12. Phylogenomics

EEOB563

Spring 2025

> ”The stream of heredity makes phylogeny: in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream.” George G. Simpson, 1945.

1 Preliminary considerations

The term ”phylogenomics” was first used by Jonathan Eisen in application to predicting gene function. Soon it was ”kidnapped,” to describe large-scale phylogenetic studies. The ultimate goal of phylogenomics is to reconstruct the evolutionary history of species through their genomes. The distinction between a ”phylogenetic” and a ”phylogenomic” analysis is somewhat murky. In most cases, both types of studies use only a small fraction of the genome. Nevertheless, access to genomic data could potentially alleviate previous problems of phylogenetics due to sampling effects by expanding the number of characters from a few thousand to tens of thousands. With this increase, the emphasis of phylogenetic inference is shifting from the search for informative characters to filtering non-informative ones and the development of better reconstruction methods for using genomic data. Despite holding considerable promise, the phylogenomic approach also has potential problems that stem from the limitations of current phylogenetic reconstruction methods.

2 Methods of phylogenomic inference

Phylogenomic reconstruction methods can be divided roughly into sequence-based methods and methods that are based on whole-genome features (Fig 1). While the latter methods are intuitively appealing, their application remains limited. As a consequence, methods based on multiple-sequence alignment, for which there is an extensive methodological background, currently remain the methods of choice.

2.1 Sequence-based methods

2.1.1 Number of characters vs. number of taxa

Sequence-based phylogenomic methods are based on comparisons of primary sequences, and phylogenetic trees are inferred from multiple-sequence alignments (MSA). Usually, MSA are performed first and phylogenetic inference later, although some approaches combine these two steps. A long-standing debate in phylogenetics was whether the greatest improvement in accuracy results from an increased number of characters (sequence length) or species. While phylogenomics is a clear increase in the number of characters, an appropriate taxonomic representation remains essential.

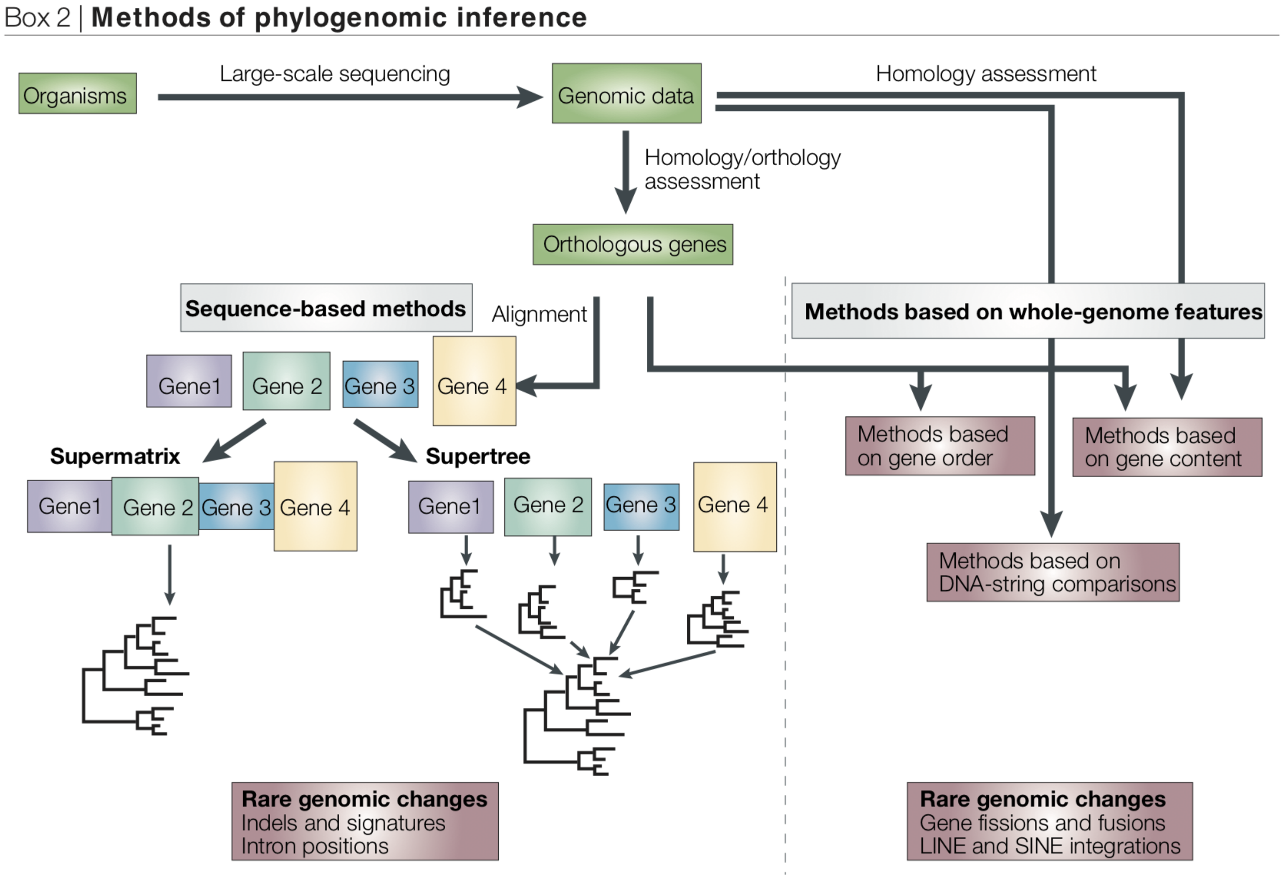

Figure 1: Different methods of phylogenomic inference circa 2005. From Delsuc et al. Nat. Rev. Genet. 6: 361–375

Several high-profile genome-scale studies with taxon-poor phylogenomic datasets produced erroneous results. It's also important to note that assembling phylogenomic datasets rich in both species and genes is usually associated with a large percentage of missing data, as some genes are not represented in all species. The effects of missing data on phylogenetic inference are not fully understood, but appear to be relatively minor.

2.1.2 Concatenation, Summary, Co-estimation

Three approaches can be used for phylogenomic reconstruction. In the **"supermatrix"** approach, alignments of individual genes are concatenated together and analyzed using standard sequence-based methods. The sequences of genes that are not represented for some species are coded as missing data. In the **"supertree"** approach, trees are build from individual gene alignments and combined together using consensus-like methods. One such method that used to be popular in the parsimony framework was the matrix representation using parsimony (MRP). Supertree methods may be considered a special case of "summary methods" that also include shortcut coalescence method, which take a similar two-step approach and estimate a species tree from a collection of gene trees. While the steps are similar, the shortcut coalescence method are based on multi-species coalescence (MSC) theory and take branch lengths into account. Finally **full MSC methods** attempt to co-estimate gene and species trees. These methods are highly computationally demanding and hence limited in the size of analyzed datasets.

2.2 Methods based on whole-genome features

2.2.1 Gene content, gene order, and conserved synteny

Unlike classical sequence-based approaches, methods that are based on gene content and gene order do not rely on a multiple-sequence alignment step. However, they do still depend on homology assessment. Changes in gene content and gene order within genomes result in characters with billions of possible states, as compared with only four states for nucleotide sequences. As a result, they are less prone to homoplasy by convergence or reversal, and might therefore potentially represent good phylogenetic markers, as long as they contain enough phylogenetic information. Although they use different character types to those that are used in sequence-based approaches, these methods nevertheless use standard tree-reconstruction algorithms. One of the methods that became popular in the last few years is based on conserved synteny (without regard to gene order). In particular, these studies found 29 groups of genes whose chromosomal linkages are conserved across bilaterians, cnidarians, and sponges (Simakov et al. 2022). Furthermore, they proposed that fusion-with-mixing of such groups provides a polarized and practically irreversible change that should be informative for phylogenetic reconstructions. A follow-up study used such fusion-with-mixing events to support the placement of Ctenophora as the sister group to the rest of the Metazoa (animals) (Schultz et al. 2023). However the latter study remains controversial due to the small number of synteny groups between animals and outgroups.

2.2.2 The DNA string approach

Finally, another approach derived from whole-genome features, which is not dependent on homology or orthology assessment, is based on the distribution of oligonucleotides (DNA strings) in genomes. This approach is based on the observation that each genome has a characteristic ”signature” with regard to these strings; these are defined, for example, as the ratio between observed dinucleotide frequencies and those expected if neighboring nucleotides were chosen at random. The few methods currently used for this approach show that it is possible to extract phylogenetic signal using this oligonucleotide ”word usage” (e.g., link) However, DNA string methods remain marginal in phylogenetic reconstruction and potentially suffer from saturation problem when a more divergent taxa are included. However, one current application of such methods is contamination detection in genomic data using k-mers.

2.3 Rare Genomic changes

Genomes can also be studied using the traditional methodology used by comparative morphologists by looking for shared complex characters – known as rare genomic changes (RGCs) – that have a very low probability of being the result of convergence. As well as gene order, such RGCs include intron positions, insertions and deletions (indels), retroposon (SINE and LINE) integrations, and gene fusion and fission events.

3 Future challenges of phylogenomics

3.1 Stochastic vs. Systematic errors

Because it uses many characters, phylogenomics leads to a drastic reduction in stocastic or sampling errors associated with the finite length of single genes in traditional phylogenetic analyses. It is not, however, immune to systematic errors, which are dependent on data quality and inference methods. The emergence of phylogenomics therefore brings the field full-circle to the roots of molecular phylogenetic analysis, with potential pitfalls in the form of tree-reconstruction artifacts, which were among the earliest issues faced by phylogenetics.

3.1.1 Misleading effects of inconsistency

The use of large datasets generally results in a global increase in the resolution of phylogenetic trees, as measured by standard statistical indices such as bootstrap percentages or posterior probability values. However, obtaining a strongly supported tree does not necessarily mean that it is correct. These statistical indices only assess sampling effects, and give an indication of tree reliability that is conditional on the data and the method. So, if the method does not correctly handle properties of the data, an incorrect tree can receive strong statistical support.

3.2 Sources of inconsistency

There are several causes of model inadequacy, as several simplifying assumptions are generally made. These include the independence of evolutionary changes at different sites and the homogeneity of the nucleotide-substitution process. For example, compositional biases can result in the artifactual grouping of species with similar nucleotide or amino-acid compositions, because most methods assume the homogeneity of the substitution process and the constancy of sequence composition (stationarity) through time. Moreover, variations in the evolutionary rate among species can cause the well-known and widespread long-branch attraction (LBA) artefact. Here, high evolutionary rates increase the chance of convergence and reversal, leading to the artifactual grouping together of fast-evolving species.

3.3 Reducing the perils of inconsistency

In the pre-genomic era, the most straightforward way of detecting an erroneous phylogenetic result was the observation that incongruent trees are obtained from different genes. However, when whole-genome information is used, an erroneous result that is due to inconsistency is difficult to ascribe as only one tree is produced. As increasingly large datasets are analyzed, the probability increases that strongly supported but erroneous groupings remain undetected. Therefore, using the most accurate tree-reconstruction methods available is of the utmost importance. The use of the most complex models will also reduce the probability of becoming inconsistent, as they will fit the data better. However, despite the fact that simulation studies indicate that probabilistic methods are relatively robust to violations of the model's primary assumptions, this might not hold in extreme cases. More realistic models of sequence evolution are therefore needed. Research in this area is continuing, with the most recently developed models relaxing the assumption of independence among sites by taking into account the occurrence of context-dependent, multiple-nucleotide and structurally constrained substitutions. Likelihood models that relax the assumptions of homogeneity and stationarity have also been designed to handle sequences with heterogeneous composition. Focusing on the rarest substitution events is another way of improving phylogenetic inference.

3.4 Computational challenges

3.4.1 The problem of tree space

The problem of three space is the same as for single-gene phylogenetic analysis, but multiplied by the extra time that is required to analyze longer sequences. Just a reminder, there are more possible unrooted topologies for 53 taxa than there are atoms in the Universe. Imagine that we are running a Bayesian analysis for a few million generation and let's assume that every generation we sampled a new tree. How much progress did we make in sampling the Universe-worth of trees? A cubic millimeter of air contains about $5 * 10^{16}$ atoms, 10,000,000,000 times more than the number of trees we sampled!

3.4.2 Divide and conquer

Given the immense size of tree space, analysis of taxa-rich datasets makes phylogenetic analysis computationally problematic. The resolution of large phylogenetic problems can be tackled by using "divide-and-conquer" strategies. These methods break the dataset down into smaller subsets (that is, a fraction of the species), infer optimal trees for these subsets, and finally combine these trees into a larger tree. Thus the combination of the supermatrix and supertree approaches is needed for reconstructing phylogenomic trees with thousands of species to eventually obtain a full picture of the tree of life.

3.5 Environmental considerations

Phylogenomic research by definition involves the analysis of large and complex data sets and, therefore, requires the use of large-scale computational resources. Although such research has enabled major advances in our understanding of organismal biology and evolution, it also entails large energy consumption, causing increase greenhouse gas (GHG) emissions associated with outdoor air pollution and climate change. The yearly electricity usage of data centers and high-performance computing facilities (200 TWh; Jones 2018) already exceeds the consumption of countries such as

Ireland or Denmark (Primary Energy Consumption by World Region 2021) and it rises rapidly, especially with the introduction of AI. As biologists, we should be aware of the environmental impact of our research and be willing to adapt our practices to minimize it. This includes a rational choice of computational tools, avoiding unnecessary redundancy, a careful study design, use of simpler models of evolution, and recycling of previous results, among others.