Genome evolution and the evolution of exon-shuffling — a review

László Patthy *
Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest, Hungary

Received 4 February 1999; received in revised form 5 May 1999; accepted 1 June 1999

Abstract

Recent studies on the genomes of protists, plants, fungi and animals confirm that the increase in genome size and gene number in different eukaryotic lineages is paralleled by a general decrease in genome compactness and an increase in the number and size of introns. It may thus be predicted that exon-shuffling has become increasingly significant with the evolution of larger, less compact genomes. To test the validity of this prediction, we have analyzed the evolutionary distribution of modular proteins that have clearly evolved by intronic recombination. The results of this analysis indicate that modular multidomain proteins produced by exon-shuffling are restricted in their evolutionary distribution. Although such proteins are present in all major groups of metazoa from sponges to chordates, there is practically no evidence for the presence of related modular proteins in other groups of eukaryotes. The biological significance of this difference in the composition of the proteomes of animals, fungi, plants and protists is best appreciated when these modular proteins are classified with respect to their biological function. The majority of these proteins can be assigned to functional categories that are inextricably linked to multicellularity of animals, and are of absolute importance in permitting animals to function in an integrated fashion: constituents of the extracellular matrix, proteases involved in tissue remodelling processes, various proteins of body fluids, membrane-associated proteins mediating cell–cell and cell–matrix interactions, membrane associated receptor proteins regulating cell–cell communications, etc. Although some basic types of modular proteins seem to be shared by all major groups of metazoa, there are also groups of modular proteins that appear to be restricted to certain evolutionary lineages.

In summary, the results suggest that exon-shuffling acquired major significance at the time of metazoan radiation. It is interesting to note that the rise of exon-shuffling coincides with a spectacular burst of evolutionary creativity: the Big Bang of metazoan radiation. It seems probable that modular protein evolution by exon-shuffling has contributed significantly to this accelerated evolution of metazoa, since it facilitated the rapid construction of multidomain extracellular and cell surface proteins that are indispensable for multicellularity. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Genome compactness; Introns; Modular proteins; Modules; Metazoan evolution

1. Introduction

Shortly after the discovery of split genes, it was realized that the existence of introns may have dramatic consequences on protein evolution (Gilbert, 1978). It was pointed out that recombination within introns could assort exons independently, and middle repetitious sequences in introns may create hotspots for recombin-ation to shuffle the exonic sequences. The presence of introns in most eukaryotic protein-coding genes and their absence from prokaryotes was explained by two types of hypotheses. The ‘introns early’ hypotheses assumed that introns and RNA splicing are the relics of the RNA world and the ‘genes in pieces’ organization of the eukaryotic genome is the original, ancestral form (Darnell, 1978; Doolittle, 1978; Darnell and Doolittle, 1986; Gilbert, 1986). According to this view, eukaryotes retained introns and the genetic plasticity of the primitive ancestors of all cells. On the other hand, bacteria gained increased efficiency by eliminating their introns. Supporters of the introns-early hypotheses assume that the introns of all protein-coding genes reflect the assembly of these genes from pieces; that exons do indeed correspond to building blocks (α-helices, β-sheets, etc.) from which all the genes were assembled by intronic recombination (Gilbert and Glynias, 1993).
In contrast with this, the ‘introns late’ theories suggest that the prokaryotic genes resemble the ancestral ones and that the introns were inserted later in genes of eukaryotes (Crick, 1979; Orgel and Crick, 1980). Cavalier-Smith, 1985; Cech, 1985; Sharp, 1985). In fact, it is now obvious that the exon-intron structure of eukaryotic protein-coding genes is not static: introns are continually inserted into (as well as removed from) genes. The actual mechanisms of insertion, propagation of some self-splicing introns have been analyzed in detail and the mechanisms responsible for the insertion of splicosomal introns are also becoming clear (Dujon, 1989; Lambowitz and Belfort, 1989; Perlman and Butow, 1989; Mori and Schmelzer, 1990; Belfort, 1991; Lambowitz, 1993; Mueller et al., 1993; Grivell, 1994; Patthy, 1995).

Since introns themselves are subject to evolution, it is clear that exon-shuffling has been evolving parallel with the evolution of introns. We have argued previously that the introns suitable for exon-shuffling appeared at a relatively late stage of evolution; therefore, exon-shuffling could not play a major role in the construction of ancient proteins (Patthy, 1987, 1991a,b). The self-splicing introns of the RNA world that could be present at the time the first proteins were formed are practically unsuitable for exon-shuffling by intronic recombination: such self-splicing introns encode an essential function, therefore their sequence is not tolerant to intronic recombination (Patthy, 1987, 1991a,b). Exon-shuffling could become significant only with the appearance of splicosomal introns: these introns play a negligible role in their own excision, therefore intronic recombination is less likely to produce recombinant introns that are deficient in splicing. Furthermore, the nonessential parts of splicosomal introns could accommodate large segments of middle repetitious sequences, further increasing the chances of intronic recombination. Since splicosomal introns evolved relatively recently from group II self-splicing introns (Cech, 1986; Jacquier, 1990; Cavalier-Smith, 1991; Coperino and Hallick, 1993; Saldanha et al., 1993; Sharp, 1994) and are restricted in their evolutionary distribution (Cavalier-Smith, 1991; Palmer and Logsdon, 1991; Logsdon, 1991) exon-shuffling could play a major role only in the construction of ‘younger’ proteins (Patthy, 1987; 1991a,b, 1994, 1995, 1996).

In the present review I wish to emphasize that the significance of exon-shuffling increased parallel with the evolution of less compact genomes. The basis of this correlation is that the number and size of introns and the proportion of repetitive sequences in introns increases parallel with the decrease of genome compactness, therefore the chances of exon-shuffling by intronic recombination are increased. Analysis of the evolutionary distribution of proteins that were clearly assembled from modules by intronic recombination suggests that exon-shuffling became significant at the time of the appearance of the first multicellular animals, and that the rise of exon-shuffling could in fact contribute to the explosive nature of metazoan radiation.

2. Results and discussion

2.1. Introns and evolution of genome compactness

In the past few years, the complete sequences of the genomes of several Eubacteria (Escherichia coli, Bacillus subtilis, Haemophilus influenzae, Borrelia burgdorferi, Mycoplasma pneumoniae, Mycoplasma genitalium, etc.), Archaea (Methanococcus jannaschii, Archaeoglobus fulgidus, etc.), a unicellular eukaryote (Saccharomyces cerevisiae), and a multicellular animal (Caenorhabditis elegans) have been determined, and significant progress has also been made on the genome of a protozoan parasite (Plasmodium falciparum), the plant, Arabidopsis thaliana, the fruitfly, Drosophila melanogaster and various vertebrates (e.g., the fish Fugu rubripes, mouse, human). As a result of these studies it has become clear that there are clear correlations between genome size, genome compactness and the proportion of the genome that is occupied by introns and repetitive elements.

At one extreme we find the small, compact genomes of Eubacteria and Archaea (with limtingly high coding density) which are practically devoid of introns and contain very little repetitive sequences. Vertebrate genomes represent the other extreme: they have large genomes with a drastically reduced coding-density, their genomes are rich in repetitive sequences and their genes are characterized by a high intron to exon ratio.

Studies on the complete genomes of H. influenzae (Fleischmann et al., 1995), M. genitalium (Fraser et al., 1995), M. pneumoniae (Himmelreich et al., 1996, 1997), B. subtilis (Kunst et al., 1997), E. coli (Blattner et al., 1997), Borrelia burgdorferi (Fraser et al. 1997) have revealed that the genomes of these eubacteria are very compact. The major part of their genomes (about 90%) is dedicated to protein-coding genes (830–976 genes per Mbp) and they do not contain large quantities of intergenic DNA, introns or repetitive elements, reflecting the fact that there is strong selective pressure to eliminate nongenic DNA.

A limitingly high coding density also seems to hold for Archaea, genes covering about 90% of their compact genomes. For example, in the case of the genome of M. jannaschii (Bult et al., 1996) coding density is similar to that of eubacteria: 1022 genes per Mbp. Similarly, in the genome of another Archaea, A. fulgidus (Klenk et al., 1997) there are 1107 genes per Mbp, genes covering 92.2% of the genome.

The genome of the unicellular eukaryote, S. cerevisiae is significantly less compact than those of eubacterial or
archaeal prokaryotes: open reading frames occupy ‘only’ 68% of the yeast genome (Bassett et al., 1996; Clayton et al., 1997; Dujon, 1997; Oliver, 1997). Thus, compared with eubacterial or Archaeal genomes, where ORFs occupy about 90% of their genomes, yeast seems to be under weaker selective pressure to reduce genome size. (Whereas in Archaea and Eubacteria there are about 830-1100 genes per Mbp, in the case of yeast this number is only 446.) Nevertheless, introns of yeast protein-coding genes are still few (233), and short: less than 4% of the genes contain introns, and introns account for less than 1% of the entire genome. Compared with higher eukaryotes, the yeast genome is compact due to the short size of intergenic regions, that introns are few and small, and that repetitive sequences are infrequent.

The genome of the protozoan parasite, Plasmodium falciparum is 30 Mb, about twice as large as that of yeast. Analysis of chromosome 2 of P. falciparum revealed that it contains 0.95 Mbp and encodes 209 protein-coding genes (Gardner et al., 1998). The estimated gene density is thus less than half of that found in the case of yeast: 221 genes per Mbp. As compared with the yeast genome, the decrease in genome compactness is also reflected in a significantly increased proportion of introns: 43% of the protein-coding genes contain at least one intron.

The 100-120 Mbp genome of A. thaliana, a small crucifer weed, is significantly less compact than the yeast genome. Analysis of chromosome 2 of P. falciparum revealed that it contains 0.95 Mbp and encodes 209 protein-coding genes (Gardner et al., 1998). The estimated gene density is thus less than half of that found in the case of yeast: 221 genes per Mbp. As compared with the yeast genome, the decrease in genome compactness is also reflected in the fact that the majority of Arabidopsis genes contain introns. A. thaliana introns are usually small (<200 bp), their average size is 146 bp (Hebgaard et al., 1996; The EU Arabidopsis Genome Project, 1998) as compared with 446 gene per Mbp for yeast. This decrease in genome compactness is also reflected in the fact that the majority of Arabidopsis genes contain introns. A. thaliana introns are usually small (<200 bp), their average size is 146 bp (Hebgaard et al., 1996; The EU Arabidopsis Genome Project, 1998). Consistent with decreased genome compactness, repeat classes already account for a significant proportion (about 19%) of genomic sequence.

Sequencing of the 97 Mbp C. elegans genome — containing approx. 19 000 genes — has essentially been completed, permitting the recognition of some basic characteristics of the genome of this simple, multicellular animal (Blumenthal and Speth, 1996; Blaxter, 1998; Chalfie, 1998; Chervitz et al., 1998; Ruvkun and Hobert, 1998; The C. elegans Sequencing Consortium, 1998). This genome is also less compact than that of yeast: whereas in yeast there are 446 genes per Mbp, in the case of C. elegans this value is only 196. The majority of C. elegans protein-coding genes have introns and most have multiple introns. Coding sequence accounts for 27% of the genome and introns (on average five per gene) account for another 26%. The C. elegans genome is significantly more compact than the large (3000 Mbp) human genome, which is estimated to have only about 20 genes per Mbp. The compactness of the C. elegans genome (relative to vertebrate genomes) is due to the fact that introns and intergenic distances are significantly shorter. More than half of the C. elegans introns are shorter than 60 bp, the most common length being only 48 bp. This tendency may be best illustrated by the fact that the size of C. elegans genes is usually much smaller than that of homologous vertebrate genes and they usually contain fewer introns. For example, the nematode α1(IV) and α2(IV) collagen genes are about 9 kb long with only 11 and 19 introns, whereas the homologous human type α2(IV), α5(IV) and α6(IV) collagen genes are 100-200 kb long with 46, 50 and 44 introns (Sibley et al., 1993; Oohashi et al., 1995), the C. elegans osteonectin gene spans 3.6 kb and has five introns, whereas the mammalian homolog spans 26 kb and has nine introns (McVey et al., 1988; Schwarzbauer and Spencer, 1993). As a reflection of the compactness of the C. elegans genome, tandem repeats account for only 2.7% of the genome, and inverted repeats account for 3.6%.

The genome of D. melanogaster is about 170 Mbp and contains about 12 000 protein-coding genes (Rubin, 1998). Although the estimated gene density (71 genes per Mbp) is significantly lower than in yeast, the genome is more compact than most vertebrate genomes. The D. melanogaster genes are rather compact, 50% of the introns are <100 bp. In general, D. melanogaster genes tend to have shorter and fewer introns than their vertebrate homologs. For example, the fly laminin A gene is 14 kb long with 14 introns, whereas the human homolog occupies more than 200 kb and has 63 introns (Kuschke-Gullberg et al., 1992; Zhang et al., 1996). Recent studies on the genome of the sea squirt Ciona intestinalis seem to confirm the hypothesis that simple chordates had a gene number very similar to invertebrates (Simmen et al., 1998). For a genome of about 162 Mbp C. intestinalis has only about 15 500 genes, a number very close to the 12000-19 000 genes of D. melanogaster and C. elegans. The gene density of this invertebrate chordate (97 genes per Mbp) is about half of that in C. elegans, only about 13% of the genome codes for protein (Simmen et al., 1998).

The inverse relation between genome compactness and the proportion of intronic and repetitive sequence may also be illustrated by comparison of the genome of the pufferfish, F. rubripes with those of other vertebrates. The Fugu genome is only 400 Mbp, about 7.5 times smaller than that of human, although these organisms have about the same number of genes: the F. rubripes genome is thus 7.5 times more compact than the human genome (Elgar et al., 1996). Fugu genes usually have the same exon-intron organization as their mammalian homologs but, despite this conservation, the size of the Fugu introns is usually drastically shorter than in the
mammalian homologs, the majority of introns being between 60 and 150 bp. An illustrative example is the Huntington’s disease gene, which is ‘only’ 22 kb in Fugu as compared with the 180 kb human gene (i.e. an eightfold difference in size). Importantly, the exon–intron organization of this gene is identical to that of the human homolog: reduction occurred at the expense of noncoding regions and introns. The gene structure of complement component C9 of the pufferfish also illustrates these features of the Fugu genome (Yeo et al., 1997). The 11 exons of the Fugu C9 gene span 2.9 kb of genomic DNA, whereas the 11 exons of human C9 span 90 kb, representing a 30-fold difference in size at the expense of intron size. The compactness of the Fugu genome (relative to the human genome) is due to the fact that it contains significantly less nongenic DNA.

There are no abundant classes of dispersed repeats in Fugu and all forms of repetitive DNA (including telomeric repeats, ribosomal RNAs) constitute less than 10% of the Fugu genome. The average gene density in F. rubripes is quite similar to that in C. elegans: 150 genes per Mb (in human this value is 20).

In summary, the evidence from genome projects suggests that prokaryotes are characterized by small compact genomes with little space for intergenic DNA, introns or repetitive sequences. In eukaryotes multiple replication origins have permitted the evolution of larger, less compact genomes which can accommodate increasing amounts of intergenic regions, introns and repetitive sequences. The inverse relationship between genome compactness and intron number and intron size of protein-coding genes is valid not only for entire genomes, but seems to hold even for different isochores of a given genome. In vertebrate genomes, genes are not evenly distributed in different isochores: for example, the most GC-rich H3 isochore of the human genome contains about 28% of the genes, although H3 accounts for only 3–5% of the genome. Conversely, only 34% of the human genes are found in the GC-poor L1 + L2, although they represent 62% of the genome (Mouchiroud et al., 1991; Bernardi, 1993). In other words, GC-poor isochores appear to be gene deserts as compared to the most GC-rich isochores: the gene density is four to eight times higher in H3 than in H1 + H2 and about 10-20 times higher than in L1 + L2 (Mouchiroud et al., 1991; Bernardi, 1993). Significantly for our present discussion, protein-coding genes of GC-poor isochores contain more introns per kb of coding sequence than those in GC-rich isochores, and the intervening sequences are, on average, three times longer for genes located in L1 + L2 than for those found in H3 isochores (Duret et al., 1995). The ratio of intervening sequence to coding sequences is thus strikingly different for L1 + L2 and H3: genes in H3 are on average 2.5 times more compact than those in L1 + L2 (Duret et al., 1995). The genes of genomic regions of greater compactness thus have fewer and shorter introns than genes located in the least compact genomic regions.

These observations thus suggest that the evolutionary forces controlling genome compactness have a direct influence on the frequency and size of introns, as well as the relative abundance of repetitive elements in genomes. It may thus be predicted that the evolution of less compact genomes was paralleled by an increase in the potential for intronic recombination and exon-shuffling. To test this prediction we have analyzed the evolutionary distribution of protein-coding genes constructed by exon-shuffling via intronic recombination.

### 2.2 Identification of cases of exons-shuffling

In order to establish a case of modular protein evolution by exon-shuffling, one has to show (1) that two unquestionably homologous modules (i.e., modules derived from a common ancestor) are present in otherwise nonhomologous protein environments and (2) that the transposition of the module was mediated by intronic recombination. In this review I will discuss only cases where both these criteria are satisfied.

It must be emphasized that frequently it is not a trivial task to prove homology of modules or to prove a role of introns in module-shuffling. Introns-early versions of the exon-shuffling theory that assume that all primordial proteins were assembled by intronic recombination from short exons that encoded various secondary structural elements (Dorit et al., 1990) face the formidable task of proving that in two otherwise unrelated proteins two structurally similar α-helices, membrane spanning domains or β-sheets, are similar due to common ancestry rather than due to convergence. Obviously, such basic structural units were invented independently several times during evolution (Pathy, 1991a).

Even if criterion (1) is satisfied, we cannot take it for granted that exon-shuffling was responsible for module-shuffling. Although exon-shuffling by intronic recombination is by far the most powerful mechanism of modular protein evolution, this does not mean that it is the only way to exchange domains among protein-coding genes (Pathy, 1996). Some recent examples show how modular proteins of bacteria may be constructed without the assistance of introns. For example, a modular protein of Peptostreptococcus magnus has been shown to be the product of a recent intergenic recombination of two different types of streptococcal surface proteins (de Chateau and Bjorek, 1994). The recent transfer of a fragment of a prokaryotic gene to another shows that introns are not absolutely essential for exchange of gene fragments. These studies have also shown that gene rearrangements by exomic recombination may be facilitated by the presence of special recombinogenic DNA sequences in intermodule linker regions, and that antibi-
Fig. 1. Schematic representation of the structures of the genes of some modular proteins assembled from the class 1-1 C-type lectin-module (LN), epidermal growth factor-module (E), complement B-type module (B), immunoglobulin-module (Ig), link protein-module (LK), follistatin-module (FS) and Kunitz-type inhibitor module (INH). The numbers indicate the position and phase class of the introns. Phase 1 introns found at the boundaries of modules are highlighted in red. The black boxes indicate signal peptide domains, vertical black bars represent transmembrane domains. The boxes designating the different domains are drawn to scale.

To provide unambiguous evidence for exon-shuffling, one has to show that shuffling of the module was mediated by flanking introns that belonged to the same phase, i.e., the module was “symmetrical” in accordance with the phase-compatibility-rules of exon-shuffling (Patthy, 1987). Many protein-coding genes produced by exon-shuffling could be recognized by a striking correlation between the exon-structure of the genes and the domain-organization of proteins (Patthy, 1985, 1987, 1991b). As a consequence of the phase-compatibility rule, the introns of these modular proteins also show a marked intron-phase bias: e.g., in the genes of modular proteins produced by exon-shuffling of class 1-1 mod-

Fig. 2. Comparison of the exon–intron structures of genes of some modular vertebrate proteins with those of invertebrate homologs. (A) Comparison of the structure of the Drosophila and human tolloid genes. (B) Comparison of the structure of the genes of the Caenorhabditis and human netrin receptors. The numbers indicate the position and phase class of the introns. Phase 1 introns found at the boundaries of modules are highlighted in red. The black boxes indicate signal peptide domains, vertical black bars represent transmembrane domains. The boxes designating the different domains are drawn to scale.
ules, the introns that participated in the assembly process are all phase 1 (Patthy, 1987, 1991a). The genes of selectins (e.g., granule membrane protein 140 of platelets, endothelial-leukocyte adhesion molecule 1, lymphocyte homing receptor), interleukin-2 receptor, factor XIIIb subunit, cartilage link protein, follistatin, lipoprotein-associated coagulation inhibitor provide typical examples of such a correlation: their C-type lectin-, growth factor-, complement B-type modules, immunoglobulin-, link-, follistatin modules, Kunitz-type inhibitor modules are encoded by discrete class 1-1 exons, i.e., all the intermodule introns are phase 1 (Fig. 1). The fact that the structure of the gene of a modular protein conforms to these rules may actually be used as evidence that it has evolved by exon-shuffling (Patthy, 1988).

There are many cases where the correlation between modular structure of a multidomain protein (consisting of class 1-1 modules) and exon-intron structure of its gene is less perfect, since some of the ‘expected’ introns are missing from the module boundaries (Patthy, 1994; 1995; 1996). The most plausible explanation for the weaker correlation between exon–intron structure of the gene and domain structure of the protein is that the original genomic organization has been obscured by removal and insertion of introns (Patthy, 1994). As a result of continual intron insertion and intron removal, the correlation between modular protein structure and gene structure may get weaker and weaker, and may eventually lead to an exon–intron organization that has little or nothing to do with the one that existed at the time of the formation of the gene. Since erosion of the original genomic structure progresses with time, it is not surprising that old genes usually show less perfect correlation with protein structure than recently assembled ones. This point may be illustrated by the exon–intron structures of laminin genes. Laminins are among the oldest modular proteins unique to metazoa, inasmuch as they are already present in Hydrozoa (Sarras et al., 1994), and their domain organization was practically unchanged during subsequent evolution. Consistent with the great age of these genes, many of the original phase 1 introns are already missing from the boundaries of the class 1-1 laminin B-type modules of Drosophila and vertebrate laminin B1 and B2 chain genes (Vuolteenaho et al., 1990; Chu et al., 1991; Kallunki et al., 1991; Gow et al., 1993).

It must be pointed out that the absence of the expected introns from some module boundaries of proteins is frequently interpreted as evidence that these proteins did not evolve by exon shuffling. Such a conclusion is unjustified. If a modular protein is composed of various class 1-1 modules (EGF-like modules, complement B-type modules, C-type lectin modules, laminin B-type modules, CUB modules, etc.) that have clearly been shown in most other cases to be duplicated, joined to each other and inserted into new locations by recombination in phase 1 introns, then we may rightfully assume that they also arose by exon shuffling, even though the ‘original’ introns are already missing from their genes.

In this respect, the tolloid genes provide illustrative examples. The tolloid gene of D. melanogaster encodes a modular astacin-type metalloprotease containing five class 1-1 complement C1r type (CUB) modules and two class 1-1 EGF-like (G) modules (Fig. 2). The 3.4 kb gene of Drosophila is rather compact, contains only six short (49–142 bp) introns (Chilas and O’Connor, 1994); only two of the module-boundaries are marked by phase 1 introns, and four of the ‘expected’ intermodule phase 1 introns are missing (Fig. 2). The gene (6 kb) of a related protease (BP10) from sea urchin blastula with a different set of six introns also failed to show a clear correlation between modular structure of the protein and exon–intron structure of the gene, therefore it was suggested that “it is unlikely that exon-shuffling had a role in the evolution of the astacin/EGF/CUB subfamily of proteases” (Lhomond et al., 1996). Studies on the genomic organization of the mammalian tolloid gene suggest that this conclusion is unwarranted. The gene of human Bone Morphogenetic Protein 1 (BMP1) encodes a protein with a domain structure identical to that of the Drosophila Tolloid protein, but its gene structure is strikingly different from that of the Drosophila homolog (Takahara et al., 1995). Consistent with the differences in genome compactness of fly and human, the human tolloid gene is more than tenfold larger (46 kb) since it contains a greater number and significantly longer introns than the fly homolog. Importantly, all boundaries separating the class 1-1 CUB and G modules are marked by the ‘expected’ phase 1 introns (Fig. 2). It thus appears that in this respect the human tolloid gene is more similar to the original gene structure than the fly homolog. The difference in gene structure may be explained by assuming that in the Drosophila lineage (characterized by a more compact genome) there was greater selective pressure to eliminate introns than in the chordate lineage (characterized by a less compact genome).

There are other cases which suggest that the gene-structures of chordate homologs are more likely to have preserved the original introns. As another example we may mention the case of human and C. elegans netrin receptor genes. The gene of the human netrin receptor DCC (Deleted in Colorectal Cancer) encodes a transmembrane protein with an extracellular part containing four class 1-1 immunoglobulin (Ig) modules and six class 1-1 fibronectin type III (FN3) modules (Fig. 2). The human DCC gene contains 28 introns and spans approx. 1400 kbp: it is the largest tumor suppressor gene identified to date (Cho et al., 1994). Significantly,
with one exception, the phase 1 introns have been preserved at the boundaries separating the class 1-1 Ig and FN3 modules (Fig. 2). In contrast with this, in the case of the *Caenorhabditis* homolog, 6 of the 11 expected phase 1 introns are missing from the module boundaries (Fig 2). This difference in genomic organization may be also explained by assuming that in the *Caenorhabditis* lineage (characterized by a more compact genome) there was greater selective pressure to eliminate introns than in the chordate lineage (characterized by a less compact genome). Importantly, the gene of the *C. elegans* homolog of DCC/netrin receptor is dramatically more compact (4.7 kb) than the human homolog (1400 kb), and it has fewer and shorter introns (Chan et al., 1996).

In summary, since the genomic organization of a gene does not necessarily reflect the structure that existed at the time of its formation, in many cases the original structure can be reconstructed only through a complex analysis of the evolutionary history of the constituent domains. So far, our analyses have identified more than five dozen class 1-1 module-types that have been used to build class 1 modular proteins by exon-shuffling. As has been discussed previously, there are much fewer class 0-0 or class 2-2 modules (Patthy, 1994, 1995). According to the modularization hypothesis, the explanation for this is that the formation of class 1-1 modules is strongly favoured (Patthy, 1994, 1995, 1999).

### 2.3. Evolutionary distribution of proteins produced by exon-shuffling

Analysis of the evolutionary distribution of modular proteins produced by exon-shuffling may permit a clear definition of the time of their formation, and thus may provide an insight into changes in the significance of exon-shuffling. Thanks to various genome projects, sufficient amount of sequence information has already accumulated on plants, protists, fungi and diverse groups of animals to perform a meaningful analysis. Using the collection of modules that were clearly spread by exon-shuffling, we have searched databases to define the evolutionary distribution of modular proteins that were constructed from these modules. In the present work, modular proteins produced by exon-shuffling were defined as those which contain at least one of the modules spread by exon-shuffling and this module is joined to at least another protein-domain. Note that according to this definition, a protein consisting of a single module and a membrane anchoring segment and/or signal peptide is not counted as a modular protein, but is considered to represent the protomodule stage of the modularization pathway (Patthy, 1994, 1995, 1999).

The search for modular proteins has brought many examples from all major groups of metazoa. The fact that proteins composed of class 1-1 modules familiar from vertebrates have been found in sponges, hydrozoa, nematodes, molluscs, arthropods, and echinoderms, etc., indicates that this machinery of exon-shuffling was available before the divergence of these metazoan phyla. There can be no doubt that the mechanism of the construction of these modular proteins was the same as that of vertebrate genes, since there are several cases of invertebrate genes where the class 1-1 modules are flanked by the original phase 1 introns (Patthy, 1994, 1995). It is noteworthy that modular proteins assembled from class 1-1 modules have already been found in sponges and hydrozoa, although only a tiny fraction of their genes have been sequenced so far. In contrast with this, there is practically no evidence for related modular proteins in plants or fungi. In addition to metazoa, a few modular proteins homologous with class 1-1 modules were found in some animal viruses and parasitic prototiza (Patthy, 1994, 1995). However, it seems likely that these viruses and protists acquired the modular proteins from their multicellular hosts by horizontal gene transfer; these proteins assist them in the invasion process. Thus the surface protein of *P. falciparum* containing four EGF-domains facilitates infection by binding to receptors on mosquito epithelial cells, the thrombospondin-homolog proteins of malarial parasites assist in their entry into hepatocytes by mediating their adherence to these cells, the vaccinia virus protein containing four complement B-type modules could counteract host immune defences by interfering with the complement cascade. The evolutionary distribution of modular proteins that have clearly evolved by exon-shuffling is thus consistent with the suggestion that exon-shuffling became significant at the time of metazoan radiation, parallel with the spread of large splicosomal introns.

Our analysis has also shown that the vast majority of metazoan modular proteins produced by exon-shuffling is extracellular or they are extracellular parts of membrane-associated proteins (Patthy, 1995). A major group of these modular proteins consists of various constituents of the extracellular matrix (laminins, modular collagens, fibronectin, etc.), modular metalloproteinases involved in the remodelling of the extracellular matrix (e.g., type IV collagenases, bone morphogenetic proteins). Another large group of modular proteins contains a variety of membrane associated or transmembrane proteins with extracellular parts constructed from modules: receptor tyrosine kinases, receptor tyrosine phosphatases, proteins involved in cell-cell or cell-matrix interactions, complement receptors, LDL-receptor, cytokine receptors, etc. A third major group of modular extracellular proteins comprises various plasma proteins: the modular proteases of the blood coagulation, fibrinolytic and complement cascades, the
non-protease factors regulating blood coagulation and complement activation, immunoglobulins, etc.

In addition to these extracellular or membrane-associated modular proteins, there are some intracellular modular proteins that have also evolved by exon-shuffling in the metazoan lineage (Pathy, 1995). For example, there is clear evidence that myomesin/skelemin of the contractile apparatus has been constructed by exon-shuffling from class 1-1 immunoglobulin and fibronectin type III modules. In the gene of this protein, the start and end of each immunoglobulin and fibronectin type III domain are still defined by phase I introns (Steiner et al., 1999).

There seem to be striking differences in the distribution of modular proteins in different metazoan lineages. Although modular receptor tyrosine kinases have already been found in most metazoan lineages, including sponges (Schäcke et al., 1994) and there is molecular evidence for the presence of laminins from most metazoan phyla including Hydrozoa (Sarras et al., 1994), there are modular proteins that appear to be restricted to certain evolutionary lineages. For example, the rich variety of modular proteases of blood coagulation, fibrinolysis, complement activation and other extracellular proteolytic cascades (Fig. 3) have probably been formed in the chordate lineage, since proteases with a similar domain organization are missing from non-chordate genomes, including the completely sequenced genome of C. elegans. Conversely, most basic constituents of the extracellular matrix, the majority of proteins involved in cell-cell or cell-matrix interactions are present in vertebrates as well as in arthropods and nematodes, suggesting that they were formed prior to the divergence of these lineages. It is worth pointing out that the genes of modular proteins that appear to be restricted to the chordate lineage provide the most convincing examples of a correlation between domain organization of a modular protein and exon-intron structure of their gene. It seems likely that this is due to the fact that these genes are relatively young and that in the chordate lineage there was no significant pressure to eliminate their introns.

It seems possible that many of the modular proteins found so far only in the C. elegans genome have been formed in the nematode lineage. Even though we cannot exclude the possibility that similar modular proteins will be found in other genomes, it seems to be clear that there are significant differences in the population of modular proteins of vertebrates vs. C. elegans. This point may be illustrated by the fact that different class 1-1 modules are not used with equal probability in these two groups (Fig. 4). Although in both groups the EGF-like module is found in the highest percentage of modular proteins, there are also striking differences in the recurrence of class 1-1 modules (Fig. 4). (The results of this analysis must be treated with some caution, since the C. elegans proteins found in current databases were predicted with a rather high error rate.)

In principle, differences in the recurrence of different modules-types may reflect differences in the time of the appearance of the module (those protomodules that were formed earlier had more time to spread by exon-shuffling), or may reflect differences in the chances of the survival of the newly formed modular protein-coding gene. Survival of the newly formed gene requires that the protein it encodes should be able to fold efficiently into a stable multidomain protein, therefore folding-efficiency of the protein-module is a critical aspect of its evolutionary success. The significance of this requirement may be illustrated by the fact that the majority of modules used in the construction of multidomain proteins display remarkable folding autonomy: the isolated domains can fold very efficiently (Pathy, 1993). Obviously, folding autonomy of domains in multidomain proteins is of utmost importance since this minimizes the influence of neighboring domains and ensures stability in the extracellular environment. Folding autonomy can ensure that folding of the domain is not endangered when inserted into a novel protein environment. It thus seems likely that the most widely used modules have been selected for folding autonomy and fold stability.

It is interesting to note in this respect that the modules used most frequently in the construction of modular proteins are not random representatives of the protein universe. Analysis of protein structures of the Brookhaven Protein Databank (Martin et al., 1998) shows that about half of the non-homologous proteins found in current databases were derived from modules-types may reflect differences in the frequency of use, or may reflect differences in the chance of survival of the newly formed gene.

In the chordate lineage, the structural distribution of disulphide-bonded extracellular domains are in general shifted towards the mainly \( \beta \) class, about another 25% is mainly \( \alpha \) (Fig. 5A). Structural classification of the class 1-1 modules shows a significantly different distribution: more than 60% belong to the mainly \( \beta \) class and only about 10% are mainly \( \alpha \) (Fig. 5B). If the differences in recurrence (Fig. 4) of modules is also taken into account, there is a further significant shift in favour of modules belonging to the mainly \( \beta \) class (Fig. 5C), suggesting that structural features (folding autonomy and stability) play a significant role in controlling the spread of module-types. The fact that the structural distribution of disulphide-bonded extracellular domains are in general shifted towards the mainly \( \beta \) proteins (Martin et al., 1998) indicates that such proteins are more stable in the extracellular environment than other fold classes. It is noteworthy in this respect that as a reflection of their adaptation to the oxidative milieu of the extracytoplasmic space, many of the class 1-1 modules are rich in disulfide bonds.

In summary, the clearcut examples of proteins assembled by exon-shuffling from modules are restricted to metazoas, with practically no counterparts in prokaryotes, plants, protists or fungi. This observation is consistent with our earlier suggestion that the exon-shuffling
Fig. 3. Architecture of some modular proteases of the blood coagulation, fibrinolysis, complement activation cascades and other modular members of the trypsin-family. The SERPRO box represents the trypsin-domain-like serine protease domain, the colored boxes represent the various class I-1 modules from which these proteases were assembled: kringle-module (K), epidermal growth factor-module (G), finger-module (F), fibronectin type II module (FN2), preactivation peptide module (PAP), scavenger receptor module (SC), vitamin K-dependent calcium-binding module (C), contact factor module (CF), complement B-type module (B), complement Clr/Cls module (CUB), von Willebrand type A module (vWA), LDL-receptor module (LDL), meprin-module (MAM), frizzled receptor module (FRZ). The black boxes indicate signal peptide domains, vertical black bars represent transmembrane domains. The boxes designating the different domains are drawn to scale.
2.4 Significance of exon-shuffling in metazoan evolution

It must be emphasized that most modular proteins produced by exon-shuffling are associated with, and are absolutely essential for, multicellularity of metazoa. For example, the appearance of the constituents of extracellular matrix is inextricably linked to the appearance of the first multicellular animals. The constituents of the extracellular matrix, membrane-associated proteins mediating cell-cell and cell-matrix interactions, receptor-proteins regulating cell-cell communications (such as receptor tyrosine kinases, receptor tyrosine phosphatases) are of absolute importance in permitting cells to function in an integrated fashion.

The fact that the overwhelming majority of the constituents of the extracellular matrix, cell adhesion proteins, receptor proteins were constructed from modules underlines the extreme importance of exon-shuffling in metazoan evolution. As outlined above, this powerful evolutionary mechanism has become significant relatively late during evolution at the time of metazoan radiation. It seems probable that the rise of exon-shuffling has contributed significantly to this accelerated evolution of metazoa.

In summary, it seems that the evolution of less compact genomes, the evolution and spread of spliceosomal introns, the creation of mobile modules have reached a critical point sometime before the Cambrian period, and this led to a dramatic increase in the efficiency of modular protein evolution by exon-shuffling. Increased efficiency of exon-shuffling was critical for the rapid creation of diverse multidomain proteins that are essential for multicellularity of metazoa.

References


Chandrasekaran, A., 1996. Small is beautiful: comparative genomics with the pupa of twintrons: potential relationships with nuclear pre-mRNA development and genomic sequence. EMBO J. 11, ...