Multiple alignment and phylogeny of biological sequences

Peter Sestoft
sestoft@dina.kvl.dk
Department of Mathematics and Physics, KVL

2004-02-05

Literature:

Lesk: *Introduction to bioinformatics*, Oxford U P 2002, page 187–207.

Optional reading:

Claverie and Notredame: *Bioinformatics for Dummies*, Wiley 2003, chapter 9 and 13.

Thompson, Higgins, Gibson: Clustal W: Improving the sensitivity . . . , *Nucleic Acids Research* 22 (1994) 4673–4680.

Notredame, Higgins, Heringa: T-Coffee: A novel method. . . , *Journal of Molecular Biology* 302 (2000) 205–217.

Gupta, Kececioglu, Schäffer: Improving the practical space and time efficiency. . . , *Journal of Computational Biology* 1995

Durbin et al: *Biological Sequence Analysis*, Cambridge University Press 1998, sections 6.1-6.4 and 7.1-7.5.

Multiple alignment

A multiple alignment of seven globins (part):

```
HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKG ...
HBB_HORSE    ... DSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKG ...
HBA_HUMAN    ... PHF-DLS-----HGSAQVKGHGKKVADALTNAVAHVDD-----MPN ...
HBA_HORSE    ... PHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPG ...
GLB5_PETMA   ... PKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSM ...
MYG_PHYCA    ... DRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEA ...
LGB2_LUPLU   ... SFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDA ...
                    :   .       . .:: *.   :   .
```

Why multiple alignment?

- Identify (protein) families and find characteristic conserved motifs
- Relate conserved motifs to physical structure and biological activity
- Predict physical structure and biological activity from motifs
- Improve precision of database searches (PSI-Blast)

The score of a multiple alignment

Let K sequences $s^1, s^2, \ldots, s^K$ be given.

A multiple alignment a has K rows, where row k contains precisely the symbols of s^k, and possibly some gaps.

The multiple alignment induces a pairwise alignment $a^{k,l}$ between s^k and s^l for every pair (s^k, s^l) of sequences.

The alignment $a^{k,l}$ is obtained by deleting all other rows of a and then disregarding columns that have only gaps.

The score $S(a)$ of a multiple alignment a is usually computed as the 'sum of pairs' score:

$$S(a) = S(a^{1,2}) + S(a^{1,3}) + \cdots + S(a^{1,7}) + S(a^{2,3}) + \cdots + S(a^{6,7})$$

Here $S(a^{k,l})$ is the score of the best pairwise alignment between s^k and s^l.

This 'sum of pairs' score has no good statistical justification — unlike the scores for pairwise alignment.

Multiple alignment of 7 sequences induces $\frac{7 \cdot 6}{2} = 21$ pairwise alignments

```
HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKG ...
HBB_HORSE    ... DSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKG ...

HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKG ...
HBA_HUMAN    ... PHF-DLS-----HGSAQVKGHGKKVADALTNAVAHVDDMPN ...

HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKG ...
HBA_HORSE    ... PHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDDLPG ...

HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN---LKG ...
GLB5_PETMA   ... PKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDDTEKMSM ...

HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKG ...
MYG_PHYCA    ... DRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEA ...

HBB_HUMAN    ... ESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKG ...
LGB2_LUPLU   ... SFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDA ...

.....

MYG_PHYCA    ... DRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEA ...
LGB2_LUPLU   ... SFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDA ...
```

Algorithms for multiple alignment

- Simultaneous multiple alignment (example: the MSA program).

  Produces the 'true' optimal alignment.

  Very time and space consuming; feasible only for few sequences which are highly similar, or short.

- Progressive multiple alignment (examples: the ClustalW and T-Coffee programs).

  A heuristic method that may or may not produce the true optimal alignment.

  But it is much faster than simultaneous multiple alignment.

- Manually edit a multiple alignment (*very* boring).

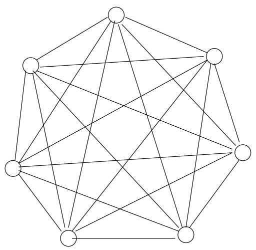
  Example: Jalview, `http://www.jalview.org/`

  Can be downloaded and installed locally.

Building a multiple alignment

A circle represents a sequence.

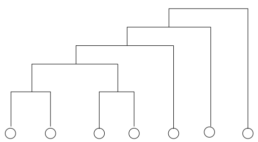
Simultaneous



MSA

Progressive

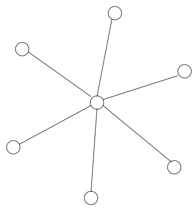
(tree)



ClustalW

Progressive

(star)



PSI-Blast

Simultaneous multiple alignment

In principle it is easy: use dynamic programming, filling in a matrix F, just as for pairwise alignment.

In pairwise alignment (two sequences), the F matrix is two-dimensional.

In a multiple alignment of K sequences, the F matrix will be K-dimensional.

Problem: Assume we have K sequences of length $\ell_1, \dots, \ell_K$.

Then the size of the F matrix will be

$$\ell_1 \cdot \ell_2 \cdot \ldots \cdot \ell_K$$

When $K = 6$ and $\ell_1 = \ell_2 = \cdots = \ell_K = 200$, this is 64,000,000,000,000 numbers.

This requires 256,000,000 MB storage, corresponding to 1 million PCs.

Far too large for present-day and foreseeable computers.

The MSA program

MSA (Multiple Sequence Alignment, Kececioglu et al, 1989, 1995) implements simultaneous multiple alignment.

The 1995 version of MSA has many algorithmic improvements over naive dynamic programming.

MSA fills in only a small part of the K-dimensional F matrix, a 'band' around the 'diagonal' (like Blast).

The size of this 'band' is determined dynamically, by a form of branch-and-bound optimization.

Therefore MSA may easily handle six closely related sequences of length ≥ 200.

But still, adding a single more distantly related sequence can make it hopelessly slow.

This is because the 'band' needs to be much 'wider' when the sequences are not closely related.

In fact the *problem* of optimal multiple sequence alignment is *NP-hard*.

It is very unlikely that there is an acceptably efficient general algorithm that solves it.

Computer experiments with MSA: time and space consumption

Computing multiple alignment of some bcII class B beta-lacamases; length approximately 260 aa.

Number of seqs	Time/s	Space/MB	Comments
4	<1	1	4 closely related
5	8	4	5 closely related
6	30	6	6 closely related
6+1	$>$4380	$>$272	6 closely related + 1 more remotely (program had to be killed)

The practical solution: Progressive alignment (Feng and Doolittle 1987)

Given K sequences to align (globally), do not align all sequences at the same time. Instead:

- Compute the $\frac{K(K-1)}{2}$ pairwise distances between all pairs (s^k, s^l) of sequences.
  This 'distance' can be computed from the score $S(a^{kl})$ of the global pairwise alignment of s^k and s^l.
- Using the distances, build a tree (as in phylogeny) to guide the progressive multiple alignment process.
- To build a multiple alignment, start with closely related sequences, add more distantly related sequences later.

Problem: the result is sensitive to the order in which sequences are added.

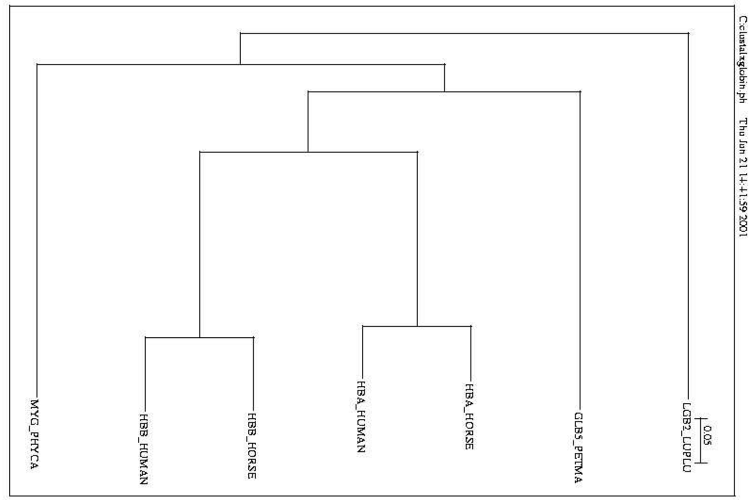
The order should follow the phylogeny, but the phylogeny is not known until the alignment is finished . . .

To align the two first sequences, use ordinary pairwise alignment by dynamic programming (explained last week).

To align a new sequence s to an existing alignment a, compute the pairwise alignment between a and every sequence in a. Then choose the best of these alignments, inserting gaps in the other sequences of a as necessary. Always preserve existing gaps in a.

To align an alignment a_1 to another alignment a_2, compute a pairwise alignment between every sequence in a_1 and every sequence in a_2. Then choose the best of these alignments, inserting gaps in the other sequences of a_1 and a_2 as necessary. Always preserve existing gaps in a_1 and a_2.

An example guide tree for the globins (built by ClustalW)



ClustalW/ClustalX (1994): Further refinements

The pairwise distances used to build the guide tree are weighted by sequence similarity, so that closely related sequences are given a lower weight. This is to compensate for sequence selection bias.

The guide tree is built using Neighbour Joining to build the guide tree; an older version of ClustalW used UPGMA.

Neighbour Joining and UPGMA ('unweighted pair group method using arithmetic averages') are two methods for building a clustering/phylogenetic tree, given a distance matrix. Neighbour Joining produces better results.

Use 'harder' substitution matrices for the early alignments (closely related sequences), e.g. GONNET120.

Use 'softer' substitution matrices for the later alignments (distantly related sequences), e.g. GONNET350.

Gap opening cost is increased near existing gaps (to encourage fewer, longer gaps), and near hydrophobic amino acids (because they are usually inside the protein, where insertions and deletions are less likely to survive).

Gap extension cost is decreased when aligning a new sequence to a gapped stretch of an existing alignment.

The graphical user interface ClustalX permits some visualization and very limited sequence editing.

T-Coffee (2000): More precise multiple alignments

In ClustalW, an early mistake in progressive alignment cannot be corrected later. Consider these sequences:

```
1  GARFIELD THE LAST FAT CAT
2  GARFIELD THE FAST CAT
3  GARFIELD THE VERY FAST CAT
4  THE FAT CAT
```

The sequences will be aligned (1,2), then ((1,2),3), then (((1,2),3),4), giving the wrong (left) alignment:

```
GARFIELD THE LAST FA-T CAT          GARFIELD THE LAST FA-T CAT
GARFIELD THE FAST CA-T ---          GARFIELD THE FAST ---- CAT
GARFIELD THE VERY FAST CAT          GARFIELD THE VERY FAST CAT
-------- THE -----FA-T CAT          -------- THE ---- FA-T CAT
```

T-Coffee (Notredame, Higgins, Heringa 2000) improves on ClustalW/ClustalX as follows:

- Build a 'library' of all pairwise local and all pairwise global alignments between the sequences.
- For each matching pair of amino acids from two sequences, record a weighted constraint.
  For instance, C in sequence 2 is likely to match C in sequence 1, sequence 3 and sequence 4.
- Use the information from the library when doing the progressive multiple alignment.

In the exercises you will see that this makes a difference in real life.

Servers, and other multiple alignment tools

- ClustalW is available for download at `ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/`
- There is a ClustalW server at EBI `http://www.ebi.ac.uk/clustalw/`
- To edit ClustalW trees, and to include them in Word documents, use Treeview:
  `http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`
- T-Coffee server at `http://igs-server.cnrs-mrs.fr/Tcoffee/`
- Multalin, looks similar to ClustalW, at `http://www.toulouse.inra.fr/multalin.html`
- AMPS (Alignment of Multiple Protein Sequences) = multalign
  `http://www.compbio.dundee.ac.uk/Software/Amps/amps.html`
- Pileup/GCG from Genetics Computer Group (commercial)
- Many more tools, including manual editing tools, are listed at `http://www.expasy.org/tools/`
  Example: Jalview, `http://www.jalview.org/`
  Example: CINEMA, `http://bioinf.man.ac.uk/dbbrowser/CINEMA2.1/`

Motifs and multiple alignments

A *motif* is a section of a multiple alignment where many sequences agree.

There are many ways to represent motifs:

Database	Representation	URL
BLOCKS	Blocks	www.blocks.fhcrc.org/
PROSITE patterns	Regular expressions	www.expasy.org/prosite/
PROSITE profiles	Profiles (position-specific scores)	www.expasy.org/prosite/
Pfam	Hidden Markov Models	www.sanger.ac.uk/Pfam/
Interpro	Integration of other databases	http://www.ebi.ac.uk/interpro/

Usually based on high-quality sequences from SwissProt.

A BLOCKS block, or a PROSITE pattern or profile, or a Pfam HMM, (partially) characterizes a protein family.

Given a PROSITE pattern or profile, one can find protein sequences that match the profile.

Given a sequence, one can find a matching PROSITE pattern or profile, and hence the protein family.

BLOCKS motifs

A BLOCKS motif is an ungapped multiple alignment of related sequences.

BLOCKS motif for class B beta-lactamases (pattern IPB001018A):

```
BLAB_AERHY|P26918  (  87) PVLEVINTNYHTDRAGGNAYWKSIGAKVVSTRQTRDL  34
BLAB_BACFR|P25910  (  91) KVTTFIPNHWHGDCIGGLGYLQRKGVQSYANQMTIDL 100
BLAB_SERMA|P52699  (  87) KIKGSISSHFHSDSTGGIEWLNSRSIPTYASELTNEL  31
BLA1_XANMA|P52700  ( 154) ANADRIVMDGEVITVGGIVFTAHFMAGHTPGSTAWTW  72
BLA2_BACCE|P04190  ( 108) RVTDVIITHAHADRIGGIKTLKERGIKAHSTALTAEL  35
BLAB_BACCE|P14488  ( 107) RVTDVIITHAHADRIGGMKTLKERGIKAHSTALTAEL  43
BLA2_BACSP|P10425  ( 108) RVTDVIITHAHADRIGGITALKERGIKAHSTALTAEL  41
O51899             ( 154) ASADRIIMDGEVVTVGGIAFTAHFMPGHTPGSTAWTW  59
O67103             (  88) PVIYAIVTHYHLDHWYGAKTFKDKGAKVIAHEKLKEF  72
O67667             (  90) PIRFLVVTHYHTDHFYGAKAFREVGAEVIAHEWAFDY  89
O68919             (  76) PVRLLVNTHFHGDHSFGNGILGKDAVIVAHERTRTEM  82
Q02057             (  62) PGRTVVNTHFHGDHAFGNQVFAPGTRIIAHEDMRSAM 100
...
```

The number to the right is the weight of the sequence (100 = most distant from consensus).

The BLOSUM substitution matrices were created from BLOCKS motifs.

PROSITE patterns

A PROSITE pattern is *regular expression* describing a motif.

Example: A PROSITE pattern for class B beta-lactamases, signature 1 (accession number PS00743):

```
[LI]-x-[STN]-[HN]-x-H-[GSTA]-D-x(2)-G-[GP]-x(7,8)-[GS]
```

Begins with `L` or `I` followed by 1 aa followed by `S` or `T` or `N` followed by `H` or `N` followed by 1 aa followed by `H` etc.

Ends with `G` or `P` followed by 7 to 8 aa followed by `G` or `S`.

All seven matches in SwissProt release 40.7 for the PROSITE pattern PS00743:

```
BLA1_XANMA/102-122     LlSHaHADhaGPvaelkrrtG
BLA2_BACCE/113-132     IiTHaHADriGGiktl-kerG
BLA2_BACSP/113-132     IiTHaHADriGGital-kerG
BLAB_AERHY/92-111      InTNyHTDraGGnayw-ksiG
BLAB_BACCE/112-131     IiTHaHADriGGmktl-kerG
BLAB_BACFR/96-115      IpNHwHGDciGGlgyl-qrkG
BLAB_SERMA/92-111      IsSHfHSDstGGiewl-nsrS
```

Drawback of PROSITE patterns: either too specific (many false negatives) or too lenient (many false positives).

PROSITE profiles, or weight matrices

Part of PROSITE profile for the SH3 domain (PS50002); occurs in many intracellular or membrane proteins:

```
/M: SY='D'; M=-4,6,-24,8,3,-27,3,-8,-25,-2,-23,-16,4,-7,-2,-7,1,-8,-20,-29,-19,0;
/M: SY='D'; M=-8,13,-27,16,7,-26,11,1,-31,-6,-26,-19,11,-13,-2,-9,2,-10,-27,-28,-16,2
/M: SY='W'; M=-20,-40,-50,-40,-30,10,-20,-29,-20,-20,-20,-20,-40,-30,-20,-20,-40,-30,
/M: SY='W'; M=-16,-32,-32,-33,-26,11,-23,-20,-11,-19,-7,-9,-31,-29,-18,-18,-30,-19,-1
/M: SY='R'; M=-11,-5,-27,-5,4,-14,-21,-8,-15,9,-10,-6,-5,-15,2,10,-9,-7,-13,-18,-5,2;
/M: SY='G'; M=15,-16,-14,-20,-20,-19,17,-23,-11,-18,-12,-8,-11,-20,-19,-21,0,-8,0,-23
/M: SY='R'; M=-12,-2,-26,-1,12,-23,-20,-2,-22,16,-17,-9,-1,-14,10,23,-5,-7,-17,-24,-1
/M: SY='N'; M=-8,-2,-10,-6,-9,-11,-18,-5,-9,-9,-7,-6,2,-19,-9,-7,-3,-3,-7,-29,-8,-10;
/M:         M=-8,-5,-22,-7,-4,-14,-15,-6,-8,-4,-6,-3,-3,-16,-5,-4,-7,-6,-7,-26,-9,-5;
/I:         I=-4; MI=-12; MD=-12; IM=-12;
/M: SY='R'; M=-1,0,-12,-1,-2,-12,-4,-6,-11,1,-10,-7,1,-8,-2,3,3,3,-6,-16,-8,-3; D=-2;
/I:         I=-4; MI=0; MD=-12; IM=0; DM=-12; E1=-50;
/M: SY='T'; M=-2,2,-13,-1,-2,-11,-7,-9,-11,-5,-10,-8,3,-10,-4,-7,7,12,-7,-21,-9,-3; D
/I:         I=-4; MI=-12; IM=-12; DM=-12;
/M: SY='G'; M=-5,0,-23,0,-7,-26,28,-13,-30,-9,-23,-16,4,-19,-11,-9,-2,-13,-23,-26,-23
/M: SY='Q'; M=-10,3,-28,4,15,-28,-14,1,-24,13,-21,-10,3,-8,17,11,-2,-8,-23,-25,-13,15
/M: SY='E'; M=-8,-6,-25,-6,10,-18,-21,-8,-13,4,-13,-7,-6,-10,8,7,-3,-3,-11,-22,-11,8;
/M: SY='G'; M=0,-10,-30,-10,-19,-29,67,-20,-39,-20,-28,-19,-1,-20,-19,-20,-1,-20,-29,
```

`M` is match state with position-specific score for the 20 aa + BZ in the order ABCDEFGHIKLMNPQRSTVWYZ.

`I` is insert state; `MI` is gap-open score; etc.

The scores are computed from multiple alignments (not obvious how).

Searching for sequences matching a given profile

Some of the 287 sequences SwissProt 40.7 matched by profile PS50002:

```
...
VINE_HUMAN/380-439     .....VDKNWLEGEHH-................-GRLGIFPANYVEVLPA
VINE_HUMAN/454-515     .....VNENWYEGRITG................TGRQGIFPASYVQVSRE
VINE_HUMAN/617-671     .....CDDGWFVGVSRR................TQKFGTFPGNYVAPV--
VINE_MOUSE/444-503     .....VDKNWLEGEHH-................-GRLGIFPANYVEVLPA
VINE_MOUSE/523-579     .....VNEHWYEGRITG................TGRQGIFPASYVQINRE
VINE_MOUSE/679-733     .....CDDGWFVGVSRR................TQKFGTFPGNYVAPV--
ITN1_HUMAN/740-806     ..sqtGEPGWLGGELK-................-GKTGWFPANYAEKIPE
ITN1_HUMAN/913-971     .....-QDMWWFGEVQ-................-GQKGWFPKSYVKLISG
ITN1_HUMAN/1002-1060   .....-DGDWWTGTVG-................-DKAGVFPSNYVRLKDS
ITN1_HUMAN/1074-1138   .....NPGGWWEGELQArg..............kKRQIGWFPANYVKLLSP
ITN1_HUMAN/1155-1214   .....EDPDWWKGEVN-................-GQVGLFPSNYVKLTTD
YKA7_CAEEL/165-226     .....VDAEWFYGENQN................QRTFGIVPSSYLDIKIP
YKA7_CAEEL/245-304     .....VNEEWLEGECF-................-GRSGIFPSQFVDCPNL
YKA7_CAEEL/329-388     .....IDAQWLLAECR-................-GQQGMVPKTFLGPPYG
YKA7_CAEEL/405-467     .....VDDYYYKGKLEAf...............kTLPAGILPKNIVKLEN-
```

Note: `G` column highly preserved (high scores); others rather poorly preserved.

Yet the PROSITE profile rather accurately characterizes a family of proteins.

PSI-Blast: Position-Specific Iterated Blast

The sensitivity of a Blast database search can be considerably improved as follows:

(1a) Given a query sequence x of length L, do a Blast search against x using e.g. BLOSUM62; go to (2)

(1b) Given a position-specific score matrix (profile) of length L, make a Blast search against that matrix.

(2) Create a multiple alignment of the significant hits found.

(3) Create a position-specific score matrix (profile) from this multiple alignment.
This is a $20 \times L$ matrix, giving position-specific scores (for each amino acid).

(4) Start over at (1b), using the position-specific matrix instead of the original query sequence.
Iterate until satisfied with the result.

Each new iteration may find additional significant hits, not found in previous iterations.

PSI-Blast builds the multiple alignment by relating each hit only to the query sequence, not to the other hits.

Risk: if an irrelevant sequence is mistakenly included among the hits, then the next iterations may drag in more and more completely irrelevant sequences.

This risk has recently been reduced: PSI-Blast version 2.2 (April 2001) is considerably more discriminating.

See Schäffer et al: Improving the accuracy of PSI-BLAST [. . .], *Nucleic Acids Research* 29 (14) 2001.
Available at `http://nar.oupjournals.org/`

A little phylogeny

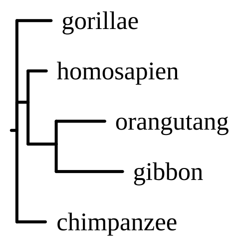
A phylogeny describes the evolutionary relationship between organisms (Greek phylon = race).

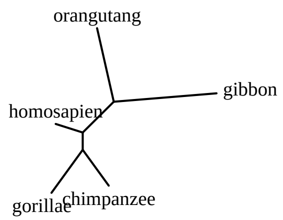
Sequence-based phylogeny was suggested by Zuckerkandl and Pauling (1962).

Orthologous genes serve the same purpose in different organisms, and are derived by speciation only.

Paralogous genes serve different purposes and are derived by duplication.

A phylogeny is usually shown as a tree, rooted or unrooted:





Different ways to make phylogenetic trees

The starting point for a phylogeny is a multiple alignment of sequences (possibly from Clustal or T-COFFEE).

- Clustering based on pairwise evolutionary distance (ca. 1963).

  Compute all $\frac{n(n-1)}{2}$ pairwise distances between the n observed sequences.

  The evolutionary distance between two sequences may be computed from a pairwise alignment.

  There are several ways to construct a tree from a set of pairwise distances:

  - UPGMA = unweighted pair group method using arithmetic averages (Sokal and Michener 1958)
  - Neighbour-joining (Saitou and Nei 1987)
  - Others (Fitch-Margoliash 1967, ...)

- Parsimony (Eck and Dayhoff 1966; Fitch 1971)

  Find the tree and ancestor sequences that require the fewest mutations to produce the observed sequences.

  (Unsuitable for protein: ignores the statistics of the genetic code)

- Probabilistic methods: Maximum likelihood (Felsenstein 1981)

  Find the most probable evolutionary tree and branch lengths for the observed sequences.

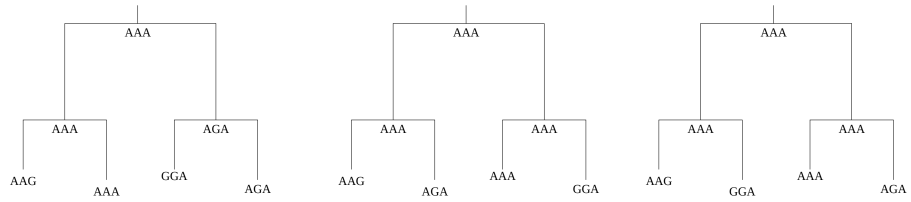
Phylogeny by parsimony

Given a multiple alignment of sequences, for example (from Durbin et al. 1998):

```
AAG
AAA
GGA
AGA
```

Find the tree with fewest mutations, the sequences as leaves, and hypothetical ancestors as internal nodes.

The left tree has 3 mutations and is best; the others have 4 mutations:



Parsimony ignores the variable in mutability in coding DNA (codon third position), and in protein.

Making distance-based phylogenetic trees with Clustal

After making a multiple alignment, a phylogenetic tree can be made (Clustal's guide tree isn't one).

Select `Trees | Draw NJ Tree` and choose a name for the phylogenetic tree file (type `.ph`).

Pairwise sequence distances are computed from the multiple alignment.

These distances are used to build a tree by Neighbour Joining.

Tips on Treeview

Treeview can be used to edit and display (phylogenetic) trees created by Clustal.

It can also produce images of trees that are suitable for inclusion in MS Word.

Displaying Clustal's bootstrap values in Treeview requires some foresight. In Clustal, do:

```
Trees | Output Format Options | Bootstrap Labels on | Nodes
```

before doing

```
Trees | Bootstrap N-J Tree
```

Bootstrap: How reliable is the tree?

Usually some branches in a phylogenetic tree are poorly determined.

Two estimate which are and which aren't, one uses bootstrap:

- Create many (1000) new synthetic multiple alignments as follows:

  Randomly select columns (with replacement) from the original multiple alignment.
- From each synthetic multiple alignment, create a synthetic phylogenetic tree.
- For each branch in the original tree, count how many times it appears in the synthetic trees.

The synthetic alignment to the right is created from columns 5 5 6 1 2 3 9 7 3 6 of the left one:

```
ESFGDLSTPD      DDLESFPSFL
DSFGDLSNPG      DDLDSFPSFL
PHF-DLS---      DDLPHF-SFL
PHF-DLS---      DDLPHF-SFL
PKFKGLTTAD      GGLPKFATFL
DRFKHLKTEA      HHLDRFEKFL
SFLKGTSEVP      GGTSFLVSLT
```

The program's pseudo-random number generator needs a seed, so you must choose a random number.

Phylip: Classic phylogenetic tools by Joseph Felsenstein

Download from `http://evolution.genetics.washington.edu/phylip.html`

Or use `http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`

- Clustering based on pairwise evolutionary distance:

  DNADIST or PROTDIST, followed by NEIGHBOR or FITCH or KITSCH
- Parsimony:

  DNAPARS or PROTPARS
- Maximum likelihood:

  fastDNAml

After computing a phylogeny (and a treefile), use Phylip to draw the trees:

- Use DRAWGRAM to draw a rooted tree
- Use DRAWTREE to draw an unrooted tree

The phylogenetic utility of different kinds of sequences

- Protein (amino acid) sequences can be used to determine distant evolutionary events.

  The short-term non-functional changes in the third codon position are hidden.
- Protein-coding DNA sequences can be used to determine recent evolutionary events for closely related organisms.

  The short-term non-functional changes in the third codon position are visible.
- Some RNA-coding DNA sequences can be used to study the phylogeny of ancient organisms.

  For instance, 16 S rRNA is part of the protein-synthesizing machinery (the ribosome).

  This machinery is ancient, universal and essential, so the 16 S rRNA sequence is well conserved across long evolutionary distances.

  16 S rRNA was used by Woese et al. to establish the domains of Archaea, Eukaryota and Eubacteria (1977).
- Mitochondrial DNA (in humans) is inherited only from the mother and therefore permits 'single inheritance' phylogeny despite sexual reproduction.

Cladistics

Cladistics is a particular school of phylogeny founded by Willi Hennig; `http://www.cladistics.org/`

A *cladogram* is a tree with organisms only at the leaves, and each branch is a clade.

A *clade* is a monophyletic taxon: a group of organisms which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor.

Cladistics determines the evolutionary relationships based on *derived* similarity.

There seems to be some disagreement (or perhaps, a religious wars) around phylogeny, and in particular cladistics.

For instance, Felsenstein's Phylip webpage says:

> *Inspiration to keep going came from* ***Cladistics magazine****, for endless negative reviews and disparaging comments. If they liked it I'd worry.*