Published in final edited form as:
J Mol Evol. 2013 June ; 76(6): 351–352. doi:10.1007/s00239-013-9566-z.

Phylogenetic models of molecular evolution: next-generation data, fit and performance

David Posada
Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain.
dposada@uvigo.es

Next-generation omics techniques have clearly changed the prospects in molecular evolution. Today we have access to massive amounts of all types of molecular data, and the possibility of obtaining much more at a reasonable cost. More data is good news, but indeed it will require well-planned collection, tractable models, new analytical methods and additional computational power. In other words, we will need, more than ever, to be practical. In the field of phylogenomics, concerned as it is with the reconstruction and interpretation of evolutionary relationships from genome-scale data, computational pragmatism will be essential in order to efficiently analyze large number of sequences for hundreds or thousands of loci. First of all, phylogenomic data is error-prone and its collection should be carefully planned to avoid artifacts and to maximize the amount of information we can extract from it. Fortunately, recent advances in target enrichment techniques have allowed the 'easy' capture of thousands of conserved (Faircloth et al. 2012) or rapidly evolving (Lemmon et al. 2012) homologous regions for multiple species, facilitating the construction of large data sets. In the near future, these techniques will inevitably become cheaper and better.

Once genome and sequence alignments are in place, we need to think about proper models to describe the data. Statistical models of nucleotide substitution, codon substitution or amino acid replacement –models of molecular evolution *sensu lato*– have been traditionally used to compute probabilities of change among DNA or protein sequences given a certain amount of time and/or mutation rate and therefore to understand and interpret past evolutionary events. One of the first models of molecular evolution was proposed together with Charles R. Cantor by Thomas Jukes in 1969, who started this editorial column in JME in the late 80's. Since then, increasingly realistic (or at least we think so) models of evolution have been developed under the implicit assumption that they should result in more reliable phylogenetic inferences. For example, new models have been proposed that consider site interdependence, context-dependent changes, insertions, deletions and genomic rearrangements, and there is a growing trend to consider molecular structure and function, 'bringing back molecules into molecular evolution' (Wilke 2012).

While in general it is true that more realistic models offer a better statistical fit to the data, in my opinion it is also true that their performance is usually not carefully benchmarked. For example, while amino acid replacement models that consider biochemical profiles fit the data better and can affect phylogenetic inference on selected, particular empirical data sets, as far as I know their phylogenetic accuracy has not been systematically evaluated using

Europe PMC Funders Author Manuscripts

computer simulations. Still, the majority of phylogenetic analyses are based on relatively simple models that work pretty much at the sequence level. The reason for this gap between theory and practice is analytical efficiency, as sophisticated models can be computationally too expensive. However, in recent years there have been significant computational advances like better algorithms, faster processors, multicore computing and other parallelization approaches (e.g., Ayres et al. 2012; Price et al. 2010), that will facilitate more complex analyses.

In fact, the large amount of next-generation data available today can be seen as an opportunity or as a intimidation. The former because with large data sets we will be able to fit more complex models with more confidence and conceivably better performance. The latter because the amount of data could be overwhelming. For example, Chan and Ragan (2013) have argued that we should abandon model-based phylogenomic approaches in favor of purely informatic, more scalable alignment-free methods. However, such approaches lack any biological intuition –they altogether discard the concept of homology– and in my opinion they will not be able to generate any insight. So far, in phylogenetics we have learned nothing new from them. I am convinced that we do have the proper conceptual and technical tools to carry out high-quality evolutionary analyses for omic data sets, but, in order to couple them with the massive amount of data, a compromise is required between model complexity and performance. Besides biological realism, which indeed is important, we should always characterize the performance of new computational models and methods ‘in the light of’ the available computational power.

References

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol. 2012; 61:170. [PubMed: 21963610]

Chan CX, Ragan MA. Next-generation phylogenomics. Biol Direct. 2013; 8:3. [PubMed: 23339707]

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 2012; 61:717. [PubMed: 22232343]

Lemmon AR, Emme SA, Lemmon EM. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol. 2012; 61:727. [PubMed: 22605266]

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one. 2010; 5:e9490. [PubMed: 20224823]

Wilke CO. Bringing molecules back into molecular evolution. PLoS Comp Biol. 2012; 8:e1002572. [PubMed: 22761562]