Probabilistic models of eukaryotic evolution: time for integration

Nicolas Lartillot

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard Lyon 1, F-69622 Villeurbanne Cedex, France

In spite of substantial work and recent progress, a global and fully resolved picture of the macroevolutionary history of eukaryotes is still under construction. This concerns not only the phylogenetic relations among major groups, but also the general characteristics of the underlying macroevolutionary processes, including the patterns of gene family evolution associated with endosymbioses, as well as their impact on the sequence evolutionary process. All these questions raise formidable methodological challenges, calling for a more powerful statistical paradigm. In this direction, model-based probabilistic approaches have played an increasingly important role. In particular, improved models of sequence evolution accounting for heterogeneities across sites and across lineages have led to significant, although insufficient, improvement in phylogenetic accuracy. More recently, one main trend has been to move away from simple parametric models and stepwise approaches, towards integrative models explicitly considering the intricate interplay between multiple levels of macroevolutionary processes. Such integrative models are in their infancy, and their application to the phylogeny of eukaryotes still requires substantial improvement of the underlying models, as well as additional computational developments.

1. Introduction

For multiple reasons, reconstructing the global evolutionary history of eukaryotes is a challenging task. First, the combination of a very ancient adaptive radiation, mutational saturation, uneven rates of molecular evolution, heterogeneous nucleotide compositions across lineages and distant outgroups, represents a dangerous cocktail of factors jointly increasing the risk of reconstruction artefacts. Second, the evolutionary history of single genes does not directly follow the history of the organisms bearing them [1]. This is also true in Bacteria and Archaea, where lateral gene transfers also play an important role. However, the evolutionary history of eukaryotes also has its own specific features. In particular, multiple endosymbiotic events have taken place during the history of the kingdom [2,3]. These events have resulted in the coexistence of multiple genetic compartments in the eukaryotic cell. They also formally introduced reticulations in the evolutionary history of eukaryotes, while inducing highly correlated patterns of gene loss and gene transfer across multi-gene families.

From a methodological point of view, eukaryotic evolution therefore combines some of the most challenging problems of current phylogenetic research. More globally, molecular phylogenetics has always been a difficult task, in particular at large evolutionary scale, where systematic errors, or tree reconstruction artefacts, have been the rule rather than the exception (reviewed in [4,5]). Such challenging empirical problems represent a very stimulating context for new statistical developments. In this respect, the history of molecular phylogenetics clearly illustrates how the development of statistical practice can both be stimulated by the empirical questions and, in turn, have a decisive influence on the way these problems are addressed. Thus, the introduction of probabilistic inference in phylogenetics was in part motivated by the problem of systematic errors in tree reconstruction. The original description of the long-branch attraction (LBA) artefact was provided as a proof of the statistical inconsistency of maximum parsimony [6], as well as an argument in favour of
the use of statistically well-behaved probabilistic methods. Three years later, the first likelihood approach based on an explicit stochastic model of nucleotide substitution was introduced [7]. As can be seen across much of the phylogenetic literature published since then, likelihood and Bayesian methods are not immune from artefacts, which in their case betray fundamental problems in the design of the underlying stochastic models of sequence evolution. Accordingly, much effort has been devoted to the development of better nucleotide substitution and amino acid replacement models (e.g. [8–10]). Meanwhile, statistical practice itself has evolved, in particular with the introduction of Bayesian inference [11–14]. In turn, Bayesian inference and Monte Carlo methods have opened the way to new approaches for designing more flexible models, with hierarchical structures or using non-parametric priors.

The aim of this article is to discuss some of these statistical developments, especially those that are of higher relevance in the context of current research on eukaryotic evolution. Developments along two main directions will be considered: first, the introduction of improved models of sequence evolution, accounting for modulations of the substitution process across sites and across lineages (§2); second, the development of models explicitly accounting for the evolutionary history of multi-gene families by duplication, loss and horizontal gene transfers (§3). Thus far, these two lines of applied statistical research have proceeded in parallel, without much cross-talk between them. Ultimately, however, the two perspectives should be combined. From this fusion, a more general integrative modelling framework could emerge, representing a unified statistical framework for reconstructing the patterns and understanding the processes of the global macroevolutionary history of eukaryotes (§4).

2. Models of sequence evolution

(a) Systematic errors in phylogenomics

The first eukaryotic phylogenies were obtained based on single universal genes, most notably ribosomal RNA and protein-coding genes such as elongation factors or tubulins (e.g. [15–17]). It was quickly realized, however, that the phylogenetic signal contained in single genes is insufficient for resolving large-scale phylogenies, typically resulting in stochastic errors (i.e. errors due to small sample size) in the estimated trees. To overcome this limitation, subsequent phylogenetic reconstructions have most often relied on large sets of genes analysed in conjunction. After some developments using targeted genes specifically amplified across a broad set of taxonomic groups (e.g. [18]), the domain of molecular phylogenetics across eukaryotes was revolutionized by the use of expressed sequence tags (reviewed in [5]).

For the most part, multi-gene data have been analysed under the so-called supermatrix paradigm [19]. Typically, sets of putatively orthologous sequences are identified and concatenated into a supermatrix, to which standard tree reconstruction methods are then applied. The use of concatenated sets of genes almost entirely alleviates the problem of stochastic errors. However, as is now well recognized, this does not imply that the resulting phylogenies, even if statistically well supported, are accurate. In a context where genetic sequences are mutationally saturated, systematic errors (i.e. wrong trees with high statistical support) are frequent, most often taking the form of long-branch attraction artefacts [6]. Early versions of the phylogeny of life [15], showing fast-evolving lineages (such as microsporidia, diplomonads or Entamoeba) emerging early in the evolution of eukaryotes, represent a typical example of the LBA phenomenon [20].

In the context of probabilistic methods, which are normally consistent under correct model specification, tree reconstruction artefacts are caused by the violation of one or several assumptions of the model by the empirical data. This observation clearly emphasizes the need to develop better models of sequence evolution. In this respect, heterogeneities in the process of sequence evolution across positions and across lineages turn out to be two particularly important aspects of the substitution process that need to be properly modelled.

Across lineages, the problem can be easily understood and has been pointed out early on (e.g. [21]). Similar compositional biases increase the probability of convergent accumulation of identical nucleotides or amino acids in otherwise unrelated species. Statistical models assuming compositional stationarity of the substitution process across the phylogeny will not correctly anticipate the high probability of convergent evolution induced by these compositional effects and will therefore tend to artefactually group species with similar composition.

A more subtle problem is induced by heterogeneities of the amino acid replacement process across sites. The protein alignments typically used in phylogenetics are characterized by strong site-specific selective constraints. These constraints are generally stable over large phylogenetic distances, being mostly induced by long-term structural conservation. A typical position of a multiple sequence alignment of orthologous genes at large evolutionary scale is thus either strictly conserved or has undergone repeated substitutions within a restricted subset of amino acids. The small number of amino acids acceptable at a typical fast-evolving position in turn implies a high probability of convergent evolution towards the same amino acid at that position in unrelated lineages. Models that do not correctly account for site-specific amino acid propensities and that tend to anticipate a broad spectrum of acceptable amino acids at each position of the alignment will underestimate the probability of convergent evolution induced by this phenomenon. In practice, this appears to be a major cause of LBA artefacts when reconstructing deep phylogenies using classical models relying on empirical amino acid replacement matrices [22].

(b) Towards more flexible models of sequence evolution

Over the past 10 years, several models accounting for site- and time-qualitative heterogeneities of the sequence evolutionary process have been developed. From a methodological point of view, these developments have illustrated an important distinction between two main statistical and computational strategies.

On one side, some developments have relied on what could be called parametric models. Such models are characterized by a well-defined likelihood function depending on a fixed, relatively small, number of parameters. The likelihood can be evaluated, either analytically or numerically, under arbitrary parameter configurations, making it relatively straightforward to either maximize the likelihood or combine it with a prior and sample from the posterior distribution. Parametric models have represented the most common type of model
classically considered in many domains of phylogenetics and statistical evolutionary genetics. The fact that they are easily implemented both in a maximum likelihood and in a Bayesian context may have given the impression that the choice between these two paradigms is merely a philosophical one, mostly depending on one’s personal stance with respect to the status of priors and the interpretation of probabilities.

On the other side, more sophisticated models, involving several levels of distributions, non-parametric priors, continuous-time processes, among other features, have been progressively introduced and explored in phylogenetics. Such models do not always have a clearly defined number of parameters. Instead, as much as possible, they tend to replace what would be naturally seen as parameters from a classical perspective by random variables produced by a higher level stochastic process. For that reason, they may be generically referred to as non-parametric models. Specific examples include the Brownian relaxed molecular clock [23], the use of Gaussian processes for modelling correlations of the rate of evolution across the three-dimensional structure of proteins [24], or the Dirichlet process models, further discussed in §2c. Computationally speaking, non-parametric modelling schemes generally require integration over complex distributions, most often using Monte Carlo methods. In part for that reason, such models are typically and most conveniently formulated in Bayesian terms. Thus, the choice of Bayesian inference, in this specific context at least, may be motivated primarily by considerations about flexibility in model design, rather than by philosophical preferences.

These two alternative modelling strategies, parametric or non-parametric, have their respective theoretical and practical merits. Briefly, classical parametric models are more compact and more easily implemented. Since they allow for direct numerical evaluation and maximization of the likelihood, they are particularly convenient in the context of model comparison and hypothesis testing. On the other hand, as will be further argued in §2c, in terms of model design, they tend to be overly rigid. Conversely, non-parametric approaches offer much more flexibility for accommodating the fine-grained patterns that are present in empirical sequences. On the other hand, they are computationally more demanding, requiring extensive Monte Carlo developments.

(c) Modelling variation across sites

The relative merits of the parametric and the non-parametric modelling strategies can be illustrated by contrasting the solutions that have been proposed for modelling two types of heterogeneities across sites, either in the overall rate of substitution (slow-evolving versus fast-evolving sites) or in more qualitative aspects of the substitution process, such as nucleotide or amino acid propensities.

Heterogeneity in substitution rate across sites is usually captured by formalizing site-specific rates as random effects. These random effects are integrated over a parametric distribution, usually a gamma distribution, whose variance is controlled through its shape parameter [25]. In practice, the gamma distribution is discretized into a small number of quantiles, typically four [8], and the likelihood at each site is averaged over these discrete categories. In this way, an efficient solution to the problem of modelling across-site variation in substitution rate can be achieved, while staying within the classical framework of parametric models—in this case, using only one extra parameter and with not much more than a fourfold increase in computational time.

Why could not this approach be used for modelling variation in nucleotide or amino acid propensities across sites? One major problem is the relatively large dimensionality of the effect to be modelled as site-specific 4- or 20-dimensional vectors, making discretization approaches much less convenient. Another problem, also related to the dimensionality of the effect, is the difficulty finding a parametric distribution that would correctly describe, even qualitatively, the empirical distribution of nucleotide or amino acid propensities across sites.

An alternative approach, still within the limits of the classical parametric paradigm, relies on finite mixture models. Mixture models attempt to capture variation across sites through summation over a small number of nucleotide or amino acid propensity vectors. Much effort has been spent in this direction [26–31]. However, a series of empirical observations suggest that finite mixtures are not sufficient. First, the empirical fit shows a substantial increase with the number of components even when this number is high [28,32], indicating that even relatively rich mixtures do not capture a sufficient fraction of the available variation. Forcing the number of categories to a relatively small number is possible. In some cases, compact empirical mixtures appear to bring a real improvement in the robustness of phylogenetic reconstructions [33]. However, it is not clear whether they would be sufficient in more complex situations. Finite mixtures also tend to result in highly rugged landscapes in the mixture space, characterized by a large number of sharp local maxima, making it difficult to implement standard numerical optimization or MCMC.

Fundamentally, the true distribution of amino acid propensities (and probably of many other complex random effects) across sites does not appear to be correctly described by a simple parametric distribution, nor by a small number of sharp peaks. Instead, it looks like a complex and multi-modal continuum, which apparently requires to be modelled using more sophisticated approaches. In this direction, Bayesian non-parametric methods have represented a particularly convenient and flexible solution.

Formally, the starting point of non-parametric Bayesian inference is to model variation across observations as random effects drawn from some unknown distribution, about which no strong parametric assumption is made—hence the non-parametric nature of the model. As is usual in a Bayesian framework, a prior is invoked to express this uncertainty. Here the prior is thus over the space of all possible distributions of random effects across sites.

Technically, a prior having as its support the space of all possible distributions would be impossible to define (intuitively, the space of all distributions is just too large). On the other hand, since any distribution can be approximated with arbitrarily high accuracy using a sufficiently fine-grained mixture of point masses, one can use instead a prior over a much smaller space of discrete distributions, or infinite mixtures. Even if the true distribution is not discrete, for sufficiently large datasets, this should result in good statistical consistency properties. The Dirichlet process, which is the non-parametric prior most often used in Bayesian inference, represents such a prior over all infinite mixtures [34].

Infinite mixtures are still too complex to be explicitly instantiated in a Markov chain Monte Carlo (MCMC) framework: it would take an infinite amount of time to completely
specify them. However, for finite datasets, and once each observation (each site of the alignment) is allocated to a component of the mixture, the number of occupied components is finite. In the case of the Dirichlet process, the exact composition of the rest of the mixture, i.e. the infinite part made of all unoccupied components, can be analytically integrated away, such that, in practice, it is possible to work only on the marginal distribution over the finite occupied part of the mixture. Sampling a configuration of this occupied part from the prior can be done using the so-called Chinese restaurant algorithm. Simple MCMC algorithms sampling from the posterior distribution have also been developed [35,36]. These MCMC algorithms essentially explore all possible ways to cluster the sites of the alignment, or more generally the observations gathered in the dataset, such that the total number of clusters is not fixed a priori.

Altogether, Bayesian non-parametric models based on Dirichlet processes therefore take a dual form. On the practical side, they look like simple MCMC clustering algorithms. On the conceptual side, and through the average over all clusterings visited during the MCMC, they effectively implement a Bayesian non-parametric device averaging over essentially arbitrary distributions of random effects across sites. Dirichlet processes can in principle be used in order to model any multi-dimensional real-valued aspect of the model that one might want to consider as variable across observations. In practice, they have been applied to variation in equilibrium frequency vectors over the 20 amino acids [37], and then generalized to other aspects of the substitution process, including the substitution rate [38], the non-synonymous to synonymous rate ratio [39], or the amino acid fitness profiles in the context of mutation–selection codon models [40].

In all cases, this appears to result in a substantial improvement of the fit of the model. In the context of phylogenomic analyses based on expressed sequence tag (EST) data, models using Dirichlet processes to account for variation in nucleotide or amino acid propensities across sites lead to more robust and more accurate reconstructions [22,33,41–46].

(d) Modelling variation across lineages

Concerning compositional variation across lineages, parametric models, assuming branch-specific parameters (typically, branch-specific equilibrium frequency vectors) have been considered and implemented in a maximum-likelihood framework [47,48]. Dimension reduction methods have been proposed more recently to limit the number of branch-specific parameters invoked by the model [49]. Alternatively, models considering partitions of branches into a small set of categories, each characterized by its vector of equilibrium frequencies, have been developed [50]. Partitioning branches raise the question of how to choose among all possible partitions. Here again, MCMC approaches turn out to be the most convenient practical choice, automatically averaging over all partitions by sampling from their posterior distribution. The resulting time-heterogeneous models have been shown to be more accurate than classical time-homogeneous and time-reversible models, in the context of several key phylogenetic problems, in particular concerning the position of eukaryotes relative to Archaea [51,52].

The time-dependent substitution models mentioned thus far proceed from a parametric perspective, in the sense that they associate sets of parameters (typically, equilibrium frequency vectors over nucleotides or amino acids) to the branches of the tree, possibly clustering branches into a small number of partitions. Such branchwise models are a pragmatic and computationally convenient approach to model variation through time. However, they are conceptually problematic. Branches do not represent a natural segmentation of evolutionary times. Instead, they are defined only retrospectively, based on which lineages have turned out to survive and to be sampled. Thus, from a forward-modelling perspective, there is absolutely no reason why parameters of the substitution process should be seen as constant along each branch, but distinct from one branch to the next. Instead, variation through time should ideally be modelled in terms of continuous-time processes. Some developments in this direction have already been proposed, either using compound Poisson processes [53,54], or multivariate Brownian processes [55]. These developments have not yet reached the stage where they could be routinely used for addressing large-scale phylogenetic reconstruction problems. The Brownian models have thus far been exclusively applied to comparative questions under a fixed pre-defined topology [56,57]. Clearly, more computational developments are needed concerning these modelling strategies. Nevertheless, in the long term, these developments are conceptually important, illustrating how it is possible to move away from a parametric perspective, in favour of a more flexible non-parametric paradigm fundamentally formulated in terms of hierarchies of stochastic processes.

(e) Phylogenomics using supermatrices: where do we stand?

Altogether, the development of models of sequence evolution accounting for pattern heterogeneity both through time and along the sequence has globally resulted in significantly more accurate phylogenomic reconstructions. On the other hand, current models and methods are far from perfect. Artefacts are still often observed even under the best model configurations currently available. Models jointly accounting for site- and time-heterogeneities are certainly needed. Simple versions have already been developed [58]. However, they are not yet really applicable on large datasets. Mechanical insights could be very useful here, as a guide for deriving empirically better fitting amino acid replacement processes [40,59]. Finally, current implementations are computationally very intensive, obviously reaching their limits at a time when the size of the datasets considered in phylogenomics is exploding.

In spite of these limitations, the extent to which modern phylogenomics has been able to reconstruct the phylogeny of eukaryotes is clearly impressive. A global view of the phylogeny of eukaryotes does seem to emerge, progressively organizing the diversity of eukaryotic species into large supergroups (reviewed in [60]). Still, artefacts have played such an important role in molecular phylogenetics over the past 20 years that some caution is warranted. After all, the current picture of eukaryotic evolution is the result of analyses of supermatrices of tens of thousands of aligned positions, many of which are mutationally saturated and therefore contain mostly noise—noise whose statistical distribution is poorly described even by the most sophisticated models of sequence evolution currently available. In this context, it is difficult to rule out the possibility that the observed phylogenetic
resolution, however consistent across studies, may in fact be entirely contributed by the slow build-up, across very large datasets, of residual systematic errors. A similar argument can be made concerning many other phylogenetic questions currently explored using phylogenomic approaches, most notably the deep phylogeny of metazoans [43,44,61,62].

Critical here, in order to settle these questions in a more definitive manner, would be some independent evidence based on shared molecular events other than point substitutions. In this direction, the phylogenetic distribution of ultra-conserved elements [63], retroposons [64], or the analysis of synteny and gene rearrangements [65], have already been explored. Alternatively, and as will now be discussed in §3, the evolutionary history of gene families could represent a promising source of phylogenetic signal.

3. Gene trees and species trees

The evolution of multi-gene families may be seen as one of the most critical blind spots of the supermatrix paradigm. In supermatrix analyses, gene duplications, losses and transfers are mostly considered as a nuisance, the focus being on (putatively) orthologous genes. This is potentially problematic for several reasons. First, focusing on sets of orthologous genes amounts to throwing away a large fraction of the protein-coding sequences of a typical eukaryotic genome. More critically, phylogenetic reconstructions can be very sensitive to the inclusion, in the supermatrix, of even very few paralogues misidentified as orthologues [66–68]. Conversely, patterns of duplication and loss of genes may be informative about the phylogenetic history of the organisms. A typical example is the duplication of haemoglobin resulting in the alpha and beta chains, representing a shared derived character of vertebrates except lampreys. Thus, moving away from the supermatrix paradigm and properly accounting for the evolution of multi-gene families would appear to be the next important move to make in phylogenomics.

Starting with the work of Arvestad et al. [69], increasingly sophisticated statistical models of gene family evolution have been proposed [70–76]. More recently, these models have been used to reconstruct species phylogenies [73,77]. Briefly, the idea is to rely on a hierarchical model consisting of two integrated levels of stochastic processes: a first level describing the process of diversification of gene copies by duplication, loss and transfer along the species tree; and a second level representing the evolution of genetic sequences by point substitution along the gene trees produced by the first level. Conditioning this hierarchical model on a set of multiple sequence alignments across independent gene families then allows joint estimation of the species tree and of the underlying gene trees (reviewed in [78]). In the following, such models will be broadly referred to as gene-species reconciliation models.

In addition to duplications, losses and transfers, incomplete lineage sorting (ILS) is an additional cause of discrepancy between gene trees and species trees. ILS is typically addressed using probabilistic models based on the multi-species coalescent [79–81]. In their general structure, multi-species coalescent models are similar to the gene-species reconciliation models accounting for gene duplications, losses and transfers. In fact, ILS and diversification of gene families can be simultaneously considered in one single modelling framework [82].

Reconciliation approaches represent an example of what could be called integrative modelling, in the sense that multiple layers of evolutionary processes are integrated in the model. An important practical consequence of this integration is that multiple levels of parameter estimation are performed simultaneously: in the present case, simultaneous estimation of gene and species trees [79,80,83]. This is in sharp contrast to more classical stepwise approaches, such as supertree methods (reviewed in [19]) or, in the case of incomplete lineage sorting, so-called shortcut coalescent methods [84,85]. Such stepwise approaches typically first reconstruct gene trees and then combine those trees into a consensus species tree. One main problem with stepwise approaches is the propagation of errors along the chain of inference. Trees estimated based on single-gene sequence alignments are generally dominated by stochastic error, so that using them as a starting point for the subsequent consensus step, without accounting for the uncertainty associated with their estimation, is problematic (as pointed out by Gatesy & Springer [86], in the context of shortcut coalescent approaches). Integrative models, by contrast, optimally integrate multiple sources of information, while duly taking into account the uncertainty prevailing at each level. In the context of gene-species reconciliation methods, this means two things. First, the species tree is reconstructed using not just a (noisy) point estimate of each gene tree. Instead, it is informed by the entire likelihood function induced by the sequence alignment over the space of all possible gene trees. Second, and reciprocally, the estimated gene trees automatically integrate some refinements induced by the estimated species tree [77,87].

Apart from gene-species reconciliation models, other recent successful applications of the idea of integrative modelling include total evidence dating [88], the molecular comparative method, jointly inferring substitution rates, divergence dates and quantitative traits [56] and statistical epidemiology using viral phylogenies [89,90]. In all these cases, the intermediate level (gene trees, viral phylogenies, divergence times) is not sufficiently robustly inferred based on the primary signal coming from empirical data, either because of limited sequence length or, in the case of molecular dating, because of intrinsic lack of identifiability of times and rates. Integrating out the intermediate level through the use of a hierarchical model represents a principled way to deal with this uncertainty.

Because of their complex hierarchical structure, gene-species reconciliation models are difficult to implement and to apply to phylogenetic estimation. As in the case of the non-parametric models of sequence evolution, Monte Carlo methods, combined with dynamic-programming algorithms, have to be recruited for that purpose (first introduced in a fully integrative form in [70]). Even so, current computational approaches seem to reach their limit. In particular, proposing efficient Monte Carlo updates of the species tree is difficult, as this also requires the proposal of correlated moves on gene trees, making them well adapted to the new species tree. Several computational shortcuts or statistical alternatives have been proposed to circumvent these rather challenging MCMC mixing issues (reviewed in [78]).

4. Achieving the integration

Gene-species tree reconciliation models represent a recent key development in statistical evolutionary genetics. They are
very likely to become one of the methods of reference in the near future. On the other hand, presently, the approach does not seem to be completely ready for routine application to large-scale phylogenetic problems. This is in part due to the computational limitations mentioned in the previous paragraph. However, there are other potential problems, concerning the models themselves, which are still inadequate in several respects. In the following, two possible model developments are briefly outlined: first, how reconciliation methods should integrate the more flexible models of sequence evolution discussed in §1, to avoid reconstruction artefacts; second, and more specifically concerning eukaryotes, how to correctly account for the specific patterns of gene transfer induced by endosymbioses.

(a) Models of sequence evolution and of gene diversification

Thus far, there has been an obvious disconnect between the two threads represented by the progress made on models of sequence evolution, on the one hand (§1), and the parallel development of explicit gene-species reconciliation models, on the other hand (§2). Currently, reconciliation methods use simple models of sequence evolution, in particular, ignoring qualitative modulations of the substitution process across sites and through time. Given how inadequate these simple models have turned out to be in the context of supermatrix analyses, particularly at large evolutionary scale, there are good reasons to believe that they are equally problematic in the context of gene-species reconciliation methods.

Typically, systematic errors at the level of gene trees will tend to create an excess of discordance between the estimated gene trees and the species tree, which will in turn translate into an overestimation of transfer, duplication and loss rates [77]. What is perhaps less clear is to what extent errors in gene tree reconstruction induced by poor models of sequence evolution will have an impact on the estimation of the species tree, compared, for instance, with other more direct violations of the model of gene family diversification (e.g. inadequate description of the variation in rates of gene duplication, loss and transfer, both between species and across gene families).

In principle, the hierarchical nature of the model makes it relatively easy to combine these independent statistical developments, simply by using current models of sequence evolution at the lower level of the hierarchy. In practice, however, doing this will suffer from the current limitations already encountered in the supermatrix paradigm, which does not work well for large data matrices.

Interestingly, in the context of a reconciliation model, the site- and time-modulations of the sequence evolutionary process can be modelled in a more fine-grained manner, compared to what is currently done in a supermatrix context. In particular, these modulations can now be made explicitly dependent on the specific history of each gene copy (such as already developed in [87]). In the case of transfers, the substitution process followed by a given gene sequence should in principle depend on the organism in which this gene copy resides at any given time. Accounting for such host-dependent modulations of the substitution process could be done by attaching distinct parameter values to each branch of the species tree or, as suggested above (§1), by invoking continuous-time stochastic processes along the branches of the species phylogeny. Doing so could potentially result in a better resolution of the gene-specific histories. This is but one example of the subtle modulations that can be more finely captured once the two problems of sequence evolution and gene diversification are jointly considered.

(b) Endosymbiotic transfers

Another challenge, in the context of eukaryotic evolution, is to correctly account for the specific patterns of gene diversification and genome evolution induced by endosymbiotic gene transfers (EGTs) [91]. EGTs are a fundamental aspect of the evolutionary history of eukaryotes. They could even be considered as the hallmark of eukaryotic evolution at the genetic level, being at the origin of mitochondria, primary and secondary plastids, plus additional more restricted events (reviewed in [2]). In each case, the endosymbiosis was followed by the loss of many genes from the donor, combined with a massive gene transfer to the nucleus of the host. In the case of secondary endosymbiotic events, at least three genetic compartments (plastid, nucleus of the donor and nucleus of the host) are involved. Correctly inferring and interpreting the patterns of EGTs is therefore an essential aspect of the reconstruction of the global evolutionary history of eukaryotes.

Thus far, patterns of EGT have been empirically analysed using stepwise approaches, based on gene trees that were first reconstructed using standard phylogenetic methods (e.g. [92]). As mentioned above, however, such stepwise approaches are fraught with the difficulties resulting from the stochastic errors accumulated at the level of gene tree estimation [93,94]. In the present case, the problem is made even more intractable by the presence of secondary transfers, duplications or losses, blurring an already noisy picture of gene family evolution [95].

Integrative models, in contrast, would provide a robust methodological framework for integrating over the uncertainty about the evolutionary history of gene families. However, this requires some adaptations of the reconciliation models, to account for the specific patterns of gene loss and transfer induced by endosymbiosis. Indeed, one key problem of current reconciliation models is that they consider only one-at-a-time gene transfers, such that the donor and the acceptor are each time a different pair of species. This one-at-a-time property leads to a convenient factorization of the likelihood, essentially making the evolutionary histories independent across gene families, conditional on the species tree. However, gene transfer and gene loss associated with endosymbioses are strongly correlated—simultaneous and massive right at the time of the endosymbiosis, followed by a long series of events occurring between a donor and a host that are tightly linked even long after endosymbiosis. Such patterns are not adequately described by models assuming conditional independence across gene families—at least, not conditional on a species tree.

There is one simple way to make EGTs conditionally independent given the evolutionary history of species, thus making the problem amenable to current probabilistic modelling strategies: simply by representing endosymbioses as reticulation events, in what would then be a species network, instead of a species tree. This reticulated structure could then be used as the backbone on which to model the independent history of each gene family. In this direction, recent developments of reticulated models of evolution suggest interesting new modelling ideas [96], which could easily be adapted to the
specific needs of eukaryotes, in particular, by accounting for the inherent asymmetry of endosymbiotic events.

(c) Towards fully integrative models of eukaryotic evolution

The modelling developments just suggested in §4b are essential for addressing the specific problems raised by eukaryotic evolution. Conversely, such developments would open new possibilities concerning highly interesting questions about the macroevolutionary history of eukaryotes. In particular, the question of the origin of eukaryotes, and the role of endosymbiosis in this origin, can be naturally formalized in the context of such an integrated and reticulated model. Model comparison, using Bayes factors or reversible-jump Monte Carlo, could be used to formalize the question of how many prokaryotic species have contributed to the original eukaryote and where these contributors branch in the archaeal and eubacterial trees. As another example, the substitution process undergone by a given gene critically depends on the molecular evolutionary regime prevailing not just in the specific organism but also in the specific genetic compartment in which this gene resides at any given time. Integrative approaches accounting for host- and compartment-specific substitution patterns, again in a reticulated context, should thus be able to make optimal use of this information scattered across the sequence alignments of multi-gene families. Relying on such models would not only improve the estimation of the global history of eukaryotes, but it would also touch upon several interesting comparative questions. In particular, there are important connections to explore here, with recent speculative ideas about the evolution of genome content and architecture as a function of the life history of the organisms and the population genetic regimes prevailing in their multiple genetic compartments [97]. Such connections can be empirically tested by recruiting the molecular comparative method [55] in this integrative framework.

5. Conclusion

In many respects, most of the ideas explored above are not new. Many have been repeatedly discussed in the literature. Some of them, in particular concerning the question of whether eukaryotic evolution should be represented by a tree or by a network, have been at the core of long-standing philosophical disputes and controversies [1]. However, the empirical leverage for addressing these questions has thus far been insufficient because, fundamentally, the statistical methods used did not match the challenge.

The main argument of the present discussion is that it is now, or will soon be, possible to formalize these fundamental questions about the evolutionary history of eukaryotes in a much more robust statistical context. The crucial step is to move away from stepwise approaches and classical parametric models and to fully embrace more flexible non-parametric and hierarchical modelling strategies. Ultimately, the hierarchy of macroevolutionary processes should be modelled in one single integrated probabilistic framework, to optimally investigate the full range of their interactions.

Of course, the computational challenges raised by the prospect of developing such integrative models are formidable. At first sight, it is not totally clear how bluntly integrating current site- and time-heterogeneous models of sequence evolution with current models of gene family diversification, while introducing additional complexities to fully account for the role of endosymbioses, will lead to anything other than just compounding the computational limitations already faced by all of these current models—given that each of them cannot even correctly handle the genome-wide datasets over hundreds of taxa that are now considered in recent phylogenomic studies.

On the other hand, the problem could be turned on its head and could perhaps be reconsidered as follows. Although Bayesian inference using MCMC has been a revolution, current implementations are nearly exclusively reliant on simple algorithmic strategies, such as standard MCMC methods based on the Metropolis–Hastings algorithm [98], which was developed more than 60 years ago. Perhaps this simple MCMC paradigm is precisely what is fundamentally reaching its computational limits, and this, simultaneously in all current applications in phylogenetics. Therefore, in order to proceed, what is more fundamentally needed is a quantum jump-like advance in the development of MCMC algorithms. In this direction, new approaches, such as sequential Monte Carlo, already applied to phylogenetics [99], or particle MCMC [100], appear to have substantially better scaling properties than classical MCMC. Investing in these more sophisticated Monte Carlo approaches, combined with other algorithmic developments and relying on current parallelization opportunities, could be a key strategic move, making it possible to further integrate the multiple threads of probabilistic inference currently integrated in phylogenetics and evolutionary genetics. Possibly, approximate methods based on data resampling [101–103] could also be considered as a more practical alternative to exact likelihood or Bayesian inference.

Clearly, there is a strategic wager here, whose outcome is more than uncertain. However, the bet is probably worth a gamble, as it could lead to a viable integrative modelling paradigm, representing the ultimate approach to embrace and fully capture the complex interplay between endosymbiosis, ecology and molecular evolution characterizing the macro-evolutionary history of eukaryotes.

Competing interests. I declare I have no competing interests.

Funding. This study was financially supported by French National Research Agency grant no. ANR-10-BINF-01-01 ‘Ancestrome’.

Acknowledgements. The author wishes to thank two anonymous reviewers for their useful comments on this manuscript.

References


