

Multiple Sequence Alignment

We frequently need to compare more than two sequences. The multiple sequence alignment is one of the most widely used methods for that purpose. Unfortunately, computing the alignment of more than two sequences is a very difficult task and remains unsolved. Major issues in the multiple sequence alignment are:

1. Given scoring matrices, there is no way directly compare more than two characters.
2. Even with a method to compare multiple characters, computing an optimal alignment is too expensive in terms of computing time and space.

A Progressive Approach to the Multiple Sequence Alignment

After computing all pairwise alignments, the best aligning pair is progressively merged into one. **CLUSTALW** is the most widely used multiple sequence alignment method and it employs the progressive alignment approach.

Try `clustalw seqsfile` and a file `seqsfile.aln` will have the resulting alignment.

A Multiple Sequence Alignment

CLUSTAL W (1.74) multiple sequence alignment

```
gi|2687881|gb|AAC66392.1| -----MMAKIIGI
gi|2688007|gb|AAC66509.1| -----MIIAI
gi|2687944|gb|AAC66451.1| MSIKTVKDFSSFAGKRALVRCDFNVPLKEGSISDDTRIRAALSTIEYLKE
                                                                    :
```

```
gi|2687881|gb|AAC66392.1| SGGSGSGKTTVVS KIS-----EFIP-----EFVLISQDNYYKSVG DYEH
gi|2688007|gb|AAC66509.1| DGPSASGKSSIARELGVRLNYKFISSGHL YRIITLIAQRSLMNSC-DFIS
gi|2687944|gb|AAC66451.1| RGARIVLVSHLGRPDGKKNPKYSLKPVANRLSELLGQDVKMLSDCIGSEV
*           : :           .           :           * : . . .
```

```
gi|2687881|gb|AAC66392.1| EFSKVNFDHP-DA-FDNNLFY EHLK NLK---KNSPIDMPLYDFINHKRQL
gi|2688007|gb|AAC66509.1| EDSLLNLILENDISFNNFAFL L NGENVENQILNDKIDFQVS-FYSSYVGI
gi|2687944|gb|AAC66451.1| VNSTLQMKDGDVV LLENVRFY AEEEEKNDK-NFAKKLSENGDVFVNDAFGA
*  :::           ::* * . :: .           . :. * .
```

Phylogenetic Tree

Given a set of sequences, computing a multiple sequence alignment and a phylogenetic tree is two sides of a coin. With knowledge on the phylogenetic relationship among the sequences, we can generate more accurate multiple sequence alignment. With a accurate, biologically meaningful multiple sequence alignment, we can generate a phylogenetic tree easily.

Approaches to the Phylogenetic Tree Computation

There are four different approaches to the phylogenetic tree computation.

1. UPGMA, unweighted pair group method using arithmetic average.
2. Neighbor joining.
3. Maximum parsimony.
4. Maximum likelihood.

The PHYLIP Package

The PHYLIP Package provides a comprehensive set of tools for the phylogenetic tree computation. (see the web page on our course home page).

PROTPARS. Estimates phylogenies from protein sequences (input using the standard one-letter code for amino acids) using the parsimony method, in a variant which counts only those nucleotide changes that change the amino acid, on the assumption that silent changes are more easily accomplished.

DNAPARS. Estimates phylogenies by the parsimony method using nucleic acid sequences. Allows use the full IUB ambiguity codes, and estimates ancestral nucleotide states. Gaps treated as a fifth nucleotide state.

DNAMOVE. Interactive construction of phylogenies from nucleic acid sequences, with their evaluation by parsimony and compatibility and the display of reconstructed ancestral bases. This can be used to find parsimony or compatibility estimates by hand.

DNAPENNY. Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search. This may not be practical (depending on the data) for more than 10 or 11 species.

DNACOMP. Estimates phylogenies from nucleic acid sequence data using the compatibility criterion, which searches for the largest number of sites which could have all states (nucleotides) uniquely evolved on the same tree. Compatibility is particularly appropriate when sites vary greatly in their rates of evolution, but we do not know in advance which are the less reliable

The PHYLIP Package

DNAINVAR. For nucleic acid sequence data on four species, computes Lake's and Cavender's phylogenetic invariants, which test alternative tree topologies. The program also tabulates the frequencies of occurrence of the different nucleotide patterns. Lake's invariants are the method which he calls "evolutionary parsimony".

DNAML. Estimates phylogenies from nucleotide sequences by maximum likelihood. The model employed allows for unequal expected frequencies of the four nucleotides, for unequal rates of transitions and transversions, and for different (prespecified) rates of change in different categories of sites, with the program inferring which sites have which rates.

DNAMLK. Same as DNAML but assumes a molecular clock. The use of the two programs together permits a likelihood ratio test of the molecular clock hypothesis to be made.

DNADIST. Computes four different distances between species from nucleic acid sequences. The distances can then be used in the distance matrix programs. The distances are the Jukes-Cantor formula, one based on Kimura's 2-parameter method, Jin and Nei's distance which allows for rate variation from site to site, and a maximum likelihood method using the model employed in DNAML. The latter method of computing distances can be very slow.

PROTDIST. Computes a distance measure for protein sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix, Kimura's 1983 approximation to it, or a model based on the genetic code plus a constraint on changing to a different category of amino acid. The distances can then be used in the distance matrix programs.

The PHYLIP Package

RESTML. Estimation of phylogenies by maximum likelihood using restriction sites data (not restriction fragments but presence/absence of individual sites). It employs the Jukes-Cantor symmetrical model of nucleotide change, which does not allow for differences of rate between transitions and transversions. This program is VERY slow.

SEQBOOT. Reads in a data set, and produces multiple data sets from it by bootstrap resampling. Since most programs in the current version of the package allow processing of multiple data sets, this can be used together with the consensus tree program CONSENSE to do bootstrap (or delete-half-jackknife) analyses with most of the methods in this package. This program also allows the Archie/Faith technique of permutation of species within characters.

FITCH. Estimates phylogenies from distance matrix data under the "additive tree model" according to which the distances are expected to equal the sums of branch lengths between the species. Uses the Fitch-Margoliash criterion and some related least squares criteria. Does not assume an evolutionary clock. This program will be useful with distances computed from DNA sequences, with DNA hybridization measurements, and with genetic distances computed from gene frequencies.

KITSCH. Estimates phylogenies from distance matrix data under the "ultrametric" model which is the same as the additive tree model except that an evolutionary clock is assumed. The Fitch-Margoliash criterion and other least squares criteria are assumed. This program will be useful with distances computed from DNA sequences, with DNA hybridization measurements,

The PHYLIP Package

CONTML. Estimates phylogenies from gene frequency data by maximum likelihood under a model in which all divergence is due to genetic drift in the absence of new mutations. Does not assume a molecular clock. An alternative method of analyzing this data is to compute Nei's genetic distance and use one of the distance matrix programs.

GENDIST. Computes one of three different genetic distance formulas from gene frequency data. The formulas are Nei's genetic distance, the Cavalli- Sforza chord meas

An Example of PHYLIP Package Use

As the Phylip have many tools in one package, output from a tool can be input to another tool, which allows more complex analysis possible. In general, this kind feature is desirable allowing more flexibility for the users. Developing more generalized sequence analysis framework is a good choice for the term project.

An example phylogeny analysis:

1. Read a multiple sequence alignment using PROTDIST and create a distance matrix.
2. Input the distance matrix to NEIGHBOR and generate a phylogeny based on neighbor joining method.
3. Read the phylogeny into DRAWTREE and produce an unrooted phylogenetic tree.

The multiple sequence alignment can be generated using ClustalX.

Tree Representation

A tree can be drawn using DRAWTREE and DRAWGRAM. However, visualizing the tree is different from computing on trees.

An example of a phyogeny representation:

$((A:1.2, B:1.5):1.0, (C:2.5, D:0.5):1.2)$

\Rightarrow draw the tree manually.

To compute on trees, we need to represent the tree using a graph package; for example, GRAPH.PM for PERL and LEDA and BOOST packages for C++.

JalView and ClustalX

JALVIEW is another widely used package for the phylogenetic tree construction. It is written in **JAVA** and provides multiple methods for the phylogenetic tree construction and sequence handling.

CLUSTALX also provides functionalities similar to **JALVIEW**.

However, **PHYLIP** is more comprehensive than the two.

Blocks, Profiles, Motifs, and PSSMs.

From a multiple sequence alignment, we can extract conserved regions, which can be transformed into blocks, a profile, sequence motifs, and even a new type of scoring matrix called *position specific scoring matrix* (PSSM).

Sequence motif discovery is often carried out *without* alignment.

There are different class of sequence motif discovery algorithms, PRATT and TREISIAS.

Profiles

Profiles assigns an average of standard substitution scores from all the residues seen in the corresponding column.

Seq1 . . . VGA--HAGEY

Seq2 . . . VEA--DVAGH

Seq3 . . . FNA--NIPHK

Seq4 . . . IAGADNGAGV

Aligning the character **a** to the first column will give an average score:

$$\frac{2}{4}score(V, a) + \frac{1}{4}score(F, a) + \frac{1}{4}score(I, a).$$

To complete a *scoring scheme*, we need to define the gap penalties.

Position Specific Scoring Matrix(PSSM)

Let's assume that we have a model M for a set of seqs and a seq x of length L . Then, the probability of x given M can be written as

$$P(x|M) = \prod_{i=1}^L p(x_i)$$

However, $P(x|M)$ is length dependent and we also want compare with the random model based on the background prob.

$$Score(x) = \sum_{i=1}^L \log \frac{p(x_i)}{q_{x_i}}$$

where q_{x_i} denotes a random probability based on the background prob.

Why is it called a scoring matrix?

To answer this question, you need to understand how a scoring matrix like BLOSUM 62 is constructed. (we will come back to this topic later.)

Databases of Blocks, Profiles, Motifs, PSSMs, and HMMs.

1. **Blocks** is a database of ungapped multiple seq. alignments (conserved regions).
2. **PROSITE** is a database of sequence patterns.
3. **Pfam** is a database of profile Hidden markov Models (HMMs).
4. **PRINTS** is database of *multiple* sequence patterns.

HMMer

HMMer is a package for building and searching profile HMMs. There are many tools (see p213 in our textbook). Among them are:

- HMMALIGN reads both a seq. file and a profile HMM and create a multiple sequence alignment.
- HMMBUILD builds a profile HMM from a multiple seq. alignment.
- HMMPFAM searches a profile HMM database (Pfam) with a query sequence.
- HMMSEARCH searches a sequence database with a profile HMM.

MEME

The MEME package provides three tools.

- MEME discovers motifs in a set of *unaligned* sequences.
- MAST takes a motif discovered by MEME and uses it to search a sequence databases.
- METAMEME constructs a model from multiple MEME motifs and uses it to search a sequence database.

PSI-BLAST and PHI-BLAST

PSI-BLAST is Position Specific Iterated BLAST and works in two major steps.

- With a scoring matrix (BLOSUM62) and a query, a database is searched with a standard BLAST.
- With the set of matches from the search, a PSSM is constructed by aligning the matches to the query. Then the next round of database search is performed *with the PSSM* instead of the stand scoring matrix.
- Users can specify how many position specific iterations will be performed.

Try `blastpgp`!

PHI-BLAST is a pattern hit initiated BLAST and works the same way as PSI-BLAST except that matches are required to have the input pattern.