BMC Bioinformatics

SOFTWARE Open Access

DiscML: an R package for estimating evolutionary rates of discrete characters using maximum likelihood

Tane Kim[1,2] and Weilong Hao[1*]

Abstract

Background: The study of discrete characters is crucial for the understanding of evolutionary processes. Even though great advances have been made in the analysis of nucleotide sequences, computer programs for non-DNA discrete characters are often dedicated to specific analyses and lack flexibility. Discrete characters often have different transition rate matrices, variable rates among sites and sometimes contain unobservable states. To obtain the ability to accurately estimate a variety of discrete characters, programs with sophisticated methodologies and flexible settings are desired.

Results: DiscML performs maximum likelihood estimation for evolutionary rates of discrete characters on a provided phylogeny with the options that correct for unobservable data, rate variations, and unknown prior root probabilities from the empirical data. It gives users options to customize the instantaneous transition rate matrices, or to choose pre-determined matrices from models such as birth-and-death (BD), birth-death-and-innovation (BDI), equal rates (ER), symmetric (SYM), general time-reversible (GTR) and all rates different (ARD). Moreover, we show application examples of DiscML on gene family data and on intron presence/absence data.

Conclusion: DiscML was developed as a unified R program for estimating evolutionary rates of discrete characters with no restriction on the number of character states, and with flexibility to use different transition models. DiscML is ideal for the analyses of binary (1s/0s) patterns, multi-gene families, and multistate discrete morphological characteristics.

Keywords: Discrete character states, Gene family evolution, Birth and death, Maximum likelihood, Phylogeny

Background

Many evolutionary processes involve transitions among different discrete characteristic states, including changes in morphological characteristics [1], sequence gain and loss [2,3], gene family expansion and contraction [4], gain and loss of mobile promoters [5] and epigenetic characteristics such as methylation [6]. Evolutionary rates of discrete characters have been estimated using programs primarily developed for constructing ancestral character states such as the ACE function of the APE package [7] in R, standalone programs BayesTraits [8] and Mesquite [9]. Recently, great efforts have been made to estimate gene family turnover rates. The GLOOME program maps gain and loss rates using binary characters (or 1s/0s) [10], while Count [11], BEGFE [12], BadiRate [13], and CAFE3 [14] employ birth-and-death (BD) models to study gene family expansion and contraction.

Some of these programs have advanced (or realistic) features that are not implemented in other programs. For instance, the BayesTraits program implements a Γ-distribution for rate variation [8]. The GLOOME program allows the estimation of prior root probabilities of the character states [10,15]. The BadiRate program allows variable birth rates and death rates, and corrects for unobservable data [13]. Furthermore, many multistate characters do not necessarily evolve in a BD manner [16], and should therefore be modeled using transition rate matrices other than BD.

In order to perform accurate rate estimation on a variety of discrete characters, we have developed a unified

*Correspondence: haow@wayne.edu
[1]Department of Biological Sciences, Wayne State University, 48202 Detroit, USA
Full list of author information is available at the end of the article

program DiscML by implementing the advanced features mentioned above as well as flexible options for transition rate matrices.

Implementation

DiscML estimates the evolutionary rates of discrete characters by fitting the distribution of all character states (the data) on a given phylogeny. The data need to be in a matrix format (vector format for a single site) as required in many other phylogenetic programs in R (see examples in Additional file 1). The provided phylogeny is required to have branch lengths, as branch lengths will be used as a relative time scale in the analysis. The evolutionary rates, transition rate matrices, and additional parameters discussed below will be optimized to maximize the likelihood of the data. The optimization is achieved using the PORT routines [17] implemented in the nlminb function in R.

Implementation of rate variation in the analysis

Rate variation among the character sites has long been recognized and implemented in DNA analyses [18], but has been missing from most analyses of non-DNA discrete characters (but see [8]). DiscML considers rate variation among the character sites by implementing a discrete Γ distribution (with the option of `alpha=TRUE`).

Estimation of prior root probabilities

Most programs for the analysis of discrete characters assume only uniformly distributed prior root probabilities, e.g., $\pi_1 = \pi_2 = .. = \pi_a = \frac{1}{a}$, ($a$ is the total number of character states). DiscML allows the estimation of prior root probabilities (π_a) for different character states (with the option of `rootprobability=TRUE`).

Flexibility on both the transition model and the number of character states

DiscML is flexible on both the size and type of the transition rate matrix (Q), which can be customized by users. This option could open the door for novel evolutionary analyses on different discrete characters. Several transition rate matrices are pre-determined in DiscML: `model="ER"` (equal rates, i.e., all entries in equation 1 are equal), `model="SYM"` (symmetric, i.e., $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$, ..), and `model="ARD"` (all rates different, i.e., all entries are free to vary). `ER` and `SYM` are reversible matrices, while `ARD` matrices are irreversible.

$$Q = \begin{pmatrix} & 0 & 1 & 2 & 3 & .. \\ \hline 0 & - & \alpha_1 & \beta_1 & \delta_1 & .. \\ 1 & \alpha_2 & - & \gamma_1 & \varepsilon_1 & .. \\ 2 & \beta_2 & \gamma_2 & - & \zeta_1 & .. \\ 3 & \delta_2 & \varepsilon_2 & \zeta_2 & - & .. \\ .. & .. & .. & .. & .. & - \end{pmatrix} \quad (1)$$

Evolutionary processes can follow a birth-and-death (BD) process. The birth processes correspond to transitions from state n to state $n+1$, while the death processes correspond to transitions from state n to state $n-1$. The BD transitions can be represented as matrices containing non-zero entries only between the neighboring states (equation 2). Several pre-determined BD transition rate matrices are available: `BDER` (equal rates), `BDSYM` (symmetric, i.e., $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$, ..), `BDISYM` (symmetric, all entries except α are equal, i.e., $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2 = \gamma_1 = \gamma_2 = ..$), and `BDARD` (all rates different).

$$Q = \begin{pmatrix} & 0 & 1 & 2 & 3 & .. \\ \hline 0 & - & \alpha_1 & 0 & 0 & 0 \\ 1 & \alpha_2 & - & \beta_1 & 0 & 0 \\ 2 & 0 & \beta_2 & - & \gamma_1 & 0 \\ 3 & 0 & 0 & \gamma_2 & - & .. \\ .. & 0 & 0 & 0 & .. & - \end{pmatrix} \quad (2)$$

Finally, all transition rate matrices (Qs) are calibrated [19], i.e., each Q satisfies

$$-\sum_a \pi_a Q(a,a) = 1, \quad (3)$$

so that the evolutionary rate parameter (μ) is the average number of transition events per site per evolutionary time unit [20].

Forced reversibility and flexible irreversible options

When the prior root probabilities (π) for different character states are estimated, reversible transition matrices will no longer necessarily result in reversible evolutionary processes (because of potentially different probabilities of character states). Since it is sometimes of biological interest to assume reversibility (i.e., the expected $x \to y$ changes equal to the $y \to x$ changes), DiscML can allow forced reversibility by setting `reversible=TRUE`. In practice, reversibility is obtained by multiplying the corresponding root probabilities (equation 4) to the entries in reversible transition matrices, e.g., `ER` and `SYM`. Such a practice is conceptually the same with the general time-reversible (`GTR`) DNA substitution model [21]. In DiscML, `model="GTR"` is equivalent to the combination of `model="SYM"` and `reversible=TRUE`.

$$Q = \begin{pmatrix} & 0 & 1 & 2 & 3 & .. \\ \hline 0 & - & \alpha\pi_1 & \beta\pi_2 & \delta\pi_3 & .. \\ 1 & \alpha\pi_0 & - & \gamma\pi_2 & \varepsilon\pi_3 & .. \\ 2 & \beta\pi_0 & \gamma\pi_1 & - & \zeta\pi_3 & .. \\ 3 & \delta\pi_0 & \varepsilon\pi_1 & \zeta\pi_2 & - & .. \\ .. & .. & .. & .. & .. & - \end{pmatrix} \quad (4)$$

Similarly, when the prior root probabilities for different character states are estimated, forced reversibility can be applied to the BD related matrices (equation 5).

$$Q = \left(\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & .. \\ \hline 0 & - & \alpha\pi_1 & 0 & 0 & 0 \\ 1 & \alpha\pi_0 & - & \beta\pi_2 & 0 & 0 \\ 2 & 0 & \beta\pi_1 & - & \gamma\pi_3 & 0 \\ 3 & 0 & 0 & \gamma\pi_2 & - & .. \\ .. & 0 & 0 & 0 & .. & - \end{array}\right) \quad (5)$$

In DiscML, the default setting is `reversible=FALSE` and users have the flexibility to conduct analysis by assuming irreversible evolutionary processes. Unlike in reversible processes, the root position can greatly affect the maximum likelihood calculation in irreversible cases [22,23]. Therefore, it is only meaningful to perform irreversible analysis on a rooted tree. If the provided phylogenetic tree is unrooted, DiscML will first reroot the tree by midpoint rooting, and perform analysis on the midpoint rooted tree.

Correction for unobservable data

Some characters may contain unobservable character states, which can only be inferred indirectly from the presence of observable states of the same characters in related taxa. Ancient characters can be lost from all examined extant taxa, and result in unobservable data. DiscML provides the option of `zerocorrection=TRUE` to calculate the likelihood conditional on a pattern being observable following [24], i.e.,

$$L_+ = \frac{L}{1 - L_-}, \quad (6)$$

where L_- is the likelihood of unobservable patterns. The correction for unobservable data (shown as '+0' in Table 1) is essential for systems such as gene family data due to the complete loss of some ancient genes, but not suitable for single-site analyses and for systems in which all character states are observable (e.g., nucleotide bases).

Site and branch specific estimations

Even though the default setting of DiscML is to perform rate estimation by fitting the distribution pattern of all character sites on a phylogeny, there is an option to perform rate estimation on individual sites (`ind=TRUE`). Individual rates can be graphically displayed using `plotmu=TRUE`. Furthermore, DiscML allows branch specific rate estimation, which can be specified using '$' on branches in the provided tree file. For instance, `(((taxon1$1: 0.01, taxon2$1: 0.01)$3: 0.01, taxon3$2: 0.02)$3: 0.01, taxon4$2: 0.03);` specifies three rates, one for the branches leading to taxon1 and taxon2 (`$1`), one for the branches leading to taxon3 and taxon4 (`$2`), and one for the remaining branches (`$3`). The modified tree files are no longer in the conventional Newick format, we have developed a function `read.tree2` in DiscML to read such modified tree files.

Table 1 DiscML estimates from the gene family data in the Bacillaceae (B1, B2, B3) clades

Models	Parameters	B1	B2	B3
ER	μ	3.073	0.677	0.540
(1s/0s only)	LnL	-15150	-16467	-22229
ER+0	μ	1.887	0.463	0.388
(1s/0s only)	LnL	-13682	-15268	-21207
BDER	μ	2.490	0.590	0.485
	LnL	-20901	-22196	-29127
BDISYM	μ	2.669	0.556	0.438
	LnL	-19684	-20973	-27811
BDARD	μ	5.746	1.369	1.450
	LnL	-18254	-20073	-26578
ER	μ	2.940	0.638	0.459
	LnL	-21411	-23273	-31405
SYM	μ	2.635	0.546	0.427
	LnL	-19615	-20947	-27801
ARD	μ	5.601	1.345	1.314
	LnL	-18143	-19678	-26239
GTR	μ	3.731	0.739	0.632
(SYM+π^{REV})	LnL	-17753	-19337	-25381
ER+0	μ	2.339	0.531	0.395
	LnL	-20595	-22586	-30753
ER+π	μ	2.935	0.624	0.454
	LnL	-20070	-21783	-28771
ER+Γ	μ	3.205	0.638	0.459
	LnL	-21398	-23273	-31405
ER+0+π+Γ	μ	1.358	0.236	0.240
	LnL	-18719	-19960	-26712
ER+0+π^{REV}+Γ	μ	3.630	0.379	0.387
	LnL	-16839	-17960	-23398

The parameter μ is the estimated evolutionary rate of the characters. "1s/0s only" indicates binary analysis by converting all non-zero characters to 1s using `simplify=TRUE`, '+0' indicates the correction for unobservable data using `zerocorrection=TRUE`, '+Γ' indicates the implementation of a discrete Γ distribution using `alpha=TRUE`, '+π' indicates the estimation of prior root probabilities using `rootprobability=TRUE`, '+π^{REV}' indicates the estimation of prior root probabilities with forced reversibility using `rootprobability=TRUE` and `reversible=TRUE`.

Additional features

DiscML allows binary (1s/0s) analysis on data with more than two character states by converting all non-zero characters to 1s with `simplify=TRUE`.

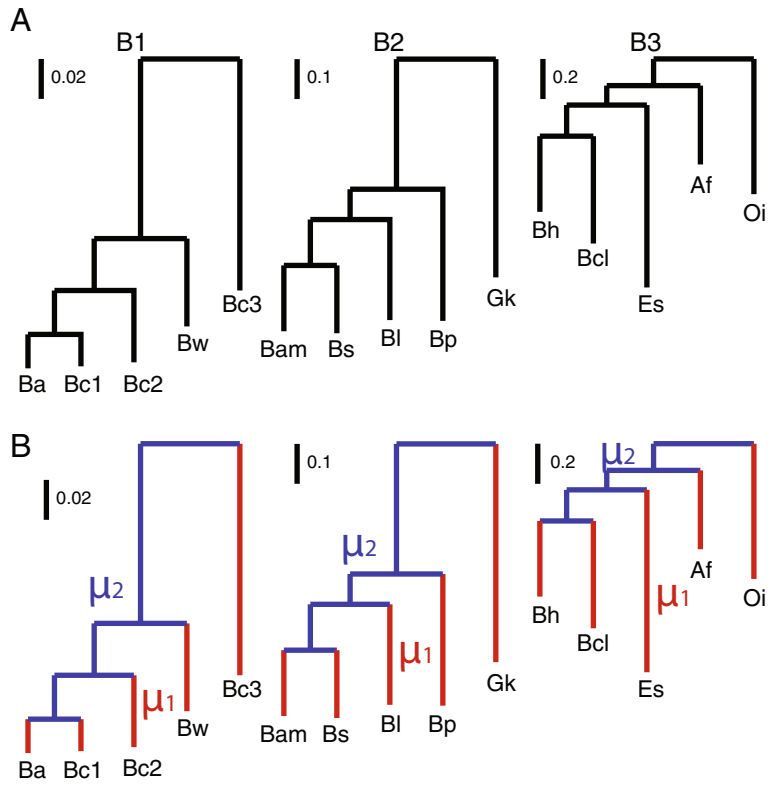

Figure 1 Phylogenetic relationship of three Bacillaceae (B1, B2, B3) clades, on which the evolutionary rates of gene families are estimated using DiscML. A, a constant rate is estimated on each phylogeny; **B**, separate rates are estimated for external branches (μ_1) versus internal branches (μ_2) on each phylogeny. These three clades were studied in our previous study on gene presence, absence, and fragments [20]. Gene families are recategorized, with gene absence and fragments as character state 0, single-copy genes as 1, and gene families with two or more members as 2.

Table 2 Computational time on an Intel Core i7 (3.4 Ghz) 16 GB RAM Dell desktop to generate the results in Table 1

Models	B1(5453)	B2(5614)	B3(6813)
ER (1s/0s only)	0 m 49 s	1 m 00 s	1 m 26 s
ER+0 (1s/0s only)	1 m 39 s	2 m 01 s	3 m 03 s
BDER	0 m 48 s	1 m 06 s	1 m 36 s
BDISYM	1 m 58 s	2 m 20 s	3 m 01 s
BDARD	7 m 54 s	6 m 58 s	8 m 28 s
ER	1 m 04 s	1 m 15 s	1 m 17 s
SYM	3 m 14 s	4 m 47 s	5 m 31 s
ARD	9 m 53 s	9 m 12 s	16 m 59 s
GTR(SYM+π^{REV})	9 m 04 s	9 m 54 s	11 m 44 s
ER+0	1 m 36 s	2 m 34 s	2 m 21 s
ER+π	2 m 41 s	3 m 13 s	4 m 40 s
ER+Γ	12 m 00 s	39 m 01 s	45 m 23 s
ER+0+π+Γ	82 m 22 s	81 m 20 s	178 m 27 s
ER+0+π^{REV}+Γ	80 m 13 s	67 m 33 s	91 m 42 s

The number of gene families is shown in parentheses for each clade. The time is shown in minutes (m) and seconds (s).

Results and discussion

DiscML was first tested using the gene family data on three Bacillaceae clades (Figure 1A, Additional file 1 and [20]). In the previous study [20], we distinguished gene fragments from gene absence and gene presence. In this study, we eliminated the character state specific for gene fragments and re-categorized gene fragments as gene absence or character state 0, single-copy genes as character state 1, and gene families with two or more members

Table 3 Separate rates on branches estimated from the gene family data in the Bacillaceae (B1, B2, B3) clades

Models	Parameters	B1	B2	B3
($\mu_1 = \mu_2$)ER	μ	2.940	0.638	0.459
	Ln*L*	-21411	-23273	-31405
($\mu_1 \neq \mu_2$)ER	μ_1	4.430	0.674	0.477
	μ_2	0.306	0.526	0.344
	Ln*L*	-21045	-23267	-31395
	2ΔLn*L*	732***	14***	20***

μ_1 is for external branches, while μ_2 is for internal branches on each tree as illustrated in Figure 1B.
***$P < 0.001$ (df=1), as 2ΔLn*L* approximately follows a χ^2-distribution.

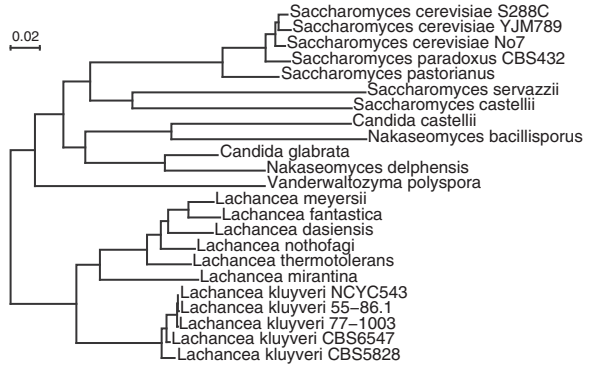

Figure 2 Phylogenetic relationship of the yeast strains in the *Saccharomyces* complex, on which the rates of mitochondrial intron gain and loss are estimated using DiscML. The phylogeny was reconstructed using the concatenated sequences of all mitochondrial protein genes after excluding the *var1* gene.

as 2 (Additional file 1), so that the application of BD models on these data is meaningful. It is worth to note that, though the number of character states is restricted to three here, DiscML is flexible and capable of analyzing a large number of character states.

The performance of DiscML is found to be reliable. For instance, the `ER+0` model with the option of `simplify=TRUE` in Table 1 is mathematically identical to the M_{00} model in [20]. The optimization in [20] was achieved using the Nelder-Mead simplex method [25], while the optimization in Table 1 was achieved using the PORT routines [17]. Importantly, the DiscML estimates are identical to the previous estimates for all three clades. As expected, the parameter-rich models consistently outperformed the nested simplistic models (e.g., LnL `BDARD` > LnL `BDISYM` > LnL `BDER`; LnL `ARD` > LnL `SYM` > LnL `ER`). Consistent with previous studies [3,20,26], rate estimates in closely related clades tend to be higher than those in distantly related clades due to the transient nature of many acquired genes (Table 1). Tested on an Intel Core i7 (3.4 Ghz) 16 GB RAM Dell desktop, the computation using DiscML is fast (Table 2). For instance, the `ER` (1s/0s only) analysis took 49 seconds (0 m 49 s) for B1 (5453 gene families), 60 seconds (1 m 00 s) for B2 (5614 gene families), and 86 seconds (1 m 26 s) for B3 (6813 gene families). Computational time increases with the complexity of transition rate matrices and the addition of estimated parameters. For instance, the `ER+0+`π`+`Γ analysis took 82 m 22 s for B1, 81 m 20 s for B2, and 178 m27 s for B3 (Table 2).

DiscML was developed to allow separate rates among branches since evolutionary rates can vary among lineages [27-29]. In the three Bacillaceae clades, we assigned separate rates between external branches (μ_1) and internal branches (μ_2) as illustrated in Figure 1B. Our results in Table 3 support the previous findings of higher gene turnover rates on external branches than those on internal branches [26,30].

It is often of interest for users to know the individual rate of each character site. Previously, we have shown that the mitochondrial intron in the 21S rRNA gene undergoes very rapid turnover in yeast [31]. In this study, we estimated the individual rates of all 17 mitochondrial introns on the yeast phylogeny (Figure 2 and Additional file 1) based on the intron distribution pattern (Additional file 1). On the plot generated by DiscML using `ind=TRUE` (Figure 3), users can visually compare the individual rates of different introns. For instance, the introns at sites 7 and 8 have faster turnover rates than the 21S rRNA intron at site 17 (Figure 3). The R commands used in the study are provided in Additional file 1.

Conclusion

We illustrated the versatility of DiscML on different types of data and analyses. With a great flexibility and

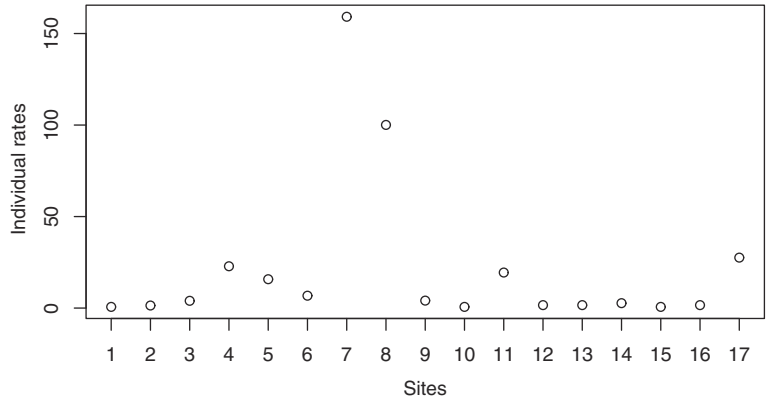

Figure 3 Plot of individual turnover rates of the 17 mitochondrial introns in yeast. Ten introns in the *cox1* gene are shown as sites 1-10, six introns in the *cob* gene are shown as sites 11-16, and one intron in the 21S rRNA gene is shown as site 17.

fast computational speed, we are confident that DiscML can be used in a variety of studies on different discrete characters.

Availability and requirements

Project name: DiscML
Project home page: http://cran.r-project.org/web/packages/DiscML/index.html
Operating system(s): Platform independent.
Programming language: R.
Other requirements: R (2.14 or newer); R-package: ape from CRAN.
License: GNU GPL

Additional file

Additional file 1: Files and commands used in the analyses of the B1 clade and the yeast clade. B1.tre is the B1 tree in the conventional Newick format. B1_pattern contains the distribution pattern of gene families in the B1 clade with gene absence and fragments as 0s, single-copy genes as 1s, and gene families with two or more members as 2s. Each column is for one genome, and each row is for one gene family. B1_2rates.tre is the B1 tree with assigned separate rates for external branches and internal branches. The rate for external branches is $1, and the rate for internal branches is $2. The yeast.tre file is the phylogenetic tree of 13 yeast strains in the conventional Newick format. The intron_pattern file contains the distribution pattern of the 17 mitochondrial introns in the 13 yeast strains. Each column is for one intron, and each row is for one strain. Data matrix in this format will need to be transformed before the analysis (see R.inputs for details). Some R commands are in R.inputs.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WH designed the package, TK wrote the codes, WH performed data analysis, TK and WH wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank two anonymous reviewers for their helpful comments, Dr. Edward Golenberg for critical reading of the manuscript, Dr. Brian Golding for suggestions during the development of DiscML, Dr. Baojun Wu for assistance in collecting yeast mitochondrial data used in Figures 2 and 3. The work was supported by funds from Wayne State University to WH.

Author details

[1]Department of Biological Sciences, Wayne State University, 48202 Detroit, USA. [2]Mathematics Undergraduate Program, Wayne State University, 48202 Detroit, USA.

Received: 26 July 2014 Accepted: 25 September 2014
Published: 27 September 2014

References

1. Lewis PO: **A likelihood approach to estimating phylogeny from discrete morphological character data.** *Syst Biol* 2001, **50:**913–915.
2. Csűrös M: **Likely scenarios of intron evolution.** In *Comparative Genomics.* Lecture Notes in Computer Science. Edited by McLysaght A, Huson DH. Berlin: Springer; 2005:47–60.
3. Hao W, Golding GB: **The fate of laterally transferred genes: life in the fast lane to adaptation or death.** *Genome Res* 2006, **16:**636–643.
4. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N: **Estimating the tempo and mode of gene family evolution from comparative genomic data.** *Genome Res* 2005, **15:**1153–1160.
5. van Passel MW, Nijveen H, Wahl LM: **Birth, death, and diversification of mobile promoters in prokaryotes.** *Genetics* 2014, **197:**291–299.
6. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR: **Patterns of population epigenomic diversity.** *Nature* 2013, **495:**193–198.
7. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics* 2004, **20:**289–290.
8. Pagel M, Meade A, Barker D: **Bayesian estimation of ancestral character states on phylogenies.** *Syst Biol* 2004, **53:**673–674.
9. Maddison WP, Maddison DR: **Mesquite: a modular system for evolutionary analysis.** 2011. Version 2.75, [http://mesquiteproject.org]
10. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T: **GLOOME: gain loss mapping engine.** *Bioinformatics* 2010, **26:**2914–2915.
11. Csűrös M: **Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood.** *Bioinformatics* 2010, **26:**1910–1912.
12. Liu L, Yu L, Kalavacharla V, Liu Z: **A Bayesian model for gene family evolution.** *BMC Bioinformatics* 2011, **12:**426.
13. Librado P, Vieira FG, Rozas J: **BadiRate: estimating family turnover rates by likelihood-based methods.** *Bioinformatics* 2012, **28:**279–281.
14. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3.** *Mol Biol Evol* 2013, **30:**1987–1987.
15. Cohen O, Rubinstein ND, Stern A, Gophna U, Pupko T: **A likelihood framework to analyse phyletic patterns.** *Philos Trans R Soc Lond B Biol Sci* 2008, **363:**3903–3911.
16. Hibbett DS: **Trends in morphological evolution in homobasidiomycetes inferred using maximum likelihood: a comparison of binary and multistate approaches.** *Syst Biol* 2004, **53:**889–903.
17. Gay DM: *Usage Summary for Selected Optimization Routines.* Computing science technical report 153. Murray Hill: AT&T Bell Laboratories; 1990.
18. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39:**306–314.
19. Yap VB, Speed T: **Rooting a phylogenetic tree with nonreversible substitution models.** *BMC Evol Biol* 2005, **5:**2.
20. Hao W, Golding GB: **Inferring bacterial genome flux while considering truncated genes.** *Genetics* 2010, **186:**411–426.
21. Waddell PJ, Steel MA: **General time-reversible distances with unequal rates across sites mixing gamma and inverse Gaussian distributions with invariant sites.** *Mol Phylogenet Evol* 1997, **8:**398–414.
22. Felsenstein J: *Inferring Phylogenies.* Sunderland: Sinauer Associates, Inc.; 2004.
23. Boussau B, Gouy M: **Efficient likelihood computations with nonreversible models of evolution.** *Syst Biol* 2006, **55:**756–758.
24. Felsenstein J: **Phylogenies from restriction sites: a maximum-likelihood approach.** *Evolution* 1992, **46:**159–173.
25. Nelder JA, Mead R: **A simplex method for function minimization.** *Comput J* 1965, **7:**308–313.
26. Hao W, Golding GB: **Uncovering rate variation of lateral gene transfer during bacterial genome evolution.** *BMC Genomics* 2008, **9:**235.
27. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution.** *Mol Biol Evol* 2002, **19:**1–7.
28. Wang HC, Spencer M, Susko E, Roger AJ: **Testing for covarion-like evolution in protein sequences.** *Mol Biol Evol* 2007, **24:**294–305.
29. Spencer M, Sangaralingam A: **A phylogenetic mixture model for gene family loss in parasitic bacteria.** *Mol Biol Evol* 2009, **26:**1901–1908.
30. Hao W, Golding GB: **Patterns of bacterial gene movement.** *Mol Biol Evol* 2004, **21:**1294–1307.
31. Wu B, Hao W: **Horizontal transfer and gene conversion as an important driving force in shaping the landscape of mitochondrial introns.** *G3 (Bethesda)* 2014, **4:**605–612.

doi:10.1186/1471-2105-15-320
Cite this article as: Kim and Hao: DiscML: an R package for estimating evolutionary rates of discrete characters using maximum likelihood. *BMC Bioinformatics* 2014 **15**:320.