Phylogenetics

GLOOME: gain loss mapping engine

Ofir Cohen¹, Haim Ashkenazy¹, Frida Belinky², Dorothée Huchon²,³ and Tal Pupko¹,³,*

¹Department of Cell Research and Immunology, ²Department of Zoology, Tel-Aviv University, Tel Aviv 69978, Israel and ³National 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 (phylectic patterns) is widely used in biology. It is assumed that the observed phylectic pattern is the result of gain and loss dynamics along a phylogenetic tree. Examples of characters that are represented by phylectic 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/

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 phylectic pattern and is equivalent to a gap-free multiple sequence alignment (MSA), in which rows correspond to species and columns correspond to binary characters. Phylectic 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 phylectic 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 phylectic patterns has progressed from the traditional parsimony (Mirkin et al., 2003) to models, in which the dynamics of gain (0→1) and loss (1→0) is assumed to follow a continuous-time Markov process (Csuros, 2006; Hao and Golding, 2006; Spencer and Sangaralingam, 2009). Probabilistic-based analysis of phylectic 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 phylectic pattern provided as a 0/1 MSA. A phylectic tree is either provided as input by the user or estimated from the phylectic 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.
The stochastic mapping approach infers for each branch and each textually and visually (Fig. 1).

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 phylectic 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 character the probability and expected number of both gain and loss events. Maximum-likelihood analyses must be corrected for such unobservable data. We allow several such corrections under the menu ‘Correction for un-observable data’.

### 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 phylectic 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 phylectic pattern. There are also several options to correct for missing data (explained in the web server under OVERVIEW->METHODOLOGY).

### REFERENCES


---

**Conflict of Interest**: none declared.