Journal of Molecular Evolution (2021) 89:146–150
https://doi.org/10.1007/s00239-020-09971-z

COMMENTARY

Quest for the Best Evolutionary Model

Rafael Zardoya[1]

Received: 15 September 2020 / Accepted: 4 November 2020 / Published online: 17 November 2020

Abstract

In the early 1980s, DNA sequencing became a routine and the increasing computing power opened the door to reconstruct molecular phylogenies using probabilistic approaches. DNA sequence alignments provided a large number of positions containing phylogenetic information, which could be extracted using explicit statistical models that described the mutation process using appropriate parameters. Consequently, an active quest started for building increasingly improved (more realistic) statistical models of nucleotide substitution. The simplest model assumed that nucleotide frequencies were in equilibrium and one single category of substitutions. Subsequent models allowed either unequal nucleotide frequencies or separate rates for transitions and transversions. The HKY85 model (Hasegawa et al. in J Mol Evol 22:160, 1985) combined elegantly both options into a single model, which became one of the most useful ones and has been the choice in many molecular phylogenetic studies ever since. The use of improved substitution models such as HKY85 allows reconstructing more accurate and reliable phylogenies, which in turn provide robust frameworks for understanding how biological diversity evolved and for performing a wealth of comparative studies in different disciplines such as ecology, biogeography, developmental biology, biochemistry, genomics, epidemiology, and biomedicine.

Keywords Molecular phylogenetics · Maximum likelihood · Evolutionary models · Transitions · Transversions

All living organisms on Earth are related by descent from common ancestors (Darwin 1859) and the main goal of systematics is to disentangle their phylogenetic relationships (Wiley and Lieberman 2011). First phylogenetic trees were reconstructed based on morphological characters (this is still the case in paleontology) using cladistics (Hennig 1966) and maximum parsimony as optimality criterion (Fitch 1971). However, morphology-based phylogenies are normally based only on a restricted number of characters (Scotland et al. 2003) because many have to be discarded if they are not functionally independent, character states not always can be defined unambiguously, and homology (similarity due to common ancestry) is difficult to ascertain between distantly related taxa. Moreover, morphological characters experiencing similar selective forces are prone to convergence, thus producing homoplasy and misleading phylogenetic inference (Wake 1991).

The discovery that protein sequences accumulated amino acid changes at a constant rate over time (the so-called molecular clock) opened the possibility of using this evolutionary information to infer phylogenetic relationships (Zuckerkandl and Pauling 1965). Molecular sequences offered a vast number of independent characters and they could be compared among all living organisms. Moreover, most mutations are neutral due to genetic random drift (Kimura 1983) leading to reduced levels of homoplasy. All these valuable features motivated that molecular sequences have superseded morphological traits as the source data for the reconstruction of robust and reliable phylogenetic trees over the years. Furthermore, it was early on suggested that probabilistic methods such as maximum likelihood, although computationally demanding, could be the most powerful approach for phylogenetic inference based on molecular sequences (Cavalli-Sforza and Edwards 1967). The maximum likelihood optimality criterion searches for the phylogenetic tree (topology plus branch lengths) that best explains the observed alignment of sequences given an explicit statistical Markov model of molecular evolution. It

Handling Editor: **Aaron Goldman**.

✉ Rafael Zardoya
rafaz@mncn.cisc.es

[1] Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (MNCN-CSIC), José Gutiérrez Abascal, 2, 28006 Madrid, Spain

provides a statistical framework to phylogenetic inference and thus allows the application of well-known statistical tools in downstream analyses. In summary, by the end of the 1960s, the theoretical foundations for molecular phylogenetics were set but only a handful of molecular sequences were available and computing power could barely handle most simple maximum likelihood analyses.

The next decade started with a plethora of studies based on immunological techniques and protein electrophoresis assessing genetic variation within populations, such as that of Lewontin (1972), who showed that much of the human genetic variation is found within local populations and rejected the use of the race concept. Moreover, the 1970s witnessed the burst of RNA sequencing (Sanger et al. 1965), which culminated in the discovery of the domain Archaea (Woese and Fox 1977). In parallel, the popularization of molecular cloning techniques using restriction enzymes and plasmid vectors (Cohen et al. 1973), together with the advent of chain-terminating sequencing (Sanger et al. 1977) provided an accurate, robust, and routine methodology to obtain DNA sequences. Thereby, at the onset of 1980s, the complete human mitochondrial genome was sequenced (Anderson et al. 1981) and maximum likelihood algorithms to reconstruct trees based on nucleotide sequences were developed (Felsenstein 1981), demonstrating that molecular phylogenetics could effectively move forward from theory to practice. Many influential studies on molecular evolution (several here cited; see also other commentaries in this anniversary issue) were published in the *Journal of Molecular Evolution* during these years.

The study of Hasegawa et al. (1985) focused on dating the divergences of orangutans, gorillas, chimpanzees, and humans. A pioneering molecular work (Sarich and Wilson 1967) based on immunological distances had estimated that the split of gorillas and chimpanzees from humans occurred about five million years ago (Ma), challenging the commonly held paleontological view at that time that this divergence could have occurred as far back as 30 Ma. A lively debate started confronting molecular and paleontological evidences, and fostered the use of different types of molecular data (DNA-hybridization, restriction enzyme cleavage sites, protein electrophoresis, amino acid sequences) to provide an accurate estimate of divergence dates within hominids. Hasegawa et al. (1985) was the first phylogenetic analysis tackling this evolutionary question that was based on nucleotide sequences (complete mitochondrial genomes) and used maximum likelihood as method of phylogenetic inference. The study inferred rather young estimates for the separation of gorillas (3.7 ± 0.6 Ma) and chimpanzees (2.7 ± 0.6 Ma) from humans, which have not been confirmed later. Recent studies using probabilistic methods and large genomic data sets provide an estimate for the human-chimpanzee split between 4.98 and 7.90 Ma depending on the calibrations and the estimates of ancestral population size (Kumar et al. 2005; Amster and Sella 2016; Moorjani et al. 2016). Similarly, a phylogenetic analysis integrating paleontological and genomic data estimated the human-chimpanzee split between 6.9–7.9 Ma (Wilkinson et al. 2011).

Despite the study clearly underestimated divergence dates between apes, Hasegawa et al. (1985) has been highly influential (> 8,000 citations) because it contained a hidden jewel. In order to use a model of evolution that could best fit the sequence data, the authors made two important decisions. First, they took into account that in a protein-coding gene, most synonymous substitutions (implying no amino acid replacement) occur in third codon positions, and thus they estimated parameters of the model independently for first plus second *versus* third codon positions. Second, it had been observed previously that in mitochondrial DNA, nucleotide composition was highly biased (G was particularly underrepresented in the L-strand), and that transitions i.e., changes between purines (A G) or between pyrimidines (C T) were more frequent than transversions, which imply changing purines into pyrimidines or vice versa. Therefore, the authors built a statistical model, henceforth named HKY85, which in the so-called Q matrix (Fig. 1) estimated separately four nucleotide frequencies as well as two instantaneous rates of substitution for transitions and transversions, respectively (Hasegawa et al. 1985).

The quest for best evolutionary models had started with the simplest model assuming equal base frequencies and one single type of mutations, and continued adding parameters that distinguished different types of mutation or unequal base frequencies (Fig. 2). The HKY85 model improved all previous models while offering a good compromise between bias and variance in the estimation of the parameters. Hence, it has been the choice in many molecular phylogenetic studies ever since. The sophistication of evolutionary models continued after HKY85, until the most complex evolutionary model possible, the general time reversible (GTR) was built (Tavaré 1986). Afterwards, it was realized that evolutionary models would need also to consider the heterogeneity

$$Q=\begin{bmatrix} -\mu\,(\kappa\pi_G+\pi_\gamma) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu\,(\kappa\pi_T+\pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu\,(\kappa\pi_A+\pi_\gamma) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu\,(\kappa\pi_C+\pi_R) \end{bmatrix}$$

Fig. 1 The Q instantaneous rate matrix for the HKY85 model. The order of the nucleotides for columns and rows are A, C, G, and T. Each (i,j) entry represents the rate at which a nucleotide i is substituted by a nucleotide j (in a Markov model this rate is equal for the change j to i; i.e., the reversibility property). The diagonal is used to constrain the row sums of the matrix to equal zero. π = nucleotide frequencies; μ = mean instantaneous substitution rates; k = transition/transversion ratios; γ = pyrimidines; R = purines

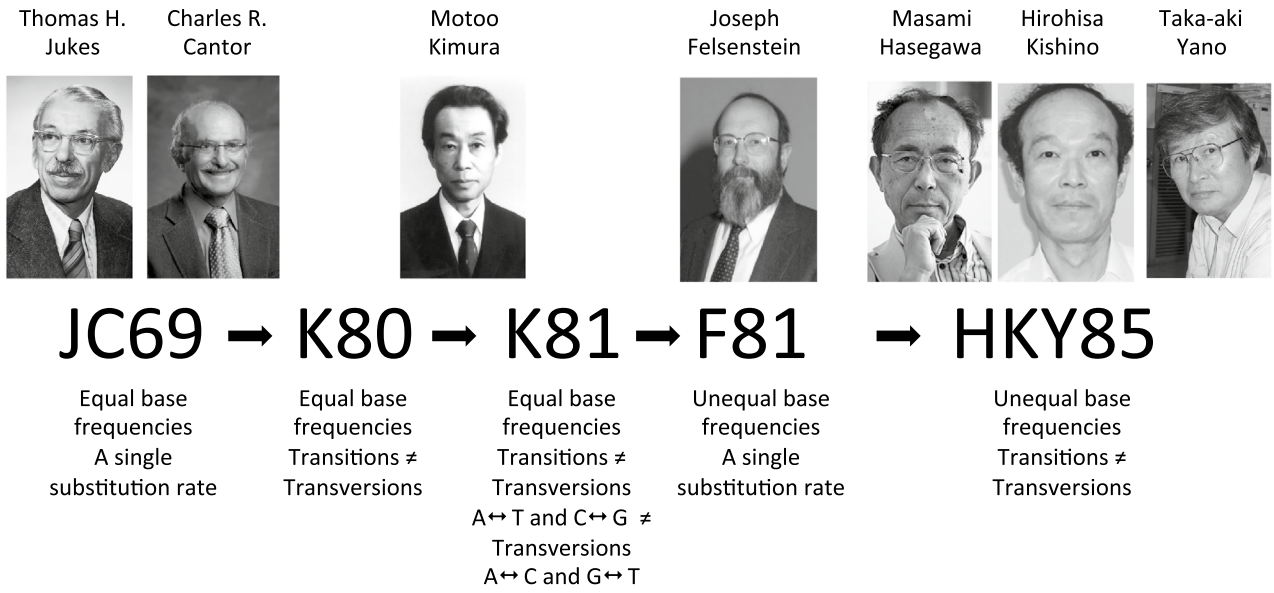

Fig. 2 The quest for the best evolutionary model. The simplest nucleotide substitution model (JK69; (Jukes and Cantor 1969) was improved in the early 1980s by adding parameters that either assumed different types of substitution (K80, K81; Kimura 1980, 1981), unequal base frequencies (F81; (Felsenstein 1981) or both (HKY85; (Hasegawa et al. 1985)

of substitution rates across the sequence, which can be incorporated into the model by estimating the proportion of invariable sites (Hasegawa and Horai 1991), the alpha parameter of a gamma distribution (Yang 1993), or both (Gu et al. 1995). Given the variety of models of nucleotide substitution available, the Akaike information criterion (Akaike 1973) has been suggested for selecting the one that best fit the data (Posada and Buckley 2004). Furthermore, the same criterion can be used to select optimal partition schemes of the data (Lanfear et al. 2014).

The build of models of amino acid replacement has followed a parallel historical development. In this case, the number of changes between the 20 amino acids makes the Q matrix really complex, and thus researchers normally have opted to use empirical matrices that summarize the frequencies of amino acid replacements observed in large data sets such as mtREV (Adachi and Hasegawa 1996), mtART (Abascal et al. 2007) and mtZoa (Rota-Stabelli et al. 2009) for mitochondrial data and JTT (Jones et al. 1992), WAG (Whelan and Goldman 2001), and LG (Le and Gascuel 2008) for nuclear data.

At the end of the 1980s, the advent of automated Sanger sequencing (Ansorge et al. 1987), the popularization of the polymerase chain reaction (Saiki et al. 1988), and the design of versatile primers to amplify genes in many different living organisms (e.g., Kocher et al. 1989) greatly accelerated the acquisition of DNA sequence data for molecular phylogenetics in the 1990s. Moreover, at the turn of the century phylogenetic methods came of age, first by the incorporation of likelihood ratio tests that started the possibility of contrasting evolutionary hypotheses (Huelsenbeck and Rannala 1997) and afterwards by the application of Bayesian inference (Yang and Rannala 1997; Huelsenbeck et al. 2001). The latter allowed the use of empirical mixture models for across-site heterogeneities (Lartillot and Philippe 2004), the implementation of relaxed molecular clocks (Drummond and Suchard 2010), and triggered a burst of phylogenetic comparative methods (Revell 2012), among other innovations.

Since the advent of high-throughput sequencing technologies in the last decade, the new field of phylogenomics has emerged, allowing the reconstruction of phylogenies based on genomic sequences and thus a vast number of characters (Lemmon et al. 2012; McCormack et al. 2012). Nonetheless, this new field is not exempt of challenges. Genomes encode numerous gene families and a first serious problem encountered is to separate unambiguously orthologs (gene copies due to speciation) from paralogs (gene copies due to duplication), as only the former can be used to reconstruct species trees. The concatenation of multiple genes renders robust phylogenetic trees, although it is computationally intensive and poses modeling challenges. Moreover, it disregards single gene tree information, which could be incongruent due to diverse evolutionary phenomena. This is particularly worrisome when inferring phylogenetic relationships among closely related taxa, and new methods of phylogenetic reconstruction based on coalescence models have been devised to account for incomplete lineage sorting,

hybridization, and recombination, although they need to be improved in the coming years as they are computationally highly demanding (Jiang et al. 2020).

The possibility of reconstructing the Tree of Life as first envisioned by Darwin (1859) is closer than ever. Moreover, as more whole genomes become available throughout the Tree of Life, phylogenetic comparative methods will pave the way to link genotype and phenotype variation, thus decisively contributing to a better understanding of the evolutionary processes and mechanisms underpinning the origin and maintenance of biological diversity (Smith et al. 2020).

Compliance with Ethical Standards

Conflict of interest The author has no conflicts of interest to declare that are relevant to the content of this article.

Human and Animal Rights and Informed Consent The research does not involve human participants and/or animals. No clinical research was conducted and thus, no informed consent was required.

References

Abascal F, Posada D, Zardoya R (2007) MtArt: a new model of amino acid replacement for arthropoda. Mol Biol Evol 24:1

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) 2nd International Symposium on Information Theory. Budapest: Akadémiai Kiadó, Budapest, pp. 267–281

Amster G, Sella G (2016) Life history effects on the molecular clock of autosomes and sex chromosomes. Proc Natl Acad Sci USA 113:1588

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res 15:4593

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550

Cohen SN, Chang ACY, Boyer HW, Helling RB (1973) Construction of biologically functional bacterial plasmids in vitro. Proc Natl Acad Sci USA 70:3240

Darwin C (1859) On the origin of species. John Murray, London

Drummond AJ, Suchard MA (2010) Bayesian random local clocks, or one rate to rule them all. BMC Biol 8:114

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406

Gu X, Fu YX, Li WH (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol 12:546

Hasegawa M, Horai S (1991) Time of the deepest root for polymorphism in human mitochondrial DNA. J Mol Evol 32:37

Hasegawa M, Kishino H, Yano T-a (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160

Hennig W (1966) Phylogenetic systematics. Univeristy of ILLINOIS PRESS, Urbana

Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310

Jiang X, Edwards SV, Liu L (2020) The Multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. Syst Biol 69:795

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8:275

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc Natl Acad Sci USA 86:6196

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB (2005) Placing confidence limits on the molecular age of the human–chimpanzee divergence. Proc Natl Acad Sci USA 102:18842

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:82

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095

Le SQ, Gascuel O (2008) An Improved general amino acid replacement matrix. Mol Biol Evol 25:1307

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst Biol 61:727

Lewontin RC (1972) The apportionment of human diversity. In: Dobzhansky T, Hecht MK, Steere WC (eds) Evolutionary biology. Springer, New York, pp 391–398

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res 22:746

Moorjani P, Amorim CEG, Arndt PF, Przeworski M (2016) Variation in the molecular clock of primates. Proc Natl Acad Sci USA 113:10607

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53:793

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol 3:217

Rota-Stabelli O, Yang Z, Telford MJ (2009) MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol Phylogenet Evol 52:268

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487

Sanger F, Brownlee GG, Barrell BG (1965) A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol 13:373

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463

Sarich VM, Wilson AC (1967) Immunological time scale for hominid evolution. Science 158:1200

Scotland RW, Olmstead RG, Bennett JR (2003) Phylogeny reconstruction: the role of morphology. Syst Biol 52:539

Smith SD, Pennell MW, Dunn CW, Edwards SV (2020) Phylogenetics is the new genetics (for most of biodiversity). Trends Ecol Evol 35:415

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 17:57

Wake DB (1991) Homoplasy: the result of natural selection, or evidence of design limitations? Am Nat 138:543

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691

Wiley EO, Lieberman BS (2011) Phylogenetics: theory and practice of phylogenetic systematics, 2nd edn. Wiley-Blackwell, Hoboken

Wilkinson RD, Steiper ME, Soligo C, Martin RD, Yang Z, Tavaré S (2011) Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst Biol 60:16

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA 74:5088

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396

Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14:717

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357