ModelFinder: fast model selection for accurate phylogenetic estimates

Subha Kalyaanamoorthy[1,2,6], Bui Quang Minh[3,6], Thomas K F Wong[1,4,6], Arndt von Haeseler[3,5] & Lars S Jermiin[1,4]

Model-based molecular phylogenetics plays an important role in comparisons of genomic data, and model selection is a key step in all such analyses. We present ModelFinder, a fast model-selection method that greatly improves the accuracy of phylogenetic estimates by incorporating a model of rate heterogeneity across sites not previously considered in this context and by allowing concurrent searches of model space and tree space.

Model-based molecular phylogenetic analysis plays a key role in comparative genomics and evolutionary biology. It allows us to annotate genomes more accurately[1]; test our understanding of the evolution of species, genomes and genes[2–6]; and determine the likely origins and dispersal routes of pathogens and agricultural pests[7,8]. Selecting an optimal model of sequence evolution (SE) is a critical step in all such analyses. Here we introduce ModelFinder, a model-selection method that combines substitution models used in other popular model-selection methods[9,10] with a flexible rate heterogeneity across sites (RHAS) model and show that its use often leads to substantial improvements in the fit between tree, model and data (**Supplementary Software** and http://www.iqtree.org/ModelFinder/).

Model selection is used to identify the best-fitting model of SE that led to the available data. Several methods for identifying optimal models of SE are available for DNA[9] and protein[10], and model selection is even possible when different models are required to analyze different sets of sites in an alignment[11]. Finding an optimal model of SE for a given sequence alignment entails finding the best-fitting substitution model and the best-fitting RHAS model. Usually, this means comparing three models of RHAS that assume either (i) all sites evolved at the same rate, (ii) some sites evolved at the same rate whilst the others were invariable (I), or (iii) RHAS follows a probability distribution like the popular discrete Γ distribution[12].

The discrete Γ distribution is parameterized using k rate categories, each comprising a rate (r_i) and a weight (w_i), where $r_i > 0$, $w_i = 1/k$, and $1 = \sum_{i=1}^{k} r_i w_i$. This parameterization imposes two constraints on the model—it is assumed that RHAS can be modeled accurately by a Γ distribution, and that the probability that a site belongs to rate category i equals $1/k$. These assumptions may be unrealistic and may bias phylogenetic estimates.

One solution to the problem of using these potentially unrealistic assumptions is to infer the weights from the data, as proposed by Yang[13]. The advantage offered by this probability distribution free (PDF) model of RHAS is that the distribution of rates of change across sites may take any shape, implying that its estimates of rates and weights under many circumstances should be more accurate than those obtained under a Γ distribution. Until now, however, the PDF model was unavailable in the context of model selection.

In order to make this model available, we developed ModelFinder, a model-selection method for alignments of nucleotides, codons, amino acids or other discrete data. ModelFinder is implemented in IQ-TREE[14] and offers many features, including the choice of comparing models of SE inferred on the same tree (default) or on different trees (advanced). When the advanced option is used, ModelFinder searches tree space for every model of SE considered and, therefore, may find superior models of SE. ModelFinder incorporates 22 and 36 substitution models for DNA and protein, respectively, and 13 models of RHAS, including the PDF model with $k = 2,\ldots,k_{\max}$ rate categories. By default, $k_{\max} = 10$, but this value can be increased if needed. Each PDF model, henceforth labeled R_k, is a family of RHAS models. The user can specify the numbers and types of models to compare. In summary, ModelFinder considers models of RHAS that are more complex than those considered by other model-selection methods[9–11].

The PDF model is more parameter rich than the discrete Γ model, so parameter estimation is a challenge for the PDF model. ModelFinder uses the expectation maximization (EM) algorithm[15] to estimate the parameters for every R_k model and another algorithm to identify the optimal value of k for the PDF model (see Online Methods). The accuracy of ModelFinder was assessed by analyzing 100 amino acid alignments generated on a 100-tipped tree (**Fig. 1a**). Alignments with 10,000 sites were generated using INDELible[16] and the LG[17] + R_5 model of SE and a bimodal distribution of RHAS. ModelFinder estimated model parameters accurately when the data were analyzed using the correct tree and model of SE (**Fig. 1b**), and ModelFinder was accurate regardless of the optimality criterion (AIC, AICc or BIC)

[1] Land & Water, CSIRO, Canberra, Australian Capital Territory, Australia. [2] Faculty of Pharmacy & Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta, Canada. [3] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna & Medical University of Vienna, Vienna, Austria. [4] Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia. [5] Bioinformatics and Computational Biology, Faculty of Computational Science, University of Vienna, Vienna, Austria. [6] These authors contributed equally to this work. Correspondence should be addressed to L.S.J. (lars.jermiin@anu.edu.au).

RECEIVED 19 SEPTEMBER 2016; ACCEPTED 5 APRIL 2017; PUBLISHED ONLINE 8 MAY 2017; DOI:10.1038/NMETH.4285

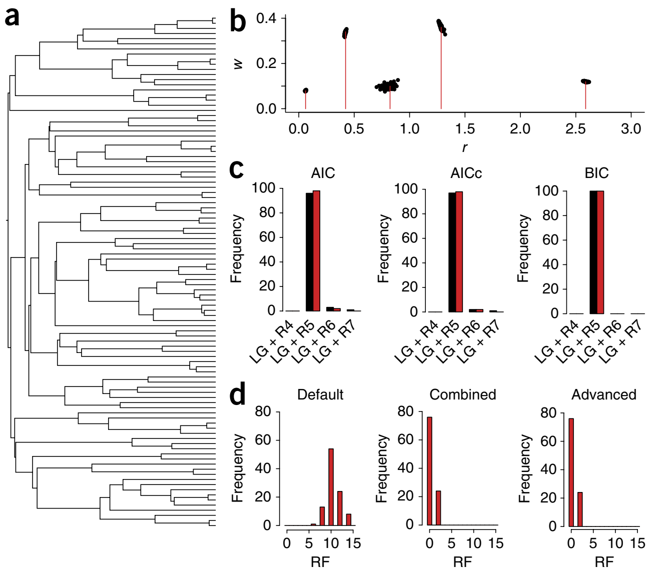

Figure 1 | ModelFinder obtains accurate phylogenetic estimates. (**a**) A rooted 100-tipped tree, with a root-to-tip distance of 0.5 substitutions per site, was used to generate the simulated data. (**b**) True values of r_i and w_i (red lines; r_i = (0.06, 0.42, 0.82, 1.28, 2.58) and w_i = (0.08, 0.34, 0.10, 0.36, 0.12)) and estimated (r_i ,w_i) values for the 100 simulated data sets (black dots). (**c**) Frequency of models of SE identified under different criteria (AIC, AICc and BIC) using the default (black) and advanced (red) search options. (**d**) Distribution of Robinson–Foulds (RF) distances between the true tree and (i) the tree found using default model search (Default), (ii) the tree found using the optimal model of SE obtained with default model search (Combined), and (iii) the tree found using the advanced model search (Advanced). The BIC optimality criterion was used in this example.

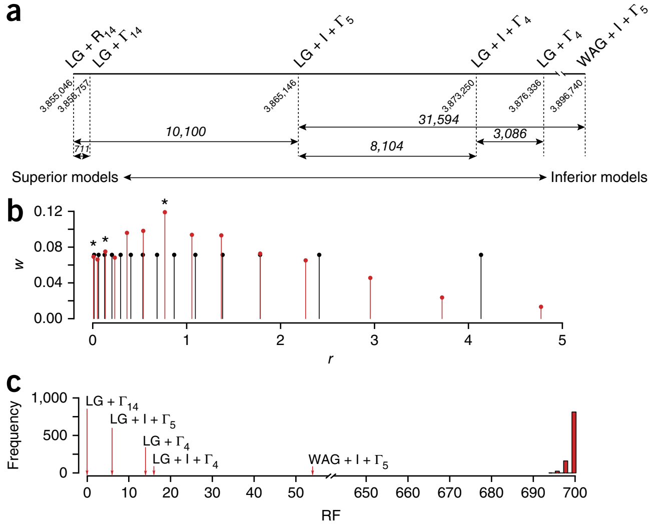

Figure 2 | Advantages provided by ModelFinder. (**a**) BIC scores of selected models of SE, given the alignment of bacterial and archaeal amino acids used by Wu *et al.*[19]. Models are listed above the thick horizontal line. Numbers along the line are BIC scores, and those in italics denote ΔBIC. (**b**) r_i and w_i values obtained under the R_{14} model of RHAS (red lines and balls) and the Γ_{14} model of RHAS (black lines and balls) for the alignment analyzed by Wu *et al.*[19]. Stars indicate local peaks in the R_{14} model. (**c**) RF distances between the most likely tree inferred under various models of SE. For comparison, a histogram with the distribution of 1,000 RF distances is included; each of these distances was obtained by comparing the most likely tree inferred under the LG + R_{14} model of SE to a randomly generated tree with the same number of leaves.

and search option (default or advanced) (**Fig. 1c**). When AIC or AICc were used, a 2–3% bias toward more parameter-rich RHAS models was found. The high success rate of BIC is noteworthy, because the optimal model of SE was inferred even when the best tree found differed from the true tree. We calculated the distribution of Robinson–Foulds (RF) distances[18] between the true tree and (i) the parsimony tree (found using the default search option), (ii) the tree inferred using the best model of SE found using the default search option, and (iii) the tree found using the advanced search option (**Fig. 1d**). The RF distances ranged from 0 to 14, implying that the trees were identical in the best cases, and that 7 of the 97 internal edges differed between the trees in the worst cases. ModelFinder is thus accurate and can identify models of SE that other model-selection methods are unable to detect.

We applied ModelFinder to an amino acid alignment that formed the basis for a genomic encyclopedia of bacteria and archaea[19]. The data were originally analyzed using the WAG + I + Γ_5 model. Using ModelFinder, both search options produced the same optimal model of SE (LG + R_{14}), but the model inferred using the advanced search option was better parameterized (BIC = 3,855,048) than that inferred using the default search option (BIC = 3,858,039). When ΔBIC > 10, there is strong evidence against the model with the higher BIC score[20]. The large difference in BIC scores (ΔBIC = 2,991) concurs with a large difference between the corresponding trees (RF = 138) and implies that the default search option relied on a suboptimal tree. Although the use of a suboptimal tree did not lead to the selection of a suboptimal model of SE in this case, it is a risk to consider when using the default search option.

We next performed a phylogenetic analysis to compare estimates for selected models, and we confirmed that the LG + R_{14} model is the best (in that its BIC score is smaller than those of the other models; **Fig. 2a**). Its superior fit is due to factors including changes in substitution model (WAG + I + $\Gamma_5 \rightarrow$ LG + I + Γ_5: ΔBIC = 31,594) and the RHAS model (LG + I + $\Gamma_5 \rightarrow$ LG + R_{14}: ΔBIC = 10,100). Other models that were considered revealed the effects of the I model of RHAS (LG + $\Gamma_4 \rightarrow$ LG + I + Γ_4: ΔBIC = 3,086) and the number of rate categories used to model the Γ distribution (LG + I + $\Gamma_4 \rightarrow$ LG + I + Γ_5: ΔBIC = 8,104). Given the latter result, we wondered whether the LG + Γ_{14} model might fit the data better than did the LG + R_{14} model, but this was not the case (ΔBIC = 711). Unlike the Γ_{14} model, the R_{14} model is trimodal and has a larger maximum/minimum rate ratio (r_{max}/r_{min} = 575 for R_{14} and 274 for Γ_{14}) (**Fig. 2b**).

Finally, we wanted to see whether the optimal tree for these data was model dependent. The RF distances between the most likely tree inferred under the LG + R_{14} model and those inferred under the other models ranged from 0 to 54, so the optimal tree for these data is clearly model dependent (**Fig. 2c**). Interestingly, although the trees inferred under the other models differ from those inferred under LG + R_{14}, they are still significantly more like the LG + R_{14} tree ($P < 0.001$) than random trees, so the other models are not too misleading. Nonetheless, the best explanation for these data is provided by the tree inferred under the LG + R_{14} model.

Table 1 | Model selection for five diverse data sets.

Data type, source and origin	Sequences	Sites	ModelFinder	BIC	Other methods	BIC	ΔBIC	RF
DNA, Lassa virus[7]	179	3,186	SYM + R_5	131,325	SYM + I + Γ_4	131,540	215	16
DNA, mitochondrial, mammals[3]	274	7,370	GTR + R_8	681,837	GTR + I + Γ_4	684,469	2,632	16
DNA, nuclear, birds[4]	200	394,684	GTR + R_8	18,891,706	GTR + I + Γ_4	18,969,054	77,348	4
Protein, plastids, green plants[5]	360	19,449	JTT + F + R_{10}	2,830,471	JTT + F + I + Γ_4	2,838,957	8,486	4
Protein, nuclear, yeast[6]	23	634,530	LG + F + R_7	25,629,204	LG + F + I + Γ_4	25,638,043	8,839	0

The numbers of sequences and sites in the alignment, optimal models of SE identified using ModelFinder and IQ-TREE's implementations of jModelTest[9] and ProtTest[10] (see Online Methods), and the differences in terms of the ΔBIC score and RF distance between phylogenetic estimates inferred using these optimal models of SE are given.

Using ModelFinder on other phylogenetic data gave similar results (**Table 1**). The best model of SE involved the PDF model of RHAS in every case, and the best tree inferred using this model often differed from that found using the best model identified using other model-selection methods. Clearly, using ModelFinder can lead to a substantial improvement (i.e., ΔBIC > 10) in the fit between tree, model and data irrespective of the source and type of data. A survey of 130 other data sets from TreeBASE[21] reinforces this conclusion (**Supplementary Table 1**)—in 122 of the cases, the fit between tree, model and data improved (in 111 cases, ΔBIC > 10); and in 118 of the cases, the tree topology changed. In addition, a better fit between tree, model and data was found using the advanced instead of the default search option in 75 of the 130 cases. In 46 of these 75 cases, the models of SE differed, and in every one of these 46 cases the optimal trees differed; hence, the advanced search option provides a substantial advantage over the default search option.

ModelFinder is fast and more flexible than other model-selection methods[9–11] and can detect models of SE that the other methods cannot (e.g., multimodal distributions of RHAS). Based on surveys of simulated and real data, ModelFinder proved accurate and often outperformed other model-selection methods in terms of the fit between tree, model and data. Fears of overparameterization have traditionally led users of model-based phylogenetic methods to avoid parameter-rich models of SE; but the use of the BIC, AIC and AICc criteria should alleviate this concern. Although the accuracy and benefits of ModelFinder were demonstrated using proteins generated under time-reversible conditions, the method is also suitable for other data that have evolved under such conditions. If, however, the data have evolved under conditions that are not reversible over time, then ModelFinder is not suitable for model selection. When data have evolved under such conditions, model selection is a challenge, because different edges in the tree may require different models of SE. In practical terms, the HAL–HAS model[22] addresses this need for nucleotides, but a similar solution for other data is not yet available.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D.Y. Wu, J.A. Eisen, P. Donoghue and A. Rokas for access to their data; E. Susko for discussions about the EM algorithm; and V. Jayaswal for constructive feedback. B.Q.M. and A.v.H. were supported by the Austrian Science Fund (FWF I-2805-B29).

AUTHOR CONTRIBUTIONS

S.K., T.K.F.W. and L.S.J. conceived the method and executed a pilot study to assess the method's likely impact on model selection. B.Q.M. and T.K.F.W. implemented the method in IQ-TREE with contributions from S.K., L.S.J. and A.v.H. S.K., T.K.F.W., L.S.J. and B.Q.M. assessed the performance and accuracy of the method. S.K., T.K.F.W. and L.S.J. carried out the analyses of simulated and real data. L.S.J., S.K., T.K.F.W., B.Q.M. and A.v.H. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Eisen, J.A. *Genome Res.* **8**, 163–167 (1998).
2. Hardy, M.P., Owczarek, C.M., Jermiin, L.S., Ejdebäck, M. & Hertzog, P.J. *Genomics* **84**, 331–345 (2004).
3. dos Reis, M. *et al. Proc. R. Soc. B* **279**, 3491–3500 (2012).
4. Prum, R.O. *et al. Nature* **526**, 569–573 (2015).
5. Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E. & Burleigh, J.G. *BMC Evol. Biol.* **14**, 23 (2014).
6. Salichos, L. & Rokas, A. *Nature* **497**, 327–331 (2013).
7. Andersen, K.G. *et al. Cell* **162**, 738–750 (2015).
8. Tay, W.T. *et al. Sci. Rep.* **7**, 45302 (2017).
9. Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. *Nat. Methods* **9**, 772 (2012).
10. Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. *Bioinformatics* **27**, 1164–1165 (2011).
11. Lanfear, R., Calcott, B., Ho, S.Y.W. & Guindon, S. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
12. Yang, Z. *J. Mol. Evol.* **39**, 306–314 (1994).
13. Yang, Z. *Genetics* **139**, 993–1005 (1995).
14. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. *Mol. Biol. Evol.* **32**, 268–274 (2015).
15. Dempster, A.P., Laird, N.M. & Rubin, D.B. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38 (1977).
16. Fletcher, W. & Yang, Z. *Mol. Biol. Evol.* **26**, 1879–1888 (2009).
17. Le, S.Q. & Gascuel, O. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
18. Robinson, D.F. & Foulds, L.R. *Math. Biosci.* **53**, 131–147 (1981).
19. Wu, D. *et al. Nature* **462**, 1056–1060 (2009).
20. Kass, R.E. & Raftery, A.E. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
21. Sanderson, M.J., Donoghue, M.J., Piel, W. & Eriksson, T. *Am. J. Bot.* **81**, 183 (1994).
22. Jayaswal, V., Wong, T.K.F., Robinson, J., Poladian, L. & Jermiin, L.S. *Syst. Biol.* **63**, 726–742 (2014).

ONLINE METHODS

ModelFinder is included in IQ-TREE version 1.5.4 and available from http://www.iqtree.org. ModelFinder complements other methods for identifying the optimal model of SE[9–11,23–30] for data comprising alignments of nucleotides or amino acids, but it differs from most of these other methods in three important ways:

1. ModelFinder considers alignments of nucleotides, codons, amino acids and other discrete data (e.g., binary and morphological data). Like the methods cited above (except for PartitionFinder[11]), ModelFinder defines the alignment as a single partition of sites.
2. ModelFinder includes the PDF model of RHAS proposed by Yang[13]; thus, the method increases the variety of models of RHAS that are considered during model selection. The PDF model has since been used elsewhere[31], but its suitability is not yet widely recognized.
3. ModelFinder allows the tree topology to vary during the search for an optimal model of SE, thus reducing the chance of entrapment in local optima during model selection. This search strategy has been used previously[28], but its suitability is not yet widely recognized.

ModelFinder uses three algorithms to search model space. Algorithm 1 (default search option), uses the following steps:

```
0. Given an alignment of characters ($\mathbf{D}$)
1. Find a reasonable tree $T$ (inferred using parsimony)
2. Obtain $L(\mathbf{D}|T,S_i,H_j)$ over $i$ and $j$—where $S_i$ is a list of substitution models, and $H_j$ is a list of RHAS models
3. Identify ($S_{\mathrm{opt}}$,$H_{\mathrm{opt}}$) using AIC, AICc or BIC (default)
   where $L(\mathbf{D}|T,S_i,H_j)$ denotes the likelihood of the data, given a tree, $T$, the $i$-th substitution model and the $j$-th model of RHAS, $S_{\mathrm{opt}}$ denotes the optimal substitution model, and $H_{\mathrm{opt}}$ denotes the optimal RHAS model.
```

Algorithm 2 (advanced search option) uses the following steps:

```
0. Given an alignment of characters ($\mathbf{D}$)
1. Obtain $L(\mathbf{D}|T_h,S_i,H_j)$ over $h$, $i$ and $j$—where $T_h$ is a list of trees (generated by IQ-TREE), $S_i$ is a list of substitution models, and $H_j$ is a list of RHAS models
2. Identify ($S_{\mathrm{opt}}$,$H_{\mathrm{opt}}$) using AIC, AICc or BIC
```

Algorithm 3 identifies the optimal PDF model of RHAS and is a key component of algorithm 1 and algorithm 2 (it is used whenever the PDF model of RHAS is considered). In the example given below, the BIC optimality criterion is used (but the AIC and AICc optimality criteria can be used if the user chooses to do so):

```
0. Given an alignment of characters ($\mathbf{D}$), a tree ($T$) and a substitution model ($S$)
1. Set $k = 2$
2. Obtain $L(\mathbf{D}|T,S,R_k)$ and $L(\mathbf{D}|T,S,R_{k+1})$
3. If $\mathrm{BIC}(L(\mathbf{D}|T,S,R_k)) > \mathrm{BIC}(L(\mathbf{D}|T,S,R_{k+1}))$
4. Increment $k$ by one unit, and go to 2
5. Else stop, and report $R_k$ as the optimal PDF model
```

In practice, algorithm 1 is invoked with this command (given here for an alignment of amino acids):

```
iqtree -s data.fst -st AA -m MF;
```

while algorithm 2 is invoked using:

```
iqtree -s data.fst -st AA -m MF -mtree.
```

IQ-TREE includes several other options (**Supplementary Table 2**) that will cause ModelFinder to conduct the search under different constraints. For example, the -m TEST and -m TESTONLY options will cause ModelFinder to operate like jModelTest[9] and ProtTest[10], while the -m TESTMERGE and -m TESTMERGEONLY options will cause it to operate like PartitionFinder[11]. However, none of these options considers the PDF model of RHAS. To do this, it is necessary to use the -m MF and -m MFP options.

When the PDF model is used, it is often necessary to optimize more than two parameters (the I + Γ_4 model is parameterized using two parameters). To ensure that these parameters are estimated as accurately as possible, we initially compared parameter estimates obtained using two parameter optimization procedures: the expectation maximization (EM) algorithm[15] (see subsection below) and the quasi-Newton BFGS algorithm[32]. We found the EM algorithm to be most accurate (data not shown).

ModelFinder is fast. For example, when benchmarking time required by the standard model-selection procedure of ModelFinder, we saw a 39- to 289-fold speedup when compared with jModelTest[9] (based on 70 alignments of DNA) and a 16- to 52-fold speedup when compared with ProtTest[10] (based on 45 alignments of amino acids).

Model selection for the alignment used by Wu *et al.*[19] (i.e., 6,597 sites and 353 species) was done using two commands:

```
iqtree -s data.fst -st AA -m MF -msub nuclear -cmax 20
and
iqtree -s data.fst -st AA -m MF -msub nuclear -cmax 20 -mtree.
```

Having found the optimal model of SE for the data, phylogenetic analyses were done under six models of SE using the following commands:

```
iqtree -s data.fst -st AA -m WAG+I+G5
iqtree -s data.fst -st AA -m LG+I+G5
iqtree -s data.fst -st AA -m LG+I+G4
iqtree -s data.fst -st AA -m LG+G4
iqtree -s data.fst -st AA -m LG+R14
and
iqtree -s data.fst -st AA -m LG+G14.
```

Each of these analyses was repeated 100 times to reduce the likelihood of the search algorithm being caught in local optima. The fact that the fit between tree, model and data varied across the 100 results for each of these models of SE indicates that this problem is an issue to consider, as done here.

Model selection for the alignments considered in **Table 1** was done using commands like those above, albeit with some variations to accommodate, for example, the type of data.

Model selection for the data considered in **Supplementary Table 1** was done using two commands:

doi:10.1038/nmeth.4285

```
iqtree -s data.fst -m MF -mtree
and
iqtree -s data.fst -m TEST.
```

The first command causes IQ-TREE to run the advanced version of ModelFinder; the second command causes IQ-TREE to run its implementation of jModelTest[9] or ProtTest[10], followed by a phylogenetic analysis under the optimal model of SE.

The PDF model is available in three other phylogenetic programs (i.e., PhyML[33], PhyTime[34], and BEAST[35]), so users of ModelFinder are not limited to using IQ-TREE to solve their phylogenetic questions.

Practical considerations. When using ModelFinder, it is important to remember that the method optimizes the likelihood of the tree and model, given the data, whenever it searches for the optimal values of parameters considered. Therefore, it is possible that the search algorithms may become trapped in local optima. To reduce the chance of this occurring, we strongly recommend model selection be repeated many times for each data set, as noted above. Doing so may entail using much more computing time, especially when long, species-rich alignments are considered, or the advanced search option of ModelFinder is used. Therefore, when the alignment is very long, we recommend the following set of strategies to reduce the amount of time used on model selection.

1. If the computational resources allow distributed computing, invoke the –nt *x* option to spread the processes over *x* threads.
2. If the data are characters encoded by a specific type of genome (e.g., mitochondrial), invoke the –msub source option to limit the search to this specific type of data.
3. If the optimal model turns out to include the R_{10} model of RHAS, we recommend the analysis be rerun with both the –cmin *x* and –cmax *y* options invoked (e.g., –cmin 8, –cmax 20). Doing so will ensure that PDF models with k = 8,9,…,20 are considered (i.e., lower values of k are ignored). The program will stop when the optimal value of k has been found, even if this value turns out to be 10.
4. Use the default search option to find the optimal model of SE. Having identified this model, use the advanced search option with the optimal substitution model selected (e.g., –mset LG) to search for the optimal model of RHAS. While there is no guarantee that this approach will identify the optimal model of SE, our experience suggests that the choice of RHAS model is highly influenced by the topology of the tree, while that of the substitution model is not.

The expectation-maximization algorithm to estimate PDF model parameters. Let $\Theta = \{w_1, \ldots, w_k, r_1, \ldots, r_k\}$ be the weights and rates of the PDF model R_k that we want to estimate. First, we initialize Θ using a discrete Γ_k model[12] (i.e., the initial values of $\widehat{w_1} = \ldots = \widehat{w_k} = 1/k$ and $\widehat{r_1},\ldots,\widehat{r_k}$ are derived from the discrete Γ distribution with k categories and a shape parameter $\alpha = 1$). This becomes the current estimate $\hat{\Theta}$. The EM algorithm iteratively performs an expectation (E-) step and a maximization (M-) step to update the current estimate until a (local) maximum in likelihood is reached.

E-step. For the i-th site in the alignment $\mathbf{D}_i$ and the j-th category, compute the posterior probability $\widehat{p_{ij}}$ of $\mathbf{D}_i$ belonging to category j based on the current estimate $\hat{\Theta}$:

$$\widehat{p_{ij}} = \frac{\widehat{w_j} L(\mathbf{D}_i \mid T,S,\widehat{r_j})}{\sum_{c=1}^{k} \widehat{w_c} L(\mathbf{D}_i \mid T,S,\widehat{r_c})} \qquad (1)$$

where $L(\mathbf{D}_i \mid T,S,\widehat{r_j})$ is the likelihood of the tree T, substitution model S and relative rate $\widehat{r_j}$ for the alignment site $\mathbf{D}_i$.

M-step. For each category j the log-likelihood function:

$$\log L = \sum_{i=1}^{N} \widehat{p_{ij}} \log L(\mathbf{D}_i \mid T,S,r_j) \qquad (2)$$

is maximized to obtain the next $\widehat{r_j}^{\text{NEW}}$, where N is the number of sites in the alignment. This can be done with standard numerical optimization such as Brent's method[36]. The weights are updated using

$$\widehat{w_j}^{\text{NEW}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{p_{ij}} \qquad (3)$$

that is, the new weight for category j is the mean posterior probability of each alignment site belonging to class j. This completes the proposal of the new estimate $\hat{\Theta}^{\text{NEW}}$. If the likelihood of $\hat{\Theta}^{\text{NEW}}$ is higher than that of $\hat{\Theta}$, then $\hat{\Theta}$ is replaced by $\hat{\Theta}^{\text{NEW}}$, and the E- and M-steps will be repeated. Otherwise, the EM algorithm stops and reports $\hat{\Theta}$ as the maximum-likelihood estimates of the PDF model R_k.

This EM algorithm allows estimation of the parameters of the R_k model, given a fixed tree T and a substitution model S. ModelFinder then iteratively estimates branch lengths of T, model parameters of S and R_k until the likelihood converges.

Software availability. ModelFinder is implemented in IQ-TREE version 1.5.4 (http://www.iqtree.org).

Data availability. Data and scripts used in this study are available from http://www.iqtree.org/ModelFinder/.

23. Posada, D. & Crandall, K.A. *Bioinformatics* **14**, 817–818 (1998).
24. Chiotis, M., Jermiin, L.S. & Crozier, R.H. *Mol. Phylogenet. Evol.* **17**, 108–116 (2000).
25. Abascal, F., Zardoya, R. & Posada, D. *Bioinformatics* **21**, 2104–2105 (2005).
26. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J. & Mclnerney, J.O. *BMC Evol. Biol.* **6**, 29 (2006).
27. Posada, D. *Nucleic Acids Res.* **34**, W700–W703 (2006).
28. Posada, D. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).
29. Santorum, J.M., Darriba, D., Taboada, G.L. & Posada, D. *Bioinformatics* **30**, 1310–1311 (2014).
30. Whelan, S., Allen, J.E., Blackburne, B.P. & Talavera, D. *Syst. Biol.* **64**, 42–55 (2015).
31. Soubrier, J. *et al. Mol. Biol. Evol.* **29**, 3345–3358 (2012).
32. Fletcher, R. *Practical Methods of Optimization* 2nd edn (John Wiley & Sons, 2000).
33. Guindon, S. *et al. Syst. Biol.* **59**, 307–321 (2010).
34. Guindon, S. *Syst. Biol.* **62**, 22–34 (2013).
35. Bouckaert, R. *et al. PLoS Comp. Biol.* **10**, e1003537 (2014).
36. Brent, R.P. *Algorithms for Minimization without Derivatives* (Prentice Hall, 1973).