BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 22 2010, pages 2914–2915
doi:10.1093/bioinformatics/btq549

Phylogenetics

Advance Access publication September 27, 2010

GLOOME: gain loss mapping engine

Ofir Cohen[1], Haim Ashkenazy[1], Frida Belinky[2], Dorothée Huchon[2,3] and Tal Pupko[1,3,*]

[1]Department of Cell Research and Immunology, [2]Department of Zoology, Tel-Aviv University, Tel Aviv 69978, Israel and [3]National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200 Durham, NC 27705-4667, USA

Associate Editor: David Posada

ABSTRACT

SUMMARY: The evolutionary analysis of presence and absence profiles (phyletic patterns) is widely used in biology. It is assumed that the observed phyletic pattern is the result of gain and loss dynamics along a phylogenetic tree. Examples of characters that are represented by phyletic patterns include restriction sites, gene families, introns and indels, to name a few. Here, we present a user-friendly web server that accurately infers branch-specific and site-specific gain and loss events. The novel inference methodology is based on a stochastic mapping approach utilizing models that reliably capture the underlying evolutionary processes. A variety of features are available including the ability to analyze the data with various evolutionary models, to infer gain and loss events using either stochastic mapping or maximum parsimony, and to estimate gain and loss rates for each character analyzed.

Availability: Freely available for use at http://gloome.tau.ac.il/

Contact: talp@post.tau.ac.il

Received on August 10, 2010; revised on September 20, 2010; accepted on September 21, 2010

1 INTRODUCTION

Numerous biological characteristics are coded using binary characters to denote presence (‘1’) versus absence (‘0’). The 0/1 matrix is termed a phylogenetic profile of presence–absence or phyletic pattern and is equivalent to a gap-free multiple sequence alignment (MSA), in which rows correspond to species and columns correspond to binary characters. Phyletic pattern representation is useful in the analysis of various types of biological data including restriction sites (Felsenstein, 1992; Nei and Tajima, 1985; Templeton, 1983); indels (Belinky *et al*., 2010; Simmons and Ochoterena, 2000); introns (Carmel *et al*., 2007; Csuros, 2006); gene families (Cohen *et al*., 2008; Hao and Golding, 2004; Mirkin *et al*., 2003) and morphological characters (Ronquist, 2004). Interestingly, even questions in fields other than biology can be addressed by this approach. For example, the evolution of human languages was studied by analyzing the phyletic patterns of lexical units (Gray and Atkinson, 2003).

Following the development of realistic probabilistic models describing the evolution of DNA and protein sequences, the analysis of phyletic patterns data has progressed from the traditional parsimony (Mirkin *et al*., 2003) to models, in which the dynamics of gain ($0 \rightarrow 1$) and loss ($1 \rightarrow 0$) is assumed to follow a continuous-time Markov process (Csuros, 2006; Hao and Golding, 2006; Spencer and Sangaralingam, 2009). Probabilistic-based analysis of phyletic patterns is currently available in programs such as RESTML (Felsenstein, 1992), MrBayes (Ronquist and Huelsenbeck, 2003) and Count (Csuros, 2010). Nevertheless, for the inference of branch-site-specific events the parsimony criterion is still the most commonly used methodology.

However, the parsimony paradigm may be misleading (Felsenstein, 1978; Pol and Siddall, 2001; Swofford *et al*., 2001; Yang, 1996), especially in characters experiencing multiple (recurrent) events along longer branches (Suzuki and Nei, 2001). Towards a more accurate inference of gain/loss events, we have recently integrated stochastic mapping approaches (Minin and Suchard, 2008; Nielsen, 2002) to accurately map gain and loss events onto each branch of a phylogenetic tree. The analysis is based on novel mixture models, in which variability in both the gain and loss rates is allowed among gene families (Cohen and Pupko, 2010). We have shown that our mixture models are robust and accurate for the inference of gene family evolutionary dynamics (Cohen and Pupko, 2010).

Here, we developed the user-friendly Gain and Loss Mapping Engine (GLOOME) web server. The main novelties of our web server are: (i) we implement probabilistic models that are not implemented elsewhere, which better capture gain/loss dynamics; (ii) we provide accurate estimates of the expectations and probabilities of both gain and loss events using stochastic mapping; and (iii) the interface via a user-friendly web server should make 0/1 analyses more accessible compared to other stand-alone programs.

2 AVAILABLE FEATURES AND METHODS

The required input is a phyletic pattern provided as a 0/1 MSA. A phylogenetic tree is either provided as input by the user or estimated from the phyletic pattern.

2.1 Evolutionary model

The available probabilistic models range from simple to more sophisticated ones that may capture the gain and loss dynamics more reliably. For details regarding the models, see Cohen and Pupko (2010). There are three options for gain and loss rates: (i) ‘Equal gain and loss’—the probability of a gain event is assumed to be equal to that of a loss event; (ii) ‘Fixed gain/loss ratio’—gain and loss probabilities may be different but the gain/loss ratio is identical across all characters and (iii) ‘Variable gain/loss ratio (mixture)’—gain/loss ratio varies among characters.

Simple models assume that a single evolutionary rate characterizes all characters. Our models further allow for character rate variation, assuming that the rate is either gamma distributed or gamma distributed with an additional invariant rate category.

*To whom correspondence should be addressed.

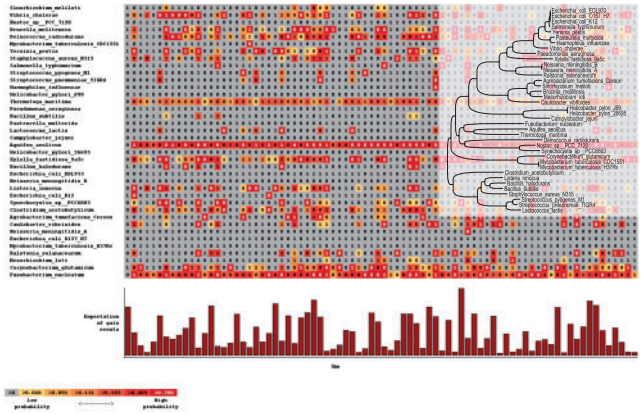

Fig. 1. Stochastic mapping inference of gain events. The size of the bar below each character indicates the sum of expected gain events over all branches. Each character in the phyletic pattern is color coded according to the probability of a gain event in this character and within the branch leading to this species (presentation is also available for loss events). (Insert) Tree with branch lengths proportional to the total number of gain and loss events.

In stationary processes, the character frequencies are equal across the entire tree. Since this assumption may not hold in certain evolutionary scenarios (Cohen *et al*., 2008), we provide the option 'Allow the root frequencies to differ from the stationary ones' to analyze the data using non-stationary models.

A column of only '0's (the character is absent in all taxa) is usually not observable in phyletic patterns. Maximum-likelihood analyses must be corrected for such unobservable data. We allow several such corrections under the menu 'Correction for un-observable data'.

2.2 Stochastic mapping

The stochastic mapping approach infers for each branch and each character the probability and expected number of both gain and loss events. These probabilities depend on the evolutionary model, the tree and its associated branch lengths. This mapping is provided both textually and visually (Fig. 1).

2.3 Parsimony

Our server allows the inference of gain and loss events under the parsimony criterion. The relative costs of gain and loss events can be modified by the user.

2.4 Additional features

In addition to the inference of gain and loss events we further provide: (i) the posterior estimation of the relative rate of each character; (ii) a separate estimation of the gain and loss rates for each character, for mixture model only; (iii) the log-likelihood of the entire tree and for each character; and (iv) the tree and its associated branch lengths estimated from the phyletic pattern, where tree topology is reconstructed using the neighbor-joining method (Saitou and Nei, 1987), from pair-wise maximum-likelihood (ML) distances. For the ML computation, we assume that the rate of gain (loss) is proportional to the frequency of 1 (0) in the data.

While the server is designed with a novice user in mind, we provide several advanced options for expert users, available under the 'Advanced' menu. For example, running times can be accelerated by changing the optimization level. Additionally, likelihood estimation of parameters can be avoided by setting their values based on character counts directly from the phyletic pattern. There are also several options to correct for missing data (explained in the web server under OVERVIEW->METHODOLOGY).

Funding: Israel Science Foundation (878/09 and 600/06, respectively to T.P. and D.H.); D.H. and T.P. are also supported by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606. O.C. and H.A. are fellows of the Edmond J. Safra program in bioinformatics.

Conflict of Interest: none declared.

REFERENCES

Belinky,F. *et al*. (2010) Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol. Biol. Evol.,* **27**, 441–451.

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

Cohen,O. and Pupko,T. (2010) Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol. Biol. Evol.*, **27**, 703–713.

Cohen,O. *et al*. (2008) A likelihood framework to analyse phyletic patterns. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **363**, 3903–3911.

Csuros,M. (2006) On the estimation of intron evolution. *PLoS Comput. Biol.*, **2**, e84.

Csuros,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.

Felsenstein,J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.*, **27**, 401–410.

Felsenstein,J. (1992) Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution*, **46**, 159–173.

Gray,R.D. and Atkinson,Q.D. (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, **426**, 435–439.

Hao,W. and Golding,G.B. (2004) Patterns of bacterial gene movement. *Mol. Biol. Evol*., **21**, 1294–1307.

Hao,W. and Golding,G.B. (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res*., **16**, 636–643.

Minin,V.N. and Suchard,M.A. (2008) Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.,* **56**, 391–412.

Mirkin,B.G. *et al*. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.

Nei,M. and Tajima,F. (1985) Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. *Mol. Biol. Evol*., **2**, 189–205.

Nielsen,R. (2002) Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.

Pol,D. and Siddall,M.E. (2001) Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics*, **17**, 266–281.

Ronquist,F. (2004) Bayesian inference of character evolution. *Trends Ecol. Evol.*, **19**, 475–481.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Simmons,M.P. and Ochoterena,H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, **49**, 369–381.

Spencer,M. and Sangaralingam,A. (2009) A phylogenetic mixture model for gene family loss in parasitic bacteria. *Mol. Biol. Evol.*, **26**, 1901–1908.

Suzuki,Y. and Nei,M. (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, **18**, 2179–2185.

Swofford,D.L. *et al*. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, **50**, 525–539.

Templeton,A.R. (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution*, **37**, 221–244.

Yang,Z. (1996) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, **42**, 294–307.