Result Verification, Code Verification, and Computation of Support Values in Phylogenetics

Alexandros Stamatakis and Fernando Izquierdo-Carrasco
The Exelixis Lab
Scientific Computing Group
Heidelberg Institute for Theoretical Studies
Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany
email: {alexandros.stamatakis,fernando.izquierdo}@h-its.org
WWW: http://www.h-its.org/english/research/sco/

Abstract

Verification in phylogenetics represents an extremely difficult subject. Phylogenetic analysis deals with the reconstruction of evolutionary histories of species, and as long as mankind is not able to travel in time, it will not be possible to verify deep evolutionary histories reconstructed with modern computational methods. Here, we focus on two more tangible issues that are related to verification in phylogenetics (i) the inference of support values on trees that provide some notion about the "correctness" of the tree within narrow limits and, more importantly, (ii) issues pertaining to program verification, especially with respect to codes that rely heavily on floating-point arithmetics. Program verification represents a largely underestimated problem in computational science that can have fatal effects on scientific conclusions.

Keywords: phylogenetics, support values, program verification, maximum likelihood, bootstrap, software bug

Introduction

The goal of the discipline that emerged in the late 1960ies [1,2] and is now called phyloinformatics consists of reconstructing phylogenetic (evolutionary) trees from morphological or molecular sequence data. The phylogenetic history of a set of *n* organisms (represented, for instance, by their DNA sequence data) is typically depicted as a fully bifurcating (strictly binary) tree topology. A tree-like model of evolution may be an over-simplification of the evolutionary processes, in particular in the presence of lateral gene transfer [3, 4], but in contrast to networks, trees are more straight-forward to visualize, understand, interpret, and to handle in a mathematical and

computational framework. The *n* extant (currently living) organisms of the input dataset (usually a multiple sequence alignment) are located at the leaves (tips) of such a tree, while the inner nodes represent extinct common ancestors. If one is willing to accept the hypothesis of a tree-like evolutionary process, two key questions arise: What is the order of magnitude for the number of potential tree topologies with *n* taxa, and based on which criteria should we chose the "best" among all potential tree topologies, that is, the tree that best fits our input data. In fact, the number of potential tree topologies increases super-exponentially with the number of organisms *n* according to: $\prod(2i-5)$ for i=3...n (this can be easily derived from the analogous formula for rooted binary trees proposed in [5]). As a consequence, for most common tree-scoring criteria such as parsimony [1,6] or likelihood [7], finding the best tree is a NP-hard optimization problem [8,9] because they require an exhaustive search of tree space. Thus, heuristics need to be deployed for finding the best-known parsimony or likelihood tree. At present, no mathematical tools are available to determine the score of the best possible tree (given a dataset), as measured by parsimony or likelihood criteria. Moreover, for trees inferred with heuristic search strategies under maximum likelihood, there does not exist a quality guarantee, for instance, that the score of the tree returned is at most 10% worse than the score of the best tree. As such one can only compare the relative performance of different search strategies on real biological datasets. By using benchmark datasets, one can empirically observe that one search strategy yields better scores than another.

Alternatively, one can assess the ability of the implementations of different models and heuristic tree reconstruction strategies such as parsimony [10], maximum likelihood [11-13], or Bayesian approaches [14,15] to recover the "true" tree on simulated datasets. This approach is best used to assess questions with respect to input dataset assembly and shape [16-22] (see [23] for a summary of accuracy assessment techniques in phylogenetics). For this type of verification, one assumes a given true tree topology generated either with real biological data or by a computer program [24]). Then, a multiple sequence alignment is generated via a Markov process by letting an ancestral sequence evolve along the branches of the tree according to some statistical model of sequence

evolution [25-27]. This will produce a multiple sequence alignment that can be used as input for phylogenetic tree reconstruction methods. The topological distances (see [28] for a tool implementing most common tree distance metrics) between the trees that were reconstructed by the inference methods and the true tree that generated the data can then be deployed to assess the ability of the methods under consideration to reconstruct the true tree. This approach however is not without caveats. For example, there is no guarantee that the implementation of the simulated alignment generation programs is correct (e.g., consider the well-documented bug history of the widely-used simulated data generation program Seq-Gen [25] at http://tree.bio.ed.ac.uk/software/seqgen/). Furthermore, the sequences along the true tree are typically generated using one of the common models of statistical sequence evolution, such as, for instance, the widely used [29] General Time Reversible (GTR [30]) model of nucleotide substitution with the Γ model of rate heterogeneity [31]. In turn, the model that was used to generate the data will then also be used to reconstruct the tree. Thus, the true model of evolution is known a priori, and all the potentially simplistic assumptions, for instance, that the model is time-reversible and that sites (alignment columns) evolve independently, are already incorporated into the simulated data.

In other words, phylogenetic tree inference on simulated data (assuming a perfect model) tends to be easier than on real data. For instance, the assessment of the original PHYML [20] code for likelihood-based phylogenetic inference was mainly based on simulated data, and showed that PHYML performed as well as other programs. An assessment using real data then revealed that other heuristic search strategies that explore the tree search space more thoroughly returned trees with significantly better (in the statistical sense) likelihood scores [32]. Thus, because real biological data is less perfect, optimization tends to be harder on real biological data than on simulated data that has been generated using a known model. Nonetheless, simulated data can be used to test the behavior of search algorithms and their ability to recover the “true” tree in a best case scenario, that is, when the model perfectly fits the data.

Until recently, the incorporation of insertion and deletion events into simulated alignments represented a problem. However, new tools such as DAWG [27] and Indelible [26] now provide for simulated data generation under a more realistic insertion/deletion model. Another potential problem that can occur on datasets with a relatively small number of alignment sites, is that the true tree is not necessarily the maximum likelihood tree [22]. This is a result of maximum likelihood being consistent in the statistical sense, that is, the reconstructed tree converging to the true tree when the number of sites goes to infinity [33]. By using simulated data experiments (where model misspecification is not an issue) and RAxML for tree reconstruction, we show that, especially on short alignments, RAxML consistently finds trees that have a better log likelihood score than the true tree. This does not necessarily mean that the inferred ML estimates on the short alignments are incorrect or of bad quality with respect to their topological distance to the true tree. However, as shown in Figure 1, the difference between the log likelihood scores of the true tree and the inferred trees as well as the topological distances between the true tree and the estimated trees decrease with alignment length (alignment width).

We generated simulated alignments using DAWG containing between 768 and 12,799 sites on a real biological tree with 1,908 organisms. We optimized ML model parameters (branch lengths, GTR, and Γ model parameters) using RAxML (-f e option) on the true tree and the inferred trees to compute the ratios between the log likelihood scores. RAxML was also used to compute the symmetric topological difference (frequently also denoted as RF distance, albeit this terminology is not entirely correct; see [33]) between the inferred trees and the true tree.

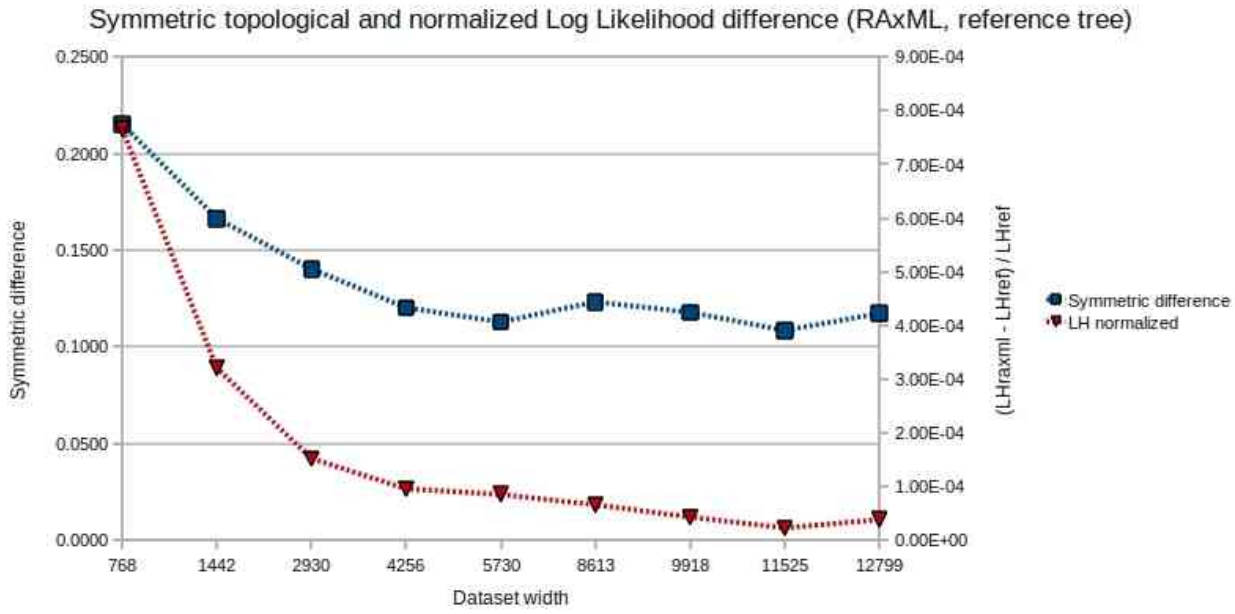

Figure 1: Relative differences between the log likelihood scores under GTR+Γ of the inferred tree (using RAxML) and the true tree as a function of the number of simulated alignment sites (denoted as "Dataset width" on the x-axis) using DAWG for a tree with 1,908 organisms. We also plot the symmetric topological differences between the true tree and the inferred trees as a function of dataset (alignment) width.

Based on the prolegomena, it is generally impossible to verify tree reconstruction programs with respect to their ability to infer the true or a reasonably correct tree because for real biological data the true tree and evolutionary model (if evolution is tree-like) is commonly not known and computational verification experiments based on simulated data may represent easy tree reconstruction problems because the simulated data has been generated according to our simplistic models. Also no guarantee exists that the implementation of simulated data generation programs is correct.

Finally, the method deployed for computing a multiple sequence alignment (MSA) can also have a significant impact on the tree reconstruction process [34] and ideally one would like to simultaneously infer trees and MSAs [35-37] which represents a hard computational problem and also a difficult modeling problem, because, at present, it is not clear which criterion to use for assessing multiple alignment optimality. In phylogenomics, issues pertaining to a lack of a broadly

accepted criterion for orthology assignment (see [38-40] for alternative approaches), gappy data [41] (sequence data for every organism and every gene under study is typically not available), and the discordance between gene trees and the species tree [42, 43] lead to the introduction of additional errors at the stages of input data assembly and alignment as well as the actual phylogenetic inference step.

Thus, given all of the aforementioned problems, verifying phylogenetic inference approaches is practically impossible. However, phylogenetic inference methods may be verified empirically, that is, if they can be successfully deployed to improve our living conditions, by using phylogenetic information to develop new drugs [44] or to disentangle [45] and predict viral outbreaks. Therefore, in the following we will briefly address two more tangible verification problems: the inference of support values on trees and the verification of codes for phylogenetic inference.

Inference of Support Values

The inference of support values on unrooted binary tree topologies intends to solve the following problem: Given a tree with 5 organisms A, B, C, D, E as depicted in Figure 2 (one may assume that this is the best-known maximum likelihood tree), we want to determine a degree of confidence for the organisms being split into the subgroups that are induced by cutting the tree at an inner branch. In fact, we want to compute such a confidence value for every inner (internal) branch of the tree. Such splits of the taxon set into two disjoint taxon sets that are induced by inner branches of the tree are also termed non-trivial bipartitions. A trivial bipartition is a cut/split of the tree at a branch leading to a terminal organism, for instance, at the branch leading to organism A in Figure 2. Those bipartitions are called trivial because they do not provide any information about the tree topology, that is, the bipartition A|BCDE is contained in all possible 15 unrooted tree topologies for 5 organisms. Note that, the set of all *n-3* non-trivial bipartitions of an unrooted binary tree (AB|CDE and ABC|DE in the example) with *n* taxa suffices to fully characterize the tree topology. In other words, the list of non-trivial bipartitions and the tree topology are equivalent representations of the

same mathematical object. A common misconception among biologists is that support values are assigned to nodes of the tree, rather than to inner branches.

A list of non-trivial bipartitions as induced by a set of trees (e.g., a collection of plausible trees that have been sampled using Bayesian methods [14,15,46] or a collection of bootstrap replicate trees; see below) and their respective frequency of occurrence therein can then, for instance, be used to assign confidence values to the inner branches of a best-known ML tree or to reconstruct a consensus tree. It is important to be aware that, given all of the aforementioned limitations regarding verification in phylogenetics, those bipartition support values do not provide an indication of whether a true evolutionary split of the taxa has been recovered. Nonetheless, bipartition support values are commonly interpreted as a measure for the correctness of bipartitions.

To date, the most widely used techniques are: Bootstrapping [47], likelihood ratio-tests [13,48] (a modification of the standard likelihood ratio test [49]), and posterior probabilities as obtained by MCMC-based Bayesian methods [50] (see [51,52] for reviews on support values).

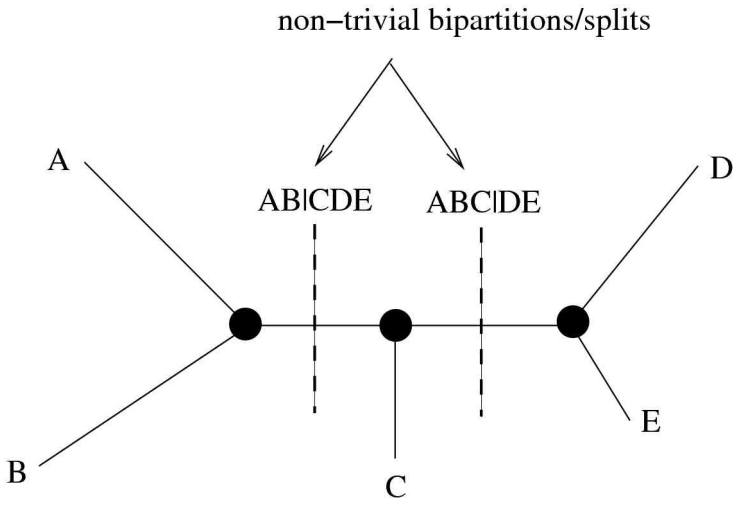

Figure 2: An unrooted tree of 5 organisms with two non-trivial bipartitions/splits AB|CDE and ABC|DE for which one can infer support values.

Bootstrapping

The general bootstrap procedure was introduced by Brad Efron in 1979 [53] as a means to infer the

variability of an unknown distribution for an estimator *T* using computer-based methods. The phylogenetic bootstrap procedure (see Fig. 3) was proposed in 1985 by Joe Felsenstein [47]. The underlying idea of the phylogenetic bootstrap is to assemble a certain number *r* of so-called bootstrap replicate alignments by randomly drawing sites (with replacement) from the original multiple sequence alignment. So, each bootstrapped alignment will contain exactly the same number of sites as the original alignment, but exhibit a slightly different site composition. Then, a tree is reconstructed for each of the *r* replicates using the estimator *T* (the tree reconstruction algorithm of choice), such that one obtains a set of *r* (potentially) distinct tree topologies. Those trees, that can be represented as a list of bipartitions, can then be used to build a strict or majority rule consensus tree as originally proposed in [47]. From the *r* replicates, one builds a -in most cases multifurcating- tree containing all bipartitions that occur in all *r* (strict) or more than *r/2* (majority rule) trees in the set of replicates (see Fig. 4 for an example).

Alternatively, one can draw bipartition support values on the best-known ML tree obtained from the original alignment. To do this, one counts how frequently a bipartition of the best-known ML tree occurs in the *r* bootstrap trees and assigns the frequency of occurrence to the bipartition.

A persistent problem is that the tree inference problem on each bootstrap replicate is also NP-hard. Hence, unlike in general applications of the bootstrap, our estimator *T* (whose quality is unknown) is only an approximation of the exact estimator *T'* (an algorithm that finds *the* optimal ML tree).

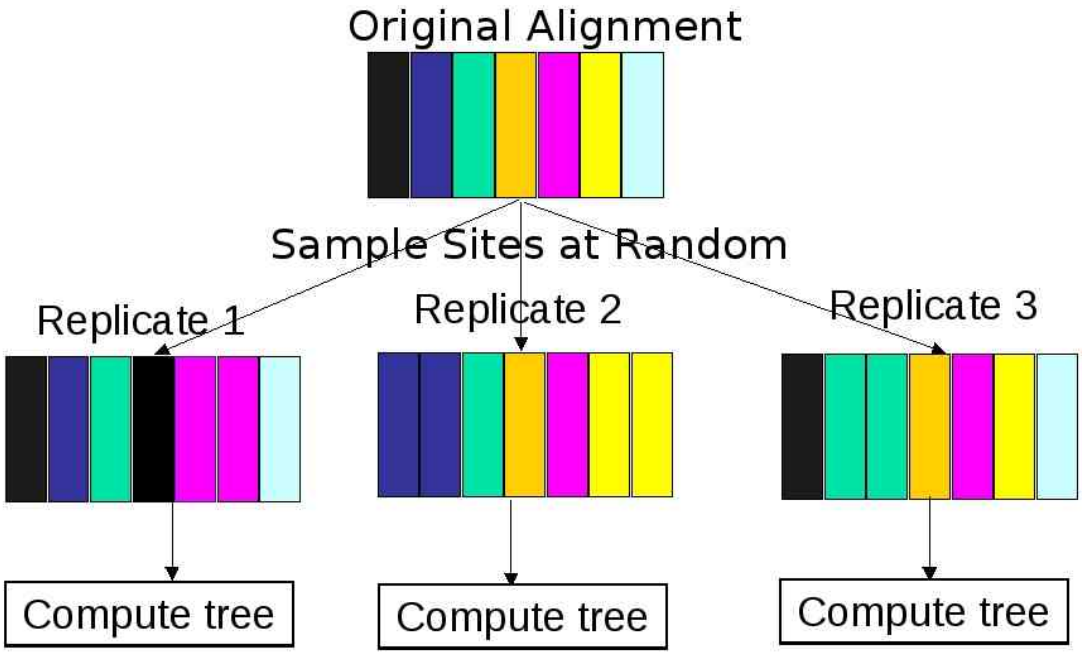

Figure 3: Outline of the phylogenetic bootstrap procedure.

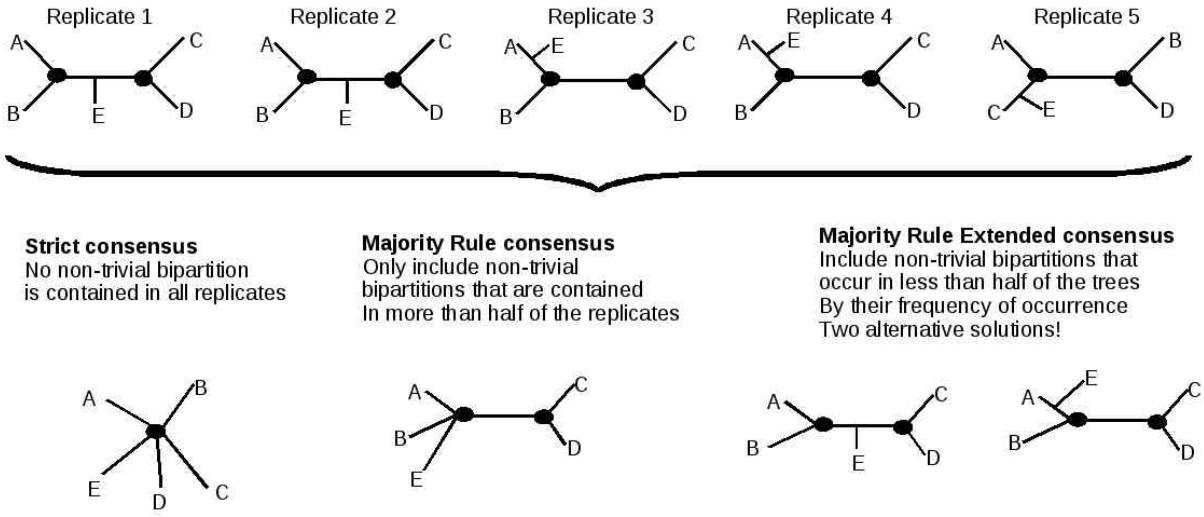

Figure 4: Outline of strict, majority rule, and majority rule extended consensus trees for a collection of 5 input trees.

One major disadvantage of the bootstrap is that, it is computationally expensive. Instead of conducting a small number of tree inferences, we need to compute between 100 to 1,000 replicates. In fact, it is not clear how many bootstrap replicates *r* may be required to obtain reliable support values, albeit some theoretical [54] and practical [55] suggestions exist. The number of required replicates appears to be closely related to the behavior of the estimator *T* on the specific dataset being analyzed [55]. If the phylogenetic signal in the data is strong, relatively few replicates (*r* =

100) are required and the bootstrap procedure will return a large number of identical trees. If the signal is weak (pointing towards multiple local optima in the likelihood landscape), a large number of replicates (r=1000) is required [55] (each replicate returning a different tree) to obtain stable majority rule trees (i.e., consensus trees that do not change if more replicates are computed). At present this is only an empirical observation that should be further investigated.

Likelihood-Ratio Tests

Approximate likelihood ratio tests have been proposed [48] as a fast alternative to bootstrapping and implemented in programs such as PHYML v3.0 [13], FastTree v2.1 [56], and RAxML v7.2.8 [11]. In principle they work as follows: Given a maximum likelihood tree that is locally optimal with respect to the application of NNI (Nearest Neighbor Interchange) moves, the algorithm visits one inner branch at a time and computes three log likelihood values L_0, L_1, and L_2. L_0 is the log likelihood of the NNI-optimal tree, and L_1 and L_2 are the log likelihoods of the two alternative trees obtained by applying the two possible NNI interchanges to the branch under consideration (see Fig. 5). Then, L_0 and L_alt (the best of the two alternative values L_1 and L_2) are used to compute support statistics (see [48] for details). Instead of simply using L_0 and L_alt as originally proposed in [48], current implementations [11,13,56] use a Shimodaira Hasegawa-like test [13,57], that essentially relies on random re-sampling of the per-site log likelihood scores by using between 100 to 1,000 samples. While this method is still very fast and seems to yield support values that are mostly in agreement with other bipartition support measures [13, 56], the interpretation of bootstrap support values remains difficult (see the introduction in [48] that entails additional references, [58,59] and pp 346-354 in [33]). Any comparison of aLRT versus bootstrap support values may be debatable, and there is one key issue that may require further investigation. The NNI move that is executed to obtain the three likelihood values L_0, L_1, and L_2 (or SH-like re-sampled values for that matter) is only a local move and does not guarantee that the two best log likelihood scores used in the test for that inner branch have been found. This is especially problematic for large trees with thousands of taxa. Thus, in comparison to the standard phylogenetic

bootstrap or Bayesian posterior probabilities (see below), the immensely large tree space is explored to a significantly smaller degree for obtaining support values. This may potentially influence the quality of the support values obtained by these tests. It would therefore be desirable to assess the impact of more rigorous topological moves such as SPR (Subtree Pruning Re-Grafting) or TBR (Tree Bisection Reconnection) moves on approximate LRT support values. Therefore, it may be good practice to use at least one method (Posterior Probabilities or Bootstrapping) that deploys more rigorous topological sampling for obtaining support values in addition to approximate likelihood ratio tests.

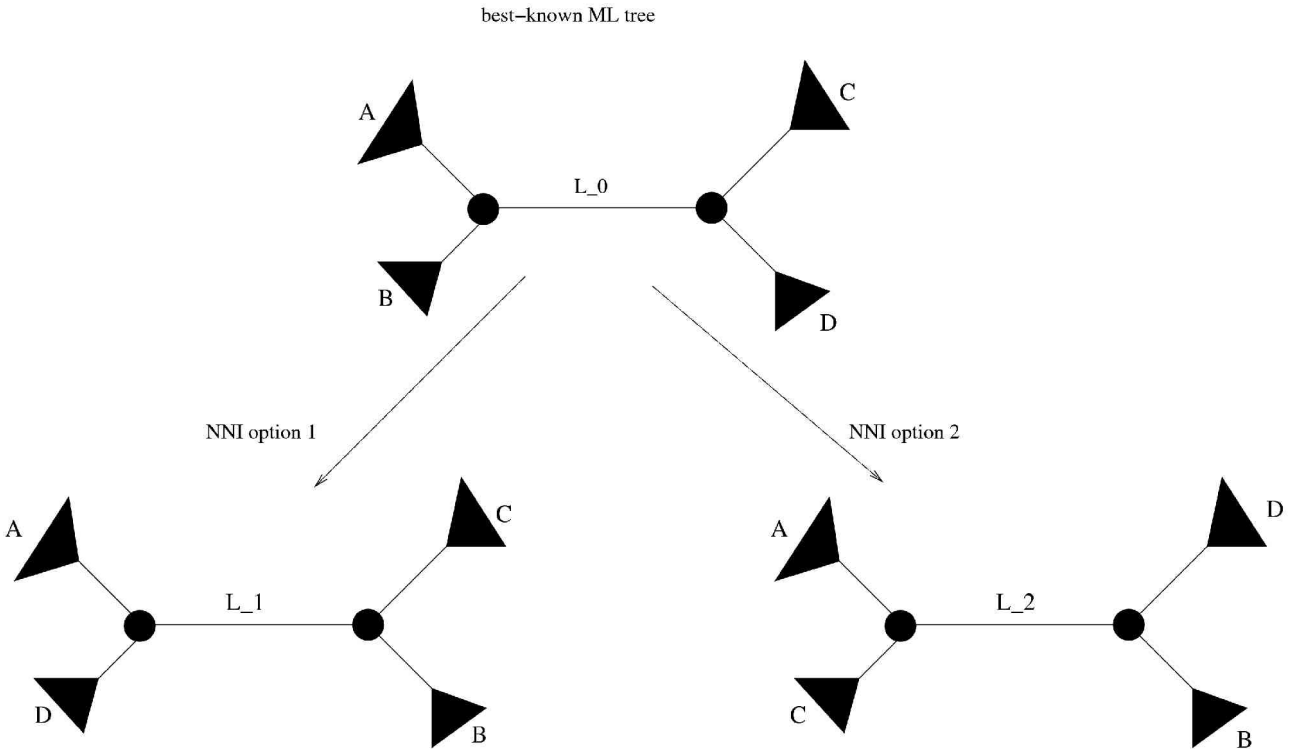

Figure 5: Outline of approximate likelihood ratio tests for obtaining bipartition support values using NNI moves on the best-known NNI-optimal maximum likelihood tree.

Posterior Probabilities

Posterior probabilities for bipartitions are obtained from Bayesian analyses, assuming that the parameter space has been sampled to a sufficient degree. The Metropolis [60] or more commonly the Metropolis-Hastings [61] Markov Chain Monte Carlo (MCMC) algorithm must have run long enough to converge, and the chosen prior probabilities must be reasonable. Bipartition support

values can then be obtained, for example, by building consensus trees from the posterior set of trees that have been sampled (using the bipartitions and posterior probabilities of the set of trees). The art of developing MCMC algorithms consists in designing proposal mechanisms that are able to efficiently (i.e., without spending many generations) escape local optima, avoiding early stopping of Bayesian analyses when chains reach apparent stationarity. While those problems can be alleviated by executing several chains, conducting multiple runs, or using convergence analysis tools [62], there is still no guarantee that the tree search space has been sampled to a sufficient degree. In the worst case scenario, the topology proposals for one or more chains will almost always get rejected such that the same tree topology is sampled for slightly different model parameters (e.g., alpha shape parameter of the Gamma model of rate heterogeneity [31], GTR model parameters [30], branch lengths) that can then potentially lead to very high posterior probabilities for incorrect tree bipartitions (see also [63]). However, promising ideas to improve the proposal mechanism exist, such as the TBR-biased proposal in MrBayes 3.2, that allows the chain to more easily escape from situations of apparent stationarity (John Huelsenbeck, pers. comm. August 2010). Clearly, Bayesian methods have advantages with respect to implementing more complex models such as mixture models [64,65] (but see [66]) or heterotachy [67] that are more straight-forward to integrate into a Bayesian (using a Gibbs sampler [see review in [68]] or reversible jump MCMC algorithms [69]) than into a maximum likelihood framework. On the other hand, ML searches are more targeted in that they strive to find the absolute optimum in tree space and the search strategy may better fit the difficult NP-hard optimization problem. However, for ML there is also no guarantee that the search has recovered the global optimum. Finally, it has also been proposed to use bootstrapped posterior probabilities [70]; the authors also suggest that bootstrap proportions and plain posterior probabilities without bootstrapping should not be directly compared. See also [23,33,48,70,71, 72, 73] for discussions and further references on the interpretation and comparison of bootstrap and posterior probability values.

Program Verification

Program verification represents a largely underestimated problem in disciplines, such as phylogenomics, that increasingly rely on scientific computing. Assuming that our methods and models are correct, in most cases, there is no guarantee that implementations are correct. While we are not aware of any catastrophic bugs in likelihood-based phylogeny programs, we will outline some of the potential pitfalls.

While theoretical tools for program verification such as for instance the Hoare calculus (also called Hoare logic or Hoare Rules) and software tools for program verification (e.g.,[74,75]) exist, they are rarely used in practice for academic software development (most popular tools for phylogenetic inference are freely available academic software) because of lack of time.

Likelihood-based programs face a substantial additional challenge, because they rely on floating point (also called machine numbers) arithmetics, whose correctness -if possible at all- is even more difficult to demonstrate [76], because simple mathematical rules for the real numbers such as (a + b) + c = a + (b + c) do not hold for floating point arithmetics.

In fact, the problem is so difficult, that it required, for instance, an entire PhD thesis [77] to demonstrate that an implementation of elementary functions such as the exponential function, called millions of times in any phylogenetic likelihood implementation, returns correctly rounded results until the last bit under double-precision arithmetics.

Moreover, it is not guaranteed that the same code, for instance RAxML [11], will return exactly the same numerical results, or even the same tree topology, if executed on different computer architectures, if executed sequentially or in parallel, or if compiled with different compilers. This behavior results from compilers optimizing assembly code by assuming that the mathematical laws for real numbers hold for machine numbers. Additionally, many computer architectures allow for denormalized floating point values (i.e., floating point numbers that have more bits than the IEEE 754 standard), mainly to prevent numerical underflow.

The use of denormalized floating point numbers also has serious implications on the assessment of

supercomputer systems using floating-point intensive benchmark codes [78] because execution times can heavily depend upon the input data. This has been observed for the phylogenetic likelihood function during the development of the short read phylogenetic placement algorithm in RAxML [79] where execution times for the likelihood function on data of the same size varied by 50% depending on the input data.

It would be possible to demonstrate the correctness of phylogenetic analysis programs that use parsimony or codes for reconstructing consensus trees because they operate on discrete entities, such as natural numbers, graphs, and trees.

As a concrete example, for code verification related erroneous conclusions in phylogenetics, a paper was published and later withdrawn, that assessed (using simulated data) the extent to which bootstrap-based bipartition support values were correlated with recovering true bipartitions. The paper stated that there was no correlation. However, a careful reader detected an error which pointed to a bug in the scripts that were used to analyze the bootstrap replicates and compute correlations. After fixing the bug, the conclusions of the paper were inversed and a strong correlation between bootstrap support values for bipartitions and their occurrence in the true simulated tree was observed. Unfortunately, the revised paper was not published to document the degree to which computational science and the hypotheses that are based on computational analyses rely on the correctness of software. Another example is, that the empirical base frequencies for C and G in the likelihood implementation of RAxML [11] were inversed until 2006, that is the frequency of C was used for G and vice versa. As a consequence, RAxML returned topologically distinct trees when the difference between the empirical base frequencies of C and G was large.

To this end, extreme caution should be exercised with increasingly complex analysis pipelines that rely on a plethora of programs: sequence assembly, orthology assignment, multiple sequence alignment, tree building, post-analysis, biogeographical analyses, and intermediate scripts to parse and adapt data formats. An example of the impact of program errors on scientific results is the retraction of five papers (published -among other journals- in Science and PNAS) on protein

crystallography because of a bug in a data analysis program [80].

Conclusions

We have discussed issues related to the generally impossible task of verifying models, algorithms, and software for phylogenetic inference and provided a brief overview over common techniques for obtaining support values on trees. There exist some promising empirical verification results using fast-evolving organisms [73,81], which indicate that our reconstruction methods may be reasonable and that support values may be used for assessing whether a true bipartition has been recovered. However, this still represents anecdotal evidence and does not allow for extrapolation to larger time-scales (deeper phylogenies) or different types of organisms. Consider, for instance, the on-going debate about the "correct" phylogeny of the metazoa [82,83,84,85,86]. As phylogenetics increasingly rely on computational methods, we believe that practitioners, should cease using Bioinformatics programs as black boxes and be aware of the important and potentially fatal effects software errors can have on the hypotheses they develop. While computational phylogenetics are coming off age, and widely used programs such as PhyloBayes, Beast, MrBayes, PHYML, RAxML, are becoming ever more complex from the software engineering point of view, more time should be dedicated to test and verify such codes (see also the discussion in [87] about the use of an incorrect hastings ratio in several Bayesian programs).

Key Points

We intend to emphasize the following points:

Firstly, verification in phylogenetics is impossible, because the true evolutionary history of organisms is generally unknown. Secondly, given that the models are correct, there is no guarantee that the optimal tree, according to the model can be found, because most optimization problems associated with tree reconstruction are NP-hard. Thirdly, we review the most common techniques for inference of support values and provide extensive references to the respective literature and on-going debate about the interpretation of support values, while emphasizing that high support values

do not induce that a true bipartition has been recovered. Finally, and most importantly, we address issues pertaining to code verification of scientific applications, especially with respect to floating-point intensive applications, and provide examples for the fatal effects erroneous codes can have on modern science.

Acknowledgements

This work is entirely funded by the German Science Foundation (DFG). We would like to thank Barry G. Hall for sharing the history of the bug in the perl script with us (retraction of paper on bootstrap correlation with true trees), Florent de Dinechin for discussions on floating point arithmetics, and John Huelsenbeck for providing the reference to the mathematical error in the hastings ratio. We would also like to thank Bernhard Misof, Karen Meusemann and Stephen A. Smith for providing helpful comments on this manuscript.

Biographical Notes

Alexandros Stamatakis: Alexandros (Alexis) Stamatakis is currently the group leader of the scientific computing group at the Heidelberg Institute for Theoretical Studies (HITS gGmbH). He received a PhD in Computer Science from the Technical University of Munich in 2004. His research focus is on tools, algorithms, emerging parallel architectures, and high performance computing for evolutionary biology. Alexis is the main developer of RAxML, a code for large-scale phylogeny reconstruction under Maximum Likelihood.

Fernando Izquierdo-Carrasco: In 2009 Fernando completed a Master of Science in Bioinformatics at the University of Hamburg, Germany. Since 2010 he is a PhD student working with Alexis in the Exelixis Lab that is part of the scientific computing group at HITS. His research focus is on algorithms and tools for the inference of trees with tens of thousands of taxa under Maximum Likelihood.

HITS - Heidelberg Institute for Theoretical Studies: HITS (Heidelberg Institute for Theoretical Studies) is a private, non-profit research institute. It emerged in January 2010 from the EML

Research institute by change of name. HITS continues and substantially extends the successful scientific agenda of the EML Research institute. HITS will accommodate about 10 research groups from various fields of the natural sciences, including Mathematics and Computer Science. It is a highly inter-disciplinary organization, and is intended to closely collaborate with research labs world-wide. HITS gets its base funding from the Klaus Tschira Foundation, which was established in 1995. The institute is jointly managed by Dr. h.c. Klaus Tschira and Prof. Dr.-Ing. Andreas Reuter.

References

[1] Camin J.H, Sokal R.R. A method for deducing branching sequences in phylogeny. Evolution 1965; 19:311-326.

[2] Kluge A.G, Farris J.S. Quantitative phyletics and the evolution of anurans. Systematic Zoology 1969; 18:1-32.

[3] Kunin V., Goldovsky L., Darzentas. N., Ouzounis C.A. The net of life: Reconstructing the microbial phylogenetic network. Genome. Res. 2005; 15:954-959.

[4] Creevey C.J., Fitzpatrick D.A., Philip G.K., Kinsella R.J., O'Connell M.J., Pentony M.M., Travers S.A., Wilkinson M., McInerney J.O. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B 2004; 271(1557):2551-2558.

[5] Cavalli-Sforza L.L., Edwards A.W.F. Phylogenetic Analysis: Models and estimation procedures. Evolution 1967; 21: 550-570.

[6] Farris J.S. Methods for computing Wagner trees. Systematic Zoology 1970; 18: 374-385.

[7] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution 1981; 17(6):368-376.

[8] Foulds L.R., Graham R.L. The Steiner problem in phylogeny is NP-complete. Advances in Applied Mathematics 1982; 3:43-49.

[9] Roch S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006; 3(1):92-94.

[10] Goloboff P.A. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 1999; 15(4):415-428.

[11] Stamatakis A. RAxML-VI: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006; 22(21):2688-2690.

[12] Zwickl D. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Thesis, The University of Texas at Austin, 2006.

[13] Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 2010; 59(3):307-321.

[14] Ronquist F., Huelsenbeck J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003M 19(12):1572-1574.

[15] Lartillot N., Lepage T., Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 2009; 25(17): 2286-2288.

[16] Wiens J.J. Does adding characters with missing data increase or decrease phylogenetic accuracy? Systematic Biology 1998; 47(4):625-640.

[17] Hillis D.M., Huelsenbeck J.P., Cunningham C.W. Application and accuracy of molecular phylogenies. Science 1994; 264(5159):671-677.

[18] Strimmer K., von Haeseler A. Accuracy of neighbor joining for n-taxon trees. Systematic Biology 1996; 45(4):516-523.

[19] Yang Z. Phylogenetic analysis using parsimony and likelihood methods. Journal of Molecular Evolution 1996; 42(4):294-307.

[20] Guindon S., Gascuel O. A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52(5):696-704.

[21] Zwickl D.J., Hillis D.M. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 2002; 51(4):588-598.

[22] Nei M., Kumar S., Takahashi K. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proceedings of the National Academy of Sciences of the United States of America 1998; 95(21):12390-12397.

[23] Hillis D.M. Approaches for Assessing Phylogenetic Accuracy. Sytematic Biology 1995; 44 (1): 3-16.

[24] Huelsenbeck J.P., Lander K.M. Frequent Inconsistency of Parsimony Under a Simple Model of Cladogenesis. Systematic Biology 2003; 52(5):641-648.

[25] Rambaut A., Grass N.C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 1997; 13(3):235-238.

[26] Fletcher W., Yang Z. INDELible: a flexible simulator of biological sequence evolution. Molecular biology and evolution 2009; 26(8):1879-1888.

[27] Cartwright R.A. DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 2005; 21(Suppl 3): iii31-iii38.

[28] Puigbo P., Garcia-Vallve S., McInerney, J.O. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 2007; 23(12):1556-1558.

[29] Ripplinger J., Sullivan J. Does choice in model selection affect maximum likelihood analysis? Systematic Biology 2008; 57(1):76-85.

[30] Lanave C., Preparata G., Saccone C., Srio G. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution 1984; 20:86-93.

[31] Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 1994; 39(3):306-314.

[32] Stamatakis A., Ludwig T., Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 2005; 21(4):456-463.

[33] Joe Felsenstein. Inferring Phylogenies pp 269-274 Sinaur Associates Inc. 2004.

[34] Ogden T.H., Rosenberg M.S. Multiple sequence alignment accuracy and phylogenetic inference. Systematic Biology 2006; 55(2):314-328.

[35] Redlings B.D., Suchard M. Joint Bayesian estimation of alignment and phylogeny. Systematic Biology 2005; 54(3):401-418.

[36] Löytynoja A., Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 2008; 320(5883):1632-1635.

[37] Fleissner R., Metzler D., von Haeseler A. Simultaneous statistical multiple alignment and phylogeny reconstruction. Systematic Biology 2005; 54(4): 548-561.

[38] Sennbald B., Lagergren J. Probabilistic orthology analysis. Systematic Biology 2009; 58(4):411-424.

[39] Li L., Stoeckert C.J., Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research 2003; 13(9):2178-2189.

[40] Smith S. A., Beaulieu J., Donoghue M.J. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol Biol 2009; 9:37.

[41] Hartmann S., Vision T.J. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evolutionary Biology 2008;8:95.

[42] Maddison P.W. Gene Trees in Species Trees. Systematic Biology 1997; 46 (3): 523-536.

[43] Degnan J.H, Rosenberg N.A. Discordance of species trees with their most likely gene trees. Plos Genetics 2006; 2(5):e68.

[44] Hill A.W., Guralnick R.P., Wilson M.J.C., Habib F., Janies D. Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza. Infection, Genetics and Evolution 2009; 9(2):169-178.

[45] Salzberg S.L., Kingsford C., Cattoli G., Spiro D.J., Janies D.A., Aly M.M., Brown I.H., Couacy-Hymann E., De Mia G.M., Dung do H., Guercio A., Joannis T., Maken Ali A.S., Osmani A., Padalino I., Saad M.D., Savi V., Sengamalay N.A., Yingst S., Zaborsky J., Zorman-Rojs O., Ghedin E., Capua I. Genome Analysis Linking Recent European and African Influenza (H5N1) Viruses. Emerg Infect Dis. 2007; 13(5): 713–718.

[46] Larget B., Simon D.L. Markov chain Monte Carlo algorithms for the Bayesian analysis of

phylogenetic trees. Molecular Biology and Evolution 1999; 16(6):750-759.

[47] Felsenstein J. Confidence limits on phylogenies: An Approach using the bootstrap. Evolution 1985; 39: 783-791.

[48] Anisimova M., Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Systematic Biology 2006; 55(4): 539-555.

[49] Stuart A., Ord J.K., Arnold S. Kendall's advanced theory of statistics. New York: Oxford University Press; 1999.

[50] Li S. Phylogenetic reconstruction using Markov chain Monte Carlo. Ph.D. dissertation. Columbus: The Ohio State University; 1996.

[51] Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M., Phylogenetic inference. In: Hillis D. M., Moritz C., Mable B. K., editors. Molecular systematics. Sunderland, Massachusetts: Sinauer Associates; 1996. p. 407-514.

[52] Siddall M.E. Measures of support. In: DeSalle R., Giribet G., Wheeler W., editors. Techniques in molecular systematics and evolution. Basel: Birkhäuser Verlag; 2002. p. 80-101.

[53] Efron B. Bootstrapping methods: Another look at the jackknife. The Annals of Statistics 1979; 7(1):1-26.

[54] Hedges B.S. The Number of Replicates Needed for Accurate Estimation of the Bootstrap P Value in Phylogenetic Studies. Mol. Biol. Evol 1992; 9(2):366-369.

[55] Pattengale N.D., Alipour M., Bininda-Emonds O.R.P., Moret B.M.E., Stamatakis A. How many bootstrap replicates are necessary? Journal of Computational Biology 2010; 17(3):337-354.

[56] Price M.N., Dehal P.S., Arkin A.P. FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments. PloS ONE 2010; 5(3):e9490.

[57] Shimodaira H., Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and evolution 1999; 16:1114-1116.

[58] Susko E. Bootstrap support is not first-order correct. Systematic Biology 2009; 58(2):211-223.

[59] Susko E.. First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. Molecular Biology and Evolution 2010;

27(7):1621-1629.

[60] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. Equation of state calculations by fast computing machines. Journal of Chemical Physics 1953; 21:1087-1092.

[61] Hastings W.K. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970; 57:97-109.

[62] Nylander J.A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. AWTY(are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 2008; 24(4):581-583.

[63] Yang Z. Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. Molecular Biology and Evolution 2007; 24(8):1639-1655.

[64] Pagel A., Meade A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 2004; 53(4):571-581.

[65] Lartillot N., Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution 2004; 21(5):1095-1109.

[66] Wang H.C., Li K., Susko E., Roger A.J. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evolutionary Biology 2008; 8(1):331.

[67] Kolaczkowski B., Thornton J.W. A mixed branch length model of heterotachy improves phylogenetic accuracy. Molecular biology and evolution 2008; 25(6):1054-1066.

[68] Casella G., George E.I. Explaining the Gibbs sampler. American Statistician 1992; 46(3):167-174.

[69] Green P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995; 82(4):711-732.

[70] Douady C.J., Delsuc F., Boucher Y., Doolittle W.F., Douzery E.J.P. Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability. Mol Biol Evol 2003; 20(2):248-254.

[71] M.E. Alfaro, S. Zoller, F. Lutzoni. Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. Mol. Biol. Evol. 20(2):255-266, 2003.

[72] Wilcox T.P., Zwickl D.J., Heath T.A., Hillis D.M. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 2002; 25(2):361-371.

[73] Hillis D.M., Bull J.J. An empirical test of bootstrapping as a method for assesing confidence in phylogenetic analysis. Syst. Biol. 1993; 42(2):182-192, 1993.

[74] Henzinger T.A., Jhala R., Majumdar R., Sutre G. Software verification with BLAST. Proceedings of the 10th international conference on Model checking software, pp 235-239, Springer, 2003.

[75] Vardi M.Y., Wolper P. An automata-theoretic approach to automatic program verification. Proceedings of the First Symposium on Logic in Computer Science 1986; 322-331.

[76] Monniaux D. The pitfalls of verifying floating-point computations. ACM Transactions on Programming Languages and Systems (TOPLAS) 2008; 30(3):1-41.

[77] Defour D. Fonctions élémentaires: algorithmes et implémentations efficaes pour l'arrondi correct en double précision. PhD thesis, École Normale Supérieure de Lyon, Lyon, France, 2003.

[78] Björndalen J., Anshus O. Trusting floating point benchmarks-are your benchmarks really data independent?Applied Parallel Computing. State of the Art in Scientific Computing 2010; pp 178-188, Springer.

[79] Berger S.A, Krompaß D., Stamatakis A.. Performance, Accuracy, and Web-Server for Evolutionary Placement of Short Sequence Reads under maximum-likelihood, submitted 2010. Technical report available at http://arxiv.org/abs/0911.2852v1

[80] Miller G. A Scientist's nightmare: Software problem leads to five retractions. Science 2006; 384(5807):1856-1857.

[81] Pomeranz Krummel D.A, Altman S. Verification of phylogenetic predictions in vivo and the

importance of the tetraloop motif in a catalytic RNA. Proc. Natl. Acad. Sci. USA 1999; 96:11200-11205, 1999.

[82] Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith, S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 2008; 452(7188):745-749.

[83] Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguna J., Bailly X., Jondelius U. Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proceedings of the Royal Society B 2009; 276(1677):4261-4270.

[84] Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Queinnec E., Da Silva C., Wincker P., Le Guyader H., Ley S., Jackson D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. Phylogenomics revives traditional views on deep animal relationships. Current Biology 2009; 19(8):706-712.

[85] Schierwater B., Eitel M., Jakob W., Osigus H.J., Hadrys H., Dellaporta S.L., Kolokotronis, S.O., DeSalle R. Concatenated analysis sheds light on early metazoan evolution and fuels a modern" urmetazoon" hypothesis. PloS Biol 2009; 7(1):e20.

[86] Pick K.S., Philippe H., Schreiber F., Erpenbeck D., Jackson D.J., Wrede P., Wiens M., Alie A., Morgenstern B., Manuel M., Wörheide G. Improved phylogenomic taxon sampling noticeably affects non-bilaterian relationships. Molecular Biology and Evolution 2010; on-line doi: 10.1093/molbev/msq089

[87] Holder M.T., Lewis P.O., Swofford D.L., Larget B. Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics. Systematic Biology 2005; 54(6):961-965.