Molecular Phylogenetics
An introduction to computational methods and tools for analyzing evolutionary relationships
Abstract

Molecular phylogenetics applies a combination of molecular and statistical techniques to infer evolutionary relationships among organisms or genes. This review paper provides a general introduction to phylogenetics and phylogenetic trees, describes some of the most common computational methods used to infer phylogenetic information from molecular data, and provides an overview of some of the many different online tools available for phylogenetic analysis. In addition, several phylogenetic case studies are summarized to illustrate how researchers in different biological disciplines are applying molecular phylogenetics in their work.

Introduction to Molecular Phylogenetics

The similarity of biological functions and molecular mechanisms in living organisms strongly suggests that species descended from a common ancestor. Molecular phylogenetics uses the structure and function of molecules and how they change over time to infer these evolutionary relationships. This branch of study emerged in the early 20th century but didn’t begin in earnest until the 1960s, with the advent of protein sequencing, PCR, electrophoresis, and other molecular biology techniques. Over the past 30 years, as computers have become more powerful and more generally accessible, and computer algorithms more sophisticated, researchers have been able to tackle the immensely complicated stochastic and probabilistic problems that define evolution at the molecular level more effectively. Within past decade, this field has been further reenergized and redefined as whole genome sequencing for complex organisms has become faster and less expensive. As mounds of genomic data becomes publically available, molecular phylogenetics is continuing to grow and find new applications. [4, 10, 17, 20, 22]

The primary objective of molecular phylogenetic studies is to recover the order of evolutionary events and represent them in evolutionary trees that graphically depict relationships among species or genes over time. This is an extremely complex process, further complicated by the fact that there is no one right way to approach all phylogenetic problems. Phylogenetic data sets can consist of hundreds of different species, each of which may have varying mutation rates and patterns that influence evolutionary change. Consequently, there are numerous different evolutionary models and stochastic methods available. The optimal methods for a phylogenetic analysis depend on the nature of the study and data used. [5, 19, 20]

Molecular Evolution: Beyond Darwin

Evolution is a process by which the traits of a population change from one generation to another. In On the Origin of Species by Means of Natural Selection, Darwin proposed that, given overwhelming evidence from his extensive comparative analysis of living specimens and fossils, all living organisms descended from a common ancestor. The book’s only illustration (see Figure 1) is a tree-like structure that suggests how slow and successive modifications could lead to the extreme variations seen in species today. [11, 27]
Figure 1. Evolution Defined Graphically. The sole illustration in Darwin’s *Origin of the Species* uses a tree-like structure to describe evolution. This drawing shows ancestors at the limbs and branches of the tree, more recent ancestors at its twigs, and contemporary organisms at its buds. [34]

Darwin’s theory of evolution is based on three underlying principles: variation in traits exist among individuals within a population, these variations can be passed from one generation to the next via inheritance, and that some forms of inherited traits provide individuals a higher chance of survival and reproduction than others. [11]

Although Darwin developed his theory of evolution without any knowledge of the molecular basis of life, it has since been determined that evolution is actually a molecular process based on genetic information, encoded in DNA, RNA, and proteins. At a molecular level, evolution is driven by the same types of mechanisms Darwin observed at the species level. One molecule undergoes diversification into many variations. One or more of those variants can be selected to be reproduced or amplified throughout a population over many generations. Such variations at the molecular level can be caused by mutations, such as deletions, insertions, inversions, or substitutions at the nucleotide level, which in turn affect protein structure and biological function. [11, 22]

What is a Phylogeny?

According to modern evolutionary theory, all organisms on earth have descended from a common ancestor, which means that any set of species, extant or extinct, is related. This relationship is called a phylogeny, and is represented by phylogenetic trees, which graphically represent the evolutionary history related to the species of interest (see Figure 2). Phylogenetics infers trees from observations about existing organisms using morphological, physiological, and molecular characteristics.

![Phylogeny of Mammalia](image)

Figure 2. Phylogeny of Mammalia. This phylogenetic tree shows the evolutionary relationships among six orders of Mammalian species (taxa). Taxa listed in grey are extinct.

The “tree of life” represents a phylogeny of all organisms, living and extinct. Other, more specialized species and molecular phylogenies are used to support comparative studies, test biogeographic hypotheses, evaluate mode and timing of speciation, infer amino acid sequence of extinct proteins, track the evolution of diseases, and even provide evidence in criminal cases. [19]
Understanding Phylogenetic Trees

Before exploring statistical and bioinformatic methods for estimating phylogenetic trees from molecular data, it’s important to have a basic familiarity of the terms and elements common to these types of trees. (See Figure 3.)

![Figure 3. Basic elements of a phylogenetic tree.](image)

Phylogenetic trees are composed of branches, also known as edges, that connect and terminate at nodes. Branches and nodes can be internal or external (terminal). The terminal nodes at the tips of trees represent operational taxonomic units (OTUs). OTUs correspond to the molecular sequences or taxa (species) from which the tree was inferred. Internal nodes represent the last common ancestor (LCA) to all nodes that arise from that point. Trees can be made of a single gene from many taxa (a species tree) or multi-gene families (gene trees). [1, 10]

A tree is considered to be “rooted” if there is a particular node or outgroup (an external point of reference) from which all OTUs in the tree arises. The root is the oldest point in the tree and the common ancestor of all taxa in the analysis. In the absence of a known outgroup, the root can be placed in the middle of the tree or a rootless tree may be generated.

Branches of a tree can be grouped together in different ways. (See Figure 4.)

![Figure 4. Groups and associations of taxonomical units in trees.](image)

A monophyletic group consists of an internal LCA node and all OTUs arising from it. All members within the group are derived from a common ancestor and have inherited a set of unique common traits. A paraphyletic group excludes some of its descendents (for examples all mammals, except the marsupialia
taxa). And a *polyphyletic* group can be a collection of distantly related OTUs that are associated by a similar characteristic or phenotype, but are not directly descended from a common ancestor. [1, 17]

**Trees and Homology**

Evolution is shaped by homology, which refers to any similarity due to common ancestry. Similarly, phylogenetic trees are defined by homologous relationships. *Paralogs* are homologous sequences separated by a gene duplication event. *Orthologs* are homologous sequences separated by a speciation event (when one species diverges into two). *Homologs* can be either paralogs or orthologs. [1, 11, 22]

Molecular phylogenetic trees are drawn so that branch length corresponds to amount of evolution (the percent difference in molecular sequences) between nodes. [1, 19]

![Figure 5. Understanding paralogs and orthologs.](image)

Paralogs are created by gene duplication events. (See Figure 5.) Once a gene has been duplicated, all subsequent species in the phylogeny will inherit both copies of the gene, creating *orthologs*. Interestingly, evolutionary divergence of different species may result in many variations of a protein, all with similar structures and functions, but with very different amino acid sequences. Phylogenetic studies can trace the origin of such proteins to an ancestral protein family or gene. [1, 22]

![Figure 6. Mirror Phylogenies. Gene A and Gene A¹ are paralogs, whereas all instances of Gene A are orthologs of each other in different Canid species.](image)

One way to ensure that paralogs and orthologs are appropriately referenced in a phylogenetic tree, and guard against misrepresentation due to missing or incomplete taxonomic information is to generate mirror phylogenies (see Figure 6) in which paralogs serve as each other’s outgroup. [1, 4, 19, 22]

**Estimating Molecular Phylogenetic Trees**

Molecular phylogenetic trees are generated from character datasets that provides evolutionary content and context. Character data may consist of biomolecular sequence alignments of DNA, RNA, or amino acids, molecular markers, such as single nucleotide polymorphisms (SNPs) or restriction fragment length polymorphisms (RFLPs), morphology data, or information on gene order and content. Evolution is modeled as a process that changes the *state* of a character, such as the type of nucleotide (AGTC) at a
specific location in a DNA sequence; each character is a function that maps a set of taxa to distinct states. [1, 19] Note that most of the examples in this paper use DNA sequences as character data, but trees can be accurately estimated from many different types of molecular data.

Figure 7. Evolution of a DNA Sequence

Figure 7 illustrates how a molecular sequence might evolve over time as a result of multiple mutations that results small, but evolutionarily important changes in a nucleotide sequence. At the protein level, these changes may not initially affect protein structure or function, but over time, they may eventually shape a new purpose for a protein within divergent species. [10, 19, 22] OTUs can be used to build an unrooted phylogenetic tree that clearly depicts a path of evolutionary change.

Steps in Phylogenetic Analysis

Although the nature and scope of phylogenetic studies may vary significantly and require different datasets and computational methods, the basic steps in any phylogenetic analysis remain the same: assemble and align a dataset, build (estimate) phylogenetic trees from sequences using computational methods and stochastic models, and statistically test and assess the estimated trees. [4, 19, 20]

Assemble and Align Datasets

The first step is to identify a protein or DNA sequence of interest and assemble a dataset consisting of other related sequences. For example, to explore relationships among different members of the Notch family of proteins, one might select DNA sequences for Notch1 through Notch4, in different species, such as human, dog, rat, and mouse, then perform a multiple sequence alignment to identify homologies. [1, 10, 13, 19, 20]

There are a number of free, online tools available to simplify and streamline this process. DNA sequences of interest can be retrieved using NCBI BLAST or similar search tools. When evaluating a set of related sequences retrieved in a BLAST search, pay close attention to the score and E-value. A high score indicates the subject sequence retrieved with closely related to the sequence used to initiate the query. The smaller the E-value, the higher the probability that the homology reflects a true evolutionary relationship, as opposed to sequence similarity due to chance. As a general rule, sequences with E-values less than $10^{-5}$ are homologs of a query sequence. [10]

Once sequences are selected and retrieved, multiple sequence alignment is created. This involves arranging a set of sequences in a matrix to identify regions of homology. Typically, gaps (one or more spaces in the alignment) are introduced in one or more sequences to represent insertions or deletions in the molecular code that may have occurred over time. Effective multiple sequence alignment hinged on gap analysis—determining where to insert gaps and how large to make them. There are many websites and software programs, such as ClustalW, MSA, MAFFT, and T-Coffee, designed to perform multiple sequence on a given set of molecular data. ClustalW is currently the most mature and most widely used. [1, 10, 19]
Building Phylogenetic Trees

To build phylogenetic trees, statistical methods are applied to determine the tree topology and calculate the branch lengths that best describe the phylogenetic relationships of the aligned sequences in a dataset. Many different methods for building trees exist and no single method performs well for all types of trees and datasets. The most common computational methods applied include distance-matrix methods, and discrete data methods, such as maximum parsimony and maximum likelihood. [4, 17, 20]

There are several software packages, such as Paup*, PAML, PHYLIP, that apply most popular methods. [4] Paup* is a commercially available program that implements a wide variety of methods for phylogenetic inference, including maximum likelihood analysis for DNA data using different models. Paup* also includes a set of exact and heuristic methods for searching optimal trees.

PAML (Phylogenetic Analysis by Maximum Likelihood) is open-access set of programs for phylogenetic analysis and evolutionary model comparison. PAML includes many advanced models—DNA- and AA-based models as well as codon-based models that can be used to detect positive selection. Many of the programs in PAML can model heterogeneity of evolutionary rates among sequence sites using $\gamma$ distributions, and evolutionary dynamics of different sequence regions (concatenated gene sequences).

PHYLIP is another large suite of open-access programs for phylogenetic inference that estimates trees using numerous methods, including pairwise distance, maximum parsimony, and maximum likelihood. The maximum likelihood programs can handle a few simple stochastic models and have good tree searching capabilities. PHYLIP is generally considered good educational software for novice phylogeneticists.

Distance-Matrix Methods

Distance matrix methods compute a matrix of pairwise “distances” between sequences that approximate evolutionary distance. Distance-based methods tend to be in polynomial time and are quite fast in practice. These methods use clustering techniques to compute evolutionary distances, such as the number of nucleotide or amino acid substitutions between sequences, for all pairs of taxa. They then construct phylogenetic trees using algorithms based on functional relationships among distance values. There are several different distance-matrix methods, including the Unweighted Pair-Group Method with Arithmetic Mean (UPGMA), which uses a sequential clustering algorithm; the Transformed Distance Method, which uses an outgroup as a reference, then applies UPGMA; the Neighbor-Relations Method, which applies 4-point condition to adjust the distance matrix, then applies UPGMA; and the Neighbor-Joining Method, which arranges OTUs in a star, the finds neighbors sequentially to minimize total length of tree. [4, 17]

The following section on the UPGMA method provides a more detailed example of how distance-matrix methods work.

UPGMA Method

UPGMA produces rooted trees for which the edge lengths can be viewed as times measured by a molecular clock with a constant rate. This method uses a sequential clustering algorithm to identify two OTUs that are most similar (meaning they have the shortest evolutionary distance and are most similar in sequence) and treat them as a single new composite OTU. This process is repeated iteratively until only two OTUs remain. The algorithm defines the distance (d) between two clusters $C_i$ and $C_j$ as the average distance between pairs of sequences from each cluster:
This sequential clustering process is visually described in Figure 8. In this example, the two most homologous sequences are 1 and 2. They are clustered into a new composite parent node (6), and the branch lengths \( t_1 \) and \( t_2 \) are defined as \( 1/2d_{1,2} \). The next step is to search for the closest pair among remaining sequences and node 6. Pair 4 and 5 are identified and clustered into a new parent node (7), and the branch length for \( t_4 \) and \( t_5 \) is calculated. \[4, 17\]

![Figure 8. Sequential clustering of sequences using the UPGMA method.][17]

In this interactive process, parent node 8 is created from pairs 7 and 3, and parent node 9 is created by clustering nodes 6 and 8. \[4, 17\] Thus, all sequences are clustered into a single evolutionary tree. The total time \( t_9 \) can be calculated as:

\[
D_{6,8} = \frac{1}{6} (d_{1,3} + d_{1,4} + d_{1,5} + d_{2,3} + d_{2,4} +d_{2,5})
\]

**Discrete Data Methods**

Discrete data methods examine each column of a multiple sequence alignment dataset separately and search for the tree that best represents all this information. Although distance-based methods tend to be much faster than discrete data methods, they typically yield little information beyond the basic tree structure. Discrete data analyses, on the other hand, are information rich. These methods produce a separate tree for each column in the alignment, so it is possible to trace the evolution for specific elements within a given sequence, such as catalytic sites or regulatory regions. \[10, 17, 19, 20\]

Commonly used discrete data methods include maximum parsimony, which searches for the most parsimonious tree that requires the least number of evolutionary changes to explain differences observed, maximum likelihood, which requires a probabilistic model for the process of nucleotide substitution, and Bayesian MCMC, which also requires a stochastic model of evolution, but creates a probability distribution on a set of trees or aspects of evolutionary history. \[17, 19, 20\]

Discrete data methods are generally considered to produce the best estimates of evolutionary history. However, these methods can be computationally expensive, and it can take weeks or months to obtain a reasonable level of accuracy for moderate to large datasets with 100 or more OTUs. \[19\]
Maximum Parsimony

Among the most widely used tree-estimation techniques, maximum parsimony applies a set of algorithms to search for the tree that requires the minimum number of evolutionary changes observed among the OTUs in the study. For example, Figure 9 lists four sample sequences from which phylogenetic trees could be inferred using maximum parsimony.

<table>
<thead>
<tr>
<th>Site</th>
<th>Seq</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>

Figure 9. Sample sequences for a maximum parsimony study [17]

Maximum parsimony algorithms identify phylogenetically informative sites, meaning the site favors some trees over others. Consider the sequences in Figure 9: Site 1 is not informative, because all sequences at that site (in column 1) are A (Adenine), and no change in state is required to match any one sequence (1-4) to another. Similarly, Site 2 is not informative because all three trees require one change and there is no reason to favor one tree over another. Site 3 is not informative because all three trees require two changes. (See Figure 10).

Figure 10. Site 3 trees all require one evolutionary change. [17]

Site 4 is not informative because all three trees require three changes. No one tree can be identified as parsimonious. (See Figure 10)

Figure 11. Site 4 trees all require three evolutionary changes. [17]

Site 5 is informative because one tree requires only one nucleotide change, whereas the other two trees require 2 changes. In Figure 12, the first tree on the left, which requires only one nucleotide change, is identified as the maximum parsimony tree.

Figure 12. Site 5 trees vary in the number of evolutionary changes required. [17]
Maximum Likelihood

The maximum likelihood method requires a probabilistic model of evolution for estimating nucleotide substitution. This method evaluates competing hypotheses (trees and parameters) by selecting those with the highest likelihood, meaning those that render the observed data most plausible. The likelihood of a hypothesis is defined as the probability of the data given that hypothesis. In phylogeny reconstruction, the hypotheses are the evolutionary tree (its topology and branch lengths) and any other parameters of the evolutionary model. [17, 20]

The likelihood calculations required for evolutionary trees are far from straightforward and usually require complex computations that must allow for all possible unobserved sequences at the LCA nodes of hypothesized trees. This method specifies the transition probability from one nucleotide state to another in a time interval in each branch. For example, for a one-parameter model with rate of substitution \( \lambda \) per site per unit time, the probability that the nucleotide at time \( t \) is \( i \) is:

\[
P_i(t) = \frac{1}{4} + \frac{3}{4} e^{-\lambda t/3}
\]

The probability that the nucleotide at time \( t \) is \( j \) is:

\[
P_j(t) = \frac{1}{4} - \frac{1}{4} e^{-\lambda t/3}
\]

To set up a likelihood function, given \( x \) as the ancestral node and \( y \) and \( z \) as internal nodes, the probability of observing nucleotides \( i, j, k, l \) at the tips of the tree is computed as:

\[
P_{il(t_1+t_2+t_3)}P_{yk(t_2+t_3)}P_{yz(t_2)}P_{zi(t_3)}P_{zj(t_3)}
\]

For the ancestral node (root) \( x \), the probability of having nucleotide \( l \) in sequence 4 is calculated as:

\[
P_{xl(t_1+t_2+t_3)}
\]

Because \( x, y, \) and \( z \) can be any one of four nucleotides (ACGT), it is necessary to sum over all possibilities to obtain the probability of observing the configuration of nucleotides \( i, j, k, l \), in sequences 1, 2, 3, 4, for a given hypothetical tree (see Figure 13.). This likelihood probability is calculated as:

\[
h(l,j,k,l) = [\sum g_x P_{xl(t_1+t_2+t_3)}] [\sum P_{yk(t_2+t_3)}] [\sum P_{yz(t_2)}] [\sum P_{zi(t_3)}] P_{zj(t_3)}
\]

The appropriate likelihood function depends on the hypothetical tree and the evolutionary model used. (See Figure 13.) [17]

---

**Figure 13.** Different types of model trees for the derivation of the maximum likelihood function. [17]
Stochastic Models of Evolution

Evolutionary changes in molecular sequences result from mutations, some of which occur by chance, others by natural selection. Rates of change can also differ among OTUs, depending on several factors ranging from GC content to genome size. To accurately estimate phylogenetic trees, assumptions must be made about the substitution process and those assumptions must be stated in the form of a stochastic evolutionary model. These probabilistic models are used to rank trees according to likelihood: \( P(\text{data}|\text{tree}) \). From a Bayesian perspective, they rank trees according to a posterior probability: \( P(\text{tree}|\text{data}) \). [17, 20]

The objective of probabilistic models is to find likelihood or posterior probability of a particular taxonomic feature, then define and compute:

\[
P(x^*|T,t) \quad \text{Where } x^* \text{ is } x^j \text{ for } j=1\ldots n, T \text{ is a tree with } n \text{ leaves with sequence } j \text{ at leaf } j, \text{ and } t^* \text{ are tree edge lengths.} [17]
\]

A few popular stochastic models of evolution include the single parameter Jukes-Cantor (JC) method, Kimura 2-parameter (K2P), Hasegawa-Kishino-Yano (HKY), and Equal-Input. Some software programs, such as Paup*, will automatically use a default model for the tree estimation method chosen. The JC method is the easiest one to comprehend, because it assumes that if a site changes its state, it changes with equal probability to the other states. This is not very realistic, however, as some sites are known to evolve more rapidly than others, and some sites may be invariable and not allowed to change at all. Determining how best to select the appropriate model is a topic of another paper (or papers) as there is no one model that incorporates all mutation rules and patterns across different species and macromolecules. [4, 17, 20]

Hidden Markov Models

Profile hidden Markov models (HMMs) are a form of Bayesian network that provides statistical models of the consensus structure of a sequence family. Gary Churchill at The Jackson Lab was the first evolutionary geneticist to propose using profile HMMs to model rates of evolution. Many software packages and web services now apply HMMs to estimate phylogenetic relationships. [8]

In the HMM format, each position in the model corresponds to a site in the sequence alignment. For each position, there are a number of possible states, each of which corresponds to a different rate of evolution. In addition, transitions between all possible rate-states at adjacent positions. Transition probabilities capture any tendency for patterns of rates to occur in successive sites. [2, 4]

Assessing Trees

Tree estimating algorithms generate one or more optimal trees. This set of possible trees is subjected to a series of statistical tests to evaluate whether one tree is better than another – and if the proposed phylogeny is reasonable. Common methods for assessing trees include the Bootstrap and Jackknife Resampling methods, and analytical methods, such as parsimony, distance, and likelihood. To illustrate how these methods are used, consider the steps involved in a bootstrap analysis.

Bootstrap Analysis

A bootstrap is a statistical method for assessing trees that takes its name from the fact that it can “pull itself up by its bootstraps” and generate meaningful statistical distributions from almost nothing. Using bootstrap analysis, distributions that would otherwise be difficult to calculate exactly are estimated by repeated creation and analysis of artificial datasets. In a Non-parametric bootstrap, artificial datasets
generated by resampling from original data. In a parametric bootstrap, data is simulated according to hypothesis tested. The objective of any bootstrap analysis is to test whether the whole dataset supports the tree. [1, 4, 17]

Figure 14 illustrates the basic steps in any bootstrap analysis. Sample datasets are automatically generated from an original dataset. Trees are then estimated from each sample dataset. The results are compiled and compared to determine a bootstrap consensus tree.

Figure 14. Steps in a phylogenetic tree bootstrap analysis. [1]

### Phylogenetic Analysis Tools

There are several good online tools and databases that can be used for phylogenetic analysis. These include PANTHER, P-Pod, PFam, TreeFam, and the PhyloFacts structural phylogenomic encyclopedia. Each of these databases uses different algorithms and draws on different sources for sequence information, and therefore the trees estimated by PANTHER, for example, may differ significantly from those generated by P-Pod or PFam. As with all bioinformatics tools of this type, it is important to test different methods, compare the results, then determine which database works best (according to consensus results, not researcher bias) for studies involving different types of datasets.

In addition, to the phylogenetic programs already mentioned in this paper, a comprehensive list of more than 350 software packages, web-services, and other resources can be found here:


#### PANTHER (pantherdb.org)

Protein Analysis Through Evolutionary Relationships, known by its acronym PANTHER, is a library of protein families and subfamilies indexed by function. Panther version 6.1 contains 5547 protein families.
It categorizes proteins by evolutionary related proteins (families) and related proteins with same function (subfamilies). [8, 21, 26]

PANTHER is composed of both a library and index. The library is a collection of “books” that represent a protein family as a collection of multiple sequence alignments, HMMs, and a family phylogenetic tree. Functional divergence within the tree is represented by dividing the parent tree into child trees and HMMs based on shared functions. These subfamilies enable database curators to more accurately capture functional divergence of protein sequences as inferred from genomic DNA. [25, 26]

PANTHER database entries are annotated to molecular function, biological process and pathway with a proprietary PANTHER/X ontology system, which is supposed to be easier to understand than the more global standard Gene Ontology (GO). Database entries in PANTHER are generated through clustering of UniProt database using a BLAST-based similarity score. Trees are automatically generated based on multiple sequence alignments and parameters of the protein family HMMs using the Tree Inferred from Profile Score (TIPS) clustering algorithm. Scientific curators review all family trees, annotate each tree, and determine how best to divide them into subtrees using a tree-attribute viewer that tabulates annotations for sequences in a tree. In addition, trees and subfamilies are manually cross-checked and validated by curators. [25, 26]

P-POD (ortholog.princeton.edu)

The Princeton Protein Orthology Database (P-POD) combines results from multiple comparative methods with curated information culled from the literature. Designed to be a resource for experimental biologists seeking evolutionary information on genes of interest, P-POD employs a modular architecture, based on their Generic Model Organism Database (GMOD). P-POD can be accessed from their web service or downloaded to run on local computer systems. [12]

P-POD accepts FASTA-formatted protein sequences as input, and performs comparative genomic analyses on those sequences using OrthoMCL and Jaccard clustering methods. The P-POD database contains both phylogenetic information and manually curated experimental results. The site also provides many links to sites rich in human disease and gene information. This tool may be particularly helpful for bioinformaticists and statisticians developing comparative genomic database tools and resources.

Pfam (pfam.sanger.ac.uk/)

PFam is a collection of protein families represented by multiple sequence alignments and HMMs. It contains models of protein clans, families, domains, and motifs, and uses HMMs representing conserved functional and structural domains. It is a large, widely used, actively curated mature database that has been available online since 1995. Pfam can be used to retrieve the domain architectures for a specific protein by conducting a search using a protein sequence against the Pfam library of HMMs. This database is also helpful for proteomes and protein domain architecture analysis. [6, 8, 24]

There are two versions of the Pfam database: Pfam–B is generated automatically from ProDom, using Psi-BLAST, an open access bioinformatics tool available through NCBI for identifying weak, but biologically relevant sequence similarities. Pfam-A is hand-curated from custom multiple sequence alignments. Pfam protein domain families are clustered with Mkdom2, and aligned with ProDomAlign.

ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases. Mkdom2 is a ProDom program used to make ProDom family clusters. Protein domain families in ProDom were aligned using an improved parallelized program called
ProDomAlign, developed in C++ using OpenMP. ProDomAlign is based on MultAlign, a program well suited for aligning very large sequence families with thousands of associated sequences.

As of early 2008, Pfam matched 72 percent of known proteins sequences, and 95 percent of proteins for which there is a known structure. Within the Pfam database, 75 percent of sequences will have one match to Pfam-A, 19 percent to Pfam-B. There are also two versions of Pfam-A and Pfam-B. Pfam-ls handles global alignments, and Pfam-fs is optimized for local alignments.

Interestingly, Pfam entries can be classified as “unknown,” but that doesn’t mean the protein is undocumented. Unknown entries can be proteins for which some information is known, but it has not been fully researched or cannot be adequately annotated. For example, Pfam entry PFO1816 is a Leucine-Rich Repeat Variant (LRV), which has a known structure (1LRV) available in the Protein Databank (pdb.org). LRV repeat regions, which are found in many different proteins, are often involved in cell adhesion, DNA repair, and hormone reception—but identification of an LRV within a sequence encoding a protein doesn’t specifically reveal the protein’s function.

For studies involving a large number of protein searches, it may be more convenient to run Pfam locally on a client machine. The standalone Pfam system requires the HMMER2 software, the Pfam HMM libraries and a couple of additional files from the Pfam website to be installed on the client machine. (HMMER is a freely distributable implementation of profile HMM software for protein sequence analysis.) Once the initial search is complete, researchers can go to the Pfam website to further analyze select number of sequences using additional features on website. [6, 8, 24]

TreeFam (TreeFam.org)

TreeFam is a curated database of phylogenetic trees and orthology predictions for all animal gene families that focuses on gene sets from animals with completely sequenced genomes. Orthologs and paralogs are inferred from phylogenetic tree of gene family. Release 4 contains curated trees for 1314 families and automatically generated trees for another 14351 families. [16, 23]

Like Pfam, TreeFam is a two-part database: TreeFam-B contains automatically generated trees, and TreeFam-A consists of manually curated trees. To automatically generate trees, an algorithm selects clusters of genes to create TreeFam-B “seeds” from core species with high-quality reference genome sequences, first using BLAST to rapidly assemble an initial list of possible matches, then HMMER to expand and filter probable sequence matches for each TreeFam B seed family. The filtered alignment is fed into a neighbor-joining algorithm and a tree is constructed based on amino acid mismatch distances. For TreeFam version 4, the most current release, five “clean” family trees were built for each TreeFam B seed, two using a maximum likelihood tree generated using PHYML (one based on the protein alignment, the other on codon alignment), three using a neighbor joining tree, using different distance measurements based on codon alignments. [16, 23]

Scientific curators then manually any correct errors (based on information in the literature) in automatically generated TreeFam-B trees. Curated TreeFam-B trees then become seeds for TreeFam-A trees. Clean TreeFam-A trees are build using three merging algorithms and bootstrapping to find the consensus tree of seven trees: two constrained maximum likelihood trees based on protein and codon alignment, and five unconstrained neighbor-joining trees generated using different distance measurements based on codon alignments. For both TreeFam-B and TreeFam-A families, orthologs and paralogs are inferred only from clean trees using Duplication/Loss Inference (DLI) algorithm that requires a species tree (NCBI taxonomy tree). [16, 23]
PhyloFacts (phylogenomics.berkeley.edu/phylofacts)

PhyloFacts is an online phylogenomic encyclopedia for protein functional and structural classification. It contains more than 57,000 “books” for protein superfamilies and structural domains. Each book contains heterogenous data for protein families, including multiple sequence alignments, one or more phylogenetic trees, predicted 3-D protein structures, predicted functional subfamilies, taxonomic distributions, GO annotations, and PFAM domains. HMMs constructed for each family and subfamily permit novel sequences to be classified to different functional classes. [14]

Unlike other databases mentioned in this paper, PhyloFacts seeks to correct and clarify annotation errors associated with computational methods for predicting protein function based on sequence homology. It uses a consensus approach that integrates many different prediction methods and sources of experimental data over an evolutionary tree. By applying evolutionary and structural clustering of proteins, PhyloFacts is able to analyze disparate datasets using multiple methods, identify potential errors in database annotations, and provide a mechanism for improving the accuracy of functional annotation in general. [14]

PhyloFacts can be used to search for protein structure prediction or functional classification for a particular protein sequence. Researchers may also browse through protein family books and multiple sequence alignments, phylogenetic trees, HMMs and other pertinent information for proteins of interest. This web-service also provides many links to literature and other information sources. [14]

Applied Molecular Phylogenetics

Molecular phylogenetic studies have many diverse applications. As the amount of publically available molecular sequence data grows and methods for modeling evolution become more sophisticated and accessible, more and more biologists are incorporating phylogenetic analyses into their research strategy. Here’s a sampling of how molecular phylogenetics might be applied.

Tracing the evolution of man

In one case study, molecular phylogenetic techniques were used to compare and analyze variation in DNA sequences using modern human and Neanderthal mitochondrial DNA (mtDNA). For this study, 206 modern human mtDNAs and parts of two Neanderthal mtDNAs sequences derived from skeletal remains were used to generate an initial dataset. Genetic distance was first estimated using the Jukes-Cantor single parameter model. Then the Kimura 2-Parameter model was used to distinguish between transition (replacement of one purine with another purine or one pyrimidine with another pyrimidine) and transversion (replacement of one purine with a pyrimidine or vice versa) probabilities with Kimura 2-parameter model. A phylogenetic tree representing primate evolution was generated using pairwise genetic distances between primate Hypervariable regions I and II of mtDNA. [3]

Chasing an epidemic: SARS

Using publically available genomic data, it is possible to reconstruct the progression of the SARS epidemic over time and geographically. To conduct this phylogenetic analysis, researchers used the neighbor-joining method to construct a phylogenetic tree of spike proteins in various coronaviruses and identify the viral host (a Himalyan palm civet). They then obtained 13 SARS genome sequences with documented information on the date and location of the sample. The neighbor-joining method and a distance matrix based on Jukes-Cantor model, were used to generate an epidemic tree, from which it was possible to identify the origin (date and location) of the virus by observing progression of mutations over time. [3]
Barking up the right tree

Phylogenetics is increasingly incorporated into biological and biomedical research papers. When the canine genome was published, researchers used sequence data to estimate a comprehensive phylogeny of the canid family.

![Phylogenetic Tree of the Canid family](image)

This canid family phylogenetic tree is based on 15 kb of exon and intron sequence. It was constructed using the maximum parsimony method and represents the single most parsimonious tree. A good example of how phylogenies are referenced in the literature, this tree includes bootstrap values and Bayesian posterior probability values listed above and below internodes, respectively. Dashes indicate bootstrap values below 50%. In addition, divergence time in millions of years (Myr) is indicated for three nodes. [18]

Seeing the Forest from the Trees

Molecular phylogenetics is a broad, diverse field with many applications, supported by multiple computational and statistical methods. The sheer volumes of genomic data currently available (and rapidly growing) render molecular phylogenetics a key component of much biological research. Genome-scale studies on gene content, conserved gene order, gene expression, regulatory networks, metabolic pathways, functional genome annotation can all be enriched by evolutionary studies based on phylogenetic statistical analyses. [19, 25, 27]

Molecular phylogenies have fast become an integral part of biological research, pharmaceutical drug design, and bioinformatics techniques for protein structure prediction and multiple sequence alignment. Although not all molecular biologists and bioinformaticians may be familiar with the techniques described
in this paper, this is a rapidly growing and expanding field and there is ongoing need for novel algorithms to solve complex phylogeny reconstruction problems.

References


Appendix

Website Resources

Phylogeny Programs. A University of Washington site formerly supported by the National Science Foundation. http://www.evolution.genetics.washington.edu/phylip/software.html


Cover Page

The cover image is from a phylogeny of canid species that appeared in Lindblad-Toh et al, 2005. [18]