CodonPhyML: Fast Maximum Likelihood Phylogeny Estimation under Codon Substitution Models

Manuel Gil,*[†,1,2,3] Marcelo Serrano Zanetti,[†,1,4] Stefan Zoller,[1,2] and Maria Anisimova*[,1,2]

[1]Department of Computer Science, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[3]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University Vienna, Vienna, Austria

[4]Department of Management, Technology, and Economics, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: manuel.gil.sci@gmail.com; maria.anisimova@inf.ethz.ch.

Associate editor: Claudia Russo

Abstract

Markov models of codon substitution naturally incorporate the structure of the genetic code and the selection intensity at the protein level, providing a more realistic representation of protein-coding sequences compared with nucleotide or amino acid models. Thus, for protein-coding genes, phylogenetic inference is expected to be more accurate under codon models. So far, phylogeny reconstruction under codon models has been elusive due to computational difficulties of dealing with high dimension matrices. Here, we present a fast maximum likelihood (ML) package for phylogenetic inference, CodonPhyML offering hundreds of different codon models, the largest variety to date, for phylogeny inference by ML. CodonPhyML is tested on simulated and real data and is shown to offer excellent speed and convergence properties. In addition, CodonPhyML includes most recent fast methods for estimating phylogenetic branch supports and provides an integral framework for models selection, including amino acid and DNA models.

Key words: **maximum likelihood, phylogeny inference, codon model, evolution, selection.**

Introduction

The desire to understand how molecular changes affect organism's fitness and phenotype explains the growing number of phylogenetic studies of genomic sequences, particularly protein-coding genes. Indeed, protein-coding genes are transcribed and translated into proteins, which are ultimately responsible for the inner workings of biological cells. The evolution of protein-coding genes can be studied at the level of DNA or amino acids. However, because the transcription–translation machinery reads nucleotide sequences in triplets, known as codons, modeling the substitution process at the level of codons should provide the most realistic description of protein-coding sequence evolution. Past years have seen dramatic developments in modeling evolution of protein-coding genes using Markov codon models (Anisimova and Kosiol 2009). Unlike their DNA and amino acid counterparts, codon models naturally incorporate the structure of the genetic code and can distinguish between synonymous and nonsynonymous substitutions. This means that codon models explicitly include natural selection pressure acting on proteins, typically through the nonsynonymous to synonymous substitution rate ratio ($\omega = d_N/d_S$).

Certainly, Markov codon models proved indispensible for studies of selective pressures at the protein level, and recent methodological developments further widened the spectrum of their usage to include applications to sequence alignment, studies of codon bias, and dating divergence events. Using codon models for phylogeny reconstruction is the next logical development. In most protein-coding genes, synonymous substitutions should be informative about recent divergences, whereas nonsynonymous substitutions occur at low rates and contribute to resolving deeper divergences. Ren et al. (2005) suggested that potential benefits of phylogeny inference under codon models may indeed include recovering both recent and deep nodes. Despite the common belief that synonymous substitutions quickly reach saturation, recent studies showed that synonymous substitutions often carry valuable signal even at deep divergences (Seo and Kishino 2008, 2009).

Using codon models for fast phylogeny inference is computationally challenging as the search algorithm relies on manipulation with much higher dimension matrices compared with DNA and amino acid models: For the standard genetic code with 61 sense codons, a Markov codon model is defined by a 61×61 instantaneous rate matrix. This may explain why phylogeny inference under codon models remains elusive to date. Although several implementations of codon models exist (for a comprehensive list see Anisimova 2012), none of them specifically caters for phylogeny inference. The only implementations offering a limited number of simple codon models for phylogeny estimation include MrBayes (Ronquist et al. 2012) and Beast (Drummond and Rambaut 2007), which allow Bayesian estimation, and GARLI (Zwickl 2006) and IQPNNI (Minh et al. 2005; Schmidt and von

Haeseler 2009), which perform tree estimation by the maximum likelihood (ML) criterion. However, the performance of tree inference under codon models (in terms of accuracy and speed) has not been documented for any of these programs. Thus, phylogeny inference under codon models is currently an uncharted territory in the field of computational molecular evolution.

New Approach

Here, for the first time, we present CodonPhyML—a computationally tractable implementation of codon models for phylogeny estimation using ML. CodonPhyML capitalizes on the fast ML heuristic search algorithms implemented in PhyML 3.0 (Guindon et al. 2010). Our new package boasts the largest number of codon models implemented in any package to date. CodonPhyML implements a variety of time-homogeneous and time-reversible codon models (for review on codon models see Anisimova and Kosiol 2009). These can be subdivided into parametric, empirical, and semiparametric (including models based on principal component analysis [PCA]). Most of the models can be combined with heterogeneity of evolutionary rates and selective pressures, and different ways of defining instantaneous rate matrix entries (MG-, GY-, and Yap-type) and estimating codon frequencies (e.g., F3x4, CF3x4, and by ML)—for full details see the Materials and Methods section.

We scrutinize the accuracy of our implementation with rigorous tests on simulated and real data and optimize implementation speed by adding high-performance computing and additional heuristics. In particular, we include the de facto standard libraries (BLAS/LAPACK) for linear algebra operations and exploit loop-level parallelism using OpenMP on multicore processors. Please refer to the Materials and Methods section for a detailed description of all available features.

Several fast branch support methods (Anisimova et al. 2011) are available within CodonPhyML. In addition, CodonPhyML allows a direct comparison of codon models to DNA and amino acid models by providing adjusted log likelihoods under DNA and amino acid models that are directly comparable to the log likelihoods obtained under codon models (Seo and Kishino 2008, 2009).

Results and Discussion

Verification of Correctness

Exhaustive Evaluation and Comparison on Quintet Topologies

To ensure the correctness of the log-likelihood calculation and optimization, we compared the results of optimization obtained from CodonPhyML and CodeML on simulated quintet alignments (table 1, data set S1). The observed differences between optimized log likelihoods, total tree lengths, and model parameters were very small (supplementary table S1, Supplementary Material online), often nearly identical or differing only at the third decimal. Even the highest observed difference was less than 1 (for the log likelihood under model M5, tree 10 in supplementary table S1, Supplementary Material online). Log likelihoods obtained by CodonPhyML were always at least as good as values obtained by Codeml.

Table 1. Characterization of Simulated Data Sets.

ID	No. Taxa	Seq. Length (Codons)	Kappa	Omega	
				M0	M3
S1	5	500	2	1.5	{1.50, 0.01, 0.80} {0.1, 0.5, 0.4}
S2	{25,100}	{100,1,000,10,000}	2	1.5	{1.50, 0.01, 0.50} {0.3, 0.6, 0.1}
S3	{60,120,240, 480}	500	2	1.5	{1.50, 0.01, 0.50} {0.3, 0.6, 0.1}

NOTE.—Sequences were simulated using the program evolver from the PAML package, five replicates, once under M0 and once under M3, in both cases on trees obtained with birth rate 0.2, death rate 0.05, sampling fraction 0.5, mutation rate 0.5, and branch lengths drawn from an exponential distribution with a mean of 0.1 substitutions per branch.

Statistical Consistency Property

One of the advantages of the ML estimation is the statistical consistency, that is, under the true model, the ML estimates converge to their true values with increased amount of data (Rogers 1997). To test for this property, we applied CodonPhyML with either nearest neighbor interchange (NNI) or subtree pruning and regrafting (SPR) heuristic searches under the true model (M0 or M3) to infer phylogenies for simulated 25- and 100-taxon data sets of increasing sequence length (table 1, data set S2). The following parameters were monitored: total tree length (sum of branch lengths), ω (nonsynonymous to synonymous substitution rate ratio), κ (transition to transversion rate ratio), and the topology. Figure 1 summarizes the convergence properties for large 100-taxon data sets (see supplementary fig. S1, Supplementary Material online, for results with 25 taxa). In all cases, the ML estimates moved closer to their true values with the increase of sequence length. As expected, the estimation was more precise on average when the simpler model M0 was used for both simulation and analysis. Regardless the number of taxa or applied model, with 10,000 codons, the absolute error was typically $\ll 0.1$ for all parameters (with some small exceptions for tree length because of compounding of multiple errors for branches). For example, for model M3 with an NNI search, with the increase from 100 to 1,000 and to 10,000 codons, the normalized Robinson–Foulds (RF) distance between the estimated and the true tree decreased from 0.22 to 0.02 and to 0.00, respectively (see yellow bars in fig. 1*C* and *D*). Remarkably, for large 100-taxa data sets, even with 100 codons, the ML estimates were already close to their true values. These results are very encouraging and suggest that codon models possess the expected statistical properties for their further use in phylogenetic reconstruction from protein-coding DNA.

Evaluation of Options and Heuristics

To test the correctness of our implementation and to evaluate the performance of various heuristics—all described in detail in the Materials and Methods section—we analyzed simulated data of increasing number of taxa (data set S3 in

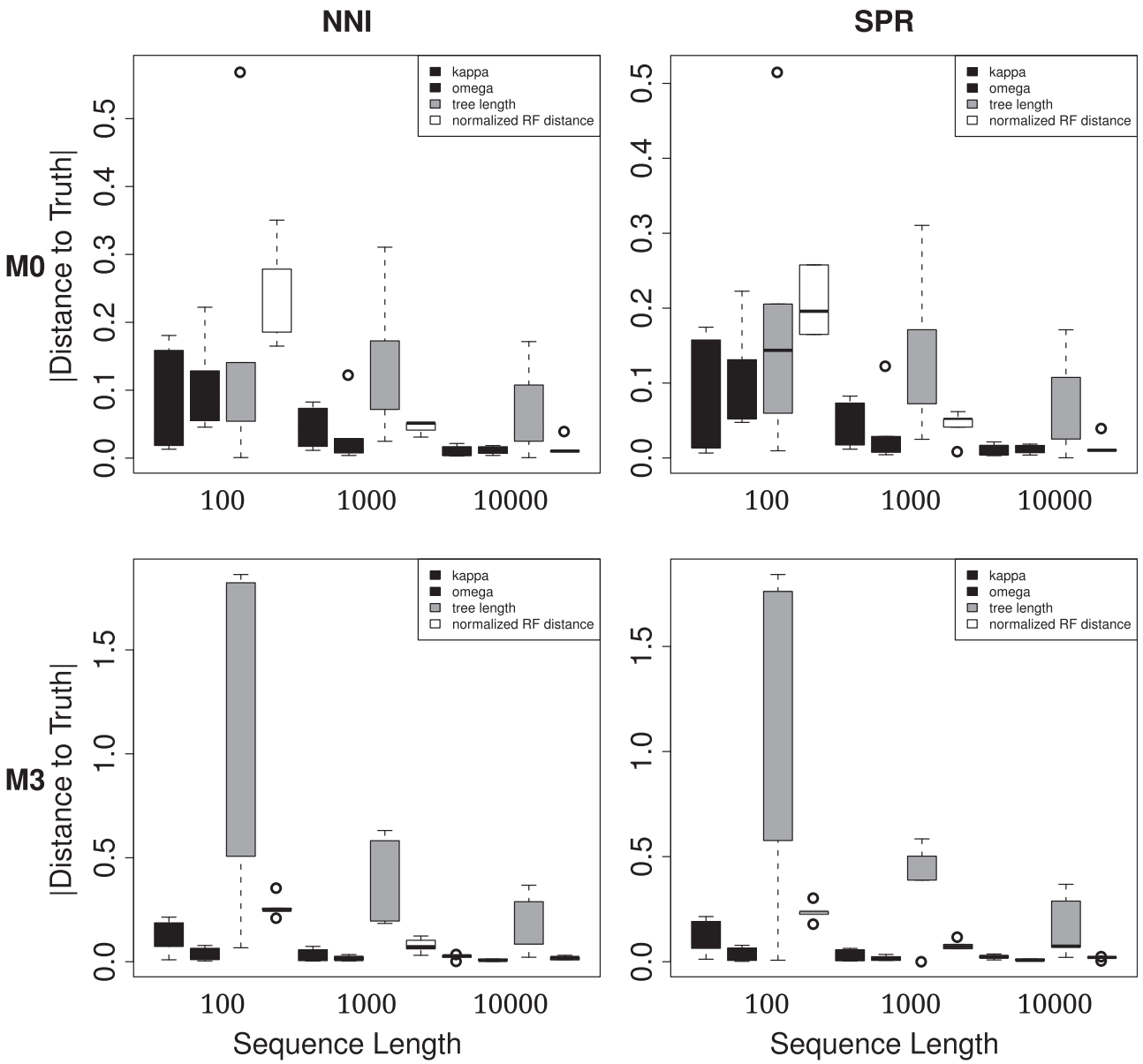

FIG. 1. Statistical consistency: asymptotic convergence of topology and parameter estimates for simulated 100-taxon data sets. With increasing sequence length, the estimates of model parameters and topology (under true model) become closer to the true values: Shown are the absolute differences between the estimates and the truth (as known from simulated data with 100 taxa, data set S2) under true models M0 or M3 and with search heuristics NNI and SPR.

table 1, 60–480 sequences of length 500 codons) under the correct models M0 and M3. Additionally to the log likelihood, we examined the parameters listed in the previous section and recorded the running times of selected combinations of heuristics. Supplementary table S2, Supplementary Material online, provides the specifications of the machines used for the computations.

The inclusion of the BLAS/LAPACK libraries and the parallelization with OpenMP were tested together. The likelihoods obtained by the combination were never significantly worse than the original implementation (supplementary figs. S2 and S5, Supplementary Material online). We achieved better speedups (execution time of the original implementation divided by the execution time when BLAS/LAPACK is used) for model M3 (2.07 with NNI and 1.77 with SPR tree rearrangement moves) compared with M0 (1.15 for NNI and 1.17 for SPR, see also supplementary fig. S3, Supplementary Material online). This was expected and is explained by the parallelization of the three rate categories of the general discrete distribution for ω. Better speedups have been observed with similar parallelization approaches when the input sequences were much longer as, for instance, in multigene alignment (Bader et al. 2006).

We devised two approximations to the ML tree. The first, referred to as +BioNJM, fully optimizes the branch lengths and substitution model parameters of the selected substitution model on a fixed BioNJ tree, without any further topology search. The second approximation, referred to as

+fixQ, performs only a rough optimization on the BioNJ tree and adds a tree search, without any further substitution model parameter optimization.

To evaluate the ability of tree search heuristics to obtain higher log likelihoods, we compared log-likelihood ranks, as in Guindon et al. (2010): For each data set, the log likelihoods obtained by different methods are sorted to determine their rank. The methods are then compared based on their average rank. Furthermore, for each data set, we picked the tree and parameter estimates from the method ranking first, that is, the best ML tree found (denoted "BestML"). BestML was used as a baseline to compare the tree and parameter estimates of the various heuristics. Because they performed differently under M0 and M3, we discuss the two separately.

Under M0, the SPR heuristic found trees with consistently but not significantly higher log likelihoods than NNI by at most 50 log-likelihood points (supplementary fig. S4, Supplementary Material online). The +Taylor option, which approximates the transition probability matrix by a second order Taylor series (but see Materials and Methods for a discussion), was tested in conjunction with NNI and performed significantly worse (at most 1,064 points on average) than the default implementation, whereas the option +fixQ, tested with NNI and SPR, leads to consistently but not significantly higher likelihoods than default by at most 12 and 1 points, respectively (supplementary fig. S2, Supplementary Material online). Accordingly, the model parameters estimated under various heuristics shows a similar difference to the ones estimated on BestML, with the exception of +Taylor (supplementary fig. S4, Supplementary Material online). As a consequence of the numerical problems (discussed earlier), its estimates of ω and κ clearly differed from the rest. For example, for the large data sets (240, 460 taxa), there were pathological cases, where κ was estimated to be 0. Surprisingly, the numerical instabilities did not affect the average topological distance. In contrast, +BioNJML leads to reasonable model parameter estimates but clearly different topologies than the ones obtained by search heuristics (supplementary fig. S4, Supplementary Material online). The likelihoods were as expected the worst but only by up to 117 points from the tree search heuristics. It is worth noting that the +BioNJML trees were on average longer than the other trees (supplementary fig. S4, Supplementary Material online). This observation is consistent with the known correlation between tree length and likelihood (Hordijk and Gascuel 2005).

Under M3, NNI and SPR converged on trees with similar log likelihoods. (supplementary fig. S5, Supplementary Material online). The +fixQ and +Taylor options often lead to lower log likelihoods compared with the original implementation (supplementary figs. S2 and S5, Supplementary Material online). Furthermore, both heuristics tended to produce biased estimates of ω and κ (compared with the estimates from BestML). In spite of the biases and similar to +Taylor under M0, these heuristics had a comparable topological distance to BestML as the trees obtained by the default settings (supplementary fig. S4, Supplementary Material online). As expected, +BioNJML was again clearly topologically different from the tree search heuristics. At the same time, it often led to higher log likelihoods and less biased estimates of ω and κ than +fixQ and +Taylor (supplementary fig. S4, Supplementary Material online). At first, this might look paradoxical; it is, however, explained by the fact that PhyML spends longer time (i.e., tries harder) optimizing substitution model parameters when tree search is disabled. The observation strongly suggests that the +fixQ heuristic could substantially be improved by a more thorough optimization of substitution model parameters before topology search.

The NNI heuristic was generally approximately 2–5 times faster than SPR under both substitution models. Under M0, NNI + Taylor and SPR + fixQ were slightly faster compared with the original implementation, whereas NNI + fixQ was slower, especially for large data sets (but in this case, it achieved consistently higher likelihoods). Under M3, +Taylor and +fixQ improved the running times of the original code, at times quite dramatically, but at the cost of less precise parameter estimates (as discussed above).

Performance on Real Data Examples

For real data (described in table 2), we observed that the algorithm typically converged faster under empirical codon models compared with either parametric or semiparametric models (e.g., see fig. 2). This trend was especially prominent for large and diverged data sets (e.g., R10–R11), where in addition, semiparametric models had much faster convergence compared with their parametric equivalents (e.g., see fig. 3). One reason why empirical and semiparametric models tend to converge faster compared with equivalent parametric models may be due to the improved model fit facilitated by their empirical elements inferred from large amounts of data. As these estimates capture some global patterns present in real data, the consequent improved model fit may be also reflected in the properties of the corresponding log-likelihood landscape.

As expected, convergence of phylogenetic search was achieved much faster for models with constant selection pressure or constant evolutionary rates over sites. On average, models with constant selection but variable rates (e.g., ECMK07 + F + Γ, M0 + Γ, and ECMK07 + Γ) took the longest to converge, with models of variable selective pressure (M3 and M5 variants) requiring less time (fig. 2). According to the AICc, the best-fitting models were often M3 and M5. Good convergence properties and model fit provided by M5 variants demonstrate clear potential of modeling variable selection using the Γ distribution compared with the general discrete distribution as in M3.

For each data set in table 2, we also compared the model fit for a set of DNA, amino acid, and codon models using the AICc scores based on the comparable log likelihoods (Seo and Kishino 2008, 2009) as computed by CodonPhyML. The phylogenies inferred under different models were compared using the RF distance.

For all data sets, using codon models instead of DNA or amino acid models resulted in topologically different

Table 2. Real Data Sets Used for Testing CodonPhyML.

ID	Protein-Coding Sequence	Organism Range	No. Taxa	Sequence Length (Nucleotides)	Phylog. Signal[a]	Average Branch Length	RF Distance[b]			Best-Fitting Model[c]
							Codon vs. AA	Codon vs. NT	AA vs. NT	
R1	Caudal-like protein activation region (PF04731)	Metazoa	8	567	0.42	0.38	0.00	0.15	0.15	LG + F + Γ
R2	PetM family of cytochrome b6f complex subunit 7 (PF08041)	Bacteria and Eukaryota	12	135	0.26	0.39	0.10	0.67	0.67	WAG + F + Γ, LG + F + Γ
R3	DKCLD (NUC011) domain (PF08068)	Archaea and Eukaryota	18	177	0.24	0.24	0.42	0.67	0.67	LG + F + Γ
R4	6-phosphofructo-2-kinase (PF01591)	Eukaryota	11	726	0.17	0.57	0.00	0.32	0.32	LG + F + Γ
R5	Intermediate filament head (DNA binding) region (PF04732)	Eukaryota	30	315	0.38	0.46	0.42	0.39	0.49	WAG + F + Γ, LG + F + Γ
R6	Zinc finger, ZZ type (PF00569)	Eukaryota	8	138	0.24	0.62	0.15	0.77	0.77	WAG + F, WAG + F + Γ
R7	Protein of unknown function (PF08004)	Archaea	6	393	0.13	1.29	0.00	0.00	0.00	ECM07 + ω_{M0} + κ + Γ
R8	Repeated sequence found in lipoprotein LPP (PF04728)	Bacteria	9	33	0.41	0.32	0.53	0.13	0.53	ECM07 + ω_{M0} + κ + Γ, ECM07 + ω_{M5} + κ + Γ
R9	Myogenic basic domain (PF01586)	Metazoa	6	345	0.58	0.46	0.00	0.00	0.00	WAG + F + Γ, LG + F + Γ
R10	7 transmembrane receptor, rhodopsin family (PF00001)	Metazoa	64	819	0.23	0.61	0.58	0.45	0.64	LG + F + Γ
R11	Homeobox domain (PF00046)	Eukaryota	179	174	0.34	0.24	0.62	0.61	0.73	LG + Γ
R12	Protein of unknown function (PF01973)	Archaea and Bacteria	23	522	0.30	0.47	0.42	0.23	0.42	WAG + F + Γ, LG + F + Γ
R13	EPH receptor A4	Mammalia	21	3,141	0.18	0.09	0.51	0.05	0.56	M0 + Γ
R14	Transcription factor 20	Mammalia	21	6,081	0.21	0.09	0.41	0.10	0.41	M3
R15	WD repeat domain 23	Mammalia	21	1,677	0.20	0.07	0.51	0.31	0.56	M3
R16	Tu translation elongation factor, mitochondrial	Mammalia	21	1,377	0.22	0.11	0.51	0.10	0.46	M3
R17	Zinc finger protein 641	Mammalia	21	1,323	0.21	0.10	0.62	0.10	0.62	M3
R18	Nucleoporin like 2	Mammalia	21	1,380	0.22	0.15	0.26	0.10	0.31	M5
R19	Gm527	Mammalia	21	930	0.16	0.05	0.72	0.31	0.72	M0 + Γ
R20	Integrin β11 binding protein 1	Mammalia	21	600	0.27	0.11	0.87	0.26	0.87	M3
R21	GALA (type III effectors)[d]	Bacterium	426	81	0.84	0.16	0.41	0.30	0.39	ECM07 + ω_{M5} + κ
R22	Lady bird early (lbe)[e]	*Drosophila*	73	429	0.37	0.005	0.97	0.45	0.97	M0 + Γ
R23	Lady bird early (lbl)[e]	*Drosophila*	72	420	0.31	0.002	0.45	0.38	0.10	M0 + Γ

NOTE.—Data sets R1–R12 are from PANDIT (Whelan et al. 2006); detailed annotations available from PANDITplus (Dimitrieva and Anisimova 2010); in parentheses shown are their respective Pfam IDs (Punta et al. 2012). Data sets R13–R20 are OMA orthologs (Altenhoff et al. 2011).

[a]The phylogenetic signal is the proportion of the total tree length that is taken up by internal branches (Phillips et al. 2001).

[b]Pairwise normalized RF distance between phylogenies inferred with best-fitting codon, amino acid (AA), and nucleotide (NT) models.

[c]Within two units to the minimum AICc.

[d]Data from Kajava et al. (2008).

[e]Data from Balakirev et al. (2011).

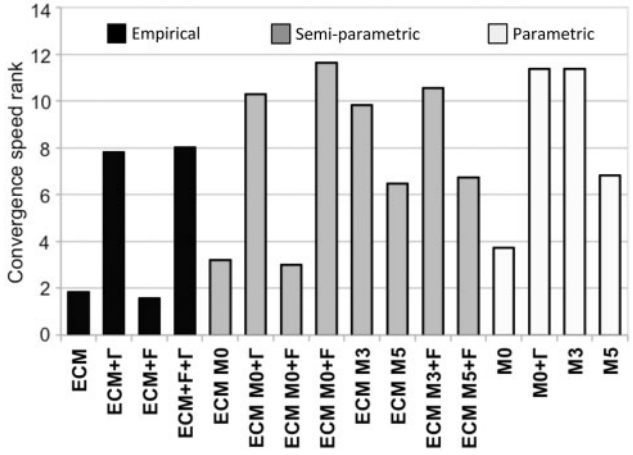

FIG. 2. Speed ranks between empirical, semiparametric, and parametric models on real data. Ranks are computed as an average over 23 real data sets (table 2). Model notations: ECM, empirical codon model of Kosiol et al. (2007); +F, frequencies are estimated empirically from data at hand; +Γ, with among-site rate variation; M0, one ω-ratio parametric model (Goldman and Yang 1994; Nielsen and Yang 1998); M3, parametric model with variable selection over sites modeled by general discrete distribution with three classes (Yang et al. 2000); M5, parametric model with variable selection over sites modeled by the discretized Γ distribution (Yang et al. 2000); ECM MX (where $X = 0$, 3, or 5), semiempirical model with additional parameters as in model MX.

phylogenies as measured by RF distance (e.g., fig. 3 and table 2). Reconstructed phylogenies also varied when different codon models were used for tree inference but to a lesser extent. However, because for a real data set the true phylogenetic relationship is typically unknown, we can judge the performance of a model only by its fit to data, for example, as measured by the AICc score (but see Dessimoz and Gil [2010] for new ideas). For many data sets, codon models provided much better fit to data compared with amino acid or DNA models (data sets R7–8, R13–23 in table 2; e.g., see fig. 3*B* and *C*). For highly divergent data sets (where synonymous substitutions already reached saturation), most of the time, there was no advantage using codon models (fig. 3*A*). Data sets R7 and R8 in table 2 appear to be exceptions to this rule, although this may also be due to estimation artifacts.

In addition, we recorded a log-likelihood improvement given by the codon model M3 with respect to trees estimated using either the amino acid model LG + Γ or DNA model HKY + Γ (supplementary table S3, Supplementary Material online). To do this, we evaluated the log likelihoods using M3: first on the fixed trees obtained by LG + Γ and HKY + Γ and then with a tree search. In most cases, a substantial improvement was observed (supplementary table S3, Supplementary Material online). This shows that candidate trees obtained by amino acid and DNA models cannot replace a tree search under a codon model.

CodonPhyML includes fast and accurate branch support methods, aLRT, SH-aLRT, and aBayes, shown to be a fast and accurate alternative for ML bootstrap (Anisimova et al. 2011). Here, we compared aLRT values (Anisimova and Gascuel 2006) for internal branches (i.e., nontrivial tree splits) inferred with codon, amino acid, and DNA models: M3, LG + Γ, and HKY + Γ. On average, branch supports computed with M3 were slightly higher compared with those computed under LG + Γ and HKY + Γ (supplementary fig. S6, Supplementary Material online). For "non-common" branches (inferred with one model but not another), the supports were low: none of aLRT values showed significant support. Correlation between pairs of models (M3 vs. HKY + Γ and M3 vs. LG + Γ) was strong, particularly for common splits (inferred with both models), see supplementary figure S6, Supplementary Material online.

Finally, we investigated how the choice of the starting tree affects the initial values of the log likelihood on the initial stage of the optimization. Presumably, starting trees leading to higher initial likelihoods should be more successful at navigating the search into a "good" parameter space (e.g., Stamatakis et al. 2005). For each real data set R1–23, we compared four ways of building initial tree from codon alignments: by maximum parsimony (MP); by the BioNJ algorithm based on the equal rate (Jukes–Cantor-like) codon model (N61) (Neyman 1971); or based on the empirical models ECMS05 or ECMK07. Once a starting tree was produced, we performed a full optimization of branch lengths and model parameters on this fixed starting tree under models GYM0 and under GYECMK07 + ω + κ + F. Using empirical models, ECMS05 and ECMK07 most often resulted in the best starting tree with respect to its optimized log likelihood (supplementary tables S4 and S5, Supplementary Material online). MP and N61 result in the best starting tree only rarely. We, therefore, recommend using the empirical models for the search of a starting tree.

Conclusions and Future Perspectives

We have presented CodonPhyML, which implements the largest variety of codon models available in any published package for ML tree inference today. It extends the established code of PhyML 3.0 (Guindon et al. 2010), supports multicore processors through OpenMP, allows model selection across the amino acid, nucleotide, or codon data abstraction, and offers new heuristics for further exploration.

The correctness of likelihood calculations and the performance of the program have been assessed on simulated and diverse real data sets. On real data, we found that codon models often provide a better fit than amino acid and nucleotide models and, particularly important, that they generate a qualitatively different class of tree topologies. Indeed, because selection on proteins (negative or positive) is a major force shaping protein-coding DNA, codon models that explicitly include selection pressure should provide qualitatively different trees in their distribution of topological shapes and branch lengths. CodonPhyML should enable us to test this premise. More generally, the availability of CodonPhyML paves the way for many other phylogenetic studies, such as exploring the utility codon models for topology inference in a systematic way.

Future work will include the development of heuristics tailored to particularities of codon models, for instance, by

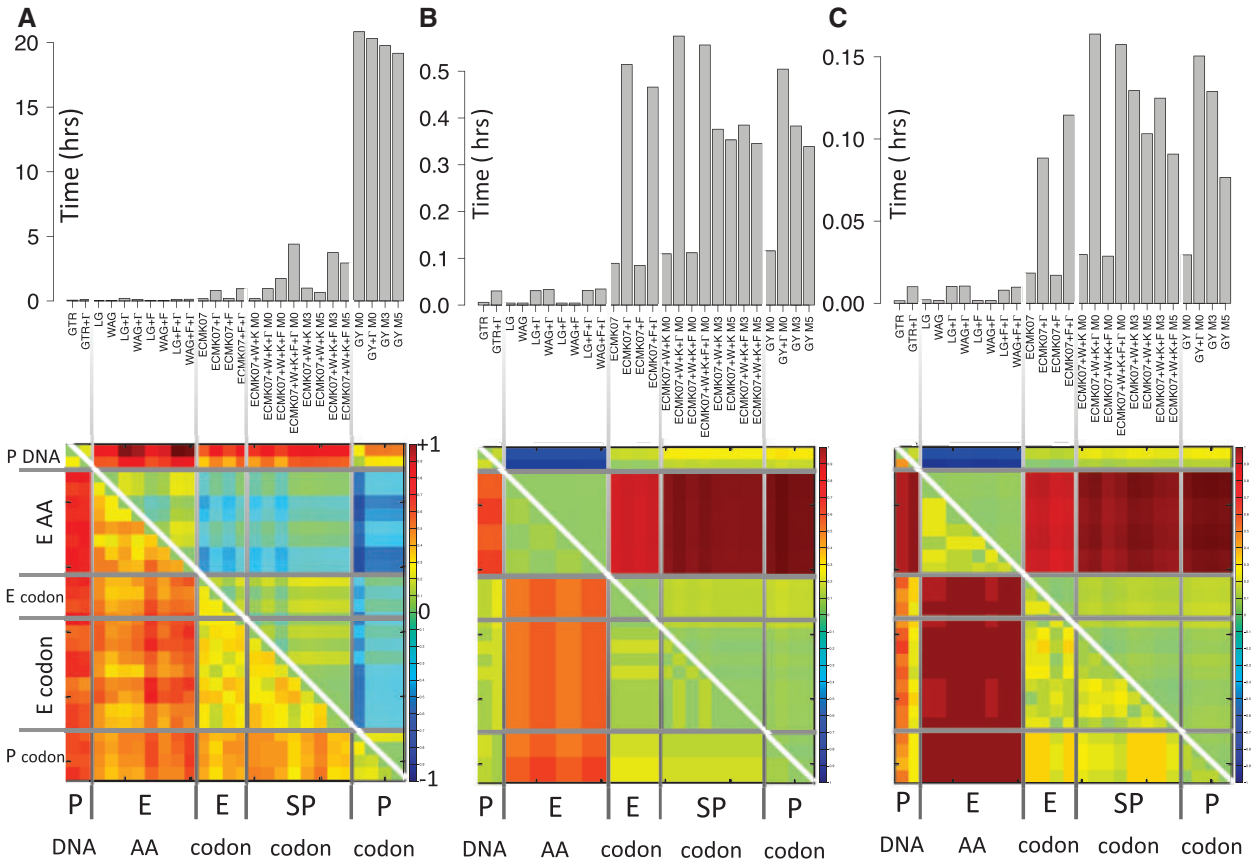

Fig. 3. Comparison of running time, tree topology, and model fit (AICc) between various empirical, semiparametric, and parametric models on three real data sets. Seven-transmembrane receptor from the rhodopsin family of metazoa (*A*), the mammalian EPH receptor A4 (*B*), and the mammalian Integrin β1 binding protein 1 (*C*). The top part of each subfigure provides the CPU time till convergence for each of the evaluated models using CodonPhyML with default parameters, in particular, without multicore support. The lower part of each subfigure shows a color-coded similarity matrix representing the normalized AICc score differences (in the upper triangle) and the normalized RF distance (lower triangle) between any pair of models tested. Both top and lower parts of subfigures utilize the same model order and exact mapping of model locations to the matrix. For readability, gray lines define partitions between different types of models: P, parametric; E, empirical; and SP, semiparametric. The comparison of models is performed from the model on the vertical scale to the model on the horizontal scale. In the upper triangle: For AICc differences, color ranges from blue (<0 or "better fit") to green (=0 or "same fit") and to red (>0 or "worse fit"). For example, amino acid models fit data better than codon models for data in (*A*) but worse for data in (*B*) and (*C*). For RF distance, because truth is unknown, the notions of "better" or "worse" cannot be defined. Thus, only the extent of differences in RF distance is shown in the lower triangle: color ranges from green (=0 or "same topology") to red (>0 or "very different topology"). For example, the choice of a codon or an amino acid model leads to significant differences in the inferred ML topology.

economizing the number of reoptimizations and exponentiations of the rate matrix during tree search (taken to the extreme by the +fixQ option presented here). Further, the evaluation of heuristics by simulations can only be regarded a first sanity check and is not necessarily generalizable to real data. Indeed, it has been observed that their behavior may be different under the two types of data, owing to the rougher log-likelihood landscapes implied by real sequences (Stamatakis et al. 2005; Guindon et al. 2010). In the future releases of CodonPhyML, we intend to include additional options: 1) to automate model selection across the three levels of data abstraction (amino acids, DNA, and codons) and 2) to facilitate a statistical comparison and ranking of candidate phylogenies obtained after different heuristic searches or under different models (e.g., the SH test, Shimodaira and Hasegawa 1999). Finally, we are currently investigating a new heuristic for topology search, based on information theoretic concepts that have been successfully used to unveil evolving topological patterns in arbitrary graph structures (Pfitzner et al. 2012).

Materials and Methods

Implemented Substitution Models

Modeling Stationary Frequencies

A Markov model of substitution is defined by a generator matrix Q describing the instantaneous substitution rates between all possible codon pairs. For a time-reversible homogenous model, each instantaneous rate of change between codons i and j can be written as a product of an exchangeability rate s_{ij} and a stationary frequency of a target codon π_j (e.g., Yang 2006, p. 33–34, 41). "GY-type" models explicitly define an instantaneous substitution rate as proportional to a target codon frequency, based on the model introduced by Goldman and Yang (1994). "MG-type" models define an

instantaneous substitution rate as proportional to the frequency of a target nucleotide, based on the model introduced by Muse and Gaut (1994). Recently, Yap et al. (2010) suggested that a better way might be modeling instantaneous frequencies proportional to the target nucleotide frequencies conditional on nucleotides observed within a codon ("Yap-type" models). This may help to partially account for context-dependent biases such as CpG bias. MG- and Yap-type models imply that a frequency of a codon can be estimated as a product of frequencies of its composing nucleotides. Moreover, it has been shown that for MG- and Yap-type models, the codon frequencies defined in this way describe the stationary distribution under the model (Lindsay et al. 2008; Rodrigue et al. 2008).

Estimating Codon Frequencies

To save the optimization time, codon frequencies are often estimated empirically from data rather than by ML. CodonPhyML provides all standard and more recent options for estimating the codon frequencies:

FxCodons: One frequency for each codon as observed in data.
F1x4: Equal nucleotide frequencies at all codon positions.
F3x4: Individual nucleotide frequencies for three codon positions.
CF3x4: Corrected version of F3x4 to account for absence of stop codons (Kosakovsky Pond et al. 2010), because by F3x4 definition equal nucleotide frequencies do not induce equal codon frequencies and vice versa.
USR: User defined (equal frequencies can be enforced with this option).
ML: By ML.

Types of Codon Models

"Parametric models" are described by a *Q*-matrix whose entries are fully defined by a set of parameters each representing certain biases of molecular sequences. For example, for a reversible codon model, the typical parameters would be codon or nucleotide frequencies and the exchangeabilities between different codons or nucleotides, described as a product of various relative rates such as transition and transversion rates, and synonymous and nonsynonymous substitution rates (d_S and d_N, respectively). One common parameter included in all current parametric codon models is the $\omega = d_N/d_S$ ratio, which is the measure of selective pressure at the protein level. Estimates $\omega > 1$ are suggestive of positive selection, whereas estimates $\omega < 1$ suggest negative (or purifying) selection. In addition to model M0, which assumes constant ω over sites, CodonPhyML implements models of variable selective pressure (Yang et al. 2000): 1) model M3 with the general discrete distribution for ω and 2) model M5 that describes ω-variation over sites using the Γ distribution with two parameters α and β.

In contrast, purely "empirical" models are described by *Q*-matrices whose entries were estimated from large amounts of data, with no free parameters. CodonPhyML implements empirical models ECMS05 based on pairwise analysis (Schneider et al. 2005) and ECMK07 based on multiple sequence alignments and optimization by expectation maximization (Kosiol et al. 2007). A user-defined empirical model can be specified using an additional input file.

Empirical models are designed to reflect the global patterns but are incapable of capturing biases specific to individual genes. CodonPhyML implements the "+F" option to account for gene-specific content bias by using character frequencies as observed in a gene under scrutiny (first proposed for empirical amino acid models by Goldman and Whelan 2002). In addition, "semiempirical" models enable estimation of important parameters, such as ω and κ, while increasing model fit through combining free parameters with empirically estimated elements, typically the empirical codon exchangeability values (Doron-Faigenboim and Pupko 2007; Kosiol et al. 2007). CodonPhyML includes semiempirical models with ω and the transition/transversion ratio $\kappa(i, j)$ as a function of counts of transitions i, and transversions j (Kosiol et al. 2007), because multiple nucleotide substitutions within one codon are allowed in empirical and semiempirical models. Note also that selective pressure estimates of the ω ratio from parametric models is not directly comparable to the ω estimates in semiparametric models. This is because the empirical exchangeabilities already take into account the global selection pressure present in the training data set. A correction is necessary to obtain a comparable estimate of selection pressure at the protein from a semiempirical model (Kosiol et al. 2007). Thus, for semiparametric models, both uncorrected and corrected estimates of the ω ratio are provided in the CodonPhyML output.

Finally, models based on PCA are also available in the current implementation (Zoller and Schneider 2010).

Site-Rate Variation

Similar to DNA and amino acid models, site-rate variation is modeled using the Γ distribution with parameters $\alpha = \beta$, so that the mean rate is 1 expected substitution per site. In the current implementation, the two options of using the Γ distribution to model selective pressure and the variation of rates over sites cannot be combined and can only be used separately due to this model being too complex for the purposes of phylogeny inference—the primary goal of this package.

Implementation Details of Tree Search and Parameter Optimization

CodonPhyML is based on the original source code of PhyML 3.0 (Guindon et al. 2010). The PhyML algorithm starts from a BioNJ (Gascuel 1997) or an MP tree and improves it using the tree rearrangement moves simultaneous NNI or/and SPR. Branch lengths are optimized locally in conjunction with the topological rearrangements. In the SPR variant, additionally, all the branch lengths are adjusted after each round of rearrangements. Periodically, the free parameters of the substitution model (defining the relative substitution rates of the Markov model and the shape parameter of the Γ distribution) are reoptimized.

The optimizations are carried out one parameter at a time, with all the others fixed, using an iterative line search

algorithm—specifically, methods based on the Brent and golden section search algorithm.

During the optimizations, the likelihood function is repeatedly evaluated, requiring the computation of the transition probability matrix $P(t)$ for different values of Q and divergence time t. PhyML computes $P(t)$ numerically from the eigenvalues and the eigenvectors of Q:

$$P(t) = \exp(Qt) = U \exp(\Lambda t)U^{-1} = U \operatorname{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \ldots, \exp(\lambda_C t)\}U^{-1},$$

where $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_C\}$ is a diagonal matrix of the eigenvalues λ_i of Q, and U is made of its eigenvectors. The decomposition is carried out with an $O(n^3)$ algorithm described in (Wilkinson and Reinsch 1971, p. 197). It is applied whenever Q is updated and reused during tree optimization for the computation of $P(t)$, which for each branch length t is roughly as expensive as one matrix–matrix multiplication. The computation of partial likelihoods, ancestral states, and the decomposition of Q is the most time-consuming parts of the PhyML algorithm (supplementary table S7, Supplementary Material online).

The main objective of CodonPhyML is to offer a variety of state of the art codon models for tree search, thereby relying on PhyML's heuristics. Opting for faster running times, however, we have extended the original code by a number of options:

- +BLAS: In PhyML, the linear algebra operations rely on the standard libraries of the C programming language. We have benchmarked their running times against the corresponding routines from the de facto standard libraries BLAS (Blackford et al. 2002) and LAPACK (Anderson et al. 1990) (supplementary table S8, Supplementary Material online). When the code was compiled without any optimization, the BLAS/LAPACK routines were considerably faster than original PhyML routines. When CodonPhyML was compiled at the highest optimization level (−O3, supplementary table S8, Supplementary Material online), only three out of seven routines were faster. Compiling with −O3 never lead to slower code than with no optimization. Consequently, we have integrated the three BLAS/LAPACK routines that were found to be faster with −O3. We advise users compiling CodonPhyML with −O3 and BLAS/LAPACK.
- +OMP: PhyML spends a considerable time in for-loops over the length of the multiple sequence input alignment to assemble the likelihood function (supplementary table S7, Supplementary Material online). The loops do not have dependencies between iterations, due to the assumption—inherent in all implemented models—that substitutions are independent among codon sites. We have exploited such loop-level parallelism with OpenMP (Dagum and Menon 1998), taking advantage of today's multicore computers. Furthermore, we have parallelized the optimization of substitution model parameters in presence of multiple discrete ω or Γ rate categories (for a similar approach see Bader et al. 2006).
- +BFGS: PhyML's strategy to optimize one parameter at the time is inefficient for strongly correlated parameters (e.g., Yang 2006, p. 134–136) such as the ω categories in model M3. We have added the multidimensional optimization algorithm BFGS (a quasi-Newton method, e.g., see Gill et al. [1981]) reusing the implementation from PAML (Yang 2007) and following the recommendation by Yang (2000) to optimize Q by BFGS and the branch lengths by a line search algorithm.
- +Taylor: With this option, the transition probability matrix is approximated by the second order Taylor series $P(t) \approx I + Qt + (Qt)^2/2$, where I denotes the identity matrix. This option can be used during the adjustments of the rate-matrix Q, so to avoid the repeated diagonalization of Q in favor of one matrix–matrix multiplication and two additions. For the tree search step, we still use diagonalization because computing $P(t)$ by $U \exp(\Lambda t)\ U^{-1}$ for each branch-length t is roughly equally time efficient but more precise than the second order Taylor series, which introduces numerical instability leading to the loss of precision (e.g., Yang 2006, p. 68–69). This problem could be avoided by the scaling and squaring method, which reduces the norm of Q by exploiting the relation $\exp(Qt) = (\exp(Qt/m))^m$, where $m = 2^k$ is set to a power of two, such that the mth power can be computed by squaring $\exp(Qt/m)$ for k times. Yang (2006, p. 68) recommended to set k to a value between 5 and 10 for small distances (and to even bigger numbers for large distances). However, k squarings are considerably slower than diagonalization.
- +Pade: This option triggers a Pade approximation with scaling and squaring for the exponentiation of the rate matrix (Higham 2005) recommended by Schranz et al. (2008) for nonreversible dinucleotide and codon models. The $O(n^3)$ algorithm implemented here corresponds to MATLAB's function expm (Moler and Van Loan 2003). Analogous to the +Taylor option, the Pade approximation is only triggered during the adjustments of the rate matrix.
- +FixQ: This option follows the rule-of-thumb "estimating using a reasonably good topology leads to reasonable parameter estimates" (Yang 1995). All the substitution model parameters are fitted once on the fixed initial topology and then kept constant during tree search. Upon convergence of the topology search, all parameters are reoptimized on the final topology. A similar approach is provided in other ML packages, for instance, RAxML (Stamatakis et al. 2005) or GARLI (Zwickl 2006). In contrast to these implementations, however, +FixQ only carries out a rough optimization.
- +CST: By default, the initial topology in PhyML is reconstructed with BioNJ, which uses a matrix of pairwise evolutionary distances. The matrix is estimated based on a nucleotide or amino acid substitution model, depending on the model selected for tree search. With the +CST option, the starting tree can be computed by either MP at the codon level or using BioNJ on distances estimated by

ML under one of three codon models: ECMS05, ECMK07, and N61 (all instantaneous rates equal).

- +BioNJML: As a baseline, we devised a very rough approximation to the ML tree as follows. A BioNJ tree based on pairwise ML distances under ECM07 is reconstructed. Then, on the resulting topology, parameters of a selected substitution model and branch lengths are fully optimized. Note that this approach differs from +fixQ in two ways: 1) the parameter optimization is carried out more thoroughly and 2) no topological search is performed.

Simulated Data Sets

To examine the performance of CodonPhyML, sequence data were generated using the program Evolver from the PAML package (Yang 2007) under M0, M3, or M5 and analyzed with the correct model. The simulation scenarios are listed in table 1. We designed three types of data sets. Simulated data of type 1 (S1) consisted of quintets of sequences and were used to compare parameter estimates obtained with CodonPhyML and Codeml of PAML on all the 15 possible topologies relating five sequences. Data set S2 had very long sequences (up to 10,000 codons) and was generated to test the correctness of CodonPhyML through verifying the consistency of ML estimators. The third data set (S3) was simulated to explore the provided options with respect to the number of sequences, ranging from 60 to 480.

Real Data Sets

We selected 23 highly diverse real data sets (table 2) and analyzed them with 2 amino acid, 2 nucleotide, and 16 codon models. The purpose of these analyses was to compare the fit of various models and the topological agreement or disagreement of the resulting trees. Model fit was assessed with the AIC (Akaike 1973, 1974) and BIC (Schwarz 1978) after casting the log likelihood of the nucleotide and amino acid models to a codon model (Seo and Kishino 2008, 2009). Topological disagreement between two trees was measured by the Robinson–Foulds distance (Robinson and Foulds 1981).

Availability

CodonPhyML is an open source project, written in C. The source code and executables for Linux, Mac OS X, and Windows (compiled with –O3 and BLAS/LAPACK) can be downloaded together with a user manual and data examples from: http://sourceforge.net/projects/codonphyml (last accessed March 13, 2013). We encourage user feedback to help us to improve the software.

Supplementary Material

Supplementary tables S1–S8 and figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Stéphane Guindon for support with PhyML code and Ziheng Yang, Olivier Gascuel, and Gaston Gonnet for discussions on code optimization strategies. Part of this work was accomplished toward the master thesis, for which M.S.Z. was awarded the ETH Zürich medal. This work was supported by the Swiss Federal Commission for Scholarships for Foreign Students grant 2006.0091 to M.S.Z. and by the Swiss National Science Foundation (SNF) grants 31003A_127325/1 to M.A., PBEZP2_140129 to M.G., and CR12I1_125298 to M.S.Z.

References

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Control.* 19:716–723.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39:D289–D294.

Anderson E, Bai Z, Dongarra J, Greenbaum A, McKenney A, Croz JD, Hammerling S, Demmel J, Bischof C, Sorensen D. 1990. LAPACK: a portable linear algebra library for high-performance computers. Proceedings of the 1990 ACM/IEEE Conference on Supercomputing. New York: IEEE Computer Society Press. p. 2–11.

Anisimova M. 2012. Parametric models of codon evolution. In: Cannarozzi G, Schneider A, editors. Codon evolution: mechanisms and models. Oxford: Oxford University Press. p. 12–33.

Anisimova M, Gascuel O. 2006. Approximate likelihood ratio test for branches: a fast, accurate and powerful alternative. *Syst Biol.* 55: 539–552.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol.* 60:685–699.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol.* 26:255–271.

Bader DA, Roshan U, Stamatakis A. 2006. Computational grand challenges in assembling the tree of life: problems and solutions. *Adv Comput.* 68:127–176.

Balakirev ES, Anisimova M, Ayala FJ. 2011. Complex interplay of evolutionary forces in the *ladybird* homeobox genes of *Drosophila melanogaster*. *PLoS One* 6:e22613.

Blackford LS, Demmel J, Dongarra J, et al. (13 co-authors). 2002. An updated set of basic linear algebra subprograms (BLAS). *ACM Trans Math Soft.* 28:135–151.

Dagum L, Menon R. 1998. OpenMP: an industry standard API for shared-memory programming. *IEEE Comput Sci Eng.* 5:46–55.

Dessimoz C, Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.* 11:R37.

Dimitrieva S, Anisimova M. 2010. PANDITplus: toward better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources. *Trends Evol Biol.* 2:e1.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. *Mol Biol Evol.* 24:388–397.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14:685–695.

Gill PE, Murray W, Wright MH. 1981. Practical optimization. London: Academic Press.

Goldman N, Whelan S. 2002. A novel use of equilibrium frequencies in models of sequence evolution. *Mol Biol Evol.* 19:1821–1831.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.

Guindon S, Dufayard JF, 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.
Higham NJ. 2005. The scaling and squaring method for the matrix exponential revisited. *SIAM J Matrix Anal Appl.* 26:1179–1193.
Hordijk W, Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* 21:4338–4347.
Kajava AV, Anisimova M, Peeters N. 2008. Origin and evolution of GALA-LRR, a new member of the CC-LRR subfamily: from plants to bacteria? *PLoS One* 3:e1694.
Kosakovsky Pond S, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One* 5:e11230.
Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol.* 24:1464–1479.
Lindsay H, Yap VB, Ying H, Huttley GA. 2008. Pitfalls of the most commonly used models of context dependent substitution. *Biol Direct.* 3:52.
Minh BQ, Vinh LS, von Haeseler A, Schmidt HA. 2005. pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics* 21:3794–3796.
Moler C, Van Loan C. 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* 45:3–49.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11: 715–724.
Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta SS, Yackel J, editors. Statistical decision theory and related topics. New York: Academic Press. p. 1–27.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.
Pfitzner R, Scholtes I, Garas A, Tessone CJ, Schweitzer F. 2012. Betweenness preference: quantifying correlations in the topological dynamics of temporal networks. arXiv:1208.0588 [physics.soc-ph].
Phillips MJ, Lin YH, Harrison GL, Penny D. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. *Proc Biol Sci.* 268:1533–1538.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Res.* 40: D290–D301.
Ren F, Tanaka H, Yang Z. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst Biol.* 54:808–818.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci.* 53:131–147.
Rodrigue N, Lartillot N, Philippe H. 2008. Bayesian comparisons of codon substitution models. *Genetics* 180:1579–1591.
Rogers JS. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst Biol.* 46: 354–357.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.
Schmidt HA, von Haeseler A. 2009. Phylogenetic inference using maximum likelihood methods. In: Lemey P, Salemi M, Vandamme AM, editors. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press. p. 181–209.
Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. *BMC Bioinformatics* 6:134.
Schwarz G. 1978. Estimating the dimension of a model. *Ann Statist.* 6: 461–464.
Schranz HW, Yap VB, Easteal S, Knight R, Huttley GA. 2008. Pathological rate matrices: from primates to pathogens. *BMC Bioinformatics* 9: 550.
Seo TK, Kishino H. 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Syst Biol.* 57:367–377.
Seo TK, Kishino H. 2009. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. *Syst Biol.* 58:199–210.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16:1114–1116.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–463.
Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.* 34: D327–D331.
Wilkinson JH, Reinsch C. 1971. Linear algebra. In: Bauer FL, editor. Handbook for automatic computation. Berlin (Germany): Springer. p. 197–199.
Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139:993–1005.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol.* 51:423–432.
Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
Yap VB, Lindsay H, Easteal S, Huttley G. 2010. Estimates of the effect of natural selection on protein-coding content. *Mol Biol Evol.* 27: 726–734.
Zoller S, Schneider A. 2010. Empirical analysis of the most relevant parameters of codon substitution models. *J Mol Evol.* 70: 605–612.
Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Austin (TX): University of Texas.