# Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues

László Bányai, László Patthy*

*Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, P.O. Box 7, H-1518 Budapest, Hungary*

Received 13 February 2004; revised 25 March 2004; accepted 26 March 2004

First published online 9 April 2004

Edited by Takashi Gojobori

**Abstract Comparison of the exon–intron structures of human, fly and worm orthologues of mosaic genes assembled from class 1-1 modules by exon-shuffling has revealed that human genes retained significantly more of the original inter-module introns than their protostome orthologues. It is suggested that the much higher rate of intron loss in the worm- and insect lineages than in the chordate lineage reflects their greater tendency for genome compaction.**



*Keywords:* Exon-shuffling; Genome evolution; Intron gain; Intron loss; Protein module

## 1. Introduction

Eukaryotic protein-coding genes usually contain multiple spliceosomal introns but the number of introns may be markedly different in orthologous genes of distant eukaryotic species due to intron insertion or intron loss. The relative contributions of intron loss and intron gain in the evolution of eukaryotic genes remain poorly understood. The primary difficulty lies in the fact that from a comparison of just two homologous genes, it is impossible to tell whether differences in their exon–intron structures are due to the loss of intron from one gene or gain of intron in the other.

When the structures of orthologous genes from several species are known, the evolutionary dynamics of introns is usually studied by parsimony analysis. However, in the case of distantly related genes parsimony trees do not always mimic the evolutionary history of the analyzed species, cautioning that intron locations are not suitable markers for phylogenetic analysis at long evolutionary distances [1].

The genes of modular proteins assembled by exon-shuffling [2–4] are unique inasmuch as there is less ambiguity as to the 'original' gene structure: intermodule introns used in the assembly process had to be present in the assembled genes at the time of their formation. In fact, many protein-coding genes produced by exon-shuffling of class 1-1 modules could be recognized by a striking correlation between the exon–intron structure of the genes and the domain-organization of proteins, as well as the presence of the 'expected' phase 1 intermodule introns [2,3,5]. Consequently, if the expected intermodule phase 1 introns are missing from some orthologues, this can be explained only by loss of these introns. To derive information on the relative rates of intron loss in the protostome and deuterostome lineages, in the present work we have compared the exon–intron structures of human, fly and worm orthologues of genes assembled from class 1-1 modules by exon-shuffling.

## 2. Results and discussion

The genes compared were those encoding modular extracellular matrix proteins, membrane-proteins, and receptor tyrosine kinases, receptor tyrosine phosphatases that were shown previously to have formed in metazoa from class 1-1 modules prior to the divergence of protostomes and deuterostomes [6]. The fact that homologous intracellular muscle proteins, titins, assembled from class 1-1 immunoglobulin and fibronectin type III domains are present both in the fly and human suggests that titin genes were also formed before the separation of the arthropod and chordate lineages [7,8].

The orthology relationship of the human, fly and worm homologues was checked using reciprocal homology searches. In each case, when a sequence from one species was used as query, from the other two species the candidate proteins gave the highest scores. Multiple alignments of the protein sequences were used to check whether orthologues from different species have identical or different domain organization. Multiple alignments of the amino acid sequences were constructed using Clustal W [9].

In some cases the domain organization of orthologues was found to show differences. This point may be illustrated by titins: the homology of *Drosophila* titin is limited to the N-terminal half of human titin [8]. As another example, we may mention hemicentin, an extracellular member of the immunoglobulin superfamily [10]. Hemicentin from worm and human have very similar domain organization in their N-terminal part consisting mainly of class 1-1 IG domains (44 in the human, 48 in the worm orthologue). The worm and human orthologues deviate in their C-terminal part: whereas human hemicentin has six thrombospondin-type 1 domains, worm hemicentin has none, human hemicentin has eight epidermal

* Corresponding author. Fax: +36-1-466-5465.
*E-mail address:* patthy@enzim.hu (L. Patthy).

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.03.088

growth factor (EGF) modules, whereas worm hemicentin has only two EGF domains. In such cases, only the aligned regions were included in the comparison.

The domain boundaries of modular proteins were defined according to SMART ([11]; http://smart.embl-heidelberg.de/) and Pfam ([12]; http://www.sanger.ac.uk/Software/Pfam/) and were further verified by comparison with consensus sequences of class 1-1 modules [13]. An updated list of class 1-1 modules (i.e. domains that were shuffled by recombination in phase 1 introns flanking these domains) is shown in Table 1.

The location of introns in the genes of the modular proteins was determined using NCBI's genomic databases, http://www.ncbi.nlm.nih.gov:80/genomes/static/euk_g.html, the Ex-Int database ([14]; http://intron.bic.nus.edu.sg/exint/exint.html) as well as data of the SMART database (http://smart.embl-heidelberg.de/). Regions separating class 1-1 modules from other domains were scanned for the presence or absence of the expected phase 1 boundary-introns; this search was limited to 20 amino acid segments outside the N-terminal and C-terminal boundaries of the class 1-1 modules.

Table 1
Class 1-1 modules found most frequently in multidomain proteins

| Module | SMART[a] | Pfam[a] |
|---|---|---|
| Fibronectin type-III module | FN3 | fn3 |
| EGF module | EGF | EGF |
| Immunoglobulin module | IG | ig |
| Complement B-type module (sushi module) | CCP | sushi |
| Thrombospondin type-I module | TSP1 | tsp_1 |
| LDL receptor type-A module | LDLa | ldl_recept_a |
| EGF-like module of laminin | EGF_Lam | laminin_EGF |
| Complement C1r module (CUB module) | CUB | CUB |
| Scavenger receptor module | SR | SRCR |
| C-type lectin module | CLECT | lectin_C |
| A-type module of von Willebrand factor | VWA | vwa |
| C-type module of von Willebrand factor | VWC | vwc |
| Kunitz-type trypsin inhibitor module | KU | Kunitz_BPTI |
| D-type module of von Willebrand factor | VWD | vwd |
| Kringle module | KR | kringle |
| Factor V/VIII type C module | FA58C | F5_F8_type_C |
| MAM module | MAM | MAM |
| Link protein module | LINK | Xlink |
| Finger module (fibronectin type-I module) | FN1 | fn1 |
| Thyroglobulin module | TY | thyroglobulin_1 |
| Fibronectin type-II module | FN2 | fn2 |
| *Frizzled* module | FRI | Fz |
| Whey protein module (WAP module) | WAP | wap |
| Calcium-binding module (gla module) | GLA | gla |
| SEA module | SEA | SEA |
| Olfactomedin domain | OLF | OLF |
| LCCL domain | LCCL | LCCL |
| Follistatin module | FOLN + KAZAL | FOLN + Kazal) |
| Follistatin module | FIMAC | |
| PAN (apple) module | PAN_AP | PAN |
| P or trefoil domain | PD | Trefoil |
| Somatomedin B-like domains | SO | Somatomedin_B |
| HYR domain | – | HYR |
| Delta serrate ligand | DSL | DSL |
| Zona pellucida (ZP) domain | ZP | Zona_pellucida |
| Ly-6 antigen/uPA receptor-like domain | LU | UPAR_LY6 |
| WSC domain | WSC | WSC |

The modules are listed in the order of decreasing frequency in non-redundant domain databases.
[a]The abbreviations correspond to those used by the SMART database ([11]; http://smart.embl-heidelberg.de/) and the Pfam database ([12]; http://www.sanger.ac.uk/Software/Pfam/).

Modules separated by less than 40 amino acid long linker regions were considered to be 'contiguous'. Class 1-1 modules that are separated by more than 40 amino acids at both their N-terminal and C-terminal boundaries from (any) class 1-1 module are defined as 'isolated'. For each gene the number of expected phase 1 introns ($E$) flanking the class 1-1 modules was calculated according to the equation $E = 2i + (c_a + 1) + (c_b + 1) + \cdots + (c_n + 1)$, where $i$ is the number of isolated class 1-1 modules, $c_a$, $c_b$, $c_n$ are the number of class 1-1 modules in contiguous segments $a, b, \ldots, n$, respectively. For example, a multidomain protein constructed from seven contiguous class 1-1 modules is expected to have eight phase 1 introns (cf. Fig. 1). The ratio of the number of observed ($O$) to expected ($E$) intermodule phase 1 introns was determined.

Our analyses have shown (Table 2) that human genes have retained significantly more of the expected phase 1 introns than their *Caenorhabditis elegans* ($P = 0.0028$) or *Drosophila melanogaster* ($P = 0.0313$) orthologues. In the case of the human genes (Table 2) 77.7% of the total expected introns are retained, whereas in the case of *C. elegans* (24.9%) and *D. melanogaster* (19.4%) a much smaller fraction of the expected introns is still present.

As a typical example, we may refer to the tolloid-like genes of worm, fly and human. These genes encode modular astacin-type metalloproteases containing five class 1-1 CUB modules and two class 1-1 EGF modules. In the case of the human BMP1 and tolloid-like genes all but one of the expected introns flanking the class 1-1 CUB and EGF modules are present, whereas in the case of the fly and worm orthologues the majority is missing (Table 2, Fig. 1).

Our observation that loss of expected introns was much more significant in the fly (80.6%) and worm (75.1%) than in human (22.3%) suggests that the rate of intron loss in general was much higher in the protostome than in the chordate lineage.

An implicit prediction of our conclusion (that intron loss was more significant in the worm and fly lineages than in the chordate lineage) is that if we can define other sets of ancestral introns, then they should also be more likely to be preserved in the human than in *D. melanogaster* or *C. elegans* genomes. The work by Rogozin et al. [1] may be relevant in this respect. These authors have compared the intron positions in 684 orthologous gene sets from eight complete genomes of animals, plants, fungi, and protests. Their observation that humans share many more introns with the plant *Arabidopsis thaliana* than with the fly or nematode is most probably a reflection of the much higher rate of intron loss in the worm and fly lineages.

It should be noted that the total number of introns (i.e. not only those flanking class 1-1 modules) is also significantly higher in the human genes than in their *D. melanogaster* ($P = 0.0074$) or *C. elegans* ($P = 0.0049$) orthologues (cf. Table 2). The number of introns in the human genes is 4.07 times higher than those in *D. melanogaster* orthologues and 1.86

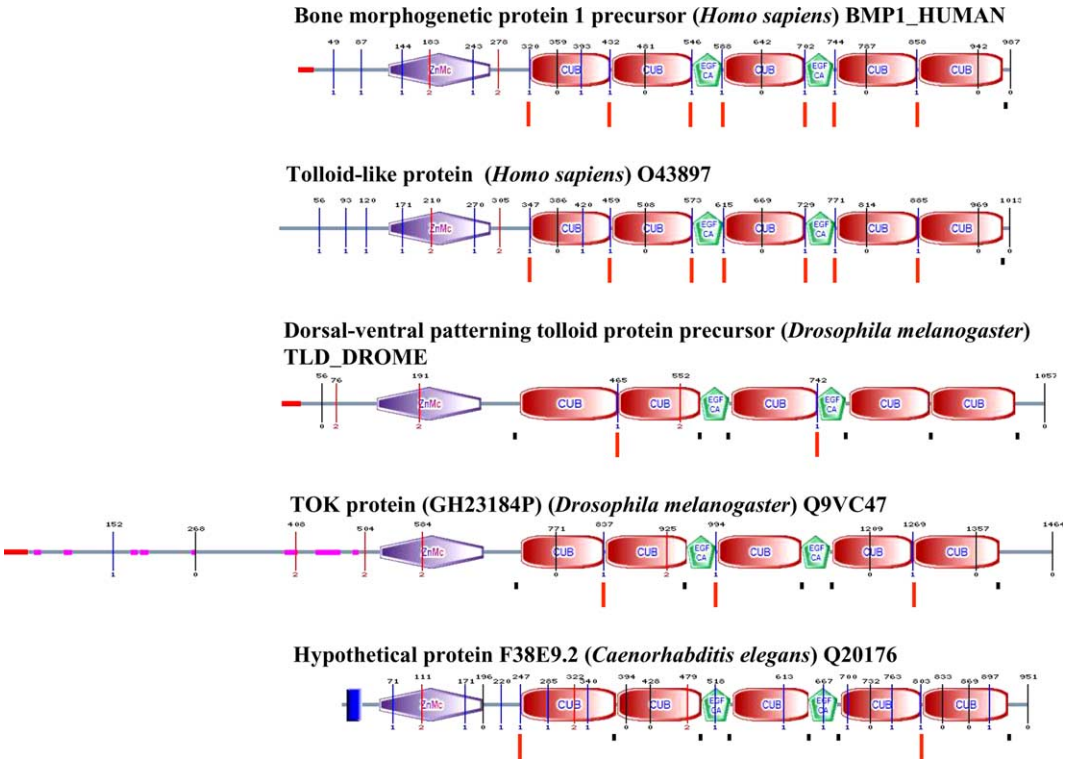


Fig. 1. Comparison of the exon–intron structures of tolloid-related genes from *Homo sapiens*, *D. melanogaster* and *C. elegans*. The figure shows SMART representations [11] with included intron positions mapped onto a ClustalW alignment of orthologous genes. The positions of the expected phase 1 intermodule introns are highlighted by vertical bars: expected intermodule introns missing are indicated by short black bars, those still present are indicated by long red bars. Note that in the case of the human genes seven of the eight expected intermodule phase 1 introns are present but only three in the gene of the fly TOK protein and only two in the gene of fly tolloid protein.

times higher than in *C. elegans* orthologues. This observation is in agreement with earlier data on 684 orthologues from these species [1]: the total number of introns in the human genes is 3.79 times higher than those in *D. melanogaster* orthologues and 2.00 times higher than in *C. elegans* orthologues.

In view of our observations on boundary introns, the most plausible explanation for the fewer introns in worm and fly is that the rate of intron-loss is much higher in protostomes than in vertebrates. This conclusion is in harmony with earlier data which suggest that the rate of intron-turnover is significantly higher in invertebrates than in vertebrates. In a comparison of the genomes of the nematodes *C. elegans* and *C. briggsae*, 263 of ~2700 introns were found to be unique to a single species. Using 70 million years as the divergence time between these two species, the rate of turnover per intron was estimated to be ~0.70 per billion years [15]. A similar analysis of *D. melanogaster* and *D. virilis* of genes has yielded an intron turnover rate of ~0.65 per billion years [15]. In contrast with such a relatively high rate of intron turnover in protostomes, turnover is much slower in vertebrates. For example, comparison of 1560 mouse–human orthologues has revealed that among 10 020 intron positions there was unequivocal evidence for only five intron losses in the mouse lineage, but no losses in human genes or intron gain in either lineage [16]. Using a divergence time of 75 million years for these species the intron turnover is ~0.003 per billion years, i.e., significantly slower than in the fly or worm lineages.

The much higher rate of intron loss in the fly and worm lineages is related to the fact that their genomes are significantly more compact than those of vertebrates. As a result of comparative genomic studies, it has become clear that there is a clear correlation between genome compactness and the frequency and size of introns in genes [5]. 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. These observations suggest that the evolutionary forces that led to the compact fly and worm genomes are also responsible for the high rate of intron loss.

As to the mechanism of intron loss, one widely accepted hypothesis assumes that it involves homologous recombination between the genomic copy of a gene and an intronless cDNA produced by reverse transcription of the processed mRNA. A key aspect of this mechanism is that only genes actively transcribed in the germline would be susceptible to intron loss, which is clearly not the case. Furthermore, since reverse transcriptases begin from the 3′ end of RNA molecules and dissociate in a length dependent manner, this mechanism is not expected to remove introns from the 5′ end of genes as efficiently as from the 3′ ends and this is expected to lead to an intron gradient along the genes. In contrast with this expectation, introns in multicellular genomes are evenly distributed throughout the genes [17]. This mechanism could be hardly efficient for the removal of introns from the 5′ part of the genes encoding giant multidomain proteins such as those studied in the present work. It seems more likely that simple genomic deletion plays a major role in intron loss – a mechanism that does not have the above limitations.

Table 2
Comparison of the exon–intron structures of genes of orthologous human, fly and worm proteins assembled from class 1-1 modules

| Protein[a] | Species | Introns | Phase 1 intermodule introns | | |
|---|---|---|---|---|---|
| | | | Expected | Observed | Observed/expected |
| *LDL-receptor-related proteins* | | | | | |
| P98164/LRP2_HUMAN | *H. sapiens* | 78 | 61 | 42 | 0.688 |
| Q9W343 | *D. melanogaster* | 11 | 58 | 6 | 0.107 |
| Q04833/LRP_CAEEL | *C. elegans* | 26 | 60 | 8 | 0.133 |
| Q07954/LRP1_HUMAN | *H. sapiens* | 88 | 62 | 49 | 0.790 |
| Q9V383 | *D. melanogaster* | 20 | 63 | 6 | 0.095 |
| ENSP00000303634 | *H. sapiens* | 17 | 9 | 9 | 1.000 |
| ENSP00000334522 | *H. sapiens* | 16 | 8 | 8 | 1.000 |
| ENSP00000321958 | *H. sapiens* | 18 | 12 | 10 | 0.833 |
| P98155/LDVR_HUMAN | H. sapiens | 18 | 13 | 11 | 0.846 |
| P01130/LDLR_HUMAN | *H. sapiens* | 17 | 12 | 10 | 0.833 |
| Q9VBN1 | *D. melanogaster* | 9 | 12 | 4 | 0.333 |
| Q9VBN2 | *D. melanogaster* | 10 | 14 | 6 | 0.357 |
| Q95ZN8 | *C. elegans* | 12 | 11 | 6 | 0.545 |
| Q8IFZ0 | *C. elegans* | 12 | 11 | 6 | 0.545 |
| *Netrin-receptors* | | | | | |
| ENSP00000261908 | *H. sapiens* | 28 | 11 | 11 | 1.000 |
| AAB37634 | *C. elegans* | 16 | 11 | 4 | 0.363 |
| Q94537 | *D. Melanogaster* | 9 | 11 | 2 | 0.182 |
| *Neuroglians* | | | | | |
| P32004/CAML_HUMAN | *H. sapiens* | 26 | 12 | 12 | 1.000 |
| Q92823/NRCA_HUMAN | *H. sapiens* | 26 | 12 | 12 | 1.000 |
| P20241/NRG_DROME | *D. Melanogaster* | 6 | 12 | 1 | 0.083 |
| *Serrate-related proteins* | | | | | |
| P78504/JAG1_HUMAN | *H. sapiens* | 25 | 19 | 15 | 0.789 |
| Q9Y219/JAG2_HUMAN | *H. sapiens* | 25 | 18 | 15 | 0.833 |
| Q9Y219/JAG2_HUMAN | *H. sapiens* | 24 | 17 | 14 | 0.824 |
| Q9VB65 | *D. melanogaster* | 13 | 19 | 6 | 0.316 |
| *UNC5 proteins* | | | | | |
| AAQ88717 | *H. sapiens* | 16 | 5 | 5 | 1.000 |
| Q9V7B5 | *D. melanogaster* | 8 | 5 | 3 | 0.600 |
| O44171 | *C. elegans* | 8 | 5 | 3 | 0.600 |
| *Attractins* | | | | | |
| O75882/ATRN_HUMAN | *H. sapiens* | 28 | 8 | 5 | 0.625 |
| Q9VB20 | *D. melanogaster* | 5 | 6 | 1 | 0.167 |
| Q19981/YC81_CAEEL | *C. Elegans* | 19 | 8 | 3 | 0.375 |
| *Semaphorins* | | | | | |
| Q13591/SM5A_HUMAN | *H. sapiens* | 20 | 8 | 4 | 0.500 |
| Q9VTT0 | *D. melanogaster* | 10 | 8 | 2 | 0.250 |
| *Anosmins* | | | | | |
| ENSP00000262648 | *H. sapiens* | 13 | 6 | 5 | 0.833 |
| Q9CCC7 | *D. melanogaster* | 5 | 6 | 1 | 0.167 |
| O62299 | *C. elegans* | 4 | 6 | 0 | 0.000 |
| *Chordins* | | | | | |
| ENSP00000204604 | *H. sapiens* | 22 | 6 | 5 | 0.833 |
| Q24025/SOG_DROME | *D. melanogaster* | 5 | 6 | 2 | 0.333 |
| *Crumbs-related proteins* | | | | | |
| P82279/CRBH_HUMAN | *H. sapiens* | 10 | 21 | 4 | 0.190 |
| Q19350 | *C. elegans* | 29 | 34 | 3 | 0.088 |
| Q9VC97 | *D. melanogaster* | 12 | 33 | 1 | 0.030 |
| *Fat-spondins* | | | | | |
| ENSP00000309297 | *H. sapiens* | 16 | 7 | 6 | 0.857 |
| Q19305 | *C. elegans* | 11 | 7 | 4 | 0.571 |
| Q9XZD0 | *D. melanogaster* | 4 | 6 | 1 | 0.167 |
| AAM68661 | *D. melanogaster* | 5 | 7 | 1 | 0.142 |
| *Fibulins* | | | | | |
| ENSP00000262722 | *H. sapiens* | 14 | 10 | 9 | 0.900 |
| ENSP00000295760 | *H. sapiens* | 16 | 12 | 11 | 0.917 |
| F56H11.1a | *C. elegans* | 14 | 10 | 6 | 0.600 |

Table 2 (*continued*)

| Protein[a] | Species | Introns | Phase 1 intermodule introns | | |
|---|---|---|---|---|---|
| | | | Expected | Observed | Observed/expected |
| *Hemicentins* | | | | | |
| ENSP00000271588 | *H. sapiens* | 106 | 45 | 44 | 0.978 |
| Q8I0L3 | *C. elegans* | 61 | 45 | 18 | 0.400 |
| *Laminins G1 and G3* | | | | | |
| ENSP00000266097 | *H. sapiens* | 27 | 13 | 4 | 0.307 |
| ENSP00000258341 | *H. sapiens* | 27 | 13 | 4 | 0.307 |
| P15215/LMG1_DROME | *D. melanogaster* | 8 | 13 | 1 | 0.077 |
| Q18823/LML1_CAEEL | *C. elegans* | 10 | 13 | 1 | 0.077 |
| *Laminins B1 and B2* | | | | | |
| ENSP00000222399 | *H. sapiens* | 32 | 15 | 7 | 0.466 |
| ENSP00000307156 | *H. sapiens* | 31 | 15 | 7 | 0.466 |
| W03F8.5 | *C. elegans* | 10 | 15 | 0 | 0.000 |
| CG7123-PA | *D. melanogaster* | 1 | 15 | 0 | 0.000 |
| *Laminin A2* | | | | | |
| ENSP00000325121 | *H. sapiens* | 63 | 19 | 4 | 0.210 |
| O45614 | *C. elegans* | 12 | 20 | 0 | 0.000 |
| Q8IP51 | *D. melanogaster* | 10 | 20 | 2 | 0.100 |
| *Laminin A5* | | | | | |
| ENSP00000252999 | H. sapiens | 70 | 24 | 14 | 0.583 |
| Q9VRW0 | *D. melanogaster* | 14 | 24 | 4 | 0.166 |
| P91904 | *C.elegans* | 14 | 24 | 2 | 0.083 |
| *Netrins* | | | | | |
| ENSP00000173229 | *H. sapiens* | 5 | 4 | 3 | 0.750 |
| O00634 | *H. sapiens* | 5 | 4 | 3 | 0.750 |
| Q9BZP1 | *H. sapiens* | 9 | 4 | 3 | 0.750 |
| Q24568/NETB_DROME | *D. melanogaster* | 6 | 4 | 3 | 0.750 |
| P34710/UNC6_CAEEL | *C. elegans* | 12 | 4 | 3 | 0.750 |
| *Osteonectin* | | | | | |
| P09486/SPRC_HUMAN | *H. sapiens* | 8 | 2 | 2 | 1.000 |
| O97365 | *D. melanogaster* | 1 | 2 | 0 | 0.000 |
| P34714/SPRC_CAEEL | *C. elegans* | 5 | 2 | 1 | 0.500 |
| *Tolloid-related proteins* | | | | | |
| P13498/BMP1_HUMAN | *H. sapiens* | 19 | 8 | 7 | 0.875 |
| O43897 | *H. sapiens* | 20 | 8 | 7 | 0.875 |
| Q20176 | *C. elegans* | 22 | 8 | 2 | 0.250 |
| P25723/TLD_DROME | *D. melanogaster* | 6 | 8 | 2 | 0.250 |
| Q9VC47 | *D. melanogaster* | 12 | 8 | 3 | 0.375 |
| *WNT inhibitory factors* | | | | | |
| AAQ88710 | *H. sapiens* | 9 | 6 | 6 | 1.000 |
| Q9W3W5 | *D. melanogaster* | 6 | 6 | 4 | 0.666 |
| *ROR receptor tyrosine kinases* | | | | | |
| ENSP00000303320 | *H. sapiens* | 8 | 4 | 4 | 1.000 |
| Q24488/ROR1_DROME | *D. melanogaster* | 7 | 3 | 0 | 0.000 |
| Q9V6K3/ROR2_DROME | *D. melanogaster* | 3 | 3 | 2 | 0.666 |
| Q9BLY1 | *C. elegans* | 8 | 4 | 0 | 0.000 |
| *DDR receptor tyrosine kinases* | | | | | |
| Q08345/DDR1_HUMAN | *H. sapiens* | 16 | 2 | 2 | 1.000 |
| Q7Z730 | *H. sapiens* | 18 | 2 | 2 | 1.000 |
| Q95ZV7 | *C. elegans* | 13 | 2 | 2 | 1.000 |
| Q18163 | *C. elegans* | 14 | 2 | 2 | 1.000 |
| *Ephrin receptors* | | | | | |
| ENSP00000275815 | *H. sapiens* | 17 | 3 | 3 | 1.000 |
| ENSP00000327688 | *H. sapiens* | 15 | 3 | 3 | 1.000 |
| Q8IMC3 | *D. melanogaster* | 8 | 3 | 1 | 0.333 |
| Q9V4E5 | *D. melanogaster* | 9 | 3 | 1 | 0.333 |
| O61460/VAB1_CAEEL | *C. elegans* | 9 | 3 | 1 | 0.333 |
| *Vascular endothelial growth factor receptors* | | | | | |
| P35968/VGR2_HUMAN | *H. sapiens* | 29 | 8 | 8 | 1.000 |
| ENSP00000261937 | *H. sapiens* | 29 | 8 | 8 | 1.000 |
| ENSP00000282397 | *H. sapiens* | 29 | 8 | 8 | 1.000 |
| Q9VLQ8 | *D. melanogaster* | 16 | 8 | 4 | 0.500 |
| Q8IPG1 | *D. melanogaster* | 16 | 8 | 4 | 0.500 |
| Q21038 | *C. elegans* | 13 | 8 | 0 | 0.000 |
| Q21041 | *C. elegans* | 14 | 8 | 0 | 0.000 |

Table 2 (*continued*)

| Protein[a] | Species | Introns | Phase 1 intermodule introns | | |
|---|---|---|---|---|---|
| | | | Expected | Observed | Observed/expected |
| *Fibroblast growth factor receptors* | | | | | |
| ENSP00000322945 | *H. sapiens* | 15 | 4 | 4 | 1.000 |
| ENSP00000263455 | *H. sapiens* | 15 | 4 | 4 | 1.000 |
| P22607/FGR3_HUMAN | *H. sapiens* | 16 | 4 | 4 | 1.000 |
| P22455/FGR4_HUMAN | *H. sapiens* | 16 | 4 | 4 | 1.000 |
| CG7223_PA | *D. melanogaster* | 0 | 3 | 0 | 0.000 |
| *LAR receptor tyrosine phosphatases* | | | | | |
| ENSP00000302753 | *H. sapiens* | 31 | 13 | 9 | 0.692 |
| Q9VIS8/Q9VIS8 | *D. melanogaster* | 16 | 13 | 7 | 0.538 |
| C09D8.1a | *C. elegans* | 27 | 13 | 8 | 0.615 |
| *Titins* | | | | | |
| Q8WZ42 | *H. sapiens* | 304 | 177 | 151 | 0.853 |
| Q9NFS3 | *D. melanogaster* | 36 | 64 | 11 | 0.172 |

[a] Proteins are identified by their Swiss–Prot or ENSEMBL (http://www.ensembl.org/Homo_sapiens/) identifiers.

## References


[1] Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G. and Koonin, E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr. Biol. 13, 1512–1517.

[2] Patthy, L. (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett. 214, 1–7.

[3] Patthy, L. (1991) Modular exchange principles in proteins. Curr. Opin. Struct. Biol. 1, 351–361.

[4] Patthy, L. (1995) Protein evolution by exon-shuffling. Molecular Biology Intelligence Unit. Springer, Berlin.

[5] Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling – a review. Gene 238, 103–114.

[6] Patthy, L. (2003) Modular assembly of genes and the evolution of new functions. Genetica 118, 217–231.

[7] Bang, M.L., Centner, T., Fornoff, F., Geach, A.J., Gotthardt, M., McNabb, M., Witt, C.C., Labeit, D., Gregorio, C.C., Granzier, H. and Labeit, S. (2001) The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. Circ. Res. 89, 1065–1072.

[8] Machado, C. and Andrew, D.J. (2000) D-Titin: a giant protein with dual roles in chromosomes and muscles. J. Cell Biol. 151, 639–652.

[9] Thompson, J.D., Higgins, D.G. and Gibson, T. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 12, 4673–4680.

[10] Vogel, B.E. and Hedgecock, E.M. (2001) Hemicentin, a conserved extracellular member of the immunoglobulin superfamily, organizes epithelial and other cell attachments into oriented line-shaped junctions. Development 128, 883–894.

[11] Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. 30, 242–244.

[12] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam Protein Families Database. Nucleic Acids Res. 30, 276–280.

[13] Patthy, L. (1987) Detecting homology of distantly related proteins with consensus sequences. J. Mol. Biol. 198, 567–577.

[14] Sakharkar, M., Passetti, F., de Souza, J.E., Long, M. and de Souza, S.J. (2002) ExInt: an Exon Intron Database. Nucleic Acids Res. 30, 191–194.

[15] Lynch, M. and Richardson, A.O. (2002) The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. 12, 701–710.

[16] Roy, S.W., Fedorov, A. and Gilbert, W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. USA 100, 7158–7162.

[17] Mourier, T. and Jeffares, D.C. (2003) Eukaryotic intron loss. Science 300, 1393.