Syst. Biol. 0(0):1–7, 2017

DOI:10.1093/sysbio/syx092

Complex Models of Sequence Evolution Require Accurate Estimators as Exemplified with the Invariable Site Plus Gamma Model

LAM-TUNG NGUYEN[1], ARNDT VON HAESELER[1,2], AND BUI QUANG MINH[1,*]

[1]*Center for Integrative Bioinformatics Vienna, Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Campus Vienna Biocenter 5, A-1030, Vienna, Austria; and* [2]*Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Waehringer Strasse 29, A-1090 Vienna, Austria*

**Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, Campus Vienna Biocenter 5 (VBC5), A-1030 Vienna, Austria; E-mail: minh.bui@univie.ac.at*

Received 8 March 2017; reviews returned 22 November 2017; accepted 23 November 2017
Associate Editor: Edward Susko

Abstract.—The invariable site plus Γ model (I+Γ) is widely used to model rate heterogeneity among alignment sites in maximum likelihood and Bayesian phylogenetic analyses. The proof that the I+ continuous Γ model is identifiable (model parameters can be inferred correctly given enough data) has increased the creditability of its application to phylogeny reconstruction. However, most phylogenetic software implement the I+ discrete Γ model, whose identifiability is likely but unproven. How well the parameters of the I+ discrete Γ model are estimated is still disputed. Especially the correlation between the fraction of invariable sites and the fractions of sites with a slow evolutionary rate is discussed as being problematic. We show that optimization heuristics as implemented in frequently used phylogenetic software (PhyML, RAxML, IQ-TREE, and MrBayes) cannot always reliably estimate the shape parameter, the proportion of invariable sites, and the tree length. Here, we propose an improved optimization heuristic that accurately estimates the three parameters. While research efforts mainly focus on tree search methods, our results signify the equal importance of verifying and developing effective estimation methods for complex models of sequence evolution. [Gamma model; invariable sites; maximum likelihood; phylogenetic inference; rate heterogeneity among sites.]

In model based phylogenetic analysis, the invariable site plus Γ model (Yang 1994; Gu et al. 1995), hereafter referred to as I+Γ, is widely used to model rate heterogeneity among sites, because it often fits the data better than the Γ model or the invariable-sites model alone (Sullivan and Swofford 1997). Thus, the I+Γ model is frequently selected by MODELTEST (Posada and Crandall 1998). The I+Γ model has two parameters: the proportion of invariable sites p_{inv} ($0 \leqslant p_{\text{inv}} < 1$) and the shape parameter α (>0) of the Γ distribution. A small α (<1) indicates strong rate heterogeneity, whereas a large α (>1) corresponds to weak rate heterogeneity. Under certain conditions p_{inv} and α compete with each other for the same phylogenetic signal. For example, $\alpha \leqslant 1$ already accounts for sites with low rates; that interferes with p_{inv} and causes a correlation between the parameters making reliable estimation of those parameters difficult (Sullivan et al. 1999; Mayrose et al. 2005). Despite this interference, it has been shown that the I+ continuous Γ model is identifiable for "all but members of the F81 family of rate matrices on any phylogeny with more than two distinct interspecies distances" (Rogers 2001; Allman and Rhodes 2008; Chai and Housworth 2011). Since the I+ continuous Γ model is identifiable, reliable parameter estimation for this model should be possible for sufficiently long multiple sequence alignments.

However, most phylogenetic software only implement the I+ discrete Γ (Yang 1994) model as an approximation to the continuous Γ model because of its computational efficiency. The discussed competition between p_{inv} and α is based on the analysis of the discrete Γ -distribution. The results have led to the suggestion to discourage the use of the I+ discrete Γ model (Yang 2006; Jia et al. 2014; Stamatakis 2014).

On the other hand, the identifiability of the I+ discrete Γ model is likely, but unproven (Chai and Housworth 2011), and it is unclear how accurately popular phylogenetic software estimate parameters of the I+ discrete Γ model.

Thus, we used simulations to assess the accuracy of the I+ discrete Γ estimators implemented in three maximum likelihood (ML) phylogenetic software: RAxML (Stamatakis 2014), PhyML (Guindon et al. 2010), IQ-TREE (Nguyen et al. 2015), and one Bayesian inference program MrBayes (Ronquist et al. 2012). More precisely, we simulated 100,000-bp long alignments along three balanced trees of 6, 24, and 96 taxa. The lengths of the alignments ensure the recovery of the correct tree topology. The three trees have uniform branch lengths of 0.1 substitutions per site except for one internal branch on the 6-taxon tree whose length equals 0.2 to allow for three distinct distances between the sequences as required for identifiability in the continuous case (Chai and Housworth 2011). We assumed the K2P model (Kimura 1980) with a transition/transversion ratio of 2.0 and the rate heterogeneity model I+ discrete Γ with four rate categories. For each tree and each pair $(p_{\text{inv}}, \alpha) \in \{0.0,\ 0.1, \ldots 0.9\} \times \{0.1, 0.5, 1.0\}$, we simulated 100 alignments using Seq-Gen (Rambaut and Grassly 1997). We used RAxML version 8.2.2, PhyML

version 20141029, IQ-TREE version 1.3.7, and MrBayes version 3.2.6 compiled with the BEAGLE library (Ayres et al. 2012) to infer the invariable proportion, the shape parameter, and the tree length from the simulated alignments. For RAxML, PhyML, and IQ-TREE, we used the default options.

For MrBayes we used the default priors, that is, uniform distribution within interval [0,1] for p_{inv}, exponential distribution with mean 1.0 for α, nonclocklike uniform Dirichlet distribution for branch lengths and Γ distribution with mean of 10 for tree lengths (Unconstrained:GammaDir(1.0,0.1,1.0,1.0)). The sequential version of MrBayes was run with four chains (one hot and three cold chains) and one million MCMC generations. One thousand four hundred and eighty-nine (16.5%) nonconvergent MrBayes runs, where the effective sample sizes (ESS) on p_{inv}, α, or tree lengths are smaller than 100, were repeated with five million generations. However, 52 of the extended reruns were stopped after 4 weeks without completing all five million generations. We note that 207 of the 1489 reruns still did not converge. MrBayes estimates are then summarized as the mean of the posterior distribution with a default burn-in of 25%.

Current Phylogenetic Programs Do Not Produce Accurate Estimates for the I+ Discrete Γ Model

Figure 1 displays the averages $\bar{\hat{\alpha}}$ of the estimated shape parameter α, the averages $\widehat{p_{inv}}$ of the estimated invariable fraction p_{inv} and the average $\bar{\hat{l}}$ of the estimated tree length l produced by PhyML, RAxML, IQ-TREE, and MrBayes for the 100 alignments simulated from each parameter combinations. A program is called *accurate* if the estimated averages $\bar{\hat{\alpha}}, \widehat{p_{inv}}, \bar{\hat{l}}$ deviate no more than 10% from the true values.

None of the tested programs estimated all parameter combinations accurately. The problem is especially pronounced for the 6-taxon alignments. For extreme rate heterogeneity ($\alpha = 0.1$) MrBayes and PhyML recovered the true α, p_{inv}, and l for 9/10 and 5/10 parameter combinations respectively, whereas the average estimates from IQ-TREE and RAxML were inaccurate. For strong rate heterogeneity ($\alpha = 0.5$), the degrees of inaccuracy observed among all programs differ unsystematically. On the one hand, IQ-TREE and MrBayes accurately estimated the parameters in four and six settings. On the other hand, RAxML and PhyML could not estimate accurately the three parameters for any of the ten parameter-combinations. For medium rate variation ($\alpha = 1.0$), only IQ-TREE produced the accurate estimates for all settings. All other programs exhibited varying degrees of inaccuracy.

For the 24- and 96-taxon alignments we observed an increase in the number of accurate estimates for all programs. These results corroborate a previous study (Sullivan et al. 1999) showing that increased taxon sampling leads to more reliable estimates. However, under extreme rate heterogeneity ($\alpha = 0.1$), only MrBayes estimated all parameter sets accurately. We note that our measure of accuracy correlates well with the Bayesian coverage probabilities, the frequency with which true parameter values are included in the 95% credible interval of the estimates (Supplementary Fig. S1 available on Dryad at https://doi.org/10.5061/dryad.4j5c7). Two hundred and seven (2.3%) nonconvergent MrBayes runs (effective sample size of α or p_{inv} are smaller than 100) partly overlap with cases where MrBayes was not accurate for 6-taxon simulations ($\alpha = 0.1$ and $p_{\text{inv}} = 0.9$; $\alpha = 0.5$ and $p_{\text{inv}} \leqslant 0.5$; $\alpha = 1.0$ and $0.2 \leqslant p_{\text{inv}} \leqslant 0.5$). Hence, nonconvergence is a predictor of difficult settings but does not fully explain the inaccuracy of MrBayes (Fig. 1).

We also observed that inaccurate estimates of α and p_{inv} could sometimes lead to tree lengths that substantially deviate from the simulated lengths. For instance, for the 96-taxon alignments simulated with $\alpha = 0.1$ and $p_{\text{inv}} = 0.8$ (expected tree length = 18.9) IQ-TREE estimated an average tree length of 177.0 that is nine times longer than the simulated tree length. The other programs also sometimes produced tree lengths that were considerably longer than the simulated ones.

In terms of computing times PhyML, RAxML, and IQ-TREE needed for all analyses 62,441, 12,563, and 7,675 CPU hours, respectively. MrBayes needed 740,681 CPU hours to complete one million MCMC generations, thus it is 96.5 times slower than the fastest ML program. We note that this is only a lower-bound for the effective time one needs to wait for MrBayes results because 16.5% of MrBayes runs did not converge after one million MCMC generations. These runs were repeated with five times more generations, that led to significantly more computations.

Multiple Local Optima on the Likelihood Surface Cause Inaccuracy

Because the tested programs performed quite differently with respect to the accuracy of parameter estimation, the number of taxa cannot be the only explanation. We suspected that the optimization heuristics as implemented in these programs drive the accuracy. Examining the likelihood surfaces for many simulated alignments revealed a common feature that the parameter space has two distinct peaks of high log-likelihoods (Fig. 2): one global close to the true parameters and one suboptimal peak with slightly lower log-likelihood ($\Delta\text{LnL} = -27$ in this example) separated by a flat valley from the true parameters. In this particular instance, MrBayes and PhyML found the true parameters whereas RAxML and IQ-TREE were trapped in the local maximum (not necessarily the case for other instances). In fact, whether the global or local optimum is detected depends on the starting values of the numerical optimization routines.

To summarize, we compared for each simulated alignment the log-likelihoods of the estimates with the log-likelihoods obtained for the true parameters.

PhyML

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
6-taxon tree (tree length = 1.0)										
0.10	0.00	0.10	0.20	0.25	0.39	0.50	0.61	0.89	0.90	0.90
	0.10	0.10	0.10	0.06	0.09	0.10	0.11	0.56	0.19	0.11
	1.00	1.00	1.00	0.92	0.98	0.99	1.17	1.71	3.68	0.76
0.50	0.19	0.18	0.17	0.17	0.25	0.36	0.43	0.53	0.68	0.69
	0.84	0.62	0.47	0.35	0.33	0.32	0.26	0.23	0.22	0.04
	1.00	1.00	1.00	0.99	0.98	0.98	0.95	0.91	0.91	18.58
1.00	0.11	0.20	0.20	0.19	0.17	0.15	0.17	0.29	0.45	0.77
	1.50	1.47	0.99	0.67	0.45	0.30	0.21	0.17	0.13	0.52
	1.00	1.00	1.00	1.00	1.00	0.98	0.94	0.90	0.80	4.54
24-taxon tree (tree length = 4.5)										
0.10	0.41	0.28	0.35	0.30	0.43	0.51	0.69	0.70	0.80	0.93
	0.55	0.30	0.26	0.10	0.13	0.11	0.18	0.10	0.10	0.17
	4.93	4.93	5.14	4.48	4.79	5.93	26.92	4.43	4.53	10.76
0.50	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	0.51	0.50	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	4.50	4.50	4.50	4.50	4.49	4.51	4.50	4.52	4.49	4.45
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.89
	1.01	1.01	1.00	1.00	0.99	0.99	1.00	0.99	1.00	0.87
	4.50	4.50	4.49	4.50	4.50	4.51	4.50	4.52	4.54	136.95
96-taxon tree (tree length = 18.9)										
0.10	0.00	0.10	0.49	0.55	0.61	0.50	0.62	0.70	0.90	0.92
	0.10	0.10	0.28	0.28	0.28	0.10	0.13	0.10	0.45	0.17
	18.83	18.91	13.90	14.00	14.06	18.69	17.95	18.27	35.95	38.45
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	18.88	18.88	18.88	18.82	18.90	18.88	18.94	18.77	18.89	18.29
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.03
	18.90	18.89	18.89	18.87	18.94	18.93	18.99	18.77	18.92	18.49

RAxML

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.71	0.74	0.77	0.79	0.82	0.84	0.86	0.89	0.90	0.92	$\widehat{p_{inv}}$
	4.63	3.85	3.08	2.36	1.69	1.15	0.67	0.56	0.24	0.22	$\tilde{\alpha}$
	1.03	1.04	1.07	1.10	1.17	1.33	1.78	1.74	4.39	1.38	$\bar{l}$
0.50	0.30	0.34	0.40	0.45	0.51	0.57	0.64	0.72	0.81	0.91	$\widehat{p_{inv}}$
	1.31	1.17	1.05	0.95	0.83	0.75	0.68	0.63	0.63	0.73	$\tilde{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	$\bar{l}$
1.00	0.16	0.22	0.29	0.37	0.45	0.53	0.61	0.71	0.80	0.90	$\widehat{p_{inv}}$
	1.89	1.74	1.59	1.45	1.35	1.25	1.17	1.14	1.10	1.05	$\tilde{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	$\bar{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.64	0.67	0.70	0.73	0.76	0.73	0.77	0.82	0.88	0.94	$\widehat{p_{inv}}$
	0.82	0.72	0.64	0.58	0.53	0.20	0.17	0.18	0.24	0.29	$\tilde{\alpha}$
	5.18	5.87	6.74	7.69	8.91	16.64	20.97	28.88	8.72	3.38	$\bar{l}$
0.50	0.01	0.11	0.21	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.51	0.51	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.51	$\tilde{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	$\bar{l}$
1.00	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.02	1.02	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	$\tilde{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.50	4.52	$\bar{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.37	0.43	0.49	0.65	0.67	0.68	0.74	0.81	0.87	0.94	$\widehat{p_{inv}}$
	0.27	0.27	0.28	0.62	0.49	0.28	0.26	0.27	0.27	0.28	$\tilde{\alpha}$
	13.85	13.88	13.94	13.43	13.79	14.08	14.55	14.32	14.20	11.90	$\bar{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\tilde{\alpha}$
	18.89	18.90	18.91	18.89	18.89	18.89	18.89	18.92	18.91	18.88	$\bar{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\tilde{\alpha}$
	18.90	18.89	18.90	18.89	18.90	18.90	18.90	18.89	18.90	18.92	$\bar{l}$

IQ-TREE

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
6-taxon tree (tree length = 1.0)										
0.10	0.71	0.74	0.77	0.79	0.82	0.84	0.86	0.89	0.90	0.92
	4.80	4.02	3.27	2.51	1.80	1.15	0.64	0.55	0.28	0.17
	1.03	1.04	1.06	1.10	1.17	1.32	1.76	1.73	3.47	1.16
0.50	0.18	0.23	0.30	0.36	0.42	0.51	0.61	0.70	0.80	0.91
	0.82	0.74	0.68	0.62	0.56	0.53	0.52	0.52	0.53	0.68
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.87
1.00	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	1.02	1.01	1.01	1.00	1.01	1.00	1.02	1.04	1.06	1.03
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98
24-taxon tree (tree length = 4.5)										
0.10	0.64	0.67	0.70	0.73	0.76	0.79	0.76	0.81	0.87	0.95
	0.82	0.72	0.64	0.58	0.52	0.29	0.14	0.14	0.14	0.30
	5.17	5.86	6.73	7.68	8.89	62.72	41.99	40.81	41.04	43.68
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.53
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.51	4.59
96-taxon tree (tree length = 18.9)										
0.10	0.36	0.43	0.59	0.65	0.70	0.75	0.79	0.80	0.87	0.93
	0.27	0.27	0.57	0.62	0.63	0.64	0.63	0.28	0.17	0.16
	13.83	13.87	13.26	13.48	13.69	14.00	14.61	19.03	177.99	74.08
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	18.89	18.90	18.90	18.88	18.92	18.95	18.99	19.09	19.23	19.57
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	18.90	18.89	18.90	18.90	18.92	18.94	18.96	18.99	19.13	19.56

MrBayes

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.07	$\tilde{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02	7.00	$\bar{l}$
0.50	0.10	0.14	0.20	0.28	0.38	0.49	0.60	0.70	0.66	0.74	$\widehat{p_{inv}}$
	0.65	0.57	0.52	0.50	0.48	0.49	0.50	0.50	0.26	0.11	$\tilde{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.51	$\bar{l}$
1.00	0.04	0.09	0.18	0.28	0.27	0.29	0.24	0.32	0.46	0.68	$\widehat{p_{inv}}$
	1.14	1.01	0.96	0.96	0.75	0.62	0.32	0.21	0.13	0.10	$\tilde{\alpha}$
	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.91	0.80	0.63	$\bar{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\tilde{\alpha}$
	4.48	4.50	4.50	4.51	4.51	4.51	4.51	4.52	4.56	4.70	$\bar{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.87	$\widehat{p_{inv}}$
	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.25	$\tilde{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	3.69	$\bar{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.66	$\widehat{p_{inv}}$
	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.10	$\tilde{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.51	4.51	2.45	$\bar{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\tilde{\alpha}$
	18.85	18.91	18.92	18.92	18.90	18.93	18.99	19.01	18.97	18.93	$\bar{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.76	0.88	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.36	0.37	$\tilde{\alpha}$
	18.88	18.90	18.91	18.90	18.90	18.90	18.90	18.93	11.91	13.89	$\bar{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.89	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.62	$\tilde{\alpha}$
	18.90	18.89	18.90	18.89	18.91	18.91	18.91	18.90	16.95	12.44	$\bar{l}$

FIGURE 1. The averages $\tilde{\alpha}$ of the estimated shape parameter α, the averages $\widehat{p_{inv}}$ of the estimated invariable fraction p_{inv} and the average $\bar{l}$ of the estimated tree length l produced by PhyML, RAxML, IQ-TREE, and MrBayes for the 100 alignments simulated from each parameter combinations. The averages are highlighted according to their differences from the true values: inaccurate (more than 25% deviation, red in online version), moderately inaccurate (10% to 25% deviation yellow in online version), and accurate (less than 10% deviation green in online version). For $p_{\text{inv}} = 0.0$ the estimated $\widehat{p_{inv}}$ is accurate if $0 \leqslant \widehat{p_{inv}} \leqslant 0.01$, moderately inaccurate if $0.01 < \widehat{p_{inv}} \leqslant 0.05$, and inaccurate if $0.05 < \widehat{p_{inv}}$.

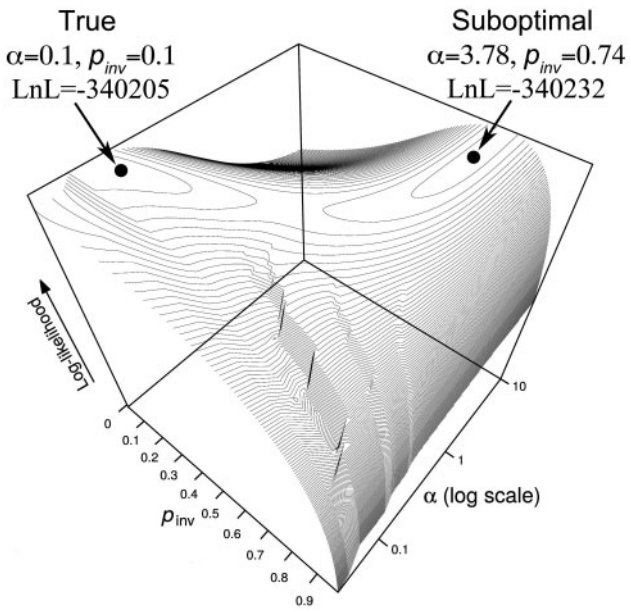

FIGURE 2. The likelihood surface for one simulated alignment as a function of α and p_{inv}.

TABLE 1. Percentage of alignments where the true simulation parameters result in higher log-likelihoods than the inferred parameters from the programs for three simulation scenarios (6-, 24-, and 96-taxon trees)

Program	6-taxon tree (%)	24-taxon tree (%)	96-taxon tree (%)
MrBayes	34.7	6.6	5.6
PhyML	60.2	21.9	45.0
RAxML	89.2	37.7	49.5
IQ-TREE	36.0	34.1	44.0

Table 1 show how often the true parameter combination produced a higher likelihood than the inferred parameters from MrBayes, PhyML, RAxML, and IQ-TREE. These fractions are particularly high for the 6-taxon tree and for the ML inference programs. Most ML phylogenetic programs use general-purpose numerical methods to find α and p_{inv} (e.g., Brent 1973). These methods are obviously not well adapted to the complex likelihood surface (Fig. 2) and explain the poor overall performance of the ML programs (Fig. 1).

EFFECTIVE OPTIMIZATION HEURISTIC PRODUCES ACCURATE ESTIMATES

As remedy, we propose an alternative optimization heuristic which employs the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) to estimate p_{inv}. We assume a discrete Γ distribution with k rate categories. Under the I+ discrete Γ model, the site rates follow a discrete mixture model consisting of $k+1$ categories with rates $r_0,..,r_k$, where $r_0=0$ represents invariable sites and $r_i>0$ $(i=1,\ldots,k)$ are the k rates determined from the shape parameter αof the discrete Γ distribution (Yang 1994). Given a tree topology, the optimization heuristic does the following:

1. Choose initial values for α and p_{inv}.
2. Optimize branch lengths by the Newton–Raphson method.
3. Optimize substitution model parameters by the Broyden–Fletcher–Goldfarb–Shanno algorithm.
4. For each alignment site D_i compute its posterior probability of being invariable ($1\leqslant i\leqslant n$, where n is the number of alignment sites):
$$P(r_0|D_i)=\frac{P(D_i|r_0)w_0}{\sum_{j=0}^{k}P(D_i|r_j)w_j},$$
where $P(D_i|r_j)$ is the likelihood of site D_i having rate r_j and $w_0=p_{\text{inv}}$, $w_j=\frac{1-p_{\text{inv}}}{k}$ $(1\leqslant j\leqslant k)$.
5. Update $p_{\text{inv}}=\frac{1}{n}\sum_{i=1}^{n}P(r_0|D_i)$.
6. Optimize α by the Brent method.
7. If the log-likelihood improvement is greater than a predefined ϵ value, go back to Step 2. Otherwise, stop the parameter optimization.

Steps 4 and 5 correspond to the E- and M-step of the EM algorithm, respectively. To avoid being stuck in local optima, we repeat this optimization procedure from ten starting values of p_{inv} evenly spaced between 0 and the fraction of constant sites observed in the alignment. The initial value of α is always set to 1.0.

We implemented the new optimization heuristic in IQ-TREE now called IQ-TREE-EM (IQ-TREE version 1.4.3) and repeated the previous simulations. Figure 3 shows that IQ-TREE-EM successfully recovered the true parameters for all but one parameter combination (6-taxon, $\alpha=0.5$ and $p_{\text{inv}}=0.0$) where the average estimates ($\tilde{\alpha}=0.59$ and $\widetilde{p_{inv}}=0.06$) slightly deviated from the true values.

Also, the percentage of instances where the estimated log-likelihoods were lower than the log-likelihood for the true parameters dropped considerably (0.06% 6-taxon tree, 0.0% 24-taxon tree, and 0.03% 96-taxon tree; compare also with Table 1).

This increase in accuracy comes at the cost of an increased total computing time by a factor of 1.3 compared to IQ-TREE.

Thus, we conclude that the inaccurate parameter estimation of the I+ discrete Γ shown for the tested phylogenetic programs is caused by ineffective optimization methods.

IMPACT ON REAL DATA

To investigate the impact of accuracy on real data for ML estimates, we analyzed 70 DNA and 45 protein

IQ-TREE-EM

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\bar{\hat{\alpha}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	0.93	$\bar{\hat{l}}$
0.50	0.06	0.10	0.19	0.27	0.38	0.48	0.59	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.59	0.52	0.50	0.48	0.49	0.48	0.48	0.52	0.52	0.54	$\bar{\hat{\alpha}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	$\bar{\hat{l}}$
1.00	0.01	0.09	0.19	0.29	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	1.05	1.00	0.98	0.98	1.02	1.03	1.03	1.04	1.00	1.02	$\bar{\hat{\alpha}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	$\bar{\hat{l}}$
24-taxon tree (tree length = 4.5)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\bar{\hat{\alpha}}$
	4.50	4.49	4.49	4.50	4.51	4.51	4.51	4.52	4.52	4.19	$\bar{\hat{l}}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\bar{\hat{\alpha}}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	$\bar{\hat{l}}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\bar{\hat{\alpha}}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.50	4.50	$\bar{\hat{l}}$
96-taxon tree (tree length = 18.9)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\bar{\hat{\alpha}}$
	18.90	18.89	18.90	18.91	18.89	18.92	18.97	18.98	19.02	18.80	$\bar{\hat{l}}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\bar{\hat{\alpha}}$
	18.90	18.90	18.91	18.89	18.90	18.90	18.90	18.92	18.91	18.86	$\bar{\hat{l}}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\bar{\hat{p}}_{inv}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\bar{\hat{\alpha}}$
	18.90	18.89	18.90	18.89	18.90	18.90	18.90	18.89	18.90	18.87	$\bar{\hat{l}}$

FIGURE 3. The averages $\bar{\hat{\alpha}}$ of the estimated shape parameter α, the averages $\widehat{\bar{p}_{inv}}$ of the estimated invariable fraction p_{inv} and the average $\bar{\hat{l}}$ of the estimated tree length l produced by IQ-TREE-EM for the 100 alignments simulated from each parameter combinations. The highlighting is explained in Fig. 1.

TreeBase alignments (Nguyen et al. 2015). We applied the GTR+I+Γ4 and LG+I+Γ4 models for DNA and protein data, respectively. Among 115 alignments, we detected 15 (5 DNA and 10 protein) alignments where the estimated α and p_{inv} by PhyML, RAxML, or IQTREE deviated more than 10% from those by IQ-TREE-EM (Fig. 4; Supplementary Table S1 available on Dryad). The estimates by PhyML and IQ-TREE deviated from those by IQ-TREE-EM only for one and two alignments, respectively. However, RAxML estimated α and p_{inv} dramatically different from IQ-TREE-EM, PhyML, and IQ-TREE for all 15 alignments. Interestingly, RAxML systematically overestimated α and p_{inv} for all 5 DNA and underestimated them for all 10 protein alignments (p_{inv} sometimes very close to zero).

DISCUSSION

Our simulations revealed a major issue for parameter estimation of the I+ discrete Γ model as implemented in phylogenetic software. Despite using very long alignments, none of the tested programs recovered the true α, p_{inv}, and tree length for all parameter combinations. Often, the estimates deviated heavily from the true values and different programs estimated different values for the same evolutionary parameters, although all programs inferred the true tree. Our further analysis of 115 TreeBase alignments showed that PhyML, IQ-TREE, and IQ-TREE-EM estimates generally agree with each other except for two alignments. However, we identified 15 (13%) alignments where RAxML systematically overestimated α and p_{inv} for

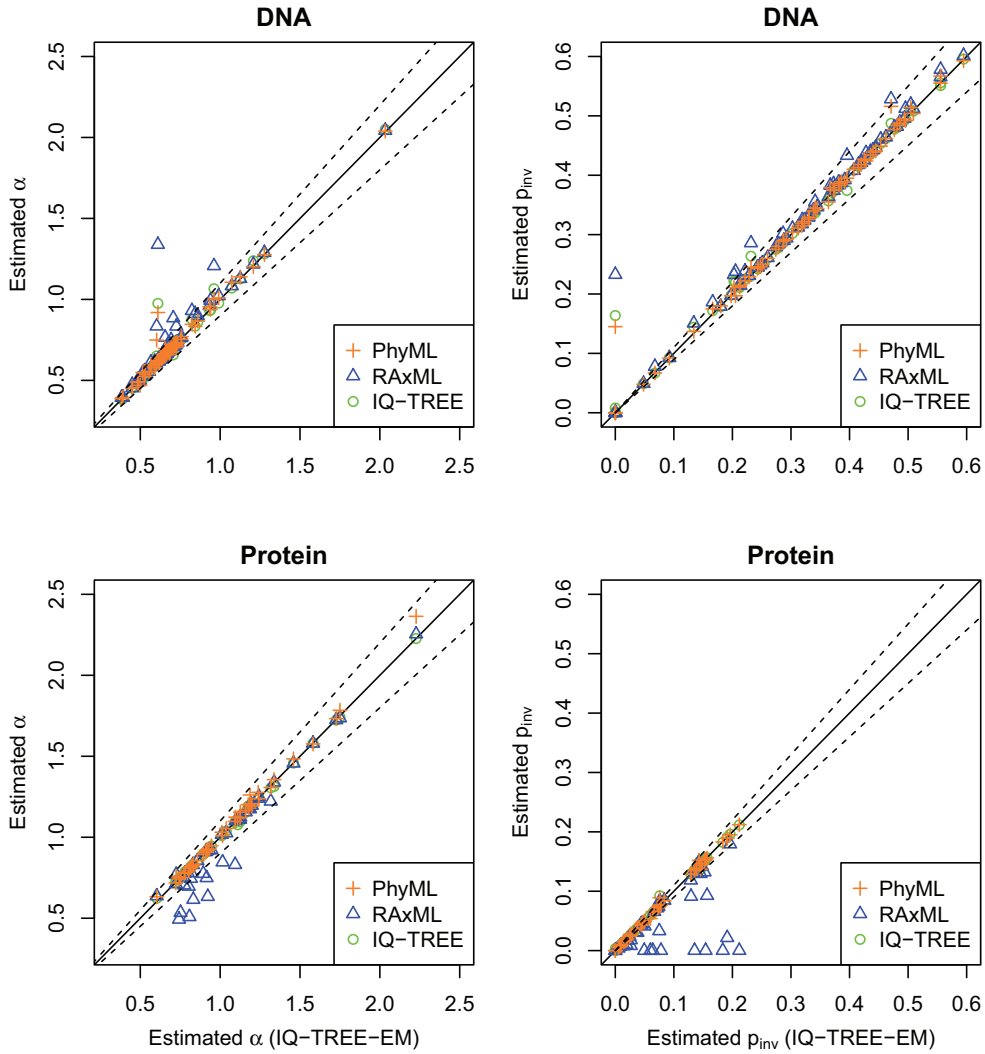

FIGURE 4. Estimation of α (left) and p_{inv} (right) for TreeBase alignments using IQ-TREE-EM (*x*-axis) and IQ-TREE (circle), PhyML (cross) and RAxML (triangle). Dashed lines show the boundaries of 10% deviation from the IQ-TREE-EM estimates. Points above the upper dashed lines indicate overestimation compared with IQ-TREE-EM, whereas points under the lower dashed lines indicate underestimation.

DNA and underestimated for protein, compared with other programs. The reasons for that behavior are unclear and deserve further analyses. While this result may not be extrapolated to other data sets, phylogenetic software should benefit from the more robust optimization described for IQ-TREE-EM.

We showed that the estimation heuristics implemented in popular phylogenetic programs causes such inaccurate estimates and the I+Γ model *per se* is not problematic. The relatively good performance of MrBayes is likely attributed to the Bayesian sampling of the parameter space but comes at the cost of excessive computing time.

With IQ-TREE-EM, we provided an alternative optimization heuristic for ML methods that allows accurate estimation of the parameters for the I+ discrete Γ model. IQ-TREE-EM combines two optimization techniques: the multiple starting point strategy and the EM algorithm. We note that the EM algorithm alone will not achieve this accuracy (Supplementary Fig. S2 available on Dryad). Therefore, while the former allows to escape local optima, the latter helps to speed-up the optimization using analytical formula for p_{inv}. This new approach effectively infers the true evolutionary parameters for long alignments. Thus, it is tempting to speculate that the GTR+I+ discrete Γ model is also identifiable as shown for the GTR+I+ continuous Γ model (Chai and Housworth 2011).

Our observations show that as models of sequence evolution become more and more complex (e.g., Dirichlet rate and other mixture models), tailored numerical optimization methods are necessary to achieve accurate estimates of evolutionary parameters. It is not enough to recover the true tree, if one wants to understand how evolutionary forces shaped contemporary genomes. The effect of wrong parameter estimates for the substitution model on the total tree length is sometimes dramatic (see Fig. 1). This may

in turn bias downstream analysis such as divergence time dating, inference of site-specific evolutionary rates, and ancestral sequence reconstruction, which are sensitive to the parameter estimates. Thus, one should critically scrutinize the heuristics implemented in popular programs. A more thorough evaluation of phylogenetic inference programs allowing for very complicated models of sequence evolution is necessary, but beyond the scope of this article.

Finally, we would like to point out that we only addressed the accurate computation of p_{inv} and α for the widely used I+ discrete Γ model. We do not discuss the biological interpretation of p_{inv}. The estimate of p_{inv} depends very much on the multiple sequence alignment at hand. p_{inv} may change if we enlarge the alignment. Thus, drawing an absolute conclusion from p_{inv} is in any case questionable.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.4j5c7.

Funding

This work was supported by the Austrian Science Fund – FWF (Grant Nos I-2805-B29 and I-1824-B22).

Acknowledgements

The authors would like to thank Heiko A. Schmidt for fruitful discussions, two anonymous reviewers, Fredrik Ronquist, and Edward Susko for constructive and helpful comments on an earlier version of the manuscript. The computational results presented have been achieved using the Vienna Scientific Cluster 3 (VSC-3).

References

Allman E.S, Rhodes J.A. 2008. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. Math. Biosci. 211:18–33.

Ayres D.L., Darling A., Zwickl D.J. Beerli P., Holder M.T., Lewis P.O., Huelsenbeck J.P., Ronquist F., Swofford D.L., Cummings M.P., Rambaut A., Suchard M.A. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol 61:170–173.

Brent R.P. 1973. Algorithms for minimization without derivatives. Englewood Cliffs, New Jersey: Prentice-Hall. p. 1–195.

Chai J., Housworth E.A. 2011. On Rogers' proof of identifiability for the GTR+ Gamma+ I model. Syst. Biol. 60:713–718.

Dempster A.P., Laird N.M., Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 75:1–38.

Gu X., Fu Y.X., Li W.H. 1995. Maximum-likelihood-estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Jia F.Z., Lo N., Ho S.Y.W. 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLos One 9:e95722.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J. Mol. Evol. 16:111–120.

Mayrose I., Friedman N., Pupko T. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q.. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp. Appl. Biosci. 13:235–238.

Rogers J.S. 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. Syst. Biol. 50:713–722.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Sullivan J., Swofford D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.

Sullivan J., Swofford D.L., Naylor G.J.P. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16:1347–1356.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 2006. Computational molecular evolution. New York: Oxford University Press. p. 113–114.