SMS: Smart Model Selection in PhyML

Vincent Lefort,[1] Jean-Emmanuel Longueville,[1] and Olivier Gascuel*[,1,2]

[1]Institut de Biologie Computationnelle, LIRMM, UMR 5506 - CNRS et Université de Montpellier, Montpellier, France

[2]Unité de Bioinformatique Evolutive, C3BI, USR 3756 - Institut Pasteur et CNRS, Paris, France

***Corresponding author:** E-mail: olivier.gascuel@pasteur.fr.

Associate editor: Tal Pupko

Abstract

Model selection using likelihood-based criteria (e.g., AIC) is one of the first steps in phylogenetic analysis. One must select both a substitution matrix and a model for rates across sites. A simple method is to test all combinations and select the best one. We describe heuristics to avoid these extensive calculations. Runtime is divided by ~2 with results remaining nearly the same, and the method performs well compared with ProtTest and jModelTest2. Our software, "Smart Model Selection" (SMS), is implemented in the PhyML environment and available using two interfaces: command-line (to be integrated in pipelines) and a web server (http://www.atgc-montpellier.fr/phyml-sms/).

Key words: **model selection, heuristic procedure, AIC and BIC criteria, web server, PhyML.**

Brief Communication

Current phylogenetic programs provide users with a wide variety of models to represent both the variability of rates across sites (RAS) and the substitution process. With proteins, a large number of substitution matrices have been inferred for various protein types (e.g., membrane and mitochondrial) and origins (e.g., mammals and viruses). To select among these many models, statistical criteria (e.g., AIC [Akaike 1973] and BIC [Schwarz 1978]) are used to find the best likelihood/model-complexity tradeoff. A simple, standard approach is to test all models and then select the best one. This forms the basis of widely used, user-friendly software programs such as ProtTest for proteins (Abascal et al. 2005).

Here, we introduce a new software tool to achieve this task: SMS, which stands for "Smart Model Selection." This tool is very simple to use, as SMS is fully integrated into the PhyML web server (fig. 1*a* and *b*; Guindon et al. 2010). SMS can also be used as a standalone application and is freely available for download (http://www.atgc-montpellier.fr/sms/). SMS uses heuristic strategies to avoid testing all models and options. These strategies are partly inspired by Posada and Crandall (1998) and Darriba et al. (2012). Notably, the latter proposed a fast method called "model filtering" to focus on the most promising substitution matrices for DNA, whereas our heuristic for proteins also ranks the matrices based on their proximity to the data being analyzed. Moreover, SMS simplifies some calculations to save computing time. This is especially relevant in a pipeline context for running extensive phylogenetic analyses, for example, to study protein families. Below, we summarize the main features of SMS and its performance compared with the exhaustive approach, as well as to jModelTest2 (Darriba et al. 2012) and ProtTest. Complete details on algorithms, benchmark data sets, and comparison results are available in Supplementary Material.

With proteins, all substitution matrices available in PhyML are also available in SMS (fig. 1*c*, 17 matrices). Moreover, users can add their own matrices. All matrices can be used with the option +F (amino-acid frequencies are estimated from the data) and −F (preestimated frequencies). SMS only has two options to model RAS: $+\Gamma$ (gamma distribution) and $+\Gamma+$I (one class of invariant sites is added). Extensive comparisons (supplementary table S4, Supplementary Material online) with 500 representative protein data sets showed that the +I option alone is rarely selected (1/500 with AIC, 4/500 with BIC), and the same holds for the $-\Gamma-$I or "none" option (3/500 with AIC, 4/500 with BIC). Protein multiple sequence alignments (MSAs) usually have few constant sites (median proportion in our data sets ≈ 3%), and we expect a high variability of site rates caused by the variability of functional and structural constraints acting along protein sequences. These results and choices are thus biologically consistent. SMS has a total of 17 (matrices) x 2 (+F/−F) x 2 (RAS) = 68 models. On average, SMS computes the likelihood value for only ~30 models. Computing time is divided by ~2 as compared with exhaustive calculations using the same models, and ~3.5 compared with ProtTest (table 1), which explores a larger set of models exhaustively (120, supplementary table S5, Supplementary Material online). Based on the user's selected criterion (AIC/BIC), the basic principle in SMS is as follows: i) using a BioNJ tree topology (Gascuel 1997), SMS estimates the branch lengths and model parameters for LG (Le and Gascuel 2008) and the two RAS options; ii) using the "most promising" RAS option with LG, SMS selects the best substitution matrix and +F/−F option; to avoid computing both +F and −F options systematically, the matrices are ranked based on the similarity of the amino-acid frequencies in the data and those preestimated in the matrix; iii) SMS selects the best "decoration" (i.e., RAS and +F/−F options) for the best matrix. The gain in computing time is explained by the fact that, for most substitution matrices, SMS performs only 1 or 2 likelihood evaluations per matrix (1.75 on average, corresponding to different decorations), compared with four for the exhaustive approach, which evaluates all decorations for all matrices.

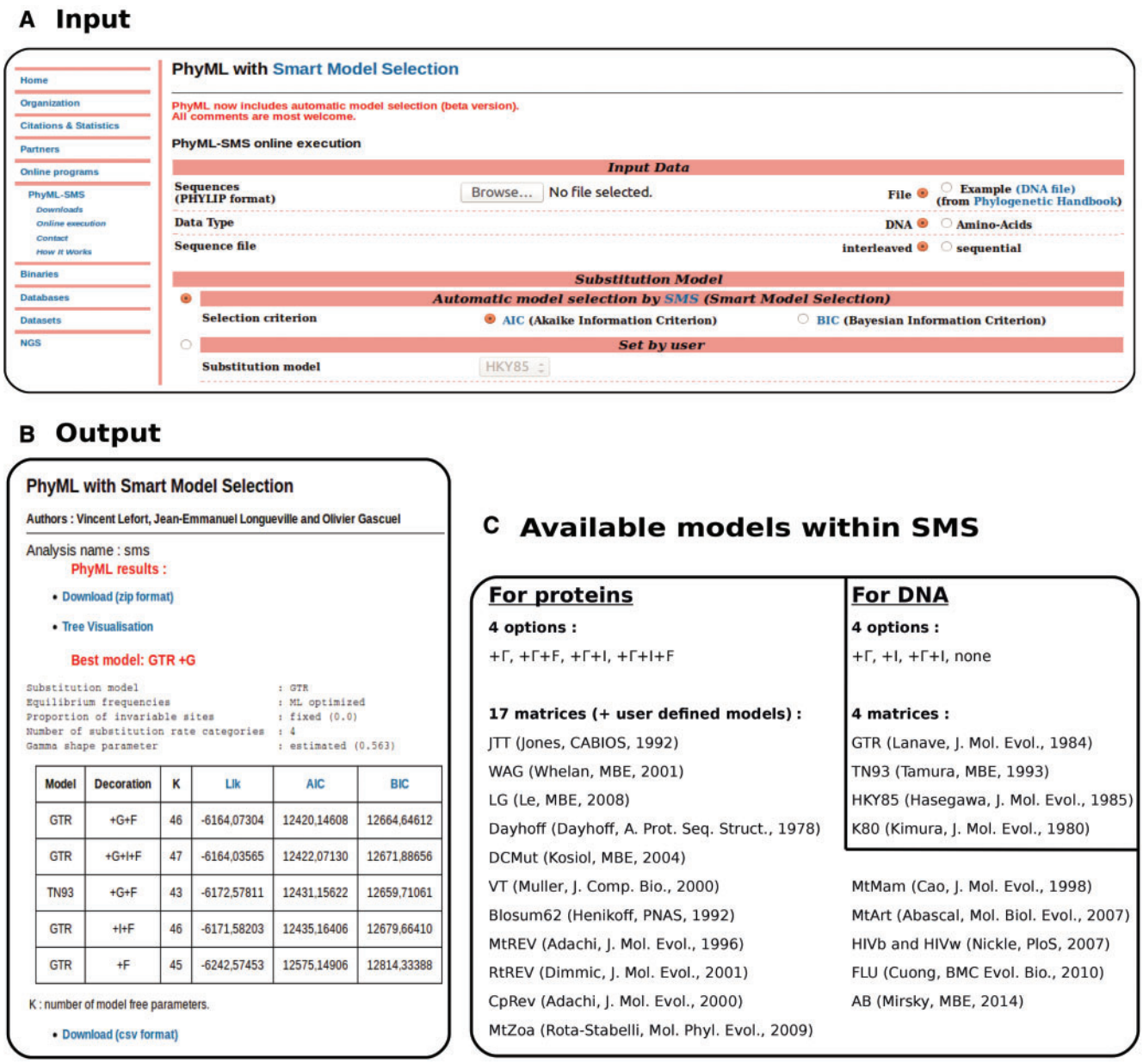

FIG. 1. Interface, input, output, models, and options. (*A*) By default, the substitution model is selected by SMS using AIC; alternatively, the user may choose BIC or select the model manually. (*B*) The output contains standard PhyML results and the model selected by SMS with detailed information. (*C*) Models and options available in SMS.

Computations with DNA are simpler than with proteins, as today's MSAs are most often large enough for GTR to be best compared to other substitution matrices. Moreover, the simplest matrices are not satisfactory because they do not account for the transition/transversion ratio and/or unequal base frequencies. Experiments with 500 representative MSAs confirmed these hypotheses, and are congruent with the large-scale study of (Arbiza et al. 2011). With AIC, GTR is best for 343/500 MSAs, whereas JC69, K80, and F81 are all best with 9/500 MSAs only (supplementary table S3, Supplementary Material online). However, with BIC, K80 is best for 48/500 MSAs. SMS thus uses four substitution matrices: GTR, TN93, HKY85, and K80, which are combined with $+I$, $+\Gamma$, $+\Gamma+I$, and "none" (all four RAS options are useful, supplementary table S3, Supplementary Material online), that is, a total of 4 x 4 = 16 models. On average, SMS computes the likelihood value of ~6 models with AIC and 7.5 with BIC, thus dividing the computing time by ~2 as compared to the exhaustive approach using the same models. Based on the user's selected criterion (AIC/BIC), the basic principle in SMS as follows: i) using a BioNJ tree topology, SMS estimates the branch lengths and model parameters for GTR and the four RAS options; ii) using the "most promising" RAS option with GTR, SMS selects the best matrix in a stepwise manner: SMS compares GTR and TN93; if GTR is better, then SMS stops and keeps GTR; otherwise, SMS compares HKY85 to TN93, and so on (remember that GTR, TN93, HKY85, and K80 are nested); iii) SMS selects the best RAS option for the best matrix. This simple approach, combined with a relatively small set of models, makes SMS nearly as fast as jModelTest2

Table 1. Method Comparison with 500 DNA, and 500 Protein Representative MSAs.

Methods	Data	Criterion	Same Model	SMS Better	SMS Worse	Δ AIC & Δ BIC per taxon per site	# PhyML Runs SMS/other	Speed Increase
SMS versus Exhaustive	DNA	AIC	486	na	14	4.6×10^{-5}	6.1/16	1.9–2.0
		BIC	476	na	24	8.0×10^{-5}	7.5/16	1.7–1.9
SMS versus Exhaustive	Protein	AIC	494	na	6	3.7×10^{-3}	29.3/68	2.2–2.1
		BIC	497	na	3	3.8×10^{-3}	30.2/68	2.1–2.0
SMS versus jModelTest2	DNA	AIC	380	85	35	-2.5×10^{-5}	6.1/7.8	1.1–0.8
		BIC	308	151	41	-1.1×10^{-4}	7.5/7.8	0.9–0.8
SMS versus ProtTest	Protein	AIC	465	14	21	-8.9×10^{-4}	29.3/120	3.7–3.4
		BIC	465	12	23	-7.5×10^{-4}	30.2/120	3.5–3.2

NOTE.—The "Exhaustive" approach uses the same set of models as SMS and evaluates all of them. "Same model": number of times (among 500 MSAs) where both methods return the same model; "SMS better": number of times where the model returned by SMS has a lower AIC/BIC value; "SMS worse": number of times where the model returned by SMS has a higher AIC/BIC value; "Δ AIC and Δ BIC per taxon per site": when both models were different, we computed the difference in AIC/BIC per taxon per site, and averaged the results over all MSAs showing a model difference (a negative/positive value means that SMS's model is better/worse in terms of AIC/BIC); "# PhyML runs": number of PhyML runs for one method versus the other; "Speed increase": for each MSA, we computed the computing time ratio of the method being compared with respect to SMS (e.g., 2 means that SMS is twice as fast), with the column displaying: i) the median value among the 500 speedup ratios for all MSAs, ii) the median value for the 50 largest MSAs (number of sites x number of taxa; see supplementary fig. S1, Supplementary Material online for additional computing time results with large MSAs).

using the fast "model filtering" option (supplementary fig. S1, Supplementary Material online).

Despite substantial gains in computing time, the results of SMS are nearly the same as those obtained with the exhaustive approach using the same models, and SMS performs well compared with jModelTest2 and ProtTest (table 1). To benchmark these methods, we used 500 DNA and 500 protein MSAs, corresponding to the first MSAs submitted to the PhyML Web server since the beta test version of SMS was made available (April 2015). No selection was performed, so these data sets are representative of the MSAs commonly used for phylogenetic analyses. Some of these MSAs are very small (e.g., 231 amino acids in total, with 11 taxa, and 231 sites); some are very large (e.g., 14,160,098 amino acids); some contain more than 1,000 taxa; and some have a huge number of sites (e.g., 52,092 nucleotidic sites). To confirm our findings, we also reused the 100 medium-size MSAs used to benchmark PhyML 3.0 (Guindon et al. 2010). The results with this second, independent set of MSAs, are fully congruent (supplementary table S6, Supplementary Material online). We launched jModelTest2 and ProtTest with fast options, since SMS was designed to be fast. Moreover, we selected the options to make these two programs as close as possible to SMS in terms of substitution matrices, RAS modeling, and equilibrium frequency estimation. The results are shown in table 1. To summarize: SMS performs well compared with the exhaustive approach, in most cases finding identical or similar models regarding AIC/BIC values, whereas the gain in computing time is quite substantial. Moreover, SMS tends to select better models than jModelTest2 with the fast "model filtering" option, and is much faster than ProtTest, thanks to tailored heuristics. The gains in AIC/BIC with SMS are partly explained by its set of substitution matrices, notably MtZoa for proteins and TN93 for DNA, which are not available in ProtTest and jModelTest2 (with default options). With proteins, SMS and ProtTest find the same model in most cases; when the models differ (35/500 MSAs), ProtTest finds a better model than SMS in ∼60% of the cases, but the average AIC/BIC difference is in favor of SMS. With DNA, the sets of models are more different than with proteins, and SMS and jModelTest2 differ for 120 and 192 MSAs with AIC and BIC, respectively; when the models differ, SMS finds a better model than jModelTest2 in ∼75% of the cases, and the average AIC/BIC difference is clearly in favor of SMS. The computing time gains of SMS with proteins are quite substantial in practice (supplementary fig. S1, Supplementary Material online). For example, ProtTest requires more than 100 h to process the largest MSA (1,151 taxa and 798 sites), whereas SMS requires ∼20 h using the same computer.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgment

This research was supported by the Institut Français de Bioinformatique (RENABI-IFB, Investissements d'Avenir).

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21(9):2104–2105.

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.

Arbiza L, Patricio M, Dopazo H, Posada D. 2011. Genome-wide heterogeneity of nucleotide substitution model fit. *Genome Biol Evol*. 3:896–908.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9(8):772.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14(7):685–695.

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(3):307–321.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*. 25(7):1307–1320.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14(9):817–818.

Schwarz G. 1978. Estimating the dimension of a model. *Ann Stat*. 6:461–464.