An idiosyncratic feature of eukaryotic gene organization is that accordingly, it is fairly certain that splicing was already present in the genomic sequences of protein-coding genes are frequently interrupted by non-coding sequences, called introns, which are excised (spliced) from the transcripts prior to translation. Fundamental constituents of the splicing machinery are present throughout main eukaryotic lineages (Collins and Penny, 2005). Intron-containing genes are spread across diverse eukaryotic phyla, and orthologous genes often have similar exon-intron organization even at large evolutionary distances (Rogozin et al., 2003). Accordingly, it is fairly certain that splicing was already present in the last common ancestor of eukaryotes (Rodríguez-Trelles et al., 2006). Gene structures changed to different extents in eukaryotic lineages (Roy and Gilbert, 2006).

Whole-genome sequencing projects have made it possible to perform large-scale phylogenetic analyses that scrutinize the evolution of exon–intron organization. Following the pioneering study by Rogozin et al. (2003), numerous results have appeared (Carmel et al., 2005; Carmel et al., 2007; Csűrös, 2005; Csűrös et al., 2007, 2008; Nguyen et al., 2005; Nielsen et al., 2004; Roy and Gilbert, 2005; Roy and Penny, 2006; Stajich et al., 2007; Sullivan et al., 2006) inferring lineage- and gene-specific features of gene structure evolution, and often describing methodological novelties. This note aims to introduce Malin, a software package developed for the analysis of eukaryotic gene structure evolution.

2 FEATURES

Malin provides a graphical user interface for various tasks commonly used to infer the evolution of exon–intron structure in multiple protein-coding ortholog sets (Fig. 1) along a fixed species phylogeny. The implemented tasks include the following:

- Identification of conserved homologous splice sites in annotated protein sequence alignments.
- Computation of primary statistics about introns in homologous sites (‘shared introns’).
- Estimation of ancestral intron content, intron losses and gains by Dollo parsimony.
- Estimation of intron loss and gain rates in a probabilistic model.
- Estimation of ancestral intron content, intron losses and gains in a probabilistic model.
- Inference of histories at individual or multiple sites.
- Error estimation for rates and histories by bootstrap.

Figure 1 illustrates the typical analysis pipeline for eukaryotic gene structure evolution (Rogozin et al., 2005). In order to infer spliceosomal introns in homologous positions, splice sites need to be projected onto coding sequences, and then homology is established in conserved regions of the protein alignments. An intron table is constructed from the projected intron annotations. The table is a binary table of intron presence and absence in homologous sites across the studied organisms: Malin can also cope with ambiguous characters. The patterns can be analyzed by Dollo parsimony (Farris, 1977) (assuming that intron gains and losses are rare events), or by probabilistic models of intron evolution. Malin works with the likelihood framework that I have elaborated (Costros, 2005; Costros et al., 2007, 2008). The corresponding probabilistic model has branch-specific intron gain and loss rates, as well as rates-across-sites variation.

Malin uses a rates-across-sites Markov model for intron evolution, with branch-specific gain and loss rates. If no rate variation is assumed across the sites, then every branch has just a gain and loss rate, with corresponding gain and loss probabilities. Briefly, an intron is lost on an edge of length $\tau$ with probability $\frac{\lambda}{\mu} \cdot \mu e^{-\mu \cdot \lambda}$ where $\lambda$ and $\mu$ are the gain and loss rates; a new intron appears in a previously unoccupied site with probability $\frac{\lambda}{\mu} \cdot \lambda (1-e^{-\mu \cdot \lambda})$. The constant rate model (Csűrös et al., 2007) is completely specified by the branch-specific gain/loss rates, and
the probability with which intron sites are occupied at the root. The rate variation model (Csörös et al., 2008) assumes that intron sites belong to discrete rate categories. Each site category is defined by a pair of loss and gain rate factors \((\alpha, \beta)\), so that the loss rates \(\lambda \alpha\) and gain rates \(\lambda \beta\) apply on each edge with prototypical rates \(\alpha\) and \(\lambda\). Malin optimizes rate factors, and can analyze the same dataset with different models simultaneously.

Malin is written entirely in Java. It can be used on any computer platform with a Java Runtime Environment (implementing J2SE 1.5 or higher), including Microsoft Windows, MacOS X and Linux. In addition, Malin is also available as an integrated application on MacOS X. The software is distributed with test data and a detailed User’s Guide. Input files commonly used formats: Newick format for alignments and the syntax used by Rogozin et al. (2003) for intron tables. Intron sites are specified in Fasta headers. Analysis results can be exported into tab-delimited text files.

The software implements previously described computational innovations (Csörös et al., 2007, 2008), including rate optimization, posterior predictions, fast evaluation of the likelihood function and estimation of statistical confidence through bootstrapping. Malin provides a feature-rich graphical user interface for the analysis tasks. Figure 1 gives an example of an alignment panel, where, in order to build an intron table, the user selects the conservation criteria (such as the minimum number of gapless positions next to an intron site) for discerning homologous sites in a set of multiple alignments.

Ideally, Malin will enable researchers to conduct phylogenetic gene structure analysis with the same ease that is currently available for molecular sequences.

ACKNOWLEDGEMENTS

I am grateful to Péter Csörös for help with the software integration in Microsoft Windows. I am greatly indebted to Liran Carmel, Eugene Koonin, Jacek Majewski, Igor Rogozin and Scott Roy for helpful advice and discussions about intron evolution.

**Funding:** This research project has been supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

**Conflict of Interest:** none declared.

**REFERENCES**

Carmel L. et al. (2005) An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. *Lect. Notes Comp. Sci.*, 3678, 35–46.

Carmel L., et al. (2007) Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.*, 17, 1034–1044.

Collins L. and Penny D. (2005) Complex splicescenal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, 22, 1053–1066.

Csörös M. (2005) Likely scenarios of intron evolution. *Lect. Notes Comput. Sci.*, 3678, 47–60.

Csörös M. et al. (2007) In search of lost introns. *Bioinformatics*, 23, 47–56.

Csörös M., et al. (2008) Extremely intron-rich genes in the absolute ancestors inferred with a flexible maximum likelihood approach. *Mol. Biol. Evol.*, 25, 903–911.

Farris J.S. (1977) Phylogenetic analysis under Dollo’s law. *Syst. Zool.*, 26, 77–88.

Niang J.-D. et al. (2005) New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.*, 1, e79.

Rodriguez-Trelles F. et al. (2006) Origins and evolution of splicescencial introns. *Annu. Rev. Genet.*, 40, 47–76.

Roy S.W. and Gilbert W. (2005) Rates of intron loss and gain: implications for early eukaryotic evolution. *Curr. Biol.*, 15, 1512–1517.

Roy S.W. et al. (2005) Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief. Bioinform.*, 6, 118–134.

Roy S.W. and Gilbert W. (2006) The evolution of splicescencal introns: patterns, puzzles and progress. *Nat. Rev. Genet.*, 7, 211–221.

Roy S.W. and Hart D.L. (2006) Very little intron loss/gain in plasmodium: intron loss/gain mutation rates and intron numbers. *Genome Res.*, 16, 750–756.

Roy S.W. and Penny D. (2006) Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. *Genome Res.*, 16, 1270–1275.

Stajich J.E. et al. (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.*, 8, R223.

Sullivan J.C. et al. (2006) A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan Nematoselloides vectensis. *Genome Inform.*, 17, 217–229.