BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–3

Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood

Miklós Csűrös

Department of Computer Science and Operations Research, University of Montréal, Montréal, Québec, Canada.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Count is a software package for the analysis of numerical profiles on a phylogeny. It is primarily designed to deal with profiles derived from the phyletic distribution of homologous gene families, but is suited to study any other integer-valued evolutionary characters. Count performs ancestral reconstruction, and infers family- and lineage-specific characteristics along the evolutionary tree. It implements popular methods employed in gene content analysis such as Dollo and Wagner parsimony, propensity for gene loss, as well as probabilistic methods involving a phylogenetic birth-and-death model.

Availability: Count 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/gene_content/count.html` It can also be launched using Java Webstart from the same site. The software is distributed under a BSD-style license. Source code is available upon request from the author.

Contact: `csuros@iro.umontreal.ca`.

1 INTRODUCTION

Some aspects of genome evolution are best captured by integer quantities. Given a phylogeny with terminal taxa $\mathcal{X}$, such a quantity forms a *numerical profile*, which extends the so-called phylogenetic profile of presence-absence (Pellegrini *et al.*, 1999; Koonin and Galperin, 2002) $\Phi\colon \mathcal{X} \mapsto \{0, 1, 2, \dots, \}$. In a typical application, $\Phi[x]$ denotes the number of genes in genome $x \in \mathcal{X}$ for a certain homolog gene family: a homolog family comprises all descendants of the same ancestral gene (Fitch, 2000) in evolutionary lineages. Such families are routinely identified by by pairwise sequence comparisons, coupled with the clustering of postulated homolog pairs (Tatusov *et al.*, 1997; Alexeyenko *et al.*, 2006). In other interesting examples, $\Phi[x]$ might be the size (Caetano-Anollés, 2005) of genome x, or a sequence length polymorphism in population x (Witmer *et al.*, 2003).

Given a phylogeny, an evolutionary character's history can be inferred by various means in order to reconstruct its state at ancestral nodes, or to estimate the tempo of evolution (Pagel, 1999). The Count software package provides a convenient graphical user interface to sophisticated computational methods in such analyses, and to the manipulation of data sets involving numerical profiles. Count was already used to study the evolution of gene repertoire in Archaea (Csűrös and Miklós, 2009), and nucleo-cytoplasmic DNA viruses (Yutin *et al.*, 2009).

2 FEATURES

Count is designed primarily to work with a data set of numerical profiles for homolog gene families. It allows for combining multiple profiles with various annotations, as found in databases of clustered homolog families such as COG (Tatusov *et al.*, 1997). Profiles can be filtered by criteria based on presence, membership count, and annotations, in order to compile winnowed data sets for further analysis.

Given an evolutionary tree T, Count computes the states $\xi[u]$ at tree nodes $u \in T$, based on each profile Φ by imposing $\Phi[u] = \xi[u]$ for all terminal taxa u. In parsimony approaches, the ancestral reconstruction minimizes a criterion based on the implied state changes $\xi[u] \to \xi[v]$ over the edges uv. Alternatively, Count works with so-called phylogenetic birth-and-death models, which consider $(\xi[u]\colon u \in T)$ as a random variable with a well-defined distribution.

Parsimony. Count implements Dollo parsimony (Farris, 1977), and Wagner parsimony (Farris, 1970). In case of the latter, it also implements an asymmetric version (Csűrös, 2008) which penalizes losses and gains differently. Count also computes Propensity for Gene Loss (Krylov *et al.*, 2003), which quantifies the frequency of loss for each family using Dollo parsimony.

Phylogenetic birth-and-death models. The probabilistic model employed in Count relies on linear birth-death-immigration processes (Kendall, 1949), commonly used to model population growth and queuing systems. In the general phylogenetic birth-and-death model, three rates are assigned to each branch: gene loss rate μ, gene duplication rate λ, and a gain rate κ. "Gain" covers multiple phenomena without specifying the origin of the gain, including de novo gene formation and lateral gene transfer. Specifically, character evolution on each edge uv with length τ is stochastically determined by a continuous-time Markov process X with $X(0) = \xi[u]$ and $X(\tau) = \xi[v]$. The process is characterized by the gain rate κ, loss rate μ and duplication rate λ: for $0 < n$, $0 \le t \le \tau$ and any $0 < \delta$,

$$\mathbb{P}\Big\{X(t+\delta) = n \;\Big|\; X(t) = n-1\Big\} = \delta\big(\kappa + (n-1)\lambda + o(1)\big)$$

$$\mathbb{P}\Big\{X(t+\delta) = n-1 \;\Big|\; X(t) = n\Big\} = \delta\big(n\mu + o(1)\big)$$

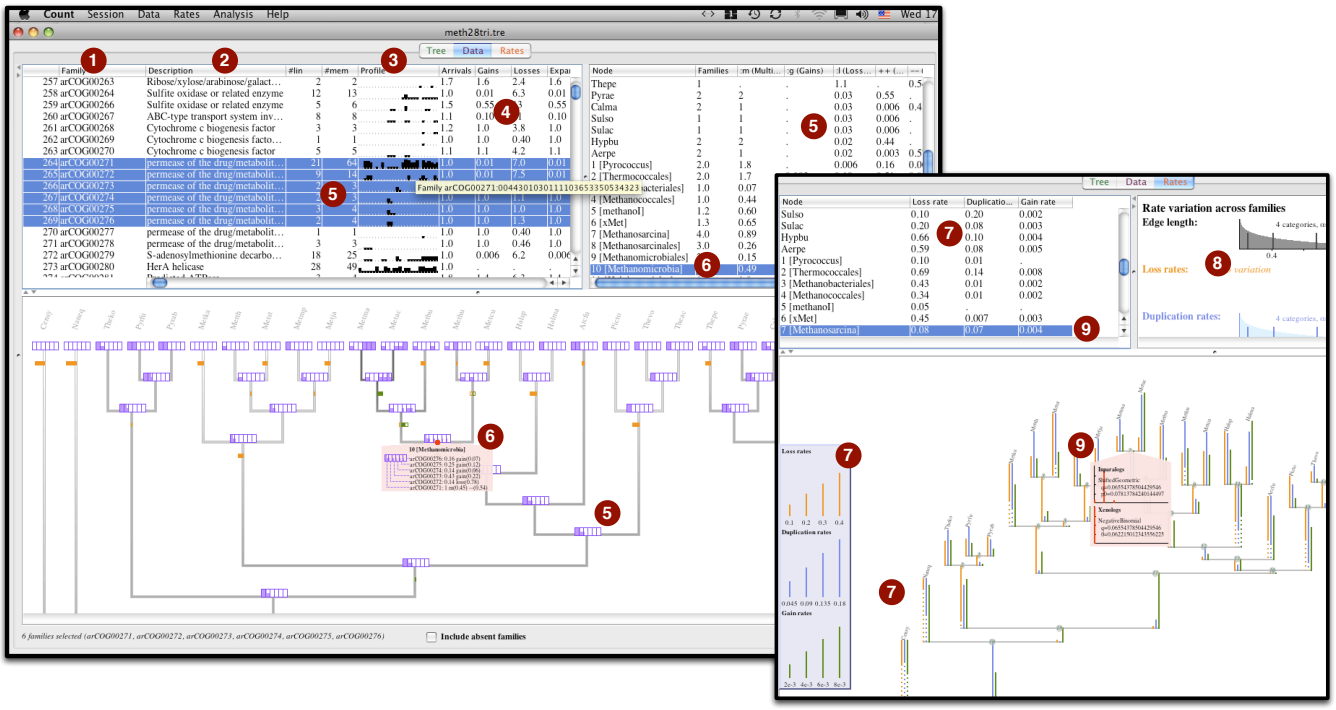

Fig. 1. Some graphical displays in Count. On the **left**, ancestral reconstruction using posterior probabilities. On the **right**, display of a phylogenetic birth-and-death model. Annotated features: (1) Data set of numerical (phylogenetic) profiles. (2) Gene family annotations loaded from separate file. (3) Small profile logo for each family (black bars show $\Phi[u]$ at terminal nodes). (4) Aggregate family-specific information on number of branches where the family was lost, gained, expanded and contracted (estimated as expectations, hence the fractional values). (5) Multiple family selection by cell content, or the mouse. Selection is reflected on the content of the top-right table and the bottom tree. The top-right table shows lineage-specific aggregate information on number of families lost, gained, expanded and contracted on each branch. The bottom tree shows the inferred probabilities for family presence and absence at ancestral nodes (filled rectangles). (6) Lineage selection by the mouse in the table row or the node of the bottom tree. The selection brings up more detailed information at the corresponding node in the bottom tree. (7) Lineage-specific rates displayed in the top-left table, and depicted on the bottom tree, along with a legend. (8) Family-specific rate variation depicted on the top-right. (9) Lineage selection by the mouse in the table row or the node of the bottom tree. The selection brings up more detailed information at the corresponding node in the bottom tree.

Less general models may forbid gain ($\kappa = 0$), or duplication ($\lambda = 0$), or even both. Paralogs evolve independently in this model, capturing the birth-and-death evolution of multi-gene families (Nei and Rooney, 2005), as opposed to concerted evolution, or events involving multiple members at a time. The standard pruning algorithm (Felsenstein, 1973) for computing likelihoods cannot be used with numerical characters, because the ancestral state space is not bounded. Adequate algorithms were proposed for $\kappa = 0, \lambda > 0$ (Arvestad *et al.*, 2004, 2009), and for $\kappa, \lambda > 0$ (Csűrös and Miklós, 2006). Count computes the likelihood using our algorithm described before (Csűrös and Miklós, 2009), which applies to the general model, and all the restricted models. Count allows for rate variation across branches and gene families. Model parameters are set by maximizing the likelihood. The optimized model can be used for ancestral reconstruction and to infer lineage-specific trends by using posterior probabilities conditioned on the profiles.

User interaction. Figure 1 illustrates the rich graphical user interface of Count. The program can work with multiple data sets and models at the same time, in order to help comparisons between different analyses. Entire work sessions can be saved, and individual analysis results can be exported into tab-delimited text files, in order to use with other programs such as spreadsheet tools. Main software components (rate optimization and ancestral reconstruction) can also be launched from the command line without invoking the graphical interface.

Implementation. Count is written entirely in Java (Java SE 6), and was tested on various computer platform, including Microsoft Windows, MacOS X, and Linux. In addition, Count is also available as an integrated application bundle on MacOS X, and a Java Webstart application. The software is distributed with test data and a detailed User's Guide.

ACKNOWLEDGEMENTS

This research project has been supported by a grant from the Natural Sciences and Engineering Research Council of Canada. I am grateful for valuable feedback on the software from Aaron Darling, Dannie Durand, Maureen Stoltzer, Gergely Szöllősi, Natalya Yutin and Yuri Wolf.

REFERENCES

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple genomes. *Bioinformatics*, **22**, e9–e15.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2004). Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In D. Gusfield, editor, *RECOMB '04: Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pages 326–335, New York, NY. ACM.

Arvestad, L., Lagergren, J., and Sennblad, B. (2009). The gene evolution model and computing its associated probabilities. *Journal of the ACM*, **56**(2), 7.

Caetano-Anollés, G. (2005). Evolution of genome size in the grasses. *Crop Science*, **45**, 1809–1816.

Csűrös, M. and Miklós, I. (2009). Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution*, **26**(9), 2087–2095.

Csűrös, M. (2008). Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. *Springer Lecture Notes in Bioinformatics*, **5267**, 72–86. Proc. Sixth RECOMB Comparative Genomics Satellite Workshop.

Csűrös, M. and Miklós, I. (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Springer Lecture Notes in Bioinformatics*, **3909**, 206–220. Proc. Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB).

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Zoology*, **19**(1), 83–92.

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

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, **22**(3), 240–249.

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics*, **16**(5), 227–231.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society Series B*, **11**(2), 230–282.

Koonin, E. V. and Galperin, M. Y. (2002). *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, New York.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, **13**, 2229–2235.

Nei, M. and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, **39**(1), 121–152.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the USA*, **96**(8), 4285–4288.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631–637.

Witmer, P. D., Doheny, K. F., Adams, M. K., Boehm, C. D., Dizon, J. S., Goldstein, J. L., Templeton, T. M., Wheaton, A. M., Dong, P. N., Pugh, E. W., Nussbaum, R. L., Hunter, K., Kelmenson, J. A., Rowe, L. B., , and Brownstein, M. J. (2003). The development of a highly informative mouse simple sequence length polymorphism (SSLP) marker set and construction of a mouse family tree using parsimony analysis. *Genome Research*, **13**, 485–491.

Yutin, N., Wolf, Y. I., Raoult, D., and Koonin, E. V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology Journal*, **6**, 223.