Published online 8 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 8 ***e47***
doi:10.1093/nar/gkn153

Exalign: a new method for comparative analysis of exon–intron gene structures

Giulio Pavesi[1], Federico Zambelli[1], Corrado Caggese[2] and Graziano Pesole[3,4,*]

[1]Dipartimento di Scienze Biomolecolari e Biotecnologie, University of Milan, Milan, [2]Dipartimento di Genetica e Microbiologia, University of Bari, Bari, [3]Dipartimento di Biochimica e Biologia Molecolare ''E. Quagliariello'', University of Bari, Bari and [4]Istituto Tecnologie Biomediche – Consiglio Nazionale delle Ricerche, Bari, Italy

Received January 7, 2008; Revised March 16, 2008; Accepted March 19, 2008

ABSTRACT

The evolution of genes is usually studied and reconstructed at the sequence level, that is, by comparing and aligning their genomic, transcript or protein sequences. However, including the exon–intron structure of genes in the analysis can provide further and useful information, for example to draw reliable phylogenetic relationships left unsolved by traditional sequence-based evolutionary studies, or to shed further light on patterns of intron gain and loss. In spite of this, no tool especially devised for this task is currently available. In this work we present Exalign, an algorithm designed to retrieve, compare and search for the exon-intron structure of existing gene annotations, that has been implemented in a software tool freely accessible through a web interface as well as available for download. We present different applications of our method, from the reconstruction of the evolutionary history of homologous gene families to the detection of as of today unknown cases of intron loss in human and rodents, and, remarkably, two never reported intron gain events in human and mouse. The web interface for accessing Exalign is available at http://www.pesolelab.it/exalign/ or http://www.beacon.unimi.it/exalign/

INTRODUCTION

The evolutionary history of the homologous members of a gene family is generally reconstructed by performing phylogenetic analyses on multiple sequence alignments of its expression products, i.e. mRNA coding regions or their encoded protein sequences. Sequence similarity analyses, typically performed at the protein level, are also used to establish homology relationships between genes in terms of orthology or paralogy. Indeed, the observation of statistically significant sequence similarity generally implies evolutionary and functional relatedness.

All these approaches, however, do not take into account the additional wealth of information provided by genomic sequences, and in particular the annotation of the exon–intron architecture of spliceosomal genes. Apart from the long-standing debate about the intron-early or intron-late hypotheses (1), the exon–intron structure of homologous genes is generally well conserved across metazoans, with estimates of intron turnover (i.e. rate of intron gain or loss) ranging between 10^{-9}/year for flies and worms, and 10^{-11}/year for mammals (2). Therefore, the additional phylogenetic signal provided by the comparison of intron position and phase may further clarify evolutionary relationships among members of the same gene family, and in some cases help drawing reliable phylogenetic relationships left unsolved by traditional sequence-based evolutionary studies. In particular, the study of the pattern of intron retention, gain and loss may detect new remote homologs for protein with very limited sequence similarity (3) and resolve deep evolutionary relationships (4).

The increasing availability of good quality data in several different species makes now possible to investigate the dynamics of gene structure and also permits genome-wide studies aimed at the detection of events of intron gain and loss. Indeed, the reconstruction of intron gain/loss events along the evolutionary history of a gene may provide valuable information to further clarify evolutionary relationships within large gene families and may help studying their possible functional implications like the generation or disruption of lineage-specific alternative splicing events (5). The dynamics of intron gain/loss has been recently reported for mammals, through the analysis of whole-genome alignments of human, mouse, rat and dog. While no evidence was found of any intron gain event, over a hundred cases of intron loss were detected, mostly in the rodent lineage, almost exclusively in highly expressed housekeeping genes (5).

*To whom correspondence should be addressed. Tel: +39 80 5443588; Fax: +39 80 5443317; Email: graziano.pesole@biologia.uniba.it

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

However, despite the recent interest and debates on these aspects of evolution no suitable tool is currently available for the automatic comparison of intron/exon structures. To our knowledge, only one method of this kind, called FINEX, was presented more than 10 years ago (6), but it seems to be no longer supported by its authors and does not take into account the possibility of intron gain/loss events. On the other hand, recent studies mainly aimed at the detection of intron gain/loss events across multiple species relied on sequence alignments [either protein or transcript alignments, as in (7) or whole-genome alignments as in (5)], and mapped the position of annotated introns within the alignments themselves.

At the structure level, a gene can be simply represented as an array of exon lengths with the additional information of the intron phase between coding exons. Once suitable matching scores have been introduced, exon length arrays can be aligned using standard dynamic programming procedures for local or global alignment, in order to perform pairwise gene structure comparisons as well as large-scale database searches. This is essentially the idea that was proposed in FINEX (6). In our approach, however, we also introduce a scoring method for matches/mismatches/gaps in alignments based on statistics on exons size distribution and intron phase, as well as a post-processing phase of the aligned sequences suitable for the automatic detection of intron gain (or loss) events. The algorithm has been implemented in a software tool called Exalign, that can be accessed through a dedicated web interface or downloaded in a standalone version. With the software and the interface, users can retrieve and compare the structure (and, if needed, nucleotide and protein sequences) of the RefSeq gene annotations for a number of different organisms, or perform BLAST-like structure similarity searches against whole genome annotations. However, gene structures (and sequences) not included in the database can be designed and added by the users themselves. The applications of the tool, as we show in the rest of the article, are various: the study the evolution of the known members of a given gene family in one or many species, taking into account not only sequence similarity, but also structural information, alternative splicings and/or intron gain and loss events; the identification and further characterization of homology and paralogy relationships, by performing whole genome or database searches based on structure similarity, which, integrated in the tool itself with comparisons at nucleotide and protein sequence level can provide a clearer and more complete picture of the evolutionary relationships between genes.

MATERIALS AND METHODS

The algorithm

A gene structure is defined by a sequence of integer numbers denoting the size of the exons, and coding exons are also annotated with their reading frame (i.e. the phase of the preceding intron, when available). Gene structure alignments are thus performed by comparing exons according to their size and reading frame. In order to define match and mismatch scores for all possible pairs of exon lengths, we first took the RefSeq gene annotations available in different organisms, and used them to compute for each organism the frequency distribution of the length of internal exons (i.e. excluding the first and the last exon).

Then, for each organism, starting from the frequency $f(l)$ of each exon length l, we defined the probability $p_{\text{match}}(\Delta l)$ of finding two exons whose length difference is at most a given Δl by chance:

$$p_{\text{match}}(\Delta l) = \sum_{e_1} \sum_{e_2}^{|e_1 - e_2| \leq \Delta l} f(e_1) f(e_2)$$

where e_1 and e_2 cover every possible exon length found in the annotations. The score corresponding to the alignment of two exons whose size difference is Δl can be thus calculated accordingly:

$$S(\Delta l) = -\log(p_{\text{match}}(\Delta l))$$

In this way, for each pair of exons in each of the organisms investigated, we have a non-negative score associated with their length difference Δl (that also includes the case of exons of equal size with $\Delta l = 0$). If two genes to be aligned come from different species, then the frequencies associated with the different exon lengths in the two species are averaged before computing the alignment scores.

Also, if two exons are translated in the same reading frame, then the corresponding alignment score should be increased. Let r_0, r_1 and r_2 be the frequencies with which the three possible frames are found in the coding exons of the species investigated [in general, frame 0—in which an exon starts with the first nucleotide of a codon—is found much more frequently, for example in nearly half of vertebrate coding exons, as also reported in (8)]. If two exons of size difference Δl have the same reading frame j (between 0 and 2), then the corresponding probability p_{match} is modified by taking into account also the probability of finding the same reading frame r_j by chance, that equals r_j^2:

$$p_{\text{match, frame}}(\Delta l, j) = r_j^2 \cdot p_{\text{match}}(\Delta l)$$

As a consequence, the score of the alignment of two exons with different reading frame has to be modified as well:

$$p_{\text{match, noframe}}(\Delta l) = \left(1 - \sum r_j^2\right) p_{\text{match}}(\Delta l)$$

In order to allow for negative scores and compute local alignments, we finally rescaled the scores by subtracting to the score associated with each length difference in each organism the expected value of the scores themselves.

The last thing left to account for is the choice of penalty values to be associated with gaps in the alignments. Given a sequence of $m-1$ elements aligned against m elements (one exon is thus aligned with a gap), there are m different choices for the placement of the gap in the alignment. Let p_{match} be the overall probability associated with the matches obtained for the remaining $m-1$ exons once the gap position has been chosen [which in turn yields an overall alignment score of $-\log(p_{\text{match}})$]. Then, the probability of finding an alignment with associated

probability p_{match} by inserting one gap in the alignment is given by $m \times p_{match}$. The corresponding log score can be thus defined by $-\log(m) - \log p_{match}$. The second term corresponds to the alignment score as previously defined, while the first is always negative and therefore accounts for the alignment of an exon with a gap. In case two gaps are aligned to a sequence, then the corresponding penalty values could be defined in a similar fashion, considering the number of possible ways of aligning two gaps to m elements. However, given the usual sequence size m (expressed as number of exons) and the small differences in the log values obtained by computing them in case of more than one gap, we approximate the penalty associated with the insertion of any gap against a sequence of length m with $-\log(m)$ (that is always negative).

Like traditional sequence alignments, structure alignments can be computed in three ways: global, local (corresponding to the standard Needleman–Wunsch and Smith–Waterman alignment algorithms, respectively), or 'glocal', in which gaps inserted at the beginning or at the end of the alignment yield no penalty (useful for comparing and identifying isoforms resulting from alternative transcription start or termination sites). Also, once two sequences have been aligned (regardless of the strategy), the algorithm tries to determine whether a possible intron gain/loss event is likely to have occurred. An intron gain/loss event in a gene, detectable by comparing its structure to one of its closest homologs, usually results in the insertion of an internal gap in the alignment, in correspondence of the extra exon found in one of the two genes, as shown in Figure 1. The idea is thus to detect these events by first aligning the structures as usual, and then by checking whether the sum of the length of two neighboring exons in a gene (one of which is aligned with a gap) matches the length of a single exon of the other (Figure 1). In this case, the two exons are merged into a single one (whose length equals the length of the exon of the other gene), the sequences are re-aligned, and in case the score of the resulting alignment is higher, a possible intron gain/loss event is noted. For computational reasons, this step is iterated for every possible pair and triplet (indicating two consecutive intron losses) of adjacent exons in the two genes to be aligned, but the same idea can be implemented by merging any number of exons. We also implemented a looser definition in the algorithm: two exons can be merged (i.e. they are likely to border a gained/lost intron) if the sum of their lengths equals approximately (the difference should be a multiple of three, not to change intron phase, and cannot exceed 15 bp) the length of an exon in the other sequence, but also if the exons preceding and following them in the alignment exactly match their counterparts in the other sequence (as shown in Figure 1).

All in all, a possible intron loss in a gene is therefore signaled by the presence of two merged exons in the gene aligned against it, or, vice versa, an intron gain in a gene can be highlighted by the presence of two merged exons in its alignment against a homolog. On the other hand, an exon aligned to a gap flanked in the alignment by exons yielding good matches (i.e. having equal or almost equal size, and same reading frame) is a clear indicator of the presence of a possible alternatively spliced isoform of the gene, with the extra exon corresponding to a cassette exon.

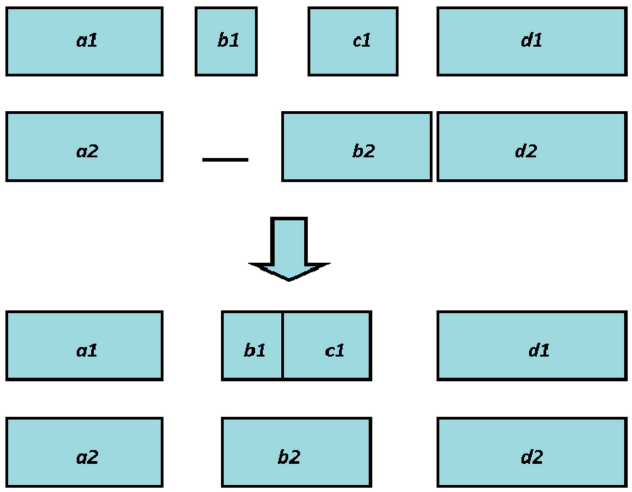

Figure 1. Detection of a possible intron gain/loss event in a structural alignment. In the 'normal alignment' (top) exon b1 is aligned with a gap, c1 with b2: but, as shown in the bottom of the figure, the sum of lengths of exons b1 and c1 equals the size of exon b2. This is thus likely to be the effect of an intron gain/loss event that took place in one of the two genes. The same principle holds also if the sum of b1 and c1 equals b2 approximately (the difference should be a multiple of three, not to change intron phase, and cannot exceed 15 bp), and the size of the exons bordering the merged ones is equal (i.e. a1 = a2 and d1 = d2).

Performing similarity searches

Once a strategy for computing structure alignments has been defined, it can be in turn easily employed to perform similarity searches against a collection of gene structures: in this case, similarity is measured according to the conservation of the structure of the genes, rather than their nucleotide or protein sequence. A query gene can be aligned to all the genes of a database (e.g. all gene annotations available for one or more different organisms), and results are sorted by alignment score. Since alignment scores are directly derived from exon length frequencies in a number of different genes for different organisms, each score can be translated back into a *P*-value and a corresponding expected value (*E*-value) can be computed, depending on the size of the gene and of the database against which the search is performed, thus providing additional information on the actual significance of the results obtained. In particular, our experiments on genome-wide comparisons of query genes of different sizes showed that the distribution of local alignments scores can be approximated by an exponential distribution with parameter $\lambda = 1$. In other words, given gene B with m exons locally aligned to all the genes of a given species (the database), and the overall number of exons of the database n, the expected number of genes E[B,S] of the database yielding a local alignment with B with score S or greater can be approximated by $E[B,S] = nme^{-S}$.

The Web interface

The Web interface allows users to access exon structure and coding sequence (CDS) annotation available for the

RefSeq genes for a number of different species by using the corresponding transcript ID. In particular, users can perform the alignment of two given genes of their choice, or a genome- or database-wide similarity search with a query gene against all the genes annotated in one or more genomes. Alignments and search results are presented in a graphical way, sorted according to the alignment scores and *E*-values computed as outlined in the previous section. Moreover, for each pair of structures the web interface also computes a local sequence alignment of the corresponding nucleotide and protein sequences by using Blast2Seq. All the parameters used by the algorithm, including Blast2Seq word sizes can be modified and fine tuned by accessing the advanced interface (see the online Help page for further information).

The structure of a gene not included in the database (defining number and size of exons) can be nevertheless designed by users, with, if available, its CDS annotation, nucleotide and protein sequences by using the 'Design your own gene' option. The gene is then saved and kept into the database, and the ID assigned to it by the interface can be used later in the main web page for alignments with other genes or as query for database searches. Other than providing users with the possibility of adding their own gene annotations, the interface will be constantly kept up to date as soon as novel genome annotations or annotation updates will become available.

An example of an alignment output by the interface is shown in Figure 2A, showing the global alignment of human and mouse OVCH2 genes. Each exon is denoted by a rectangle, whose size is proportional to the exon length (reported inside the rectangle itself). In case significant sequence similarity between the two genes has been detected as well, the results can be seen by following two links named 'BLASTN hits' and 'BLASTP hits' below the structure alignment. Clicking on them takes to the respective nucleotide and protein alignments, with respective scores and *E*-values. Exons whose nucleotides (or protein translation) have been included in the BLAST alignments are highlighted (their length appears within a box). In this way, users can immediately determine whether and where structure similarity is also supported by similarity in the transcript or protein sequences.

Rectangle/exon colors change according to score of their alignment, from dark green, indicating exact size matches with same reading frame, to light green, yellow (indicating gaps) and finally mismatch with larger exon size difference (red). The reading frame of coding exons is denoted by the colored lines in correspondence of the exons (non-coding exons have no corresponding line).

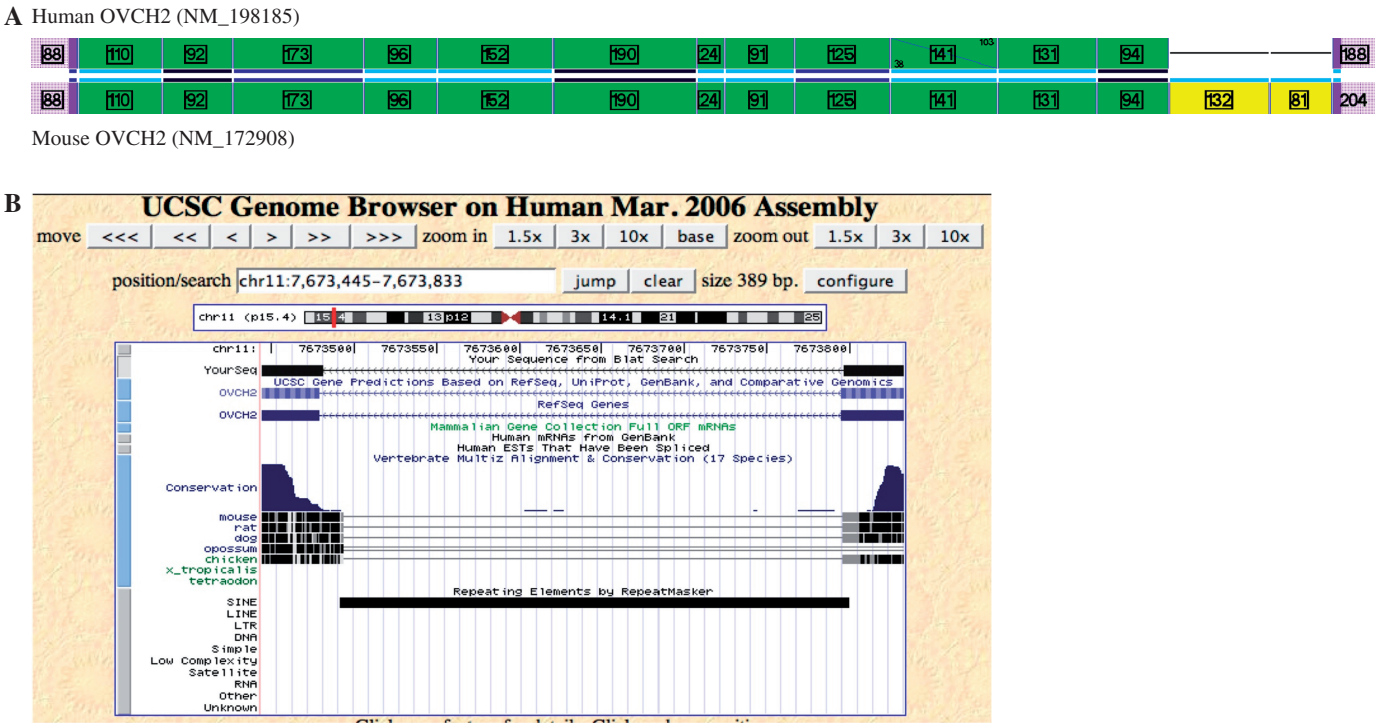

Figure 2. (**A**) Global gene structure alignment produced by Exalign on human and mouse *OVCH2* genes. Each exon is denoted by a rectangle, whose size is proportional to the exon length (reported inside the rectangle itself), and its absolute position in the complete gene structure is in the square brackets. Exons whose nucleotides (or protein translation) have produced significant BLAST alignments are highlighted (their length appears within a box). The reading frame of coding exons is denoted by the colored lines (different color for each of the three frames) in correspondence of the exons in the middle of the alignment. Untranslated exons or regions are in patterned color. The diagonal in correspondence of the 11 rectangle of the human gene shows the merging of the 11th and 12th human exons, of size 38 and 103, resulting in an exon of size 141 that perfectly matches a mouse exon. Merging yields a higher scoring alignment, thus indicating the possibility either an intron gain in human, or an intron loss in mouse. The two yellow exons of near the 3′ end of the mouse gene are aligned with gaps, indicating that probably two alternatively spliced isoforms are annotated in the two genomes. (**B**) Genomic location of the putative novel intron in the human *OVCH2* gene (located between the 11th and 12th exon), falling in correspondence of a recent ALU insertion (marked by the SINE element annotated in the UCSC genome browser).

Also, coding exons have full color, while non-coding exons (or non-coding portions of exons) are patterned. The diagonal in correspondence of the 11th rectangle of the human gene indicates that the algorithm merged two exons of size 38 and 103 (11th and 12th human exons), resulting in an exon of size 141 that perfectly matches a mouse exon and yields a better alignment, thus indicating either an intron gain in human, or an intron loss in mouse. Further examination of the genomic sequence in correspondence of the human intron led us to hypothesize an intron gain event (see 'Results' section). The two yellow exons of the mouse gene are aligned with gaps, indicating that probably two alternatively spliced isoforms are annotated in the two genomes.

RESULTS

Using Exalign to reconstruct the evolutionary history of gene structures

A typical application example of Exalign is the reconstruction of the evolutionary history of the structure of a family of orthologous genes, as for example the comparison of the Elongation Factor 2 (*EEF2*) gene structure across vertebrates. The gene encodes a translation elongation factor whose sequence is highly conserved. Human *EEF2* gene (NM_001961) has 15 exons. The human–rodents–dog comparison performed in (5) showed that—despite the high degree of conservation in the encoded proteins—the mouse gene lost a single intron, the seventh (numbering w.r.t. the sea urchin gene structure, see Figure 3), and the rat gene lost three more introns (other than 7, also 6, 11 and 14). Enlarging the analysis to other vertebrate species, using sea urchin as outgroup, revealed however a more complex scenario, summarized in Figure 3. Apparently, the ancestral gene had 16 exons (see the sea urchin structure). In fishes, introns 7 and 10 have been lost in all the species considered (zebrafish, fugu and pufferfish). Intron 12 is missing in zebrafish and fugu, but it is present in pufferfish. The size of the two exons bordering the intron, however, differs from those of sea urchin, while the overall size of exons 12 + 13 is the same in the two species. This, coupled to the fact that the two exons actually encode for homologous regions of the proteins, leads to the conjecture that intron 12 in Tetraodon, rather than being the descendant of the ancestral intron, is the effect of an intron gain event, following the loss of intron 12 in vertebrates. Similarly, intron 10 in vertebrates (other than fishes) separates exons of sizes different from the sea urchin ones, but again the overall size of exons 11 + 12 is the same in the two structures. Again, we can formulate the hypothesis of an ancient intron gain in tetrapods. On the other hand, no significant structure similarity could be detected in more distant species, like *Drosophila* or *C. elegans*, despite the high level of conservation of the EEF2 protein in these organisms.

The same approach can be extended not only to the study of a group of orthologous genes (each taken from a different species), but also to a gene family composed of any number of homologous genes, another topic that has raised recent interest (9,10). A good example is the metalloendopeptidase MET13 family. Mammals have seven genes encoding proteins belonging to this family, and the corresponding proteins are highly conserved outside the N-terminal region (from amino acid 100 circa, in correspondence to the peptidase M13 domains). The seven genes

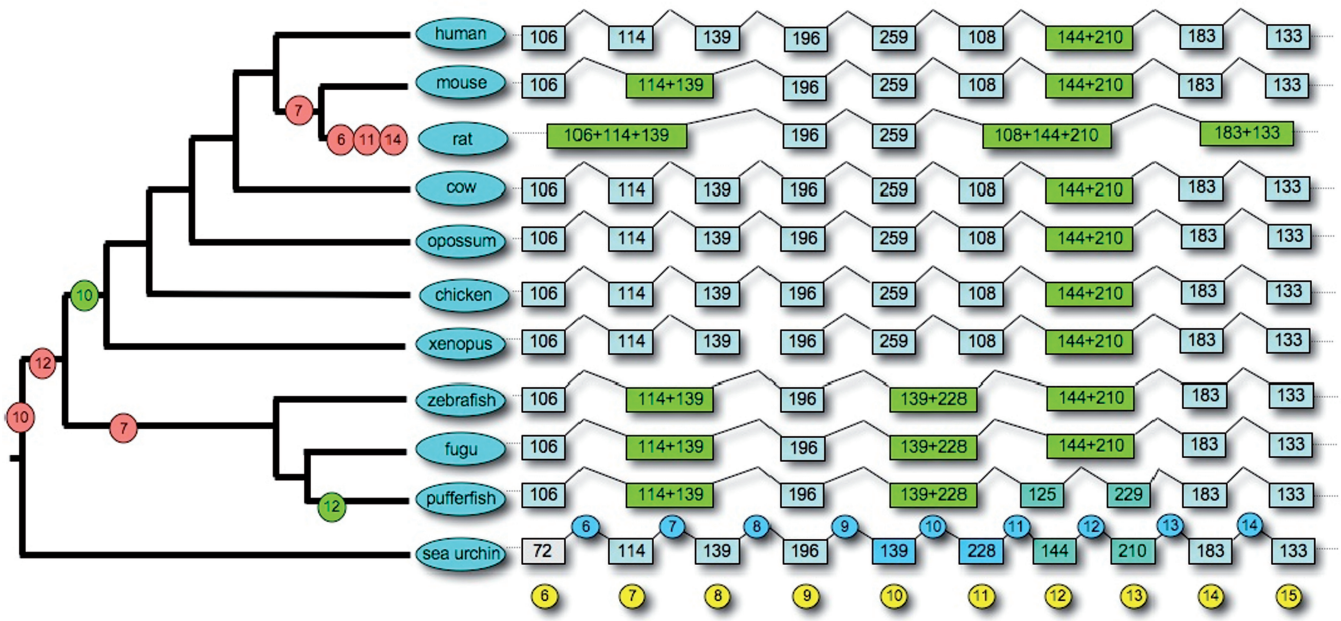

Figure 3. Structural comparison and reconstruction of the evolutionary history of *EEF2* gene structure in different species. Potential intron loss events are marked by a red circle (with the corresponding intron and exons numbered w.r.t. the sea urchin gene). Putative intron gains are marked by a green circle. The gene structures considered include: human (NM_001961), mouse (NM_007907), rat (NM_017254), cow (NM_001075121), opossum (XM_001374069), chicken (NM_205368), xenopus (NM_203924), zebrafish (NM_200458), fugu (NEWSINFRUT00000166554), pufferfish (GSTENP00020105001), sea urchin (XM_792306). See the main text for further discussion.

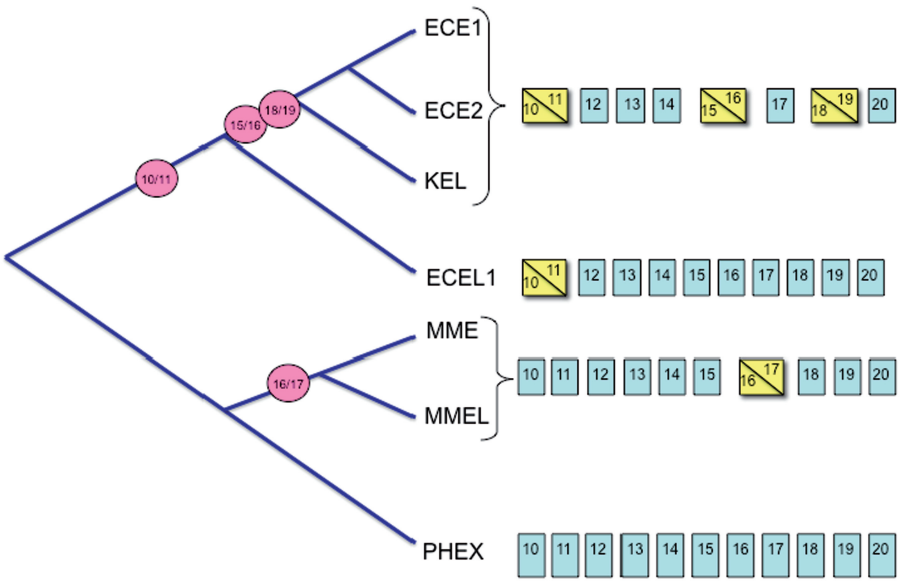

Figure 4. Phylogenetic relationships among the members of the MET13 gene family in human based on the structural comparison of the exon–intron gene structure and detected intron loss events. The paralogous genes considered are: *ECEL1* (NM_004826), *ECE2* (NM_014693), *ECE1* (NM_001397), *KEL* (NM_000420), *MME* (NM_000902), *PHEX* (NM_000444), *MMEL1* (NM_033467). Putative intron losses are marked by a pink circle where the numbers of the two flanking exons are reported.

can be easily identified, for example by performing a database search on human genes using one member of the family as query. Figure 4 shows the results obtained by performing local comparisons of the seven human paralogous (duplicated) genes against *ECEL1* (used as query to retrieve the others), which produced alignments usually starting from the 10th exon (corresponding to the exons encoding for the conserved M13 domain). It can be clearly seen that the phylogenetic relationships between the paralogous proteins (Figure 4, left) can be derived from the pattern of four specific intron loss events (Figure 4, right). Indeed, a sequence-based phylogeny obtained with the neighbor-joining method on JTT amino acid distances produces a similar tree but does not solve the sister–group relationships between the clade including *ECE1* and *ECE2* and either *KEL* or *ECEL1* (see the tree shown in Supplementary Figure 1).

Each of the seven mouse MET13 genes has a structure identical to its human homolog (defined as best reciprocal protein BLAST hit, data not shown), clearly pointing to the fact that human–mouse divergence took place after the duplications leading to the seven genes of the family.

Using Exalign to perform large-scale comparisons

As with sequence alignments, our algorithm can be used to perform an exhaustive whole-genome or whole-database comparisons (see 'Methods' section), by using each of the gene annotations available for a given organism as query against the genes of another species (or the species itself, to investigate the evolution of the structure of paralogous gene families as in the previous example). In particular, we tested the performance of the algorithm by using all human RefSeq gene structures available as query against the set of mouse RefSeq genes. The alignment strategy used was local, using as query internal exons only. In particular, we compared the results obtained with Exalign to the human–mouse orthologous gene annotations available in the HGNC/HCOP Comparison of Orthology Predictions database (11). We discarded the genes whose annotated mouse ortholog had a XM_ RefSeq annotation or a discontinued NM_ annotation. This resulted in a total of 12 581 human query RefSeq gene annotations with at least six exons (thus at least four internal exons used in the search), while shorter genes could not yield significant *E*-values (lower than one, see further on).

The results obtained are summarized in Table 1. Almost 93% of the genes had as best match the annotated mouse ortholog (and vice versa). Another 4% had as best match a duplicated mouse gene paralogous to the annotated ortholog, with virtually identical structure, thus making impossible for the algorithm, in these cases, to discriminate between the two (or more) similar mouse genes on the basis of their intron–exon structure only. All in all, only the 3% of the genes considered had as best match a gene not annotated as their ortholog in the database, or another of the same family with identical structure. But, less than half of these 'misses' could be considered significant according to the *E*-value, using a quite loose threshold of 1 (that is, less than one alignment with the same score could be expected by chance). Also quite interestingly, more than half of these potential 'false positive' hits matched nevertheless a gene of the same family (or containing a conserved protein domain), which had a structure more conserved than the annotated ortholog (in most of the cases, this in turn was the effect of truncated gene isoforms in mouse with their N or C terminal portion missing w.r.t. the human gene, nevertheless annotated as orthologous in the database). All in all, genes with a real 'false positive match' were about the 0.5% of the total, usually with a 'borderline' *E*-value

Table 1. Results of the search performed against the mouse RefSeq genes using as query all human RefSeq genes with at least 6 exons, with an annotated NM_ RefSeq mouse orthologous gene in the HCOP database (11)

Total number of genes	True positives	Identical positives	Family positives	No hit	False positives
12 581	11 656 (92.6%)	553 (4.3%)	92 (0.7%)	214 (1%)	66 (0.5%)

True positives had the annotated ortholog as best match (and vice versa); identical positives had as best match a mouse gene paralogous to the annotated ortholog with the same structure (and/or vice versa); family positives had as best hit a gene different from the orthologous one but showing significant similarity at the sequence level; no hit did not have any match with E-value lower that 1; false positive matched with *E*-value lower than 1 a gene without any significant sequence similarity with the query one.

slightly lower than 1. The most common cause of error, apart from truncated isoforms or mis-annotations of the coding sequence (leading to exons with completely different reading frame significantly lowering the alignment scores) were short (no more than four or five internal exons) genes with two or three fully non-coding exons building the 5′ or 3′ untranslated regions. These non-coding exons were usually not conserved both in size and in number, with genes having good matches only in the two or three exons actually containing the conserved coding sequences, not enough to produced a high scoring alignment with the annotated ortholog and, more important, obtain a significant *E*-value.

While *E*-values in search results, according to their definition, depend on the size of the 'database' (expressed as the total number of exons) against which a query is performed, the results of our experiments point out that they seem quite feasible in a search against a mammalian or vertebrate genome. For cases in which the number of exons of a gene is not enough a priori to obtain a significant value (i.e. genes with five or less exons), an alternative possibility is to perform the search using global (for small genes) or glocal (to detect more efficiently truncated isoforms) alignments instead of local ones. Furthermore search results can be limited by the interface to genes yielding also significant sequence similarity, either at the nucleotide or the protein level.

Using Exalign to detect intron gain/loss events

Other than providing a good feedback on the performance of the algorithm and hints to the best strategy for defining alignment scores and the respective expected values (see 'Methods' section), these human-mouse comparison, together with an analysis of all rat genes compared to mouse genes, yielded indications of as yet unreported intron gain and loss events. It should be noted that for the detection of intron gains and/or losses we limited the search to 'missing' introns yielding a novel exon equaling the size of two consecutive exons in another gene (exact merged exon match, as defined in the 'Methods' section), and located between internal and coding exons only. When the result suggested an intron gain/loss event, we submitted the genes involved to further analyses including additional species, aimed at a better characterization of the evolution of the gene and the most likely event (gain or loss). All in all, we were able to detect most of the cases reported in the human/mouse/rat/dog comparison of Coulombe-Huntington *et al*. (5), except those in which the lost introns were near external exons or in the untranslated regions of the genes, that we did not include in our comparison. The novel results are summarized in Table 2, also including one possible case of intron gain in human and another one (involving two consecutive introns) in mouse. Remarkably, we were able to identify additional intron loss events in either human or rodents (either mouse or rat) that on the basis of genomic alignments used in the genome-wide study of Coulombe-Huntington *et al*. (5) could not have been detected, mainly because of 'ambiguity' in genomic alignments, where a given human genomic region could not be aligned unambiguously to regions of the other genomes (thus omitting from the analysis most of the gene families having paralogs in a single species, as in the case of MET13). This is the case, for example, of the double intron loss in rat gene *FDPS* (NM_031840), and single intron losses in rat *ATG3* (NM_134394), *MCM7* (NM_001004203), *MRPL2* (NM_001034136), *GNB2* (NM_031037), *RPL5* (NM_031099), *RRM2* (NM_001025740). In all these cases, human,

Table 2. Summary of intron loss and gain events detected by genome-wide applications of exalign, with gene ID, mouse and rat (where available, predicted genes are in italic) RefSeq IDs, and position of the lost/gained intron (with respect to the structure of the human gene, unless otherwise specified)

Gene ID	RefSeq human	RefSeq mouse	RefSeq rat	Position
Human intron losses				
PTCD1	NM_015545	NM_133735	NM_001109665	Exon 6
PRDM10	NM_199437	NM_001080817	n.a.	Exon 17
RPH3A	NM_014954	NM_011286	NM_133518	Exon 3
Col25a1	NM_198721	NM_029838	n.a.	Exon 4
Col13a1	NM_005203	NM_007731	NM_001109172	Exon 4
Mouse (or rodent) intron losses				
PPP2R5D	NM_006245	NM_009358	*(XM_001062510)*	Intron 8
LAMA2	NM_000426	NM_008481	*(XM_219866)*	Intron 43
Rat intron losses				
FDPS	NM_002004	NM_134469	NM_031840	Introns 4–5
ATG3	NM_022488	NM_026402	NM_134394	Intron 8
MCM7	NM_005916	NM_008568	NM_001004203	Intron 3
MRPL2	NM_015950	NM_025302	NM_001034136	Intron 4
FLJ10081	NM_017991	NM_172652	NM_001034835	Intron 13
GNB2	NM_005273	NM_010312	NM_031037	Intron 5
RPL5	NM_000969	NM_016980	NM_031099	Intron 4
RRM2	NM_001034	NM_009104	NM_001025740	Intron 8
Rodent intron losses				
OSGEPL1	NM_022353	NM_028091	NM_001024787	Intron 4
RPS2	NM_002952	NM_008503	NM_031838	Intron 5
Human intron gain				
OVCH2	NM_198185	NM_172908	n.a.	Intron 11
Mouse intron gain (position w.r.t. mouse gene)				
MOSC1	NM_022746	NM_001081361	*(XM_001075480)*	Introns 3–4

mouse and other vertebrate (when available) gene structures showed the presence of the introns lacking in rat. Similarly, human gene *PTCD1* (NM_015545) also clearly lacks an intron when compared to its rodent homologs, but has not been reported so far.

An intron loss was detected in mouse for gene *PPP2R5D* (NM_009358) with respect to human, chicken, zebrafish and fugu. The absence of any annotated homologous gene in rat makes difficult to determine whether the intron loss is limited to mouse or is detectable in both rodents (predicted rat gene XM_001062510 however, seems to indicate that the loss is restricted to mouse).

Likewise, *OSGEPL1* (human NM_022353) shows an intron loss in both rodents (the structure of the zebrafish and tetraodon genes is identical to the human one), again undetectable from genomic alignments. Gene PR domain containing 10 (*PRDM10*) shows an apparent intron loss in human, detectable by the comparison to its orthologs in mouse and fugu.

Another problem deriving from genomic alignments is that small exons are often hard to be aligned correctly. This is the case of genes *RPH3A* (NM_014954), showing an intron loss in human, intron present in all the homologs in mouse, rat and cow; *LAMA2* (NM_000426), lacking an intron in mouse (no reliable annotation is available for a rat homolog, except for predicted gene XM_219866 that limits the loss to mouse); procollagen *Col13a1* (NM_0 05203) and procollagen *Col25a1* (NM_198721) both with an intron loss in human. In all these cases, we carefully checked transcript-genome alignments to see whether the additional introns present in one or more species could be the result of a wrong annotation due to mis-alignments of mRNA sequences against the genome.

Detection of intron gain events in human and mouse

Previous large-scale comparisons in mammals (2,5) did not report any intron gain event either in human or in rodents. However, a very interesting result was obtained by Exalign on the ovochymase 2 (*OVCH2*) gene (NM_1 72908 in mouse, NM_198185 in human). Exalign merged two human exons in the mouse/human alignment (Figure 2A), suggesting a possible intron loss in mouse (or rodents) not previously reported. Unfortunately, no *OVCH2* ortholog is annotated in species other than human and mouse to further address this hypothesis, apart from two predicted opossum and chicken genes (XM_00 1377916, XM_01232534) whose structure nevertheless does not include the additional human intron. However, inspection of the 318 bp long human intron in the UCSC genome browser (Figure 2B) revealed that it exactly corresponds to an Alu repeat belonging to the Y5a young subfamily (12) (Supplementary Figure 2). Following the Alu insertion, the 'intronization' of the Alu element is likely to have occurred through the activation of non-canonical 5′ and 3′ splice sites. The Alu insertion occurred after the human–chimp divergence as the Alu-like intron is missing in the chimp genome (data not shown) and interestingly a perfect 12-mer repeat can be observed at the insertion site. A possible mechanism explaining this putative intron gain is described by Giroux *et al.* (13) where a recent intron gain in maize sh2-m1 gene is described, exactly corresponding to a transposable element, and a model is proposed to explain its precise removal. However, we cannot exclude that the Alu insertion is an allelic form not present in the genomic sequence used as template for the OVCH2 mRNA collected in Genbank.

Mouse gene *MOSC1* (NM_001081361) has two additional introns (the third and the fourth in the mouse RefSeq annotation, see the alignment between the mouse MOSC1 transcript and its corresponding genome sequence in Supplementary Figure 3) of 505 and 199 bp, respectively, if compared to its ortholog in human (*MOSC1*, NM_022746) and its paralogous gene in mouse (*MOSC2*, NM_133684). Also, the human MOS C2 homolog (*MOSC2*, NM_017898) does not have the two introns (Figure 5). The same two introns are also missing from all MOSC homologs annotated in other vertebrates, namely in cow, pufferfish and fugu, and also in rat predicted genes XM_001065025 and XM_00107 5480 (annotated as putative *MOSC1* homologs). Apparently, the easiest explanation of these facts is a double intron gain in the mouse gene, or, alternatively, in rodents (no full mRNA sequence is available for rat).

DISCUSSION

Most of the eukaryotic genes contain multiple introns (8.8 on average in human genes), which are spliced out to produce mature transcripts. The increasing availability of several eukaryotic genomes, sequenced together with their transcriptomes, makes now possible the large-scale annotation and comparison of the exon–intron structure of their genes.

Previous studies have shown a remarkable conservation of the exon–intron structure of eukaryotic genes (2–4,14), also confirmed by our results, a feature that can be thus considered a reliable phylogenetic marker at long evolutionary distances. In metazoans, the comparative analysis of exon–intron structures has proved itself to be an effective strategy for the resolution of deep phylogenetic relationships (4), the detection of new and potentially very remote homologies between proteins with very limited sequence similarity (3), or vice versa between proteins without significant sequence variation in orthologous and/or paralogous sequences.

Therefore, in order to build a reliable reconstruction of the history of a gene or a gene family using all available evolutionary information we should equally consider the evolutionary pattern of nucleotide and protein sequences as well as of the exon–intron structure. However, although several methods and software are available for evolutionary and phylogenetic analyses at sequence level (15) to our knowledge there are no available tools designed to carry out pairwise or large-scale comparative analyses of exon–intron gene structures including intron gain and loss events and with a statistical assessment of the significance of the results.

Results like those we present here prove that Exalign can be an effective and reliable tool for the comparative analysis at the exon–intron structure level and study of the

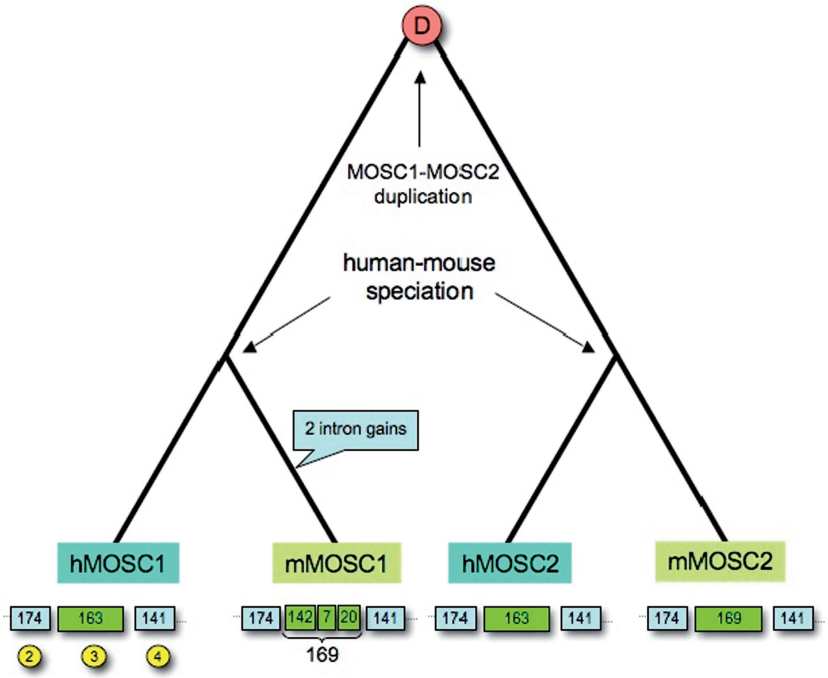

Figure 5. Evolutionary history of the *MOSC* gene family in human and mouse. Mouse gene *MOSC1* presents two additional introns that do not appear in any other annotated vertebrate MOSC gene. No full mRNA sequence for these genes is available in rat, but two predicted XM genes have the same structure of the human ones.

evolutionary history of intron turnover across lineages. Exalign can be used to obtain additional information in the inference of orthologous/paralogous relationships within gene families, a crucial issue for the functional annotation of homologous genes in different species, as well as to reconstruct the evolutionary history of the gene family. The results obtained for the metalloendopeptidase MET13 family (Figure 4) clearly show that the additional information provided by the comparative analysis of the exon–intron gene structures may clarify critical relationships between members of the same family left unsolved by sequence-based phylogenetic reconstructions.

By using Exalign we not only detected additional cases on intron loss events either in human or in rodents with respect to previous genome-wide analyses carried out with a different methodology (5), but also observed for the first time possible intron gain events in both human and mouse (results summarized in Table 1).

Exalign could also be used to detect errors in the annotations of gene structures and coding sequences, often resulting from automatic genome wide analyses. In fact, the observation of an exon–intron structure showing greater divergence in closely related species than in more distant ones may be the result of a wrong annotation or of the comparison of unrelated alternative splicing isoforms. In the latter case, Exalign can also be used to compare alternative splicing isoforms of orthologous genes for the identification of functionally related transcripts.

Finally, the combination of data derived from sequence-based evolutionary analyses and from studies on the evolution of the exon–intron structure may also reveal peculiar features such as unequal rates of intron gain and intron loss in different lineages.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the EU STREP Project 'TRANSCODE', Laboratorio Internazionale di Bioinformatica (LIBI) and Laboratorio di Bioinformatica per la Biodiversità Molecolare (Ministero dell'Università e della Ricerca), AIRC (Associazione Italiana Ricerca sul Cancro) and Telethon. Funding to pay the Open Access publication charges for this article was provided by LIBI.

Conflict of interest statement. None declared.

REFERENCES

1. Roy,S.W. and Gilbert,W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.*, **7**, 211–221.
2. Roy,S.W. and Gilbert,W. (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl Acad. Sci. USA*, **102**, 5773–5778.
3. Betts,M.J., Guigo,R., Agarwal,P. and Russell,R.B. (2001) Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? *EMBO J.*, **20**, 5354–5360.
4. Roy,S.W. and Gilbert,W. (2005) Resolution of a deep animal divergence by the pattern of intron conservation. *Proc. Natl Acad. Sci USA*, **102**, 4403–4408.
5. Coulombe-Huntington,J. and Majewski,J. (2007) Characterization of intron loss events in mammals. *Genome Res.*, **17**, 23–32.
6. Brown,N.P., Whittaker,A.J., Newell,W.R., Rawlings,C.J. and Beck,S. (1995) Identification and analysis of multigene

families by comparison of exon fingerprints. *J. Mol. Biol.*, **249**, 342–359.
7. 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.
8. Qiu,W.G., Schisler,N. and Stoltzfus,A. (2004) The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol.*, **21**, 1252–1263.
9. Babenko,V.N., Rogozin,I.B., Mekhedov,S.L. and Koonin,E.V. (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.*, **32**, 3724–3733.
10. Roy,S.W. and Penny,D. (2007) On the incidence of intron loss and gain in paralogous gene families. *Mol. Biol. Evol.*, **24**, 1579–1581.
11. Eyre,T.A., Wright,M.W., Lush,M.J. and Bruford,E.A. (2007) HCOP: a searchable database of human orthology predictions. *Brief Bioinform.*, **8**, 2–5.
12. Batzer,M.A. and Deininger,P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.*, **3**, 370–379.
13. Giroux,M.J., Clancy,M., Baier,J., Ingham,L., McCarty,D. and Hannah,L.C. (1994) De novo synthesis of an intron by the maize transposable element Dissociation. *Proc. Natl Acad. Sci. USA*, **91**, 12150–12154.
14. Rogozin,I.B., Sverdlov,A.V., Babenko,V.N. and Koonin,E.V. (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform.*, **6**, 118–134.
15. Horner,D.S. and Pesole,G. (2004) Phylogenetic analyses: a brief introduction to methods and their application. *Expert Rev. Mol. Diagn.*, **4**, 339–350.