

# Introduction

Ziheng Yang<sup>1</sup> and Philip C. J. Donoghue<sup>2</sup>

<sup>1</sup>Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT

<sup>2</sup>School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

Knowledge of absolute species divergence times is not only fascinating to evolutionary biologists in establishing the age of a species group, but also critically important to addressing a variety of biological questions. Absolute times allow us to place speciation events (such as the diversification of the mammals relative to the demise of the dinosaurs) in the correct geological and environmental contexts, and to gain a better understanding of speciation and dispersal mechanisms [1, 2]. They also allow us to characterize species richness and species diversification rates over geological periods. Estimated molecular evolutionary rates can also be correlated with life history traits, and are important to interpretation of the fast-accumulating genomic sequence data. Molecular clock methods are also employed widely in establishing the evolutionary history of viruses, including those related to human diseases.

The molecular clock hypothesis (rate constancy over time), proposed by Zuckerkandl and Pauling [3, 4], provides a powerful approach to estimating divergence times. Under the clock assumption, the distance between sequences grows linearly with time, so that if the ages of some nodes are known (for example, from the fossil record), the absolute rate of evolution as well as the absolute geological ages for all other nodes on the tree can be calculated. The past decade has seen exciting developments in clock dating methodologies, especially in the Bayesian framework, such as stochastic models of evolutionary rate change to deal with the sloppiness of the clock [5-7], flexible calibration curves to accommodate uncertain fossil information [8]. There have also been a surge of interest in probabilistic modelling of fossil presence and absence in the rock layers [9-11] and models of morphological character evolution [12] to use fossil data to generate time estimates, in analysis of either fossil data alone or in a combined analysis of data from both fossils and modern species.

However, many challenges remain, such as the relative merits of the different prior models of evolutionary rate drift (e.g., the correlated- and independent-rates models), the difference between user-specified time prior incorporating fossil calibrations and the effective

time prior used by the computer program, the partitioning of molecular sequence data in a Bayesian dating analysis, and the persistent uncertainty in time and rate estimation despite explosive increase in sequence data. Realistic models for the analysis of fossil data (either fossil occurrence data or fossil morphological measurements) are still in their infancy.

With the explosive growth of genomic sequence data, molecular clock dating techniques are increasingly being used to date divergence events in various systems. It is timely to review the recent breakthroughs in the field and highlight future directions. We thus organized a Royal Society Discussion Meeting titled “Dating species divergences using rocks and clocks”, on 9-10 November 2015, to celebrate Zuckerkandl and Pauling’s ingenious molecular clock hypothesis, to assess this fast developing field, and to identify the fundamental challenges that remain in developing molecular clock dating methodology. The meeting brought together leaders in the fields of geochronology, computational molecular phylogenetics, as well as empirical biologists who use molecular clock dating technologies to establish a timescale for some of the most fundamental events in organismal evolutionary history. This special issue is the result of that meeting.

The special issue consists of 14 reviews and original papers. In the first [13], we review molecular clock dating methods developed over the five decades, with a focus on recent developments and the Bayesian methods. The rest of the papers (13 of them) fall into three groups, (i) on features and analyses of rock and fossil data, (ii) on theoretical developments in molecular clock dating methods, and (iii) on applications of clock dating methodology to infer divergence times in various biological systems. In the first group, Holland [14] describes the structure of the fossil record. Everyone is familiar with the vagaries of fossil preservation, but the most significant ‘bias’ in the fossil record is perhaps the non-uniform nature of rock record within which it is entombed. Holland describes variations in preservation among lineages, environments, and sedimentary basins, across time and in terms of perception, and finally variation in sampling. While modern biogeography has been shaped by a reliance on the security of direct dating of tectonic events, like the opening and closure of oceans, Holland argues that the predictably non-uniform nature of the rock and fossil records is amenable to probabilistic modelling. The influential factors he has discussed may be important ‘covariates’ when one builds a model of fossil preservation and discovery. de Baets and colleagues [15] show that the high precision of radiometric dating belies the poor accuracy of the estimated age of biogeographic events, which are invariably long drawn-out episodes of tectonism that vary depending on the ecology of the clades. Nevertheless, the uncertainties associated with biogeographic calibrations can be modelled in much the same

way as in fossil calibrations and the two approaches, rather than competing, can be used in combination to constrain clade ages.

The papers in the second group explore theoretical issues of molecular clock dating or implement new models in Bayesian dating programs (e.g., MrBayes and BEAST). Note that in a modern clock-dating analysis, the calculation of the likelihood for the sequence data takes most of the CPU time but the theory is mature, with well-developed substitution models [16, 17]. Instead methodological developments have focussed on the other components of the Bayesian analysis, including the prior on divergence times (the time prior), the prior on substitution rates (the rate prior), and the model of morphological character evolution to incorporate fossil data. Rannala [18] discusses a number of issues in the construction of the time prior, such as the important distinction between the *user-specified time prior* and the *effective time prior* used by the computer program. Calibration densities for node ages specified by the user often do not satisfy the constraint that ancestors are older than descendants. Bayesian dating programs then automatically and without any warning truncate node ages that violate the constraint, altering the user-specified time prior to become the effective time prior. Rannala shows that conflicts among fossil calibrations, and between fossil calibrations and the molecular data, may lead to highly precise but grossly wrong time estimates, and warns that overly narrow posterior distributions of divergence times should be carefully scrutinized. dos Reis [19] discusses another issue with the time prior. He points out that the procedure for constructing a time prior suggested by Yang and Rannala [8] and known as the conditional reconstruction [20], which combines a birth-death model of cladogenesis with user-specified calibration densities, may generate ugly multi-moded prior densities for node ages. The difficulty is caused by the fact that phylogenetics dating analysis uses rooted trees while the birth-death model is operates on the so-called labelled histories [rooted trees with the interior nodes ranked by age, 21]. Lartillot et al. [22] explore the rate prior, in particular the independent-rates and the correlated-rates (Brownian-motion) models, for relaxing the molecular clock and allowing the rate to drift over branches on the tree. The authors propose a mixed relaxed clock model to combine the features of both models, assuming that the rate undergoes short-term independent fluctuations on the top of a Brownian long-term trend. Applied to date the divergences of mammals, the new model was found to help reduce the over-sensitivity of the posterior to the rate prior, especially when tip calibrations are used.

Drummond and Stadler [23] use the fossilized birth-death time prior [11] and a simple model of discrete morphological character evolution [24] to estimate the divergence times in

an integrated analysis of molecular sequence data for modern species and morphological characters for both modern and fossil species. They take a jackknife-style approach to estimate the age of each fossil in turn using the other dated fossils, based on two rich and well-characterized data sets. This investigation of the internal consistency of the method produced promising results, finding the posterior mean age of each fossil to be on average  $<2$  My from the midpoint age of the geological strata from which it was excavated. However, note that the credibility intervals of the posterior estimates tend to be large.

Ronquist and colleagues [25] analyzed a mammalian dataset to explore why the combined analysis of molecular and morphological data (known as total evidence dating), despite its theoretical advantages, has not closed the gap between rocks and clocks. The authors highlight that the conflict between morphology and molecules under standard models causes the dating method to generate ancient divergence time estimates. They discuss a number of influential factors, such as the inadequacy of the model of species diversification and fossil sampling used to construct the time prior (in particular, the failure to account for diversified sampling) and inadequacies in morphological models (in particular, the failure to account for correlations among characters). By assuming rapid diversification, rare extinction, or high fossil sampling rate, the authors were able to obtain highly congruent time estimates with a minimal gap between rocks and clocks. It may be an open question whether molecular time estimates should be judged by their match or conflicts with the fossil dates. Without the knowledge of the true ages, and given the general sensitivity of posterior time estimates to many aspects of the prior formulation, this question may be expected to haunt many molecular dating studies.

The sensitivity of posterior time estimates in a Bayesian dating analysis means that the details of the substitution models may also have substantial impact. Lee et al. [26] modelled the different clocks (rate-drift patterns) for different types of substitutions. For example, in mammals CpG dinucleotides have high mutation rates, which tend to be constant over calendar time, while other types of point mutations may be associated with meiosis so that the rate per year may vary if different species have different generation times. In an analysis of an intergenic region from 8 primate species, these authors found that the different groupings of substitution types affected the widths of the credibility intervals far more than the posterior means of divergence times. Scally [27] explores the mechanisms of mutation, confronting the incongruence between high rates inferred from fossil-calibrated divergence time analyses of catarrhines with the low mutation rates observed in human parent-offspring triplets. There appears to be a genuine slowing down in the evolutionary rate, obscured by incomplete

understanding of spermatogenesis. However, Scally argues that the precipitousness nature of mutation rate deceleration is perhaps an artefact of the fossil calibrations underestimating the timing of divergences among the great ape lineages. These expectations form an interesting contrast with the study of Cahill and colleagues [28], which explores the utility of a Pairwise Sequential Markovian Coalescent (PSMC) model [29] to date both the end of lineage panmixia and the cessation of gene flow among derivative lineages. The PSMC model is novel in that it accounts for the polymorphism and coalescent in extinct ancestors [30] and obviates the need for phased data. The authors' simulations suggest that the method can be used reliably for analysis of low coverage genome data. They show that while divergences among great apes and among bears show evidence of abrupt end to gene flow following the end of panmixia, more recently diverged clades like Chimpanzee and Bonobos exhibit evidence of continued gene flow post divergence.

The final three papers concern empirical application of divergence time methods to unravelling the evolutionary history of animal and plant lineages. Nicole Foley and colleagues [31] consider the timing of diversification of mammals which has achieved iconic status because of historic discord between the early applications of the molecular clock and traditional estimates based on the fossil record. Definitive records of placental mammals are limited to the Cenozoic and, hence, the fossil record has been interpreted to reflect an adaptative radiation following the demise of dinosaurs as part of the end-Cretaceous mass extinction. Molecular clock analyses have invariably indicated that placentals diverged in the Cretaceous or even earlier, though the extent of this prefossil history has been diminishing with the development of new methods, more data, more computational power, and a reinterpretation of the fossil record. Foley and colleagues suggest that a definitive timescale must await a definitive phylogeny. This appears some way off as the branching relationships involving fundamental clades like Laurasiatheria remain refractory to resolution, and interpretation of the fossil record is confounded by the same widespread convergent evolution of phenotypes that had compromised attempts at a mammalian phylogeny before the molecular revolution in systematics.

Susanne Renner and her colleagues [32] employ the Fossilised Birth-Death (FBD) model to establish a timescale for beeches and royal ferns, revealing a five-fold difference in the species turnover rate between these two clades. They attribute the low rate of turnover in royal ferns to their adaptation to low-nutrient marginal environments. This study highlights the power of the FBD model in facilitating the inclusion of fossil data that would otherwise be irrelevant to conventional node-calibrated molecular clock analyses. Finally, Lozano-

Fernandez and colleagues [33] employ the molecular clock to tackle the timing of arthropod terrestrialization. Arthropods are an ideal model for exploring this ecological transition because they have achieved this feat in a number of independent lineages. These natural experiments provide a basis for exploring the phenomena of convergence and parallelism in the physiological adaptation of a marine aquatic metabolism to the terrestrial environment. The authors show that fossil and molecular estimates of terrestrialization in arachnids are a close approximation, but the terrestrialization of myriapods significantly predates the oldest myriapid fossils and, indeed, the oldest records of plants which are assumed to terraformed the continents before the arthropod invasion. Lozano-Fernandez et al. suggest that this inconsistency may be an artefact of independent terrestrialization events in the two principal lineages of myriapods – as suggested by long-standing arguments that their tracheal systems have evolved in parallel, and not inherited from a common ancestor.

Molecular clock dating is hard when the clock is violated. The molecular sequence data contain information about the genetic distances, but not about the absolute times and absolute rates separately. In a Bayesian analysis, the resolution of distances into times and rates is achieved through the time prior and the rate prior, and through the analysis of morphological measurements of from fossil and modern species. The time-rate confounding effect combined with the well-known non-clock nature of morphological evolution means that posterior time estimates will remain sensitive to the priors and to the evolutionary models, even if whole genomes are sequenced from many species. We hope that the papers in this special issue have successfully demonstrated the challenges facing molecular dating studies, and have also highlighted areas where future methodological developments are most likely to be fruitful.

1. Stadler, T., *How can we improve accuracy of macroevolutionary rate estimates*. Syst. Biol., 2013. **62**: p. 321–329.
2. Rabosky, D.L., *Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees*. PLoS One, 2014. **9**(2): p. e89543.
3. Zuckerkandl, E. and L. Pauling, *Molecular disease, evolution, and genetic heterogeneity*, in *Horizons in Biochemistry*, M. Kasha and B. Pullman, Editors. 1962, Academic Press: New York. p. 189-225.
4. Zuckerkandl, E. and L. Pauling, *Evolutionary divergence and convergence in proteins*, in *Evolving Genes and Proteins*, V. Bryson and H.J. Vogel, Editors. 1965, Academic Press: New York. p. 97-166.
5. Thorne, J.L., H. Kishino and I.S. Painter, *Estimating the rate of evolution of the rate of molecular evolution*. Mol. Biol. Evol., 1998. **15**: p. 1647-1657.
6. Drummond, A.J., S.Y.W. Ho, M.J. Phillips and A. Rambaut, *Relaxed phylogenetics and dating with confidence*. PLoS Biol., 2006. **4**: p. e88.
7. Rannala, B. and Z. Yang, *Inferring speciation times under an episodic molecular clock*. Syst. Biol., 2007. **56**: p. 453-466.
8. Yang, Z. and B. Rannala, *Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds*. Mol. Biol. Evol., 2006. **23**: p. 212-226.
9. Tavaré, S., C.R. Marshall, O. Will, C. Soligos and R.D. Martin, *Using the fossil record to estimate the age of the last common ancestor of extant primates*. Nature, 2002. **416**: p. 726-729.
10. Wilkinson, R.D., M.E. Steiper, C. Soligo, R.D. Martin, Z. Yang and S. Tavaré, *Dating primate divergences through an integrated analysis of palaeontological and molecular data*. Syst. Biol., 2011. **60**: p. 16-31.
11. Heath, T.A., J.P. Huelsenbeck and T. Stadler, *The fossilized birth-death process for coherent calibration of divergence-time estimates*. Proc. Natl. Acad. Sci. U.S.A., 2014. **111**(29): p. E2957-E2966.
12. Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D.L. Murray and A.P. Rasnitsyn, *A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera*. Syst. Biol., 2012. **61**: p. 973-999.
13. Donoghue, P.C. and Z. Yang, *The evolution of methods for establishing evolutionary timescales*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
14. Holland, S., *The evolution of methods for establishing evolutionary timescales. The non-uniformity of fossil preservation*, 2016: p. in press.
15. De Baets, K., A. Antonelli and P.C.J. Donoghue, *Tectonic blocks and molecular clocks*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
16. Felsenstein, J., *Inferring Phylogenies*. 2004, Sunderland, Massachusetts: Sinauer Associates.
17. Yang, Z., *Molecular Evolution: A Statistical Approach*. 2014, Oxford, England: Oxford University Press. 512 pages.
18. Rannala, B., *Conceptual issues in Bayesian divergence time estimation*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
19. dos Reis, M., *Notes on the birth-death prior with fossil calibrations for Bayesian estimation of species divergence times*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
20. Heled, J. and A.J. Drummond, *Calibrated tree priors for relaxed phylogenetics and*

- divergence time estimation*. Syst. Biol., 2012. **61**: p. 138–149.
21. Edwards, A.W.F., *Estimation of the branch points of a branching diffusion process (with discussion)*. J. R. Statist. Soc. B., 1970. **32**: p. 155-174.
  22. Lartillot, N., M.J. Phillips and F. Ronquist, *A mixed relaxed clock model*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  23. Drummond, A. and T. Stadler, *Bayesian phylogenetic estimation of fossil ages*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  24. Lewis, P.O., *A likelihood approach to estimating phylogeny from discrete morphological character data*. Syst. Biol., 2001. **50**: p. 913-925.
  25. Ronquist, F., N. Lartillot and M.J. Phillips, *Closing the gap between rocks and clocks using total-evidence dating*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  26. Lee, H.-J., H. Kishino, N. Rodrigue and J.L. Thorne, *Grouping substitution types into different relaxed molecular clocks*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  27. Scally, A., *Mutation rates and the evolution of germline structure*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  28. Cahill, J.A., A.E.R. Soares, R.E. Green and B. Shapiro, *Inferring species divergence times using pairwise sequential Markovian coalescent modeling and low coverage genomic data*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  29. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences*. Nature, 2011. **475**: p. 493-496.
  30. Angelis, K. and M. dos Reis, *The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times*. Curr. Zool., 2015. **61**: p. 874-885.
  31. Foley, N.M., M.S. Springer and E.C. Teeling, *Mammal madness: Is the mammal tree of life not yet resolved?* Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  32. Renner, S.S., G.W. Grimm, P. Kapli and T. Denk, *Species relationships and divergence times in beeches: New insights from the inclusion of 53 young and old fossils in a birth-death clock model*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.
  33. Lozano-Fernandez, J., R. Carton, A.R. Tanner, M.N. Puttick, M. Blaxter, J. Vinther, J. Olesen, G. Giribet, G.D. Edgecombe and D. Pisani, *A molecular palaeobiological exploration of arthropod terrestrialisation*. Phil. Trans. R. Soc. B: Biol. Sci., 2016: p. in press.