Abstract

In this scribe note, we will attempt to describe in detail the phylogenomics approach to inferring evolutionary relationships between organisms. More analytically, we will start by explaining what created the need for phylogenomics and precursors of this approach. Thereafter, we will try to give a stricter definition on phylogenomics and describe it in greater extent. Lastly, we will focus our description to one of the most powerful phylogenomics techniques so far, SPIDIR [7].

More explicitly, in Section 1 we make a short introduction to phylogenomics, address the reasons of its necessity and describe in brief the path that led to its development. Afterwards, in Section 2, we delve to the concepts of phylogenomics in more detail and outline the main reasons that cause the weaknesses of the current tools. This last part will help us transit smoothly into Section 3, the description of SPIDIR tool which captures mainly the content of lecture 26. In this section, we will give a concise description of SPIDIR, discuss how it overcomes the known problems and mention a potential future outlook of this tool.
1 Introduction: Path to phylogenomics

The similarity of molecular mechanisms of the studied organisms suggest that all organisms on Earth evolved from a common ancestor. So, the need of depicting the relationships between organisms in a hierarchical manner has arisen a long time ago. The study of the hierarchical relationships between organisms is widely known as phylogeny. Initially, morphological characters (from both living and fossilized organisms) have been used to infer such relationships [4].

However, with the explosion of molecular data and the advent of high-throughput analysis era, scientists started thinking of using data from a microscopic rather from a macroscopic point of view. In addition, there arose the potential of a more informed scientific analysis that can be obtained by information representing genotype (molecular data) rather than phenotype (morphological data) which was used until that time.

A sceptic reader might have wondered what would be the essential need of a phylogenetic representation instead of the obvious intrusiveness of man to understand evolutionary history. Surprisingly, there is a plenty of reasons [3], some of which can be outlined as follows:

- map pathogen strain diversity for vaccines
- assist in epidemiology of infectious diseases and genetic defects
- biodiversity studies
- understanding microbial ecologies
- aid in prediction of functions of novel genes

Let’s insist a bit more on this last aim. Accurate prediction of gene function has been a lasting pursue in many areas of biological research. It is almost obvious that this could help in the decryption of underlying biological mechanisms and give rise to a new era of prevention and treatment for diseases.

The question now is what determines the gene function? Although not always the case, sequence is the prime factor of determining gene function. However, genes can become similar in sequence either as a result of convergence (or stated otherwise as homoplasy or analogy) or descent with modification from a common ancestor (homology). In essence, convergent evolution refers to species with similar sequences not coming from a common ancestor and prevails when species have similar ecological roles and natural selection has shaped analogous adaptations. For example, the wings of birds and insects which followed completely different evolutionary routes are analogous structures [2].

So, in order to be clear, phylogeny takes into account only evolutionary driven similarities, namely homologies. In consequence the two most interesting types of homology are orthology and paralogy (there are also others as xenologs, positional homologs (see Table 2 of [1]) and ohnologs). Orthologs are genes in different species that evolved from a common ancestral gene by speciation while paralogs are genes created by duplication within a genome. Normally, orthologs retain the same function in the course of evolution while paralogs tend to attain new functions although related to the original one [5]. So, as gene duplications (paralogy) are frequently characterized by functional divergence, dividing genes into groups of orthologs and paralogs can improve the accuracy of functional predictions [1].

It was Tatusov [5] who first attempted to find clusters of orthologous groups in his paper considered to be a precursor to phylogenomics. More explicitly, he named these orthologous groups as Clusters of
Orthologous Groups (COGs) and defined a COG to be comprised of *individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages*. However, clustering is only a way of classifying levels of similarity and is not an accurate method of inferring evolutionary relationships [1].

## 2 What is phylogenomics

So, after that, Eisen came with his seminal paper ([1]) introducing the concept of phylogenomics. In short, he defined phylogenomics to be the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. These methods were not only developed to infer the divergent order of species but to deal with the genome-wide orthology problem as well [6].

But what made phylogenomics to clearly outstand from the classical phylogeny methods, was the theory of gene tree evolution inside the broader framework of evolution of species corresponding to the studied genes. Namely, as it was stated explicitly in class, traditional phylogenetics focused on either species or gene trees separately. Until then, we had a many to one species to gene relationship and a one to many species to gene relationship. In other words, only ortholog or paralog trees were constructed. So, one could come up with a good story for each proposed structure. Now, for the first time it was attempted to have many to many species to gene relationships (lecture 26, page 5, 3rd slide).

This time, a much richer questions could be answered by trying to explain gene and species evolution under a unified framework [7]. On the other hand, many obstacles have arisen due to duplication, loss and horizontal gene transfer events that are not directly related to evolution, thus producing incongruent gene trees to the (supposedly true) species trees.

In more detail, it was observed that gene-tree incongruences were more prominent in short alignments, slow- and fast-evolving genes [7]. In more detail, as described in the paper, using 5154 syntenic one-to-one orthologs across 12 *Drosophila* genomes and 739 syntenic one-to-one orthologs across 9 fungal genomes, it was found that existing phylogenies recovered the correct species topology for only a small number of gene trees (24% to 42% for flies and 22% to 31% for fungi) (see lecture 26, page 5, slide 4 or better Figure 2 of [7]).

At a first thought, one could think that this inconsistency is supported by a well-hidden underlying biological theory. Indeed, biological mechanisms have been proposed for gene-tree incongruence, such as incomplete lineage sorting or pre-speciation alleles. However, these statements can barely explain all incongruencies. Not to mention, that there are multiple lines of evidence supporting that inconsistencies are due to algorithmic inaccuracies, some of which can be enumerated as follows:

1. Clear monotonic increase in recovery of congruent gene trees with the length of the corresponding genes
2. Only genes of moderate divergence gave good results (slow-evolving genes offer little evolutionary insight, while fast-evolving cannot create distinguishable topologies)
3. Simulated phylogenies with the known species topology and similar branch lengths resulted in the same alternate topologies T1-T5 at comparable frequencies
4. If alternate topologies were due to biological reasons we would expect them to be recovered with multiple methods, show high bootstrap support and have significantly higher likelihood [7].
Most methods, suffered from the so-called phenomenon of long-branch attraction. Namely, rapidly evolving lineages are inferred to be closely related regardless of their true evolutionary relationships. A major cause for this insufficiency of interpretation is that most gene tree reconstruction algorithms do not use knowledge from the (true) species tree after reconstruction phase or even worse they do not use at all. For example, in TreeFam database, human annotators have been recruited to visually inspect gene phylogenies (and use information from known species topologies) [6].

In addition, most methods assumed uniform priors in all topologies $T$ and all branch lengths $B$. Although, this could apply for topologies $T$, it is not generally the case for branch lengths $B$ due to fixed times of species divergence and well-constrained rates of acceleration within anyone lineage. This gap actually was to be filled by SPIDIR’s method. Now that we presented the short story of phylogenomics and provided the weaknesses that SPIDIR had to address, it is time to give a concise description of the tool.

3 Description of SPIDIR

First, SPIDIR works in two stages: i) learning a model of genes and species evolution based on unambiguous orthologs, and ii) using this model for gene-tree construction.

As we mentioned, SPIDIR introduces a prior over the branch lengths, $B$. We have to get a bit more into that in order to make clear how this is done.

The traditional likelihood method presented in recitation constructs a gene tree by solving the following $\arg\max$ problem:

$$\hat{B}, \hat{T} = \arg\max_{B,T} P(D, B, T) \Leftrightarrow \hat{B}, \hat{T} = \arg\max_{B,T} P(D|B, T)P(B, T)$$

where $D$ are the genes, $B$ the branch lengths and $T$ the proposed topology. In the above equation, the $P(B, T)$ will not affect the $\arg\max$ problem since it is constant (uniform priors).

However, in the case of SPIDIR the estimation of topology and branch lengths will incorporate reconciliation parameters (mappings from gene to species tree) as well as some evolution parameters, which will be discussed shortly. Namely, the formula now is:

$$\hat{B}, \hat{T} = \arg\max_{B,T} P(D, B, T, R|E) \Leftrightarrow \hat{B}, \hat{T} = \arg\max_{B,T} P(D|B, T)P(B|T, R, E)P(R|T, E)P(T|E)$$

The terms $P(R|T, E)$ and $P(T|E)$ are constant due to parsimonious reconciliation uniqueness and topology uniformity, respectively. So, in fact, the term that plays major role in finding more congruent topologies is $P(B|T, R, E)$.

Before proceeding to the brief presentation of the results, we owe the reader the clarification of notation used. More explicitly, we used symbol $R$ without defining what it is. $R$ stands for reconciliation and it is about the mapping of each gene node to its species node. In SINDIR [6], the traditional reconciliation algorithm that has been used was taken from Zmasek\(^1\) [8]. In more detail, this type of reconciliation provided

\(^1\)This algorithm was chosen for its simplicity and its $O(n)$ performance in real cases despite its $O(n^2)$ worst-case scenario
a gene node-to-species node mapping and was given by the type:

\[ R(node \, b_l) = i, \ \text{where} \ i = LCA(\bigcup_{c \in \text{children}(b_l)} R(b_c)) \]

In other words, a node is mapped to the node which is the Least Common Ancestor of the mappings of the children's node. In this setting, a reconciliation happens if a parent is mapped to the same species as one of its children.

Although the above definition of reconciliation was intuitively simple and efficient enough, it lacked the information of when a duplication event happened and in what order compared to other events. This problem was overcome by the so-called definition of generalized reconciliation. In this setting, a parameter \( k_l \in (0, 1) \) is introduced (called a duplication point) that indicates the position of a duplication \( (k_l = 1 \text{ for speciation nodes}) \). As branch of a node, we consider the branch connecting the node and its parent. So, the generalized reconciliation is defined as:

\[ R(node \, b_l) = (i, k_l), \ \text{where} \ i \ is \ a \ species, \ k_l \ a \ duplication \ point \]

Also, the mapping \( R \) of gene to species nodes implies also a mapping from of gene to species branches \( (R_b) \). The type for branch mapping is given by:

\[ R_b(branch \, b_l) = ((s_1, \ldots, s_m), (p_1, \ldots, p_m)) \]

where \( m \) is the number of species (see Lecture 26, page 7, slide 3 for further clarification).

Moreover, it is wise to mention how we arrived at the evolutionary model \( E \) characterization. More specifically, it has been found that the substitution rate of each gene \( (b_i^2) \) in each lineage can be decoupled into gene \((g)\)- and species-specific \((s_i)\) rate. The gene-specific substitution rate \( g \) is dictated by the selective pressures imposed on the function of gene, and the species-specific substitution rate \( s_i \) is dictated by the time interval and the evolutionary dynamics of each lineage \( i \) (e.g. population size, generation time, mating behavior, overall mutation rate) [7].

At this point, in order for the gene-specific rate to represent the selective constraints of a gene function, it is implicitly required that all genes under investigation (for a given function) should have exactly the same function, aka to be orthologs. In order for this requirement to be met, only congruent gene trees, namely trees having the same topology with the (true) species tree, have to be examined. Indeed, 5154 unambiguous fly sets of orthologs and 739 unambiguous fungal sets of orthologs have been used for the learning process.

Now, we can define that each (absolute) branch of gene \( i \) in species \( i \) \((b_i)\) is characterized by a gene- and species-specific rate (which are independent of each other):

\[ b_i = g \cdot s_i \]

The gene-specific rate follows a Gamma distribution: \( G = \Gamma(\alpha, \beta) \), while the species-specific rate follows a normal distribution \( S_i = N(\mu_i, \sigma^2_i) \). The aforementioned constitute the evolutionary parameters \( E \):

\[ E = (G, S) = ((\alpha, \beta), (\mu_1, \sigma_1, \ldots, \mu_n, \sigma_n)) \]

These parameters can be learned by training the model through the aforementioned unambiguous orthologous sets.

---

2 Not to be confused with \( b_l \) defined above

3 \( i \) spans from 1 to \( 2m - 2 \), where \( m \) is the number of species
For the purposes of the lecture, it may be wiser to stop here the analysis of the algorithm. For further reference, one can advise the Master Thesis of Matt Rasmussen [6], as well as the paper [7].

Lastly, it is critical to mention that this method obviously outperformed others (see lecture 26, page 1, 1st and 2nd slide) and established an innovative probabilistic framework of studying phylogenies hereafter. One potential future outlook discussed by the authors is to model jointly the deep coalescent and gene duplication/loss events\(^4\) [7].

References


\(^4\)SPIDIR does not assume that incongruences are due to deep coalescence and incomplete lineage sorting but rather to gene duplication and loss events