SMS: Smart Model Selection in PhyML

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

1. Institut de Biologie Computationnelle (IBC),
   LIRMM, UMR 5506 – CNRS & Université de Montpellier, France

2. Unité Bioinformatique Evolutive
   C3BI, USR 3756 – CNRS & Institut Pasteur, Paris, France

* Corresponding author: olivier.gascuel@pasteur.fr

Supplementary Material

SMS heuristics for proteins and DNA

The SMS software tool determines the evolutionary model that best fits the input sequences in a tree inference context. Two criteria are available, depending on the user's preferences: AIC (Akaike Information Criterion [1]) and BIC (Bayesian Information Criterion [2]). BIC penalizes free parameters more strongly than AIC. SMS does not use AICc (the corrected version of AIC) as it can be problematic with today's MSAs, where the number of taxa (t) and thus the number of free parameters ($k \geq 2t-3$ = number of branch lengths) is often close to, or even larger than, the number of sites (n). In that case, the correction term (with denominator $n-k-1$) can be very large in absolute value, and even negative, which is simply meaningless.

To reduce the computational burden, SMS uses heuristics to avoid evaluating all available models and options. Furthermore, some calculations are simplified. For example, the proportion of invariant sites is not systematically re-optimized in all settings. Below, we first provide definitions and then describe SMS heuristics for proteins and DNA sequences.

Definitions

- MSA: multiple sequence alignment (PHYLIP format).
- Model: a substitution rate matrix (e.g. GTR for DNA, JTT for proteins) + a model for rates across sites (RAS, e.g. +Γ) + the option used to define the equilibrium frequencies (only with proteins).
- Model decoration: RAS and equilibrium frequency options, that is:
    - +Γ: use of discrete gamma distribution with 4 categories to model RAS, the parameter α of this distribution is estimated from the MSA; no gamma distribution of RAS is used with –Γ (usually omitted when describing model options).
    - +I: we assume that some sites are invariant and then estimate the proportion of invariant sites in the MSA; +I alone (without +Γ) is only used with DNA, as it rarely appeared to be useful with proteins in our tests (Sup. Tab. 4); -I is usually omitted in the model option description.
    - +F: (proteins only) indicates that the equilibrium frequencies of amino-acids are estimated from the MSA (simple counting method); with -F (usually omitted), we use the equilibrium frequencies corresponding to the substitution matrix being evaluated.
- With protein data, SMS evaluates:
    - 4 decorations: +Γ / +Γ+I / +Γ+F / +Γ+I+F.
    - 17 substitution matrices (JTT, WAG, LG, etc., see Fig. 1C), with rate parameters pre-estimated from very large data sets (e.g., see [3]); moreover, users can add their own matrices (Paml format) to this list for comparison purposes.

- With DNA data, SMS evaluates:
  - 4 decorations: +I / +Γ / +Γ+I / none.
  - 4 substitution matrices: GTR, TN93, HKY85, and K80; all parameters in these matrices are estimated by maximum likelihood (ML) from the MSA, including nucleotide equilibrium frequencies.

SMS heuristic for protein data

With proteins, SMS involves four main steps: (1) fast inference of a fixed tree topology; (2) selection of the "most promising" RAS decoration with LG, to be applied to all matrices in the next step; (3) selection of the best matrix with a similarity-based heuristic to avoid evaluating both +F and –F options systematically; (4) final selection of the best decoration for the best matrix. Thanks to steps (2) and (3), SMS computes an average of only ~1.75 decorations per matrix, instead of 4 with exhaustive calculations. These four steps are detailed below.

1. Set a fixed tree topology

SMS computes a tree topology using BioNJ with LG evolutionary distances and no decoration (default PhyML option to build a first tree). Users can also supply their preferred topology. In both cases, this topology remains fixed during the next steps; only the branch lengths and model parameters are optimized by SMS for each of the models and options being evaluated.

2. Select the "most promising" RAS decoration with LG

Assuming LG with -F option and the previously-defined fixed tree topology, SMS selects the RAS decoration (+Γ, +Γ+I) that best fits the MSA, according to the user-selected criterion (AIC or BIC). This decoration is used in the next step (3) for all matrices and, when +Γ+I is selected, the proportion of invariant sites is kept fixed to accelerate parameter estimations.

3. Select the best substitution matrix

Substitution matrices for proteins are pre-estimated using very large data sets that commonly comprise thousands of MSAs. Both the relative rates and the amino-acid equilibrium frequencies attached to the matrix are estimated during this process. However, the average amino-acid frequencies obtained may be a poor fit for the MSA being analyzed. In that case, the +F option becomes relevant. To measure the closeness between the matrix equilibrium frequencies and those in the MSA, SMS computes a χ^2 distance that is used to sort the matrices.

On the one hand, the smaller the χ^2 distance, the closer the amino-acid frequencies in the MSA are to the frequencies associated with the substitution matrix; then, the +F option is likely to be of no use, and -F will produce better AIC/BIC values. On the other hand, the larger the MSA, the more likely it is that the +F option will produce better results, since, with large data sets, we have enough information to estimate a large number of parameters (i.e. here, 20 additional amino-acid frequencies). A threshold of 5,000 residues was empirically determined to distinguish large MSAs from small MSAs. To reduce the computational burden, SMS uses a strategy to minimize the number of times it evaluates both +F and -F options for the same matrix:

- Large MSAs: SMS sorts the matrices in increasing χ^2 value order, then evaluates and compares

both +F and -F options until it obtains three consecutive best +F options; subsequently, SMS no longer evaluates the -F option. Indeed, +F is expected to be the best option in most cases, except for the (hopefully few) matrices being close to the MSA with a low χ^2 value.

- Small MSAs: SMS sorts the matrices in decreasing χ^2 value order, then evaluates and compares both +F and -F options until it obtains three consecutive best -F options; subsequently, SMS no longer evaluates the +F option. Indeed, -F is expected to be the best option in most cases, except for the (hopefully few) matrices being remote from the MSA with a high χ^2 value.

4. Select the best decoration

In the previous steps, one (most matrices) to four (LG) decorations were evaluated for each of the matrices to select the best one. In this last step, SMS computes the AIC/BIC values associated with the selected matrix and the decorations that were not yet evaluated (or evaluated approximately using a fixed proportion of invariant sites).

SMS heuristic for DNA data

With DNA, SMS involves four main steps: (1) fast inference of a fixed tree topology; (2) selection of the "most promising" decoration with GTR, with this decoration then used in the next step to compare the matrices; (3) selection of the best matrix among GTR, TN93, HKY85, and K80; (4) selection of the best decoration for the best matrix. Only a few models (16 = 4 matrices X 4 decorations) are considered, as supported by our statistics with 500 representative data sets (Sup. Tab. 3). We also take advantage of the fact that GTR is usually best. On average, SMS evaluates ~6 and ~7.5 models when optimizing AIC and BIC respectively, while, with exhaustive calculations, 16 models are evaluated. Nucleotide equilibrium frequencies are estimated by ML with all models and options. Below, we detail the four SMS steps with DNA.

1. Set a fixed tree topology

SMS computes a BioNJ tree topology with HKY85 evolutionary distances and no RAS modeling (default PhyML option to build a first tree). Users can also provide their preferred topology. In both cases, this topology remains fixed during the next steps; only the branch lengths and model parameters are optimized by SMS for each of the models and options being evaluated.

2. Select the "most promising" decoration with GTR

Four decorations are evaluated (+I / +Γ / +Γ+I / none). The best decoration among four possibilities is selected and used in the next step (3) for all matrices. Moreover, when +I or +Γ+I is selected, the proportion of invariant sites is kept fixed to accelerate parameter estimations.

3. Select the best matrix

SMS compares GTR, TN93, HK85, and K80, assuming the previously selected decoration and proportion of invariant sites. These four matrices are nested: TN93 is a special case of GTR, HKY85 a special case of TN93, and K80 a special case of HKY85. Moreover, GTR is usually best (Sup. Tab. 3). Thus, SMS proceeds in a stepwise manner: it first compares GTR and TN93; if GTR is better, then SMS stops and selects GTR; otherwise, HKY85 is evaluated and compared to TN93; if TN93 is better, then SMS stops and returns TN93; otherwise, K80 is evaluated and compared to HKY85, and the best of both is selected.

4. Select the best decoration for the best matrix

In this step, SMS assumes the previously selected matrix and optimizes branch lengths and model parameters for all decorations for which these calculations were not already (or approximately) performed. SMS then returns the best matrix/decoration combination.

Data sets

1,000 representative data sets

To compare SMS with other approaches, we used 500 DNA and 500 protein MSAs, corresponding to the first data sets 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. Statistics are displayed in Sup. Table 1. Some of these data sets are very small (e.g. a total of 231 amino acids, 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). These datasets and their features are available from http://www.atgc-montpellier.fr/sms/, along with the results of the various approaches. For confidentiality reasons, taxon names have been removed, taxa are numbered, and the sites of the MSAs have been shuffled (PhyML and SMS results are unchanged).

Data	*Criterion*	*Minimum*	*1st quartile*	*2nd quartile*	*3rd quartile*	*Maximum*
DNA	# taxa (T)	**T = 6** S = 7,976 R = 47,551	**T = 27** S = 801 R = 21,353	**T = 54** S = 9,909 R = 473,500	**T = 102** S = 1,241 R = 126,574	**T = 1,093** S = 500 R = 538,138
	# sites (S)	T = 162 **S = 106** R = 17,057	T = 32 **S = 573** R = 18,336	T = 7 **S = 957** R = 6,391	T = 70 **S = 1,846** R = 115,435	T = 15 **S = 52,092** R = 781,380
	# nuc. (R)	T = 20 S = 132 **R = 2,640**	T = 58 S = 531 **R = 23,618**	T = 21 S = 2,667 **R = 54,615**	T = 240 S = 658 **R = 157,919**	T = 576 S = 7,433 **R = 4,280,950**
Protein	# taxa (T)	**T = 5** S = 555 R = 2,584	**T = 28** S = 306 R = 8,505	**T = 50** S = 299 R = 11,892	**T = 110** S = 65 R = 7,053	**T = 1,151** S = 798 R = 824,644
	# sites (S)	T = 71 **S = 17** R = 1,154	T = 198 **S = 172** R = 28,459	T = 15 **S = 395** R = 4,422	T = 39 **S = 610** R = 23,629	T = 62 **S = 230,322** R = 14,160,098
	# aa (R)	T = 11 S = 21 **R = 231**	T = 76 S = 83 **R = 6,184**	T = 159 S = 106 **R = 16,853**	T = 294 S = 141 **R = 40,724**	T = 62 S = 230,322 **R = 14,160,098**

Sup. Table 1: 500 DNA + 500 protein representative data sets. Several statistics are displayed to rank and summarize these 1,000 user-supplied data sets, which we used to compare SMS to other approaches: T is the number of taxa, S is the number of sites, and R is the number of nucleotides/amino acids (R = T X S - number of gaps and unknowns). For each of these statistics we provide the features of the MSA with minimum value, maximum value, as well as the first, second (i.e. median), and third quartiles. For example, the DNA alignment with the minimum number of taxa has T = 6, S = 7,979, and R = 47,551.

100 PhyML 3.0 data sets

To confirm our findings with previous representative MSAs and ensure that our experiments are reproducible, we also tested the medium-size data sets that were used to benchmark PhyML 3.0 [4]. These comprise 50 DNA and 50 protein MSAs extracted from TreeBase [5]. Statistics are displayed in Sup. Table 2. Globally these MSAs have less diverse sizes than the recent MSAs described in Sup. Table 1. Notably, some of the recent MSAs are extremely large, which was not the case with the previous ones, which were selected to be neither too large nor too small (see the PhyML 3.0 benchmark web site, where these data sets can be downloaded: http://www.atgc-montpellier.fr/phyml/benchmarks/).

Data	*Criterion*	*Minimum*	*1st quartile*	*2nd quartile*	*3rd quartile*	*Maximum*
DNA	# taxa (T)	**T = 50** S = 872 R = 43,600	**T = 56** S = 827 R = 46,312	**T = 68** S = 1,712 R = 116,416	**T = 99** S = 1,634 R = 161,766	**T = 191** S = 990 R = 189,090
	# sites (S)	T = 70 **S = 800** R = 56,000	T = 54 **S = 966** R = 52,164	T = 142 **S = 1,130** R = 160,460	T = 51 **S = 1,537** R = 78,387	T = 51 **S = 1,951** R = 99,501
	# nuc. (R)	T = 50 S = 872 **R = 43,600**	T = 66 S = 1,027 **R = 67,782**	T = 77 S = 1,101 **R = 84,777**	T = 143 S = 897 **R = 128,271**	T = 117 S = 1,910 **R = 223,470**
Protein	# taxa (T)	**T = 5** S = 1,006 R = 5,030	**T = 18** S = 1,561 R = 28,098	**T = 23** S = 392 R = 9,016	**T = 32** S = 129 R = 4,128	**T = 139** S = 348 R = 48,372
	# sites (S)	T = 46 **S = 68** R = 3,128	T = 30 **S = 276** R = 8,280	T = 40 **S = 430** R = 17,200	T = 30 **S = 719** R = 21,570	T = 19 **S = 1,566** R = 29,754
	# aa (R)	T = 7 S = 232 **R = 1,624**	T = 11 S = 530 **R = 5,830**	T = 38 S = 251 **R = 9,538**	T = 40 S = 430 **R = 17,200**	T = 52 S = 981 **R = 51,012**

Sup. Table 2: 50 DNA + 50 protein data sets, available from the PhyML 3.0 web site. See note to Sup. Tab. 1.

Usefulness of the models and options

Part of the computing efficiency of SMS is induced by the fact that it uses a limited number of models and options. Among all the possibilities available in PhyML, we selected those that showed to be biologically relevant and useful with the data sets currently assembled to perform phylogenetic analyses. For this purpose, we analyzed the 500 DNA and 500 protein MSAs previously described, with all models and options available from the PhyML 3.0 web server, plus the SYM substitution matrix for DNA [6], which is available in jModelTest2 [7]. Results obtained using the AIC and BIC criteria are summarized in Sup. Table 3 for DNA and Sup. Table 4 for proteins.

Criterion	*Deco./Mat.*	*GTR*	*SYM*	*TN93*	*HKY85*	*F81*	*K80*	*JC69*	**Total**
AIC	**"none"**	11	2	1	6	1		1	**22**
	+I	13	1	7	13		1		**35**
	+Γ	71	9	25	12	1	3		**121**
	+I +Γ	248	7	39	26	1	1		**322**
	Total	**343**	**19**	**72**	**57**	**3**	**5**	**1**	**500**
BIC	**"none"**	4	4		11	5	6	2	**32**
	+I	3		8	26	1	8		**46**
	+Γ	45	10	38	45	1	27	1	**167**
	+I +Γ	155	10	36	46	1	7		**255**
	Total	**207**	**24**	**82**	**128**	**8**	**48**	**3**	**500**

Sup. Table 3: Usefulness of DNA models and options. These results were obtained using our 500 representative DNA data sets. We ran PhyML with all combinations of matrices and decorations and counted the number of times each combination had the best AIC/BIC value.

With DNA data, we see that:

- All RAS decorations ("none", +I, +Γ and +Γ+I) are useful, even "none", which has the best AIC and BIC values for ~4% and ~6% of the MSAs, respectively. As expected, the most frequently selected decoration is +Γ+I (~65% and ~50%). All four decorations are thus available in SMS for DNA MSAs. It is likely that "none" and +I are mostly used with non-coding DNA, where the strength of the structural and functional constraints is less variable than in the coding regions (see results with proteins).
- The simplest matrices (F81 and JC69) are essentially useless and not available in SMS, as they are best with ~1% of the MSAs (both AIC and BIC). This was expected, since these matrices do not capture the difference between transitions and transversions.
- The other, more sophisticated matrices (GTR, SYM, TN93, HKY85, and K80), are much more useful. As expected (due to the large size of extant data sets), GTR is the most frequently selected model (~68% and ~40% with AIC and BIC, respectively), followed by TN93 (second best with AIC, ~14%), and HKY85 (second best with BIC, ~25%). SYM is clearly behind (<5% with both AIC and BIC); this was expected since SYM assumes (nearly) equal nucleotide frequencies, which is rarely the case. Thus, SYM is not available in SMS (and is not available on the PhyML 3.0 Web server). However, K80 (also assuming nearly equal nucleotide frequencies) appears to be useful with BIC (~10%), due to its simplicity (1 free parameter only). GTR, TN93, HKY85, and K80 are thus available in SMS.

With proteins, we see that:

- Among the RAS decorations, only +Γ and +Γ+I are useful and available in SMS; "none" and +I are all best with only 6 and 10 MSAs with AIC and BIC respectively, and thus are not available. Protein MSAs usually have few constant sites (median proportion ≈ 3%), and we expect a high variability of site rates caused by the variability of functional and structural constraints acting

along the protein sequences. These results and choices are thus biologically consistent.

Criterion	Model	*JTT*	*WAG*	*LG*	*Dayhoff*	*DCMut*	*VT*	*Blosum62*	*MtREV*	*RtREV*	*CpREV*	*MtZoa*	*MtMam*	*MtArt*	*HIVb*	*HIVw*	*FLU*	*AB*	Total
AIC	**"none"**	2						1											**3**
	+F													1					**1**
	+I				1														**1**
	+I+F	1																	**1**
	+Γ	19	6	42			4				1	1			2		5		**80**
	+Γ+F	21	10	33		2	8	5		3	3	4		1			2	1	**93**
	+I+Γ	26	10	117		1	3	5			1					2	1		**166**
	+I+Γ+F	30	12	74			8	3		9	4	4	1		4	1	5		**155**
	Total	**99**	**38**	**266**	**1**	**3**	**23**	**14**	**0**	**12**	**9**	**9**	**1**	**2**	**6**	**3**	**13**	**1**	**500**
BIC	**"none"**	2		1				1											**4**
	+F		2																**2**
	+I	2			1										1				**4**
	+I+F																		**0**
	+Γ	42	13	82	2	1	6	2			3	2			2	1	5		**161**
	+Γ+F	17	9	25		1	6	4		1	1	3		1	1	1	2		**72**
	+I+Γ	26	12	124	1	1	5	5			1		1				1		**177**
	+I+Γ+F	9	6	41			7	2		5	4	4			1		1		**80**
	Total	**98**	**42**	**273**	**4**	**3**	**24**	**14**	**0**	**6**	**9**	**9**	**1**	**1**	**5**	**2**	**9**	**0**	**500**

Sup. Table 4: Usefulness of protein models and options. These results were obtained using our 500 representative protein data sets. See note to Sup. Tab. 3.

- Both -F and +F options are useful and available in SMS; each is best in ~50% of the cases with AIC, and ~66% and ~34% respectively with BIC, which penalizes the number of parameters more strongly.
- Most matrices are dedicated to special types of proteins; for example, FLU is dedicated to influenza data sets, MtArt to mitochondrial MSAs from arthropods (with non-standard genetic code), and AB to proteins from immune systems. Thus, one would not expect these matrices to be selected frequently, and all matrices available in PhyML are also available in SMS. Among the general matrices, LG [3] is clearly the most useful (best AIC and BIC values for >50% of the MSAs), followed by JTT (~20%) and WAG (~8%), while Dayhoff, DCmut, VT, and Blosum62 are rarely selected (total of ~9%).

Data type	Software	Substitution Matrices	Decorations
DNA	SMS	GTR, HKY85, K80, *TN93*	+I, +Γ, +Γ +I, none
	jModelTest2	GTR, HKY85, K80, *SYM, F81, JC69*	+I, +Γ, +Γ +I, none
Protein	SMS	JTT, WAG, LG, Dayhoff, DCMut, VT, Blosum62, MtREV RtREV, CpREV, MtMam, MtArt, HIVw, HIVb, FLU, *AB, MtZoa*	+Γ, +Γ+I, +Γ+F, +Γ+I +F
	ProtTest	JTT, WAG, LG, Dayhoff, DCMut, VT, Blosum62, MtREV RtREV, CpREV, MtMam, MtArt, HIVw, HIVb, FLU	+Γ, +Γ+I, +Γ+F, +Γ+I +F, *+F, none, +I, +I +F*

Sup. Table 5: Available models in SMS and jModelTest2/ProtTest. *(differences in italics)*

Method comparisons

Methods being compared

To assess accuracy and efficiency, SMS was compared to the exhaustive approach that evaluates all matrix+decoration combinations to select the best one, instead of using heuristics to focus on the most promising combinations and save computing time. The exhaustive approach was launched with the same BioNJ tree, same sets of matrices and decorations as SMS, that is, 4 matrices X 4 decorations = 16 combinations for DNA, and 17 matrices X 4 decorations = 68 combinations for proteins. We compared the:

- Selected models (same or different).
- Difference in AIC and BIC when a different model was selected by SMS; then, the SMS model is necessarily worse than the exhaustive approach model, and we checked that the difference in AIC and BIC (per taxon per site) was acceptably small.
- Number of times PhyML was launched by SMS, that is, the number of combinations tested by SMS, to be compared to the 16 and 68 combinations tested by the exhaustive approach for DNA and proteins, respectively.
- Computing time of both methods.
- "Speed increase" brought by SMS, that is, the computing time of the exhaustive approach divided by that of SMS.

We also compared SMS to jModelTest2 [7] and ProtTest [8]. In both cases, we used fast options to run these programs, since SMS was designed to be fast. Moreover, we selected the options to make these programs as close as possible to SMS in terms of matrices and decorations, and we checked to make sure that the differences in both AIC/BIC values and computing times were not explained by the use of different PhyML versions.

Methods	Data	Criterion	Same model	SMS better	SMS worse	Δ AIC & Δ BIC per taxon-site	# PhyML runs SMS/other	Speed increase
SMS versus Exhaustive	DNA	AIC	49	na	1	2.57×10^{-6}	5.8 / 16	1.9
		BIC	48	na	2	7.62×10^{-3}	7.3 / 16	1.6
SMS versus Exhaustive	Protein	AIC	50	na	0	0	29.7 / 68	2.8
		BIC	49	na	1	7.19×10^{-3}	31.9 / 68	2.6
SMS versus jModelTest2	DNA	AIC	44	5	1	-2.24×10^{-5}	5.8 / 7.1	1.0
		BIC	28	19	3	-9.24×10^{-5}	7.3 / 7.1	0.8
SMS versus ProtTest	Protein	AIC	41	4	5	-9.89×10^{-5}	29.7 / 120	4.6
		BIC	43	2	5	9.77×10^{-5}	31.9 / 120	4.5

Sup. Table 6: Method comparison with PhyML 3.0 data sets. See note to Table 1.

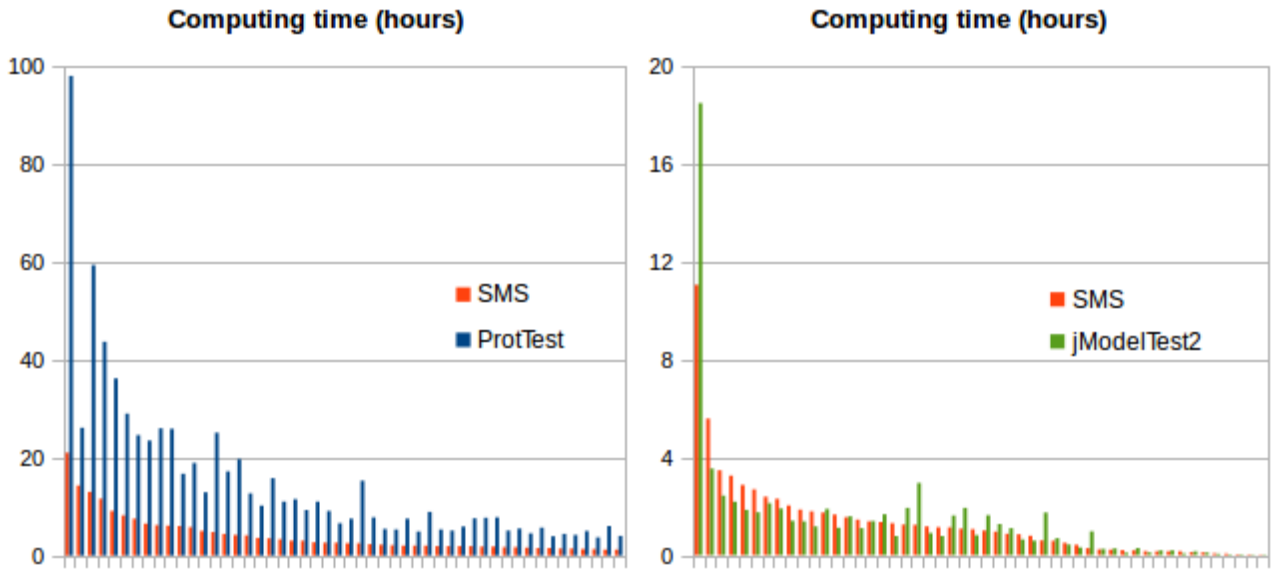
We used jModelTest2 version 2.1.10 with options:

- '-G 0.1' which corresponds to the fast "model filtering" heuristic to focus on promising models; (0.1 is the default threshold value to tune this heuristic).
- '-t BIONJ' which estimates a BioNJ tree topology separately for each of the models being tested (default involves using a unique JC69-based tree, which is presumably not accurate enough).
- '-f ' which includes matrices with unequal base frequencies (e.g. GTR and HKY85).
- '-i' which includes decorations with invariant sites (i.e. +I PhyML option).
- '-g 4' which includes decorations with 4 gamma site-rate categories (i.e. Γ4, just as SMS).

We used ProtTest version 3.4 with options:

- 'S' meaning that a fixed BioNJ tree topology with JTT is used to select the models (SMS uses LG instead), and that the branch lengths are re-optimized for each model (same as SMS).
- '-all-distributions' which includes decorations with site rate variation modeled using 4 gamma rate categories (Γ4) and/or invariant sites (+I).
- '+F' which includes models with empirical amino-acid frequency estimations.

The comparison criteria were similar to those used with the exhaustive approach. However, the set of models and options differ between these programs (Sup. Tab. 5). Thus, SMS may return a better/worse model than these programs, typically for algorithmic reasons, but also because the set of models is not the same (e.g. MtZoa with proteins, which is available in SMS but not in ProtTest, and SYM with DNA, which is available in jModelTest2 but not in SMS).



Sup. Figure 1: Computing time comparisons with large MSAs. Left: 50 largest protein alignments. Right: 50 largest DNA alignments; jModelTest2 was used with the fast "model filtering" option. All computations were performed on the same computer (single thread, Intel X5650 processor at 2.6 GHz). All MSAs were extracted from the 500+500 representative dataset. For the size of the MSAs, see Sup. Tab. 1.

Comparison results

Results are displayed in Table 1 with the 500+500 representative MSAs, and in Sup. Table 6 with PhyML 3.0 MSAs. Both sets of results are fully congruent, and the percentages and numbers below are based on the comparison of 500+500 MSAs. Computing times with the largest MSAs are displayed in Sup. Figure 1. We see that:

- With DNA, SMS most often finds the same model as the exhaustive method (~95%), and the difference in AIC/BIC is very low when the models are different. The speed increase is of ~2, as SMS evaluates ~6 (AIC) to ~7.5 (BIC) models among 16.
- With proteins, SMS finds the same model as the exhaustive approach in most cases (~99%), but when the models differ, the difference in AIC/BIC is larger than with DNA, likely because the models for proteins are more different than the SMS models for DNA. Speed increases by a factor of 2-3, as SMS evaluates ~30 models among 68 and saves computing time by not systematically re-estimating the proportion of invariant sites while searching for the best matrix.
- Comparing SMS with jModelTest2, the number of times where the models are the same is still relatively high (~76% and ~62%, with AIC and BIC, respectively), but lower than with the exhaustive approach, as the sets of models explored by SMS and jModelTest2 are different (Sup. Tab. 5). When the models differ, SMS finds a better model more often than jModelTest2 (85/35 and 151/41 with AIC and BIC, respectively), and the difference in AIC/BIC is clearly in favor of SMS. This gain is partly explained by TN93, which is a useful model (Sup. Tab. 3), available in SMS, but not in jModelTest2 (with default options). Both programs are nearly as fast as each other (Sup. Fig. 1), as expected, since both use heuristics to focus on the most promising models.

- With proteins and ProtTest, we observe that the models are generally the same (93% with both AIC and BIC). When the models differ, ProtTest tends to find a better model than SMS (14/21 and 12/23 with AIC and BIC, respectively), but the difference in the AIC/BIC value is in favor of SMS. Most of the gain obtained by SMS is derived from the use of the MtZoa substitution matrix, which is not available in ProtTest. The speed increase is in the range of 3.5 to 4.5, and thus quite substantial, especially for large MSAs (Sup. Fig. 1). This is explained by the smaller number of available models in SMS (68 versus 120), and the fact that SMS uses a heuristic approach to focus on the (~30 / 68) most promising models.

To summarize, SMS performs well as compared to the exhaustive approach, in most cases finding identical or similar models, while the gain in computing time is significant. SMS tends to select better models than jModelTest2, while it is much faster than ProtTest thanks to tailored heuristics. Gains in computing time with proteins are quite substantial in practice as, for example, ProtTest requires 105 hours to process the largest representative MSA (1151 taxa, 798 sites), while SMS only takes 21 hours.

References

[1] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. Petrov B. N. and Csaki F. (eds), Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

[2] Schwarz G. (1978). Estimating the dimension of a model. Ann. Stat. 6:461–464.

[3] Le S.Q. and Gascuel O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

[4] 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(3):307-21.

[5] Sanderson M.J., Donoghue M.J., Piel W., Eriksson T. (1994). TreeBase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Amer. Jour. Bot. 81:183.

[6] Zharkikh A. (1994). Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

[7] Darriba, D., Taboada, G.L., Doallo, R., Posada, D. (2012). JModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772.

[8] Abascal F., Zardoya R., Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics, 21(9):2104-5.