence in operons together.[39,44]

CO-TRANSCRIPTIONAL PROCESSING OF POLYCISTRONIC PRE-mRNAS IN *C. ELEGANS*

The processing of the polycistronic pre-mRNAs transcribed from the *C. elegans* operons is a subject of considerable interest since it is carried out mostly by the same machinery that processes monocistronic loci, but it must do so somewhat differently (Figure 2). It is known that the 3′ end formation

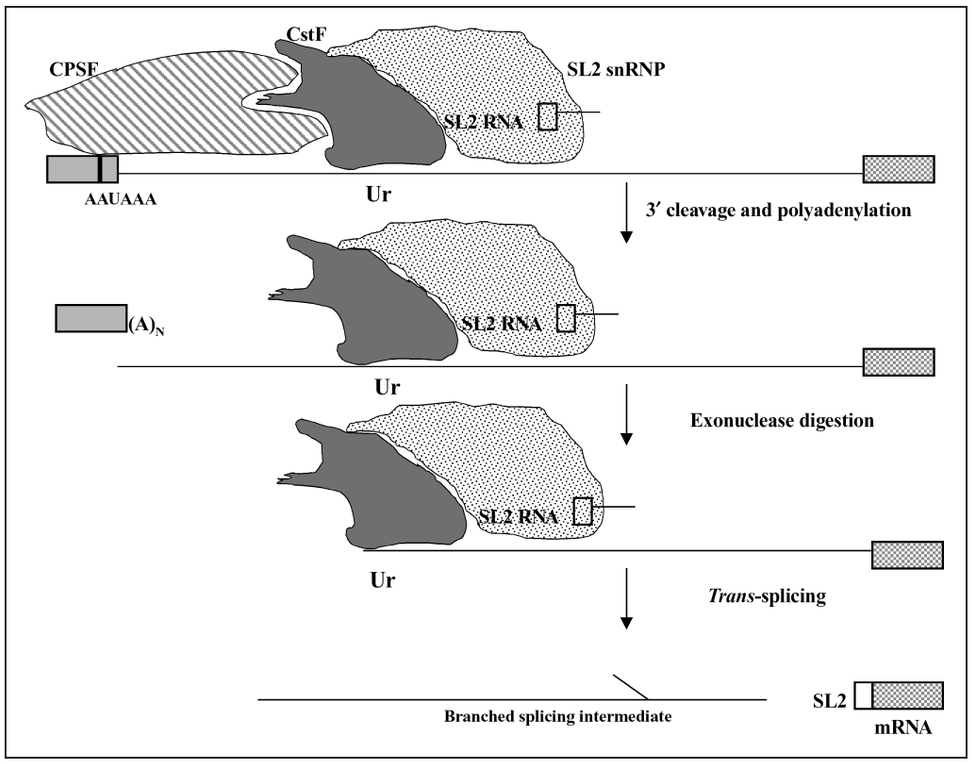

Figure 2: *C. elegans* operon processing model. The 3′ end formation proteins, CPSF and CstF bind cooperatively to signals on the pre-mRNA, AAUAAA and Ur, respectively. They specify cleavage at a site just 3′ of the AAUAAA.[45] The free 5′ phosphate remaining on the downstream RNA following cleavage is the signal for degradation by a nuclear exonuclease. Exonucleolytic cleavage is halted by the bound CstF just upstream of the Ur sequence.[9] The SL2 snRNP is brought to the reaction by a specific interaction with CstF.[46] SL2 *trans*-splicing to the 3′ splice site at the end of the inter-cistronic region completes the processing. The branched inter-cistronic RNA is discarded

Co-transcriptional processing of *C. elegans* operon pre-mRNAs involves clearage at the 3′ end of the upstream gene, partial digestion of the inter-cistronic RNA, and SL2 trans-splicing at the 5′ end of the downstream gene

machinery acting at the 3′ ends of upstream genes in operons plays a key role in the *trans*-splicing at the 5′ ends of the downstream genes. The 3′ end formation signal and *trans*-splice sites are usually only about 100 base pairs apart and the 3′ end formation signals must be intact to get efficient SL2-specific *trans*-splicing.[47] A U-rich sequence in the inter-cistronic region that probably acts as the binding site for the 3′ end formation factor, CstF, is absolutely required for the utilisation of SL2 and for accumulation of the downstream mRNAs.[48] Furthermore, a functionally important complex between CstF and the SL2 snRNP has been identified.[46] Recent results support a model in which 3′ end formation of the upstream gene occurs first, followed by 5′ to 3′ exonucleolytic cleavage of the downstream RNA, which continues until the exonuclease reaches the protein bound to the U-rich region, presumably CstF.[9] This allows downstream *trans*-splicing to occur to the SL2 bound to the CstF. It is likely that 3′ end formation and *trans*-splicing occur together in a concerted reaction.

It is known that these processes occur co-transcriptionally since it is usually difficult to detect the polycistronic precursor, even by reverse transcription polymerase chain reaction. One interesting question is how transcription termination within the operons is prevented following 3′ end formation, normally the signal for termination. Although the answer is not known, it is

tempting to speculate that it is SL2 *trans*-splicing itself that prevents termination. In the torpedo model for transcription termination, exonucleolytic cleavage following 3′ end cleavage results in termination by shortening the transcript of the elongating polymerase, thereby resulting in release of the polymerase from the template.[45] *Trans*-splicing accompanying 3′ end formation would provide a cap to the downstream mRNA and separate it from the exonuclease degrading from the site of 3′ end formation. The alternative model for transcription termination is that the RNA polymerase carrying the 3′ end formation machinery bound to the carboxy terminal domain of its largest subunit is altered in the process of cleavage so that it becomes less processive and is thus released from the template.[49] In this model, it is speculated that the fate of the polymerase is different within operons and at the 3′ ends of genes not upstream of other genes in operons, such that it does not become less processive. The presence of proteins at the U-rich region, or at the *trans*-splice site itself, could prevent the alteration of the state of RNA polymerase that normally accompanies 3′ end formation.

The mechanism by which the operons escape transcription termination following upstream genes is unknown

SL1-TYPE OPERONS

Remarkably, the *C. elegans* genome also contains relatively small numbers of a kind of operon that is processed differently from that discussed up to this point. These are termed SL1-type operons because the *trans*-splicing involves the more common snRNP, SL1, which is normally used at the 3′ ends of outrons.[50] In this kind of operon, there is no spacing between the site of 3′ end formation of the upstream gene and the site of *trans*-splicing by SL1 of the downstream gene; the two processes occur at the same site and they may be mutually exclusive. When 3′ end formation occurs first, *trans*-splicing of the downstream gene cannot occur because the *trans*-splice site is destroyed by the 3′ end cleavage. Thus, the two processes compete; each pre-mRNA can produce either the upstream or the downstream mature mRNA, but not both. If, however, the pre-mRNA is *trans*-spliced, it is possible that the free 3′ end created by *trans*-splicing could be polyadenylated, so both genes could be made from the same pre-mRNA. The mRNA for the upstream gene would presumably still have to be debranched to be functional, and it is not known whether this can occur.

Dicistronic mRNAs occur infrequently in *Drosophila* and other species

DICISTRONIC mRNAS IN OTHER SPECIES

Another kind of operon altogether is found in *Drosophila* and vertebrates and may also exist in plants. Unlike the operons that depend on *trans*-splicing to make mature monocistronic mRNAs from a multigene precursor, these operons are always composed of only two genes and the initial transcript is not processed between the genes. Instead, a dicistronic mRNA is made, transported to the cytoplasm and translated in that form. Since translation normally initiates at the first AUG in the mRNA, it is not clear how the second gene in the dicistronic mRNAs gets translated. Although the mechanism of translation has not been studied in the case of any of these operons, they presumably initiate translation on the second gene using an internal ribosome entry site or by some sort of re-initiation by the ribosomes that have just terminated translation on the first gene.

Is there something special about the genes that are co-translated with other genes? There may be. Many cases seem to involve related genes. In the tomato, γ-glutamyl kinase and γ-glutamyl phosphate reductase are expressed from a dicistronic mRNA that appears to be of bacterial origin.[23,51] In *Drosophila*, a dicistronic mRNA encodes the stoned A and B proteins, both of which are localised to nerve terminals and are involved in regulating endocytosis of synaptic vesicle proteins. This arrangement is conserved at least as far away as mosquito, but there is apparently no *stonedA* homologue in

vertebrates.[6,52] Also, in flies, the *Adh* and *Adh-related* genes are expressed from a dicistronic mRNA. In this case, two mRNAs are made — a monocistronic mRNA that encodes the ADH protein and a dicistronic mRNA that encodes both. Production of the latter is regulated at the level of 3′ end processing.[5] Phylogenetic analysis indicates that the dicistronic transcription of these two genes is the primitive state, but that it has been lost in at least some more recent *Drosophila* lineages.[53] A third *Drosophila* example is a dicistronic mRNA with two genes that encode proteins involved in meiotic recombination, mei-217 and mei-218. In this case, the two coding regions actually overlap, so it is likely that an internal ribosome entry site is responsible for initiating translation of the downstream gene.[54] In addition to these well-studied examples, the *Drosophila* genome project has identified 27 additional clusters that clearly encode dicistronic mRNAs and 17 others that probably do.[17] The author was able to identify *C. elegans* orthologues to both genes in 38 of these gene pairs and found that none of them were adjacent in worms (author's unpublished observations). This indicates that the *C. elegans* operons and the fly dicistronic clusters are not evolutionarily related. A thorough analysis of syntenies between genes in operons was recently published by Trachtulec.[55] While most members of worm and fly operons were not syntenic in other animals, this analysis uncovered several that were, suggesting an ancient association between members of these gene pairs.

Downstream genes in dicistronic mRNAs may be expressed by translational re-initiation or internal ribosome entry sites

In mammals, the clearest example is the MOCS2 operon that encodes the two subunits of molybopterin synthetase.[56] This arrangement is conserved in tunicates and arthropods.[55] In a second mammalian example, a dicistronic mRNA encodes both the growth and differentiation factor 1 (GDF-1) and a *trans*-membrane protein of unknown function (UOG-1). This arrangement is conserved between humans and mouse.[12] A third mammalian example is a dicistronic message that contains the SNRPN and SNURF genes. The former encodes the SmN protein and the latter encodes a protein of unknown function.[57] In both the GDF-1/UOG-1 and SNRPN/SNURF examples, one of the members of the gene pair is of unknown function, so whether these gene arrangements are selected to co-express genes of related function awaits further information.

ALTERNATIVE SPLICING AND BIFUNCTIONAL LOCI

A promoter can serve to produce two different and unrelated proteins by alternative splicing

In addition to operons, there are gene arrangements in eukaryotes in which a promoter can co-regulate genes that make separate mRNAs by alternative splicing. These arrangements might also be considered operons. They may also blur the definition of a gene. In these cases, a pre-mRNA may be spliced in two different ways to give rise to such different mRNAs that they are essentially products of different genes; however, the different mRNAs share one or more exons so they could also be considered alternative splice products of the same gene. For example, a very highly conserved gene pair was identified in *C. elegans*[7] that encodes two cholinergic proteins: *cha-1* encodes choline acetyl transferase and *unc-17* encodes acetylcholine transporter. The two genes share a non-coding first exon, but the entire coding region of *unc-17* is encoded in the large first intron of *cha-1*. So, any given pre-mRNA can produce the *cha-1* product or the *unc-17* product, but not both. Thus, this is an operon in the sense that the two genes are expressed from the same promoter, but they are not co-expressed, since production of one mRNA precludes the production of the other. Furthermore, this is a single gene in the sense of alternative splicing giving rise to alternative mRNAs from the same pre-mRNA, but it is two genes in the sense of one gene/one enzyme. This gene arrangement must have arisen early in the evolution of the animals and must confer a selective advantage since it is present in

vertebrates as well.[58,59] The *unc-60* locus of *C. elegans* has a very similar structure, except in this case the initiating methionine of both proteins is encoded by the common first exon.[60] Other examples of this type of gene arrangement include the fly gene for a ubiquitin/ribosomal protein fusion and a highly conserved protein of unknown function[61] and the *manx/bobcat* gene pair of ascidians.[8]

A different variation is exemplified by the *CREB* gene, which encodes a transcription factor. This gene can be alternatively spliced to include an exon that results in premature termination of the CREB protein followed by re-initiation of translation within the gene to produce short proteins that act as inhibitors of *CREB* activity.[62] Since this alternatively spliced mRNA encodes two proteins, it is equivalent to the dicistronic mRNAs discussed above.

Finally, there are two additional interesting cases that make it difficult to come up with clear definitions of either 'operon' or 'gene'. The first is the case of the *MOCS1A–MOCS1B* locus of vertebrates, which clearly produces a dicistronic mRNA, but does not produce protein from the downstream *MOCS1B* gene. Instead, the pre-mRNA is alternatively spliced to yield a bifunctional protein.[63] The second case concerns a pair of genes that are both involved in gene expression: one encodes eIF2γ, a translation initiation factor, and the other encodes SU(VAR)3-9, a heterochromatin protein. The two genes share a promoter and the first two exons, which can be spliced onto the rest of either gene. This arrangement is conserved in two other orders of holometabolic insects.[64] Is this one gene or two? Is it an operon or simply a case of alternative splicing?

Complex alternative splicing patterns of RNA expressed from a single promoter muddies the definition of the gene

CONCLUSIONS

Operons were once thought to be restricted to bacteria and archaea, but it has been known for some time that viruses and mitochondria also engage in polycistronic transcription. It is now clear that eukaryotic nuclei also contain multigene assemblages that fit the classical definition of an operon. Some are of the type where the polycistronic pre-mRNA is processed to yield monocistronic mature mRNAs by internal 3′ end formation accompanied by *trans*-splicing. This type of operon is found in trypanosomes, nematodes, flatworms and primitive chordates. The author predicts it will be found in all organisms that perform SL-type *trans*-splicing. The other type of operon is much more similar to the bacterial operons in that they produce a dicistronic mRNA. Finally, there are increasing numbers of examples of complex loci in which alternative splicing can give rise to multiple gene products which are sometimes unrelated to each other. All of these cases share one feature: multiple gene products are produced from the initiation of transcription by RNA polymerase at a single promoter. Hence, there is every reason to expect that these operons encode proteins that have a functional relationship to each other, as is the case with the bacterial and archael operons.

Acknowledgments

I am grateful to Peg MacMorris and Diego Zorio for helpful comments on the manuscript, and to Valerie Reinke for communication of results before publication.

References

1. Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960), 'Operon: a group of genes with the expression coordinated by an operator', *C. R. Hebd. Seances Acad. Sci.*, Vol. 250, pp. 1727–1729.
2. Agabian, N. (1990), 'Trans splicing of nuclear pre-mRNA', *Cell*, Vol. 61, pp. 1157–1160.
3. Spieth, J., Brook, G., Kuersten, S. *et al.* (1993), 'Operons in *C. elegans*: polycistronic mRNA precursors are processed by *trans*-splicing of SL2 to downstream coding regions', *Cell*, Vol. 73, pp. 521–532.
4. Blumenthal, T. and Spieth, J. (1996), 'Gene structure and organization in *Caenorhabditis elegans*', *Curr. Opin. Genet. Dev.*, Vol. 6, pp. 692–698.
5. Brogna, S. and Ashburner, M. (1997), 'The *Adh-related* gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger

RNA: multigenic transcription in higher organisms', *EMBO J.*, Vol. 16, pp. 2023–2031.

6. Andrews, J., Smith, M., Merakovsky, J. *et al.* (1996), 'The stoned locus of *Drosophila melanogaster* produces a dicistronic transcript and encodes two distinct polypeptides', *Genetics*, Vol. 143, pp. 1699–1711.
7. Alfonso, A., Grundahl, K., McManus, J. R. *et al.* (1994), 'Alternative splicing leads to two cholinergic proteins in *Caenorhabditis elegans*', *J. Mol. Biol.*, Vol. 241, pp. 627–630.
8. Swalla, B. J., Just, M. A., Pederson, E. L. and Jeffery, W. R. (1999), 'A multigene locus containing the *Manx* and *bobcat* genes is required for development of chordate features in the ascidian tadpole larva', *Development*, Vol. 126, pp. 1643–1653.
9. Liu, Y., Kuersten, S., Huang, T. *et al.* (2003), 'An uncapped RNA suggests a model for *Caenorhabditis elegans* polycistronic pre-mRNA processing', *RNA*, Vol. 9, pp. 677–687.
10. Davis, R. E. and Hodgson, S. (1997), 'Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*', *Mol. Biochem. Parasitol.*, Vol. 89, pp. 25–39.
11. Ganot, P., Kallesoe, T., Reinhardt, R. *et al.* (in press), 'SL RNA trans-splicing of polycistronic pre-mRNA in the chordate *Oikopleura dioica*', *Mol. Cell. Biol.*, Vol. 24, pp. 7795–7805.
12. Lee, S. J. (1991), 'Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure', *Proc. Natl. Acad. Sci. USA*, Vol. 88, pp. 4250–4254.
13. Blumenthal, T. (1998), 'Gene clusters and polycistronic transcription in eukaryotes', *Bioessays*, Vol. 20, pp. 480–487.
14. Blumenthal, T. and Gleason, K. S. (2003), '*Caenorhabditis elegans* operons: form and function', *Nat. Rev. Genet.*, Vol. 4, pp. 112–120.
15. Lawrence, J. G. (2002), 'Shared strategies in gene organization among prokaryotes and eukaryotes', *Cell*, Vol. 110, pp. 407–413.
16. Nimmo, R. and Woollard, A. (2002), 'Widespread organisation of *C. elegans* genes into operons: Fact or function?', *Bioessays*, Vol. 24, pp. 983–987.
17. Misra, S., Crosby, M. A., Mungall, C. J. *et al.* (2002), 'Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review', *Genome Biol.*, Vol. 3, p. RESEARCH0083.
18. Worthey, E. A., Martinez-Calvillo, S., Schnaufer, A. *et al.* (2003), ' *Leishmania* major chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene', *Nucleic Acids Res.*, Vol. 31, pp. 4201–4210.
19. Martinez-Calvillo, S., Nguyen, D., Stuart, K. and Myler, P. J. (2004), 'Transcription initiation and termination on *Leishmania* major chromosome 3', *Eukaryot. Cell*, Vol. 3, pp. 506–517.
20. Gao, G., Nara, T., Nakajima-Shimada, J. and Aoki, T. (1999), 'Novel organization and sequences of five genes encoding all six enzymes for *de novo* pyrimidine biosynthesis in *Trypanosoma cruzi*', *J. Mol. Biol.*, Vol. 285, pp. 149–161.
21. Zorio, D. A., Cheng, N. N., Blumenthal, T. and Spieth, J. (1994), 'Operons as a common form of chromosomal organization in *C. elegans*', *Nature*, Vol. 372, pp. 270–272.
22. Blumenthal, T., Evans, D., Link, C. D. *et al.* (2002), 'A global analysis of *Caenorhabditis elegans* operons', *Nature*, Vol. 417, pp. 851–854.
23. Garcia-Rios, M., Fujita, T., LaRosa, P. C. *et al.* (1997), 'Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase', *Proc. Natl. Acad. Sci. USA*, Vol. 94, pp. 8249–8254.
24. Gilson, P. R. and McFadden, G. I. (1996), 'The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns', *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 7737-7742.
25. Wilusz, C. J., Wormington, M. and Peltz, S. W. (2001), 'The cap-to-tail guide to mRNA turnover', *Nat. Rev. Mol. Cell. Biol.*, Vol. 2, pp. 237–246.
26. Birse, C. E., Minvielle-Sebastia, L., Lee, B. A. *et al.* (1998), 'Coupling termination of transcription to messenger RNA maturation in yeast', *Science*, Vol. 280, pp. 298–301.
27. Nilsen, T. W. (1993), '*Trans*-splicing of nematode premessenger RNA', *Annu. Rev. Microbiol.*, Vol. 47, pp. 413–440.
28. Vandenberghe, A. E., Meedel, T. H. and Hastings, K. E. (2001), 'mRNA 5-leader trans-splicing in the chordates', *Genes Dev.*, Vol. 15, pp. 294–303.
29. Davis, R. E. (1997), 'Surprising diversity and distribution of spliced leader RNAs in flatworms', *Mol. Biochem. Parasitol.*, Vol. 87, pp. 29–48.
30. Hannon, G. J., Maroney, P. A. and Nilsen, T. W. (1991), 'U small nuclear ribonucleoprotein requirements for nematode *cis*- and *trans*-splicing *in vitro*', *J. Biol. Chem.*, Vol. 266, pp. 22792–22795.
31. Denker, J. A., Zuckerman, D. M., Maroney, P. A. and Nilsen, T. W. (2002), 'New components of the spliced leader RNP required for nematode *trans*-splicing', *Nature*, Vol. 417, pp. 667–670.

32. Stover, N. A. and Steele, R. E. (2001), '*Trans*-spliced leader addition to mRNAs in a cnidarian', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 5693–5698.

33. Stein, L. D., Bao, Z., Blasair, D. *et al.* (2003), 'The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics', *PLoS Biol.*, Vol. 1, p. E45.

34. Evans, D., Zorio, D., MacMorris, M. *et al.* (1997), 'Operons and SL2 trans-splicing exist in nematodes outside the genus *Caenorhabditis*', *Proc. Natl. Acad. Sci. USA*, Vol. 94, pp. 9751–9756.

35. Lee, K. Z. and Sommer, R. J. (2003), 'Operon structure and trans-splicing in the nematode *Pristionchus pacificus*', *Mol. Biol. Evol.*, Vol. 20, pp. 2097–2103.

36. Reinke, V. (in press), 'Oogenic germline expression influences operon formation in *C. elegans*', *PLoS Biol.*

37. Treinin, M., Gillo, B., Liehman, L. *et al.* (1998), 'Two functionally dependent acetylcholine subunits are encoded in a single *Caenorhabditis elegans* operon', *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp. 15492–15495.

38. Page, A. P. (1997), 'Cyclophilin and protein disulfide isomerase genes are co-transcribed in a functionally related manner in *Caenorhabditis elegans*', *DNA Cell Biol.*, Vol. 16, pp. 1335–1343.

39. Furst, J., Ritter, M., Rudzki, J. *et al.* (2002), 'ICln ion channel splice variants in *Caenorhabditis elegans*: Voltage dependence and interaction with an operon partner protein', *J. Biol. Chem.*, Vol. 277, pp. 4435–4445.

40. Zorio, D. A. and Blumenthal, T. (1999), 'U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in *Caenorhabditis elegans*', *RNA*, Vol. 5, pp. 487–494.

41. Mazroui, R., Puoti, A. and Kramer, A. (1999), 'Splicing factor SF1 from *Drosophila* and *Caenorhabditis*: Presence of an N-terminal RS domain and requirement for viability', *RNA*, Vol. 5, pp. 1615–1631.

42. Jurica, M. S., Licklider, L. J., Gygi, S. R. *et al.* (2002), 'Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis', *RNA*, Vol. 8, pp. 426–439.

43. Kostrouchova, M., Kostrouch, Z., Saudek, V. *et al.* (2003), '*BIR-1*, a *Caenorhabditis elegans* homolog of *Survivin*, regulates transcription and development', *Proc. Natl. Acad. Sci. USA*, Vol. 100, pp. 5240–5245.

44. Eichmuller, S., Vezzoli, V., Bazzini, C. *et al.* (2004), 'A new gene-finding tool: Using the *Caenorhabditis elegans* operons for identifying functional partner proteins in human cells', *J. Biol. Chem.*, Vol. 279, pp. 7136–7146.

45. Proudfoot, N. J., Furger, A. and Dye, M. J. (2002), 'Integrating mRNA processing with transcription', Vol. 108, pp. 501–512.

46. Evans, D., Perez, I., MacMorris, M. *et al.* (2001), 'A complex containing CstF-64 and the SL2 snRNP connects mRNA 3′-end formation and *trans*–splicing in *C. elegans* operons', *Genes Dev.*, Vol. 15, pp. 2562–2571.

47. Kuersten, S., Lea, K., MacMorris, M. *et al.* (1997), 'Relationship between 3′-end formation and SL2-specific *trans*-splicing in polycistronic *Caenorhabditis elegans* pre-mRNA processing', *RNA*, Vol. 3, pp. 269–278.

48. Huang, T., Kuersten, S., Deshpande, A. M. *et al.* (2001), 'Intercistronic region required for polycistronic pre-mRNA processing in *Caenorhabditis elegans*', *Mol. Cell. Biol.*, Vol. 21, pp. 1111–1120.

49. Sadowski, M., Dichtl, B., Hubner, W. *et al.* (2003), 'Independent functions of yeast Pcf11p in pre-mRNA 3′ end processing and in transcription termination', *EMBO J.*, Vol. 22, pp. 2167–2177.

50. Williams, C., Xu, L. and Blumenthal, T. (1999), 'SL1 *trans*-splicing and 3′-end formation in a novel class of *Caenorhabditis elegans* operon', *Mol. Cell. Biol.*, Vol. 19, pp. 376–383.

51. Fujita, T., Maggio, A., Garcia-Rios, M. *et al.* (1998), 'Comparative analysis of the regulation of expression and structures of two evolutionarily divergent genes for Delta1-pyrroline-5-carboxylate synthetase from tomato', *Plant Physiol.*, Vol. 118, pp. 661–674.

52. Estes, P. S., Jackson, T. C., Stimson, D. T. *et al.* (2003), 'Functional dissection of a eukaryotic dicistronic gene: Transgenic *stonedB*, but not *stonedA*, restores normal synaptic properties to *Drosophila* stoned mutants', *Genetics*, Vol. 165, pp. 185–196.

53. Betran, E. and Ashburner, M. (2000), 'Duplication, dicistronic transcription, and subsequent evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in *Drosophila*', *Mol. Biol. Evol.*, Vol. 17, pp. 1344–1352.

54. Liu, H., Jang, J. K., Graham, J. *et al.* (2000), 'Two genes required for meiotic recombination in *Drosophila* are expressed from a dicistronic message', *Genetics*, Vol. 154, pp. 1735–1746.

55. Trachtulec, Z. (2004), 'Eukaryotic operon genes can define highly conserved syntenies', *Folia Biol. (Praha)*, Vol. 50, pp. 1–6.

56. Reiss, J., Dorche, C., Stallmeyer, B. *et al.* (1999), 'Human *molybdopterin synthase* gene: genomic structure and mutations in molybdenum cofactor deficiency type B', *Am. J. Hum. Genet.*, Vol. 64, pp. 706–711.

57. Gray, T. A., Saitoh, S. and Nicholls, R. D. (1999), 'An imprinted, mammalian bicistronic

transcript encodes two independent proteins', *Proc. Natl. Acad. Sci. USA*, Vol. 96, pp. 5616–5621.

58. Bejanin, S., Cervini, R., Mallet, J. *et al.* (1994), 'A unique gene organization for two cholinergic markers, choline acetyltransferase and a putative vesicular transporter of acetylcholine', *J. Biol. Chem.*, Vol. 269, pp. 21944–21947.
59. Erickson, J. D., Varoqui, H., Schafer, M. K. *et al.* (1994), 'Functional identification of a vesicular acetylcholine transporter and its expression from a 'cholinergic' gene locus', *J. Biol. Chem.*, Vol. 269, pp. 21929–21932.
60. McKim, K. S., Matheson, C., Marra, M. A. *et al.* (1994), 'The *Caenorhabditis elegans unc-60* gene encodes proteins homologous to a family of actin-binding proteins', *Mol. Gen. Genet.*, Vol. 242, pp. 346–357.
61. Mottus, R. C., Whitehead, I. P., O'Grady, M. *et al.* (1997), 'Unique gene organization: Alternative splicing in *Drosophila* produces two structurally unrelated proteins', *Gene*, Vol. 198, pp. 229–236.
62. Walker, W. H., Girardet, C. and Habener, J. F. (1996), 'Alternative exon splicing controls a translational switch from activator to repressor isoforms of transcription factor CREB during spermatogenesis', *J. Biol. Chem.*, Vol. 271, pp. 20145–21050.
63. Gray, T. A. and Nicholls, R. D. (2000), 'Diverse splicing mechanisms fuse the evolutionarily conserved bicistronic MOCS1A and MOCS1B open reading frames', *RNA*, Vol. 6, pp. 928–936.
64. Krauss, V. and Reuter, G. (2000), 'Two genes become one: The genes encoding heterochromatin protein Su(var)3-9 and translation initiation factor subunit eIF-2gamma are joined to a dicistronic unit in holometabolic insects', *Genetics*, Vol. 156, pp. 1157–1167.