

# DNA Sequence Analysis

Takashi Gojobori, *National Institute of Genetics, Mishima, Japan*

So Nakagawa, *National Institute of Genetics, Mishima, Japan*

Jose C Clemente, *National Institute of Genetics, Mishima, Japan*

*Based in part on the previous version of this Encyclopedia of Life Sciences (ELS) article, DNA Sequence Analysis by Takashi Gojobori and Allison Wyndham.*

Recent advances in deoxyribonucleic acid (DNA) sequencing technology have produced a massive amount of nucleotide sequences, which are stored in DNA databanks and genomic data repositories. Furthermore, comprehensive analyses of transcriptional and genomic elements have uncovered an elaborate system of gene expression that broadens our understanding of fundamental biological phenomena. The analysis of DNA data has therefore become essential to predict gene function or detect regulatory motifs through comparative studies. In this article, DNA databases, homology search tools and sequence alignment methods are surveyed. The concept of distance between genes and how to calculate this measure using DNA or amino acid sequences and introducing several commonly used techniques for phylogenetic analysis and tree evaluation are also described.

## Introduction

Deoxyribonucleic acid (DNA) sequence data contains a wealth of biologically useful information. How to extract this information has given rise to the field of DNA sequence analysis. One of the aims of sequence analysis is to reveal conserved characteristics shared by a group of related sequences, such as functional domains or active site residues, by the comparison of an uncharacterized query sequence with genes of known function. Related DNA sequences also contain phylogenetic information that can be used to infer evolutionary relationships among taxonomic groups. Pair-wise comparisons can now be performed with the same query sequence against many millions of identified sequences stored in DNA databanks. Sequencing DNA in a laboratory is rapid and simple; therefore, interest is currently focused on data analysis that utilizes statistical methods and computer analysis. The focus of this article is on bioinformatics approaches to sequence analysis, and includes explanations of database searching, sequence alignment, estimation of evolutionary distance and construction of molecular phylogenetic

**ELS subject area:** Evolution and Diversity of Life

### How to cite:

Gojobori, Takashi; Nakagawa, So; and, Clemente, Jose C (September 2009) DNA Sequence Analysis. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.  
DOI: 10.1002/9780470015902.a0001798.pub2

Advanced article

### Article Contents

- Introduction
- DNA Databases and Searching
- Multiple Sequence Alignment
- Genetic or Evolutionary Distance
- Molecular Phylogenetics
- Methods for Constructing Phylogenetic Trees
- Conclusions

Online posting date: 15<sup>th</sup> September 2009

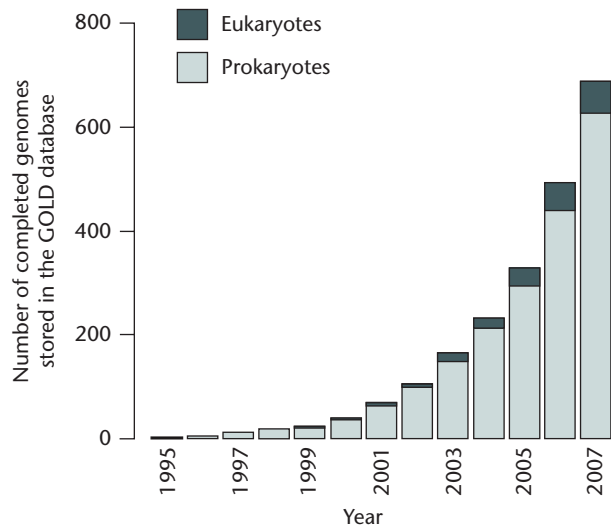
trees. **See also:** [DNA Sequencing](#); [Molecular Evolution: Techniques](#)

## DNA Databases and Searching

### Genome projects and advances in DNA sequencing technology

Recent innovations in DNA sequencing technology have greatly increased our capacity to determine massive amounts of nucleotide sequences. The genomes of 862 species (excluding viruses) have been completely sequenced (obtained from the Genomes OnLine Database, October 2008; **Figure 1**). Moreover, the number of metagenomes has also significantly increased. A metagenome is a collection of DNA sequences recovered directly from environmental samples. The objective of metagenomic analysis is to identify the sequences of organisms that are not easily cultured in a laboratory. **See also:** [Bacterial Genetics](#); [Metagenomics and Microbial Communities](#); [Sequencing the Human Genome: Novel Insights into its Structure and Function](#)

The comprehensive analysis of complementary DNAs – cDNAs, the nucleotide sequences of expressed messenger ribonucleic acids (mRNAs) – has revealed that a large amount of noncoding RNAs are expressed in eukaryotic cells and tissues, with a significant number of them being involved in the regulation of gene expression (The FANTOM Consortium, 2005). The Encyclopedia of DNA Elements (ENCODE) pilot study aimed at identifying all functional elements in 1% of the human genome, finding several overlapping genes, genes sharing exons and



**Figure 1** Growth of the number of completely sequenced genomes. Data obtained from GOLD.

alternative transcription start sites (The ENCODE Project Consortium, 2007). The traditional definition of a gene – a DNA sequence that encodes an amino acid chain – has been challenged by these and other recent results (Gerstein *et al.*, 2007), and further analysis is therefore needed to understand the complexity of the genome. **See also:** [Genome Sequence Analysis](#); [Multispot Array Technologies](#)

The technology responsible for this acceleration in DNA sequencing capacity is the so-called second-generation sequencers, such as 454 pyrosequencing, Solexa or SOLiD (454: 400 M nt per run (10 h); Solexa: 1–2 G nt per run (4 days); SOLiD: 3–5 G nt per run (8 days)). Although the read length of the sequences is relatively short (Solexa and Solid: 25–50 nt; 454: 400 nt), third-generation sequencers will hopefully provide longer reads. **See also:** [DNA Sequencing](#)

## DNA databanks and integrated genomic databases

As DNA sequencing technology advances, nucleotide sequences have accumulated with enormous speed. This vast volume of data is stored and maintained by a tripartite network of international databanks called the International Nucleotide Sequence Database Collaboration. This comprises the DNA Data Bank of Japan (DDBJ) at the National Institute of Genetics (NIG) in Japan, the European Bioinformatics Institute (EBI) at the European Molecular Biology Laboratory (EMBL) and the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) in the United States. These databanks exchange and update information daily and are searchable through the Internet. A variety of services are also provided by these organizations including

data retrieval and searching tools, curated material such as patent information, relevant literature and cross-references where possible. Any user can conduct a database search using online tools.

For genomic analysis, databases such as Ensembl and the UCSC Genome Browser are commonly used because of their user-friendly interfaces, which provide an interactive comparison of different resources, including genomic alignments with other species and domain structures of proteins. The Reference Sequence (RefSeq) database, maintained by NCBI, provides a nonredundant and well-annotated set of genome sequences. The H-investigational Database (H-invDB) collects transcriptional information, including the sequences of human cDNAs. Based on this collection, H-invDB provides curated annotation of human genes, including gene expression profiling, alternative splicing isoforms and noncoding functional RNAs (Imanishi *et al.*, 2004). In plants, the Arabidopsis Information Resource (TAIR) and the Rice Annotation Project (RAP) contain well-annotated genomic features of *Arabidopsis thaliana* and rice, including cDNAs and mutant phenotypes (Swarbreck *et al.*, 2008; Rice Annotation Project, 2008). These databases provide high-quality genomic information of model organisms. However, the Genomes OnLine Database (GOLD) and the Genome Information Broker (GIB), maintained by DDBJ, aim at a more comprehensive collection of complete genome projects, including eukaryotes, bacteria, archaea, viruses and metagenomes. Additional databases and specialized Internet-based resources useful for sequence analysis are presented in detail in regular database issues of *Nucleic Acids Research*. **See also:** [Bioinformatics](#); [Bioinformatics in Genome Sequencing Projects](#); [Biological Data Centres](#); [Genetic Databases](#); [Genetic Databases: Mining](#); [Genome Databases](#); [Genome Databases](#); [Human Genetics: Online Resources](#); [Nucleotide Sequence Databases](#); [Primary Protein and Nucleic Acid Three-dimensional Structure Databases](#); [Protein Family Databases](#); [Protein Sequence Databases](#)

## Database searching

One of the most informative methods used in sequence data analysis is similarity searching. For DNAs, similarity at the sequence level implies some structural or functional similarity between the protein products or regulatory elements of gene expression. Searching a database with an uncharacterized gene sequence can identify homologues in other species or sequence elements that encode structural domains within the protein. Searches can be conducted with either nucleotide or peptide sequences. However, detection of similarity at the nucleotide level is difficult unless the sequences are closely related. For analysis of coding DNAs, similarity searching with the translated protein sequence is more informative.

A commonly used tool for similarity searching is BLAST (Basic Local Alignment Search Tool) because of its practical balance of speed, sensitivity and selectivity

**Table 1** Homology search programmes

Programme	Query sequence	Database	Type of alignment
BLASTN	Nucleotide	Nucleotide	Gapped
BLASTX	Nucleotide translated into protein	Protein	Gapped
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Ungapped
BLASTP	Protein	Protein	Each frame gapped
TBLASTN	Protein	Nucleotide translated into protein	Each frame gapped
FASTA	Nucleotide or protein	Nucleotide or protein	Gapped
FASTX	Nucleotide translated into protein	Protein	Gapped
TFASTA	Protein	Nucleotide translated into protein	Ungapped
TFASTX	Protein	Nucleotide translated into protein	Each frame gapped
SSEARCH	Nucleotide or protein	Nucleotide or protein	Gapped
MegaBLAST	Nucleotide	Nucleotide	Gapped
PSI-BLAST	Protein	Protein	Gapped
BLAT	Nucleotide or protein	Nucleotide or protein	Gapped
BLASTZ	Nucleotide	Nucleotide	Gapped

(Altschul *et al.*, 1990). As indicated by its name, BLAST conducts a local alignment, not a global one, of two sequences using a heuristic approach similar to the Smith–Waterman algorithm. Although complete or modified versions of the Smith–Waterman algorithm have been implemented in the SSEARCH or FASTA programmes, both of them (especially SSEARCH) are significantly slower than BLAST, with only marginal gains in accuracy. There are several types of BLAST programmes depending on the purpose of search (summarized in [Table 1](#)). For a large number of query sequences, MegaBLAST is useful in terms of calculation time. PSI-BLAST broadly identifies homologous genes and it is used to obtain the protein family of a query sequence. BLAT, the BLAST-Like Alignment Tool, is commonly used to look up the location of a sequence in the genome or to determine the exon structure of an mRNA. Although the accuracy of BLAT is lower than that of BLAST, it is much faster than BLAST. BLASTZ is often utilized to align long genomic sequences such as a comparison between chromosomes. Note that the output of these search programmes depends on parameters such as ‘word size’ or ‘gap penalty’, and results might therefore change with different settings. **See also:** [Bioinformatics](#); [BLAST Algorithm](#); [Dynamic Programming](#); [Entrez and Forward Database Searching](#); [Genome Sequence Analysis](#); [Genome, Proteome, and the Quest for a Full Structure–Function Description of an Organism](#); [Mining Biological Databases](#); [Sequence Similarity](#); [Similarity Search](#); [Smith–Waterman Algorithm](#); [Sequence Alignment](#)

## Functional predictions

Similarity searching is particularly useful for predicting functions of newly identified sequences. This is based on the premise that conserved regions must be functionally or structurally important. If we search a sequence and find a region where conservation is strong, it is reasonable to

speculate that the region is functionally or structurally important and that these characteristics are shared by the similar sequences. However, several studies have questioned the relevance of sequence conservation as a functional predictor. Deletion of ultraconserved regions of 200 or more bases among various mammals yielded viable mice (Ahituv *et al.*, 2007). Also, methylation patterns at orthologous loci between human and mouse have been found to be strongly conserved despite a lack of conservation of the underlying sequences (Bernstein *et al.*, 2005). **See also:** [Evolutionarily Conserved Noncoding DNA](#); [Ultraconserved DNA Sequence Elements in the Human Genome](#)

If sequences are similar, they probably share the same ancestral sequence such as homologues of the same gene in different species. However, differentiation occurs following gene duplication, giving rise to a family of genes with related but distinct functions. In some cases, sequence similarity persists after functional similarity has been eroded. Conversely, structural and functional similarity between proteins can exist despite the complete absence of sequence similarity and such relationships will not be detected using a similarity search. **See also:** [Evolutionary Developmental Biology: Gene Duplication, Divergence and Co-option](#)

Similarity can also be restricted to one domain or a sequence motif shared by two sequences. This applies both to coding and noncoding DNAs. Noncoding DNA contains conserved regulatory elements that control gene expression, heterochromatin structure or DNA silencing. When considering a coding sequence, the protein it encodes might possess more than one functional domain. As each of the different domains is assumed to have a single and particular biological function, a protein having multiple domains is considered to exhibit multiple functions. Thus, a protein having multiple and different domains is called a mosaic protein. Accordingly, when conducting a similarity search, a query sequence might match many different

proteins so that no consistent function can be inferred. Therefore, the function of proteins or genes should be considered region by region using similarity searches. **See also:** [Chromosomes: Noncoding DNA \(Including Satellite DNA\)](#); [Proteins: Fundamental Chemical Properties](#)

Having detected a region of similarity between two sequences, the task remains to determine what the similarity means. The identity of the match sequence(s) provides a clue to this. Ideally, a specific region of the query sequence will produce matches to other sequences whose functions are both known and interrelated. Moreover, global similarity detected between two sequences indicates the existence of another member of the same gene family. A discrete region of similarity can also fall into a known class of domains or motifs, as revealed by searches of databases such as Pfam for proteins or TRANSFAC for binding sites.

At this point any available empirical data relating to the sequence will be useful for functional predictions. Furthermore, if database searches have revealed a group of related sequences that all match the query, then constructing a multiple sequence alignment is the most useful means to identify regions of functional importance.

## Multiple Sequence Alignment

The alignment of a group of related sequences can achieve two purposes; the assembly of a group of evolutionarily related sequences for the identification of conserved regions or the alignment of a group of proteins to illustrate structural characteristics (e.g. for secondary structure prediction). The focus here will be on the former purpose. Alignment of more than two sequences – either DNA or protein – is more complex than the pair-wise comparisons described earlier, particularly for divergent sequences. The aim of multiple alignments is usually to compare a group of sequences along their whole length. Therefore, most multiple alignment tools seek a global rather than a local alignment. The alignment of multiple sequences involves maximizing similarity among the DNA or amino acid sequences and allowing insertion of gaps, where a gap is a site or a series of sites that has no corresponding sites in a given sequence. When DNA sequences are aligned with each other, identification of the appropriate corresponding nucleotides is quite difficult because there are only four types of nucleotides. Alignment of amino acid sequences is easier and can be more meaningful if the intent is to compare a group of related sequences for potential functional characteristics. **See also:** [Protein Secondary Structures: Prediction](#); [Protein Structure Prediction](#); [Sequence Alignment](#)

### The alignment tools

Clustal is a commonly used programme for multiple sequence alignment. It uses a progressive algorithm to

align sequences in successively larger groups, beginning with the most closely related sequences. Using ClustalW (Thompson *et al.*, 1994), all pairs of sequences are compared and a tentative measure of similarity is derived, represented by a distance matrix. This is used to produce a phylogenetic guide tree, using the neighbour-joining (NJ) method (Saitou and Nei, 1987). The branching pattern of the tree is used to determine the most closely related pair of the sequences. A final alignment is obtained by repeating this procedure until it reaches the root of a tree. As with similarity searching, the user can select the appropriate gap penalties, word size and substitution matrix. In the case of DNA sequences, transversion-type changes can be weighted more heavily than transition-type changes, because the latter changes occur much more frequently than the former (Kimura, 1983). T-Coffee generates a multiple sequence alignment based on a similar approach to that used by ClustalW, but utilizes a system of sequence position weights (Notredame *et al.*, 2000). MAFFT is a rapid multiple sequence alignment programme that applies the Fourier transformation to reduce the calculation time of dynamic programming (Katoh *et al.*, 2002). Regardless of the software used to create an alignment, visual inspection is usually required to refine the final alignment, particularly to ensure the correct placement of gaps. **See also:** [Multiple Alignment](#)

## Genetic or Evolutionary Distance

Genetic distance is a quantitative measure of evolutionary similarity or dissimilarity between taxa. This is based on the assumption that a pair of genes descended from the same ancestral sequence will independently accumulate nucleotide changes (substitutions) over time. The numbers of nucleotide or amino acid differences among sequences (substitutions) are used to calculate genetic distance. The number of observed character differences often does not represent the real number of substitutions that have occurred, particularly between distantly related sequences where multiple mutations of the same character may have occurred over time. Furthermore, one may want to differentiate between nucleotide substitutions that cause amino acid mutations (nonsynonymous substitutions) and those that are silent (synonymous substitutions). The methods introduced later attempt to compensate for these and other factors. **See also:** [Evolutionary Distance: Estimation](#); [Molecular Phylogeny Reconstruction](#); [Mutation Rates: Data](#); [Mutations and the Genetic Code](#)

### Number of nucleotide substitutions

The number of nucleotide substitutions is estimated by making pair-wise comparisons of nucleotide sequences and by correcting for multiple substitutions at the same site. For this, one needs a model of nucleotide substitution. Random patterns of nucleotide substitution are rare; therefore parameters are used in substitution models to

take patterns that are generally observed into account. At its simplest, the one-parameter method of Jukes and Cantor (1969) assumes that the rates of nucleotide substitutions between all possible pairs of different nucleotides are equal. The number of nucleotide substitutions ( $K_n$ ) is then estimated by the equation

$$K_n = -3/4 \ln(1 - 4/3p)$$

where  $p$  is the observed proportion of nucleotide differences. There are many modifications of this formula. For example, Kimura's (1980) two-parameter method was developed under the assumption that the rates of transitions and transversions are different (which is often the case). Since then, many different methods have been developed to estimate the correct number of substitutions, each varying in its parameters and assumptions. The methods of Li *et al.* (1985) and Nei and Gojobori (1986) have also been developed to estimate the numbers of synonymous and nonsynonymous substitutions per site.

### Number of amino acid substitutions

Numbers of amino acid substitutions are estimated in a manner similar to nucleotide substitutions. Under the assumption that the substitution rates between any pair of amino acids are equal, the number of amino acid substitutions ( $K_a$ ) is given by the formula:

$$K_a = -\ln(1 - p)$$

where  $p$  represents the proportion of amino acid differences. However, the substitution rate between similar amino acids is generally much higher than that between dissimilar ones. To account for this bias, Kimura (1983) proposed the modified formula:

$$K_a = -\ln(1 - p - 1/5p^2)$$

The substitution numbers produced by this formula are very close to those produced using Dayhoff's PAM algorithm (Dayhoff *et al.*, 1978), which is used for regular amino acid substitution matrices.

### Genetic distance and phylogenetics

According to the molecular clock theory, first proposed by Zuckerkandl and Pauling (1965), sequences diverge at a constant rate. Therefore, the genetic distance between two related sequences can also be used as a measure of the time elapsed since divergence from the common ancestral sequence (Kimura, 1983). Since the proposal of the molecular clock, it has been shown that in fact the estimated substitution rates of many sequences vary considerably. Furthermore, highly divergent sequences will have accumulated multiple substitutions at the same site (referred to as saturation) such that the genuine genetic distances will become difficult to estimate. Nevertheless, genetic distance between sequences can be used as the basis to reconstruct a phylogenetic tree that describes the evolutionary relationships among taxa. Alternatively,

nucleotide or amino acid character information can be used to construct a tree. Both these approaches are discussed in the following section. **See also:** [Molecular Clocks](#); [Molecular Clock: Testing](#)

## Molecular Phylogenetics

Molecular phylogenetics is the study of molecular evolution by constructing phylogenetic trees based on DNA and amino acid sequences. As the name suggests, a molecular phylogenetic tree is a branching representation of phylogenetic relationships among sequences. The terminal nodes (leaves) of the tree represent existing sequences or species, referred to as operational taxonomic units (OTUs), whereas internal nodes represent hypothetical ancestral sequences. Ideally, the tree is bifurcating, that is, each internal node gives rise to two branches representing the products of a speciation event. Necessarily, phylogenetic reconstruction involves use of incomplete data, since the true ancestral sequences will almost always remain unknowable. Therefore, assumptions must be made during tree reconstruction in the place of ancestral data. These assumptions and the quality of the data will determine whether the final tree is reliable. Methods for constructing a phylogenetic tree can be separated into two major categories, depending on the traits used. These are character- and distance-based methods (Table 2). **See also:** [Molecular Phylogeny Reconstruction](#)

### Gene tree and species tree

When a phylogenetic tree is constructed by using gene sequences, the tree obtained (the 'gene tree') may be different from the historical evolutionary tree of the species (the 'species tree') that can be estimated utilizing information other than sequences, such as morphological differences from fossil records. This is a particular problem when paralogous genes are included in the analysis. Ideally, orthologous genes (those descended from a common ancestor) should be used to construct a phylogeny.

**Table 2** Methods for constructing a phylogenetic tree

Classification of trees	Method	Rooted or unrooted tree
Distance-based methods	UPGMA	Rooted trees
	Neighbour-joining	Unrooted trees
	Minimum evolution	Rooted or unrooted
Character-based methods	Parsimony	Rooted
	Compatibility	Rooted
	Maximum-likelihood	Rooted
	Bayesian MCMC	Rooted

However, if a duplication event has occurred during evolution that results in genes *a* and *b*, then *a* in one descendant species is the paralog of *b* in another (whereas *a* genes of all descendant species are orthologs). Tree reconstruction should involve either *a* or *b* genes, but not a mixture – otherwise the tree will assume shared ancestral nodes that are incorrect. Thus special attention in detecting paralogous genes is required when inferring a species tree from a gene tree. Similar care should be taken when considering horizontally transferred genes, a common process in prokaryotes where foreign DNA is inserted into a host genome through transformation, transduction or bacterial conjugation. In such cases, the evolutionary history of the transferred gene (known as a xenolog) is clearly different to that of the host species. Transferred genes should therefore be obviated when reconstructing species trees. **See also:** [Fossils in Phylogenetic Reconstruction](#); [Gene Duplication and Redundancy](#); [Orthologues, Paralogues and Xenologues in Human and Other Genomes](#)

## Rooted and unrooted trees

Phylogenetic trees can be rooted or unrooted. A rooted tree has a unique ancestral node from which extant OTUs evolved, whereas an unrooted tree represents a network without such an ancestral node. Tree building methods (discussed later) vary in whether they produce a rooted or unrooted tree. For example, the maximum likelihood (ML) method gives a rooted tree whereas the NJ method produces an unrooted tree. However, it is possible to identify the ancestral node in a phylogenetic tree that has been constructed by the NJ method. When considering a group of related sequences, a more distantly related sequence can be included as an 'outgroup' when a phylogenetic tree is constructed. The branching point between the outgroup and the remaining sequences is considered to be the root. However, if the outgroup is too distant to the other genes, the substitution numbers are not estimated correctly due to saturation.

## Methods for Constructing Phylogenetic Trees

### Distance-based methods

Using distance-based methods, the numbers of nucleotide and amino acid substitutions are used as evolutionary distances and should produce a correct phylogenetic tree. In reality, however, the number of substitutions is usually unknown; therefore, many methods for estimating this number have been developed. Most distance methods perform well if the sequence data are (approximately) additive, that is, the distance between two OTUs should be equal to the sum of the connecting branch lengths. However, data that include divergent sequences are often not additive, in which case these methods will

perform poorly. The unweighted pair-group method with arithmetic mean (UPGMA) and the NJ method are typical distance-based methods.

The UPGMA method is a hierarchical clustering algorithm originally developed by Sokal and Michener (1958) for constructing a tree based on the phenotypic similarities between OTUs. Topological relationships are inferred from a distance matrix in the order of decreasing similarity, and a phylogenetic tree is built in a stepwise manner. The distance between two composite OTUs is calculated as the arithmetic mean of the pair-wise distances between the constituent OTUs of the two composite OTUs. Initially, the two most similar OTUs are identified in the distance matrix; they are then treated as a new single, composite OTU. Subsequently, among the new set of OTUs the pair having the highest similarity is identified. The procedure is repeated until only two OTUs are left. UPGMA is a simple yet powerful method that will construct the correct tree if the evolutionary rates are constant in all lineages. If a molecular clock can be assumed, UPGMA guarantees the reconstruction of an *ultrametric* tree, where the branch length of any OTU to the root is the same.

The NJ algorithm of Saitou and Nei (1987) is a more commonly used distance method, which does not assume a clock and guarantees a unique tree when the distance matrix is additive, that is, for any four indices *i*, *j*, *k* and *l*, the two largest sums of distances among them ( $D_{ij} + D_{kl}$ ,  $D_{ik} + D_{jl}$ ,  $D_{il} + D_{jk}$ ) are equal. An advantage of this method is its computational efficiency. The NJ algorithm begins with a star phylogeny to compute the total number of substitutions (*S*), in which all sequences originate equidistantly from a single common root. Note that the observed number *S* will be less than the true number of substitutions, particularly for distantly related sequences that have been subject to saturation. The number of substitutions  $S_{ij}$  is then calculated for a tree in which sequences *i* and *j* are paired separately from the remaining sequences. The first pair of neighbours is chosen by calculating  $S_{ij}$  values for all pairs of sequences and choosing the smallest  $S_{ij}$ . The identified pair is treated as a composite OTU in the next step. This procedure is repeated until the star phylogeny is entirely converted into an unrooted bifurcating tree.

The minimum evolution (ME) method (Rzhetsky and Nei, 1992) is similar to the NJ method, although computationally more expensive. Rather than beginning with the closest pair of sequences (as with the NJ method), the total sum (*S*) of branch lengths is calculated for all possible branching patterns using pair-wise distance data. The branching pattern that yields the lowest *S* value is therefore the most likely tree.

Two other distance methods have been developed that allow for nontree-like characteristics among sequence data. This is particularly useful for sequences that have been subject to convergent or parallel evolution, or for examining kinship among organisms where horizontal gene transfer is common. These methods are spectral analysis, derived from a Fourier transformation method

(Hendy *et al.*, 1994), and split decomposition (Bandelt and Dress, 1992). These methods produce a group of weakly compatible splits or a spectrum of splits as a graph, instead of a tree. In this context, a split is an 'edge' within a phylogenetic tree that defines a boundary between two character states (e.g. 'ears' or 'no ears'; 'Ser' or 'Thr'). Compared with a regular phylogenetic tree, a splits graph may not be planar and it may include reticulate features. However, if the distances used to produce a splits graph are additive, split decomposition and spectral analysis will produce a graph that, to all appearances, is an unrooted, bifurcating tree. Therefore, these methods can also be used to produce a regular tree if the data support a tree-like relationship.

In general, distance methods are consistent (as data grows larger, the estimation converges to the true tree) and computationally efficient. Nevertheless, when the variation of evolutionary rates among sites is large, character-based methods are usually more accurate than distance methods.

## Character-based methods

Character-based methods utilize characters observed in a set of OTUs instead of evolutionary distances. The characters are usually nucleotides or amino acids observed in particular sites of an alignment, although morphological or behavioural characters can also be utilized. Characters with the same value (a position having the same nucleotide, or a common physical attribute) carry no phylogenetic information, since no inferences about the evolutionary history of the species under study can be made. Typical examples of character-based methods are parsimony, ML and Bayesian methods. **See also:** [Homology in Character Evolution](#)

Parsimony assumes that the tree requiring the minimum amount of evolution is the one to be preferred, and thus looks for a phylogeny that can explain the data with the smallest number of changes (reviewed in Swofford *et al.*, 1996). In this sense, parsimony is akin to Occam's razor, since the simplest solution is preferred. However, evolution may not occur in a way that minimizes evolutionary changes, particularly in the cases of convergent or parallel evolution, or reversals to an ancestral state. As a result, parsimony is reasonably good for comparisons of closely related species or genes, but less so for divergent taxa. For distant comparisons, hidden changes may have accumulated to such an extent (saturation) that the parsimony principle cannot detect all changes, leading to foreshortened branch lengths and possibly incorrect relationships. This effect, known as 'long branch attraction', is due to the inconsistency of parsimony as a statistical estimator (Felsenstein, 2004). Parsimony does not guarantee a unique solution, and in practice more than one tree can be equally parsimonious. In such cases, a consensus tree can be used to illustrate the results by showing multifurcations where various trees differ. Alternatively, methods to prune the solution space and produce a unique

tree have been proposed (Swofford and Maddison, 1987). The reconstruction of ancestral sequences using parsimony methods is straightforward and computationally efficient.

Compatibility methods (Le Quesne, 1969) assume that the most likely phylogeny is the one that maximizes the number of compatible characters, where a character is compatible with a tree if it can evolve in it without homoplasy (convergent evolution). If the characters under study are binary and no data is missing, compatibility can be calculated very efficiently in most cases by searching cliques of compatible characters. Unfortunately, nucleotide or amino acid sequences need a different approach, called *perfect phylogeny*, which has been proved to be a computationally hard problem to solve.

Although likelihood methods for phylogenies were introduced by Edwards and Cavalli-Sforza (1964), much of their current popularity is owed to the optimization introduced by Felsenstein (1981). Rather than producing a model from the sequence data, the ML method seeks the tree that is most likely to have produced the data. Given an alignment, a phylogeny with branch lengths and an evolutionary model, ML calculates the likelihood of the tree as a product of the likelihoods for each site, that is, assuming that evolution at different sites on the tree is independent. The likelihood of the tree for a site is then the sum of the probabilities of each possible scenario over all nucleotides that may have existed at the interior nodes of the tree. By assuming that lineages evolve independently, this summation can be further simplified. ML methods can be rather sensitive to the evolutionary model that has been chosen, and they are computationally expensive.

Bayesian methods (reviewed in Huelsenbeck *et al.*, 2001) have recently gained much attention in the reconstruction of phylogenetic trees in combination with Markov chain Monte Carlo (MCMC) methods. Bayesian methods are similar to ML methods, although they use a prior distribution that reflects the uncertainty about the tree being inferred. The posterior distribution can be difficult to compute, since it involves all possible hypotheses (trees). Thus MCMC is used instead to sample trees by moving randomly across the search space. Despite the passionate debate on whether priors can be agreed on or not, Bayesian methods are being increasingly used in phylogenetics with robust results.

## Tree evaluation

Once a phylogenetic tree is obtained, there are several ways to evaluate its accuracy in representing the evolutionary history of the underlying data. Consideration should also be given to the estimation of branch lengths and the comparison of tree topologies (e.g. for trees produced from the same data using different methods). The bootstrap test (Felsenstein, 1985) involves re-sampling from the data with replacement, and making samples of the same

size as the original. For each of the new samples, we infer a phylogeny (a bootstrap estimate) utilizing the desired tree reconstruction method, and the proportion of estimates in which a given cluster appears is calculated. This number represents the confidence in the cluster. However, the accuracy of bootstrap values can be over- or underestimated depending on factors such as variation in substitution rates. **See also:** [Molecular Phylogeny Reconstruction](#)

## Conclusions

DNA sequence analysis is a powerful tool for predicting gene functions and inferring evolutionary relationships among genes and species by comparative studies of nucleotide and amino acid sequences. As DNA and protein sequences have accumulated rapidly with the advancement of sequencing techniques and the progress of several genome projects, the analysis of sequence data has thus become increasingly important. This article has provided a survey of the principal DNA databanks and genome repositories, as well as a description of the tools necessary for a comprehensive analysis of sequence data.

## References

- Ahituv N, Zhu Y, Visel A *et al.* (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biology* **5**: e234.
- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.
- Bandelt HJ and Dress AW (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* **1**: 242–252.
- Bernstein BE, Kamal M, Lindblad-Toh K *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**: 169–181.
- Dayhoff MO, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed.) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 345–352. Washington, DC: National Biomedical Research Foundation.
- Edwards AWF and Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. In: Heywood VH and McNeill J (ed.) *Phenetic and Phylogenetic Classification*, pp. 67–76, Publ. No. 6. London: Systematics Association.
- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum-likelihood approach. *Journal of Molecular Evolution* **17**: 368–376.
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.
- Felsenstein J (2004) *Inferring Phylogenies*. Sunderland, MA: Systematics Association.
- Gerstein MB, Bruce C, Rozowsky JS *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17**: 669.
- Hendy MD, Penny D and Steel MA (1994) A discrete Fourier analysis for evolutionary trees. *Proceedings of the National Academy of Sciences of the USA* **91**: 3339–3343.
- Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**(5550): 2310–2314.
- Imanishi T, Itoh T, Suzuki Y *et al.* (2004) Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biology* **2**: e162.
- Jukes TH and Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed.) *Mammalian Protein Metabolism*, pp. 21–132. New York: Academic Press.
- Katoh K, Misawa K, Kuma K and Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**: 3059–3066.
- Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequence. *Journal of Molecular Evolution* **16**: 111–120.
- Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Le Quesne WJ (1969) A method of selection of characters in numerical taxonomy. *Systematic Biology* **51**: 217–234.
- Li W-H, Wu C-I and Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**: 150–174.
- Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**: 418–426.
- Notredame C, Higgins DG and Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**: 205–217.
- Rice Annotation Project (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Research* **36**: D1028–D1033.
- Rzhetsky A and Nei M (1992) A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* **9**: 945–967.
- Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406–425.
- Sokal RR and Michener CD (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **28**: 1409–1438.
- Swarbreck D, Wilks C and Lamesch P (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research* **36**: D1009–D1014.
- Swofford DL and Maddison DR (1987) Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences* **87**: 199–229.
- Swofford DL, Olsen GJ, Waddell PJ and Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C and Mable BK (eds) *Molecular Systematics*, 2nd edn, pp. 411–501. Sunderland, MA: Sinauer Associates.
- The ENCODE Project Consortium (2007) Identification and analysis of functional elements in of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.
- The FANTOM Consortium (2005) The transcriptional landscape of the mammalian genome. *Science* **309**: 1559–1563.

- Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673–4680.
- Zuckerkanndl E and Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V and Vogel VH (eds) *Evolving Genes and Proteins*, pp. 97–166. New York: Academic Press.
- Further Reading**
- Altschul SF, Boguski MS, Gish W and Wootton JC (1994) Issues in searching molecular sequence databases. *Nature Genetics* **6**: 119–129.
- Cooper GM and Brown CD (2008) Qualifying the relationship between sequence conservation and molecular function. *Genome Research* **18**: 201–205.
- Database issue (2008) *Nucleic Acids Research* **36**(1).
- DNA Database of Japan (DDBJ) [[www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)].
- European Bioinformatics Institute (EBI) [[www.ebi.ac.uk](http://www.ebi.ac.uk)].
- Fitch WM (2000) Homology. *Trends in Genetics* **16**: 227–231.
- Gojobori T, Moriyama EN and Kimura M (1990) Statistical methods for estimating sequence divergence. *Methods in Enzymology* **183**: 531–550.
- GOLD [<http://www.genomesonline.org/>].
- Graur D and Li W-H (2000) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.
- Mount DW (2004) *Bioinformatics: Sequence and Genome Analysis*, 2nd edn. Woodbury, NY: Cold Spring Harbor Laboratory Press.
- National Center for Biotechnology Information (NCBI) [[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)].
- Nei M (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics* **30**: 371–403.
- Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press.
- Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.