BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

Malin: maximum likelihood analysis of intron evolution in eukaryotes

Miklós Csűrös [a,b]

[a] Department of Computer Science and Operations Research, University of Montréal, Montréal, Québec, Canada. [b] Collegium Budapest Institute for Advanced Study, Budapest, Hungary.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Malin is a software package for the analysis of eukaryotic gene structure evolution. It provides a graphical user interface for various tasks commonly used to infer the evolution of exon-intron structure in protein-coding orthologs. Implemented tasks include the identification of conserved homologous intron sites in protein alignments, as well as the estimation of ancestral intron content, lineage-specific intron losses and gains. Estimates are computed either with parsimony, or with a probabilistic model that incorporates rate variation across lineages and intron sites.

Availability: Malin is available as a stand-alone Java application, as well as an application bundle for MacOS X, at the website http://www.iro.umontreal.ca/~csuros/introns/malin/. The software is distributed under a BSD-style license.

Contact: csuros@iro.umontreal.ca.

1 INTRODUCTION

An idiosyncratic feature of eukaryotic gene organization is that 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, 2007; Csűrös, 2005; Csűrös *et al.*, 2007; Csűrös *et al.*, 2008; Nguyen *et al.*, 2005; Nielsen *et al.*, 2004; Roy and Gilbert, 2005; Roy and Hartl, 2006; 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 (see Figure 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 if spliceosomal introns are 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 (Csűrös, 2005; Csűrös *et al.*, 2007; Csűrös *et al.*, 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 t with probability $\frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right)$ where λ and μ are the gain and loss rates; a new intron appears in a previously unoccupied site with probability $\frac{\lambda}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right)$. The constant rate model (Csűrös *et al.*, 2007)

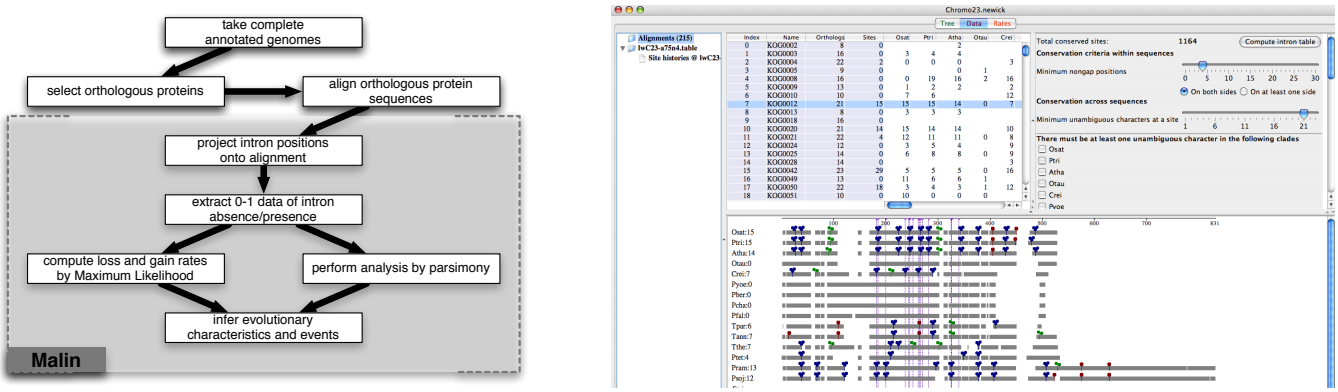

Fig. 1. Left: Typical analysis pipeline for intron evolution. MALIN can perform the tasks downstream of ortholog identification and alignment. **Right:** Alignment panel in MALIN. The intron table will be constructed from a set of multiple alignments (corresponding to the rows of the table displayed in the middle on the top), based on conservation criteria specified by the user (through the form on the upper right). The bottom half of the panel plots an illustration for the selected alignment, showing alignment gaps and projected intron sites (colored tags).

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 (Csűrös *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 (α, β), so that the loss rates $\mu\alpha$ and gain rates $\lambda\beta$ apply on each edge with prototypical rates μ and λ. MALIN optimizes rate factors, and can analyze the same data set 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 follow commonly used formats: Newick format for the possibly multifurcating species phylogeny, Fasta 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 (Csűrös *et al.*, 2007; Csűrös *et al.*, 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

This research project has been supported by a grant from the Natural Sciences and Engineering Research Council of Canada. I am grateful to Péter Csűrös 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.

REFERENCES

Carmel, L., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2005). An expectation-maximization algorithm for analysis of evolution of exon-intron structure of eukaryotic genes. *Lecture Notes in Computer Science*, **3678**, 35–46.

Carmel, L., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.*, **17**, 1034–1044.

Collins, L. and Penny, D. (2005). Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.*, **22**(4), 1053–1066.

Csűrös, M., Rogozin, I. B., and Koonin, E. V. (2008). Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum likelihood approach. *Mol. Biol. Evol.* Forthcoming.

Csűrös, M. (2005). Likely scenarios of intron evolution. *Lecture Notes in Computer Science*, **3678**, 47–60.

Csűrös, M., Holey, J. A., and Rogozin, I. B. (2007). In search of lost introns. *Bioinformatics*, **23**(13), i87–i96.

Farris, J. S. (1977). Phylogenetic analysis under Dollo's law. *Syst. Zool.*, **26**(1), 77–88.

Nguyen, H. D., Yoshihama, M., and Kenmochi, N. (2005). New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.*, **1**(7), e79.

Nielsen, C. B., Friendman, B., Birren, B., Burge, C. B., and Galagan, J. E. (2004). Patterns of intron gain and loss in fungi. *PLoS Biology*, **2**(12), e422.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2006). Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.*, **40**, 47–76.

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.

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., and Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Briefings in Bioinformatics*, **6**(2), 118–134.

Roy, S. W. and Gilbert, W. (2005). Rates of intron loss and gain: Implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci. USA*, **102**(16), 5773–5778.

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

Roy, S. W. and Hartl, D. L. (2006). Very little intron loss/gain in plasmodium: Intron loss/gain mutation rates and intron number. *Genome Res.*, **16**(6), 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**(10), 1270–1275.

Stajich, J. E., Dietrich, F. S., and Roy, S. W. (2007). Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.*, **8**(10), R223.

Sullivan, J. C., Reitzel, A. M., and Finnerty, J. R. (2006). A high percentage of introns in human genes were present early in animal evolution: Evidence from the basal metazoan Nematostella vectensis. *Genome Informatics*, **17**, 217–229.