Stochastic Evolution of Protein Sequences

Studying gains and losses over a phylogenetic tree

Petar Konovski
DCSIS, Birkbeck, University of London
petar@dcs.bbk.ac.uk
Supervisors: Prof. Trevor Fenner and Prof. Boris Mirkin

Abstract

Inferring phylogenetic trees is a general approach in the reconstruction of the evolutionary histories of organisms. In order to estimate the events over a phylogenetic tree, several criterions and algorithms are used. The research presented starts from the popular Maximum Parsimony and Maximum Likelihood criteria and the developed in Birkbeck algorithms for them, PARS and MALS. On top of that, a common approach is proposed, which opens a way to apply additional criteria, such as Minimum Entropy.
The second path of the research starts by considering the events over a branch of the phylogenetic tree as a Markov stochastic process described by the Kolmogorov forward equations. For the case considered, these equations have analytical solution. Based on the expressions for gain rate, loss rate, time and the probabilities for gain and loss in the analytical solution, a system of quasilinear equations for gain rates is formulated over the tree.

Introduction

Mapping the evolutionary events on a phylogenetic tree is a widely used approach to elucidate the inheritance of given treats in a set of species.

Maximum Parsimony

Maximum parsimony (MP) is the oldest approach to the estimate of the events over a phylogenetic tree. It aims to as explain the present-day distribution of the hereditary traits with minimal assumptions about the events which has happened in the common ancestors. Because the loss of the traits is expected to happen more often than the gain, a gain penalty coefficient $G > 1$.
An efficient algorithm, PALS, for computing the Maximum Parsimony is given in [1].

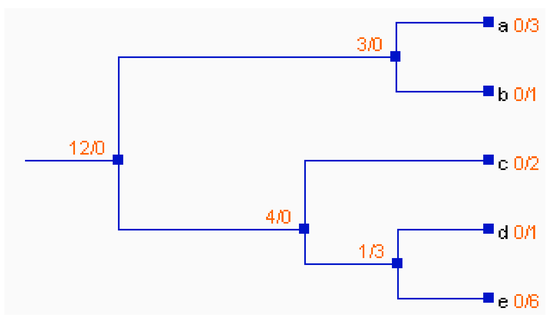

Figure 1: Maximum Parsimony reconstruction of gain/loss events

Maximum Likelihood

Maximum Likelihood is widely used in recent days and is considered more accurate than MP. The method relies on the estimation of the likelihoods for gain or loss in every node of the tree. This allows to apply the calculations on trees with variable branch lengths.
An efficient algorithm, MALS, for computing the Maximum Likelihood over a given phylogenetic tree is given in [2]. Important part of MALS is the application of iterative procedure, which starts with the probabilities for gains and losses, based on the Maximum Parsimony results and continues based on the (improved) probabilities achieved in the last step.

The entropy of a scenario over a phylogenetic tree

Considering the information content of the network events, we can derive a scoring function for ML which resembles that of the MP and allows the usage of an additive algorithm.

The next step is to consider the entropy of the events:
Let γ_N and λ_N be the probabilities for the gain or loss in node N and $\overline{\lambda_N} = 1 - \lambda_N, \overline{\gamma_N} = 1 - \gamma_N$.
The scoring function is

$$H(N) = \begin{cases} -\lambda_N log(\lambda_N) & if\ loss\ happens \\ -\gamma_N log(\gamma_N) & if\ gain\ happens \\ -\overline{\lambda_N} log(\overline{\lambda_N}) & if\ no\ loss\ happens \\ -\overline{\gamma_N} log(\overline{\gamma_N}) & if\ no\ gain\ happens \end{cases}$$

The rationale and the description of the algorithm are given in [3].

Formulation as a Markov process: The solutions of Kolmogorov's forward equations

Further in this section, we follow [4].

Let g be the gain rate and ℓ be the loss rate over the branch. The probabilities for gain and loss over a branch of length t are given by the Kolmogorov's forward equations (a system of ordinary differential equations).

Luckily, the solutions in our case can be found analytically. These are two independent solutions:

$P_{01}(t) = \frac{g}{\ell+g} - \frac{g}{\ell+g} exp(-(\ell+g)t)$ - The trait is absent and will be gained after time t.

$P_{10}(t) = \frac{\ell}{\ell+g} - \frac{\ell}{\ell+g} exp(-(\ell+g)t)$ - The trait is present and will be lost after time t.

Investigate the variations in gain and loss rate over the tree

Expressions for ℓ and t

Given the solutions mentioned above, we can pose the reverse problem:

We do not have any estimates for g, ℓ and t, but we have estimates for $P_{01}(t)$ and $P_{10}(t)$ as an output from the ML algorithm.

From this point of view, the expressions of the solutions can be considered as a nonlinear system of two equations with three unknowns. Let solve the system considering ℓ and t as unknowns and leaving g as a parameter. For simplicity, we set $P_{01}(t) \equiv p$, $P_{10}(t) \equiv q$. Now the system is:

$$\frac{g}{\ell+g} - \frac{g}{\ell+g} \exp(-(\ell+g)t) = p$$

$$\frac{\ell}{\ell+g} - \frac{\ell}{\ell+g} \exp(-(\ell+g)t) = q, \ell \geq 0, g \geq 0, t \geq 0$$

The above system is nonlinear, but luckily, it has an analytical solution:

$$\ell = \frac{g}{p}, t = -\frac{p \ln(1-p-q)}{g(p+q)}$$

These expressions give us the opportunity to proceed with the investigation of the variability of the gain rate over the tree in a more systematic way.

If we can find any estimates for g, the corresponding estimates for ℓ can be found immediately.

What can be inferred from the ML output

We found an expression for t over a single branch of the tree, depending on the gain rate and the probabilities for gain and loss. We can use it in the following way:

Let the tree has n leaves (species) and let consider the paths from the root of the tree (the common ancestor of all species considered) to every of the leaves (the extant species). Because their lengths are equal, the following system holds:

$$\left\{ \sum_{j=1}^{m_i} \frac{a_{i,j}}{g_{i,j}} = T \right\}_{i=1}^{n}$$

The system has n equations and $2n-2$ unknowns and is non-linear. Let $\overline{g}$ be the mean value of $g_{i,j}$. We may try to minimize the function

$$F(g) = (\frac{1}{(2n-2)} \sum_{i,j} |\overline{g} - g_{i,j}|^2)^{\frac{1}{2}}$$

The following additional restrictions hold: $g_{i,j} > 0$

The straightforward way is to use a general method for non-linear problems.

Another way is to make the substitution $y_{i,j} = \frac{1}{g_{i,j}}$. Then, the equations become linear and we may minimize the function

$$F(y) = (\frac{1}{(2n-2)} \sum_{i,j} |\overline{y} - y_{i,j}|^2)^{\frac{1}{2}}$$

That is a quadratic function over linear restrictions and more reliable methods optimization methods exist for it.

Forthcoming Research

Investigate the variations of gain rates over the tree.
Compare the Maximum Likelihood and the Minimum Entropy approaches.

References

[1] MIRKIN BG, FENNER TI, GALPERIN MY, KOONIN EV: *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 Evolutionary Biology, **3:2,** (2003).

[2] MIRKIN BG, CAMARGO R, FENNER TI, LOIZOU G, KELLAM P: *Aggregating Homologous Protein Families in Evolutionary Reconstructions of Herpesviruses* Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on

[3] KONOVSKI P: *A Common Approach to Finding the Optimal Scenarios of a Markov Stochastic Process Over a Phylogenetic Tree* Computational Intelligence and Bioinformatics and Computational Biology, MC Evolutionary Biology, **26:5,**32963301 (2012).

[4] ROSS SM: *Stochastic processes* New York: John Wiley & Sons, (1996)

Acknowledgements

This work a further extension of previously published results [1], [2] and the author is grateful to Prof. Boris Mirkin and Prof. Trevor Fenner from DCSIS, Birkbeck for their guidance.