

Multiple Sequence Alignment

VTISCTGSSSNIGAG—NHVKWYQQLPG
VTISCTGTSSNIGS—ITVNWYQQLPG
LRLSCSSSGFIFSS—YAMYWVRQAPG
LSLTCTVSGTSFDD—YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG—
ATLVCLISDFYPGA—VTVAWKADS—
ATLVCLISDFYPGA—VTVAWKADS—
AALGCLVKDYFPEP—VTVSWNSG—
VSLTCLVKGFYPSD—IAVEWESNG—

Induced Pairwise Alignment

S_1	S	-	T	I	S	C	T	G	-	S	-	N	I
S_2	L	-	T	I	-	C	N	G	S	S	-	N	I
S_3	L	R	T	I	S	C	S	G	F	S	Q	N	I

Induced pairwise alignment of S_1 and S_2 :

S_1	S	T	I	S	C	T	G	-	S	N	I
S_2	L	T	I	-	C	N	G	S	S	N	I

Sum-of-Pairs Scoring Function

Score of multiple alignment

$$\begin{aligned} &= \sum_{i < j} \text{score}(S_i, S_j) \\ &= \sum_{t=1}^l \sum_{i < j} \text{score}(S_{it}, S_{jt}) \end{aligned}$$

where

$\text{score}(S_i, S_j)$ = score of induced pairwise alignment

l = length of the multiple alignment

Multiple Alignment

Run-time of dynamic programming solution
 $= O(2^k n^k)$

where n = length of each sequence
 k = number of sequences

Space, $O(n^k)$, is prohibitively large!

Example: 6 sequences of length 100 $\Rightarrow 6.4 \times 10^{13}$
calculations!

Carillo-Lippman Heuristic

L = Lower bound on multiple alignment score

If $T[i_1, i_2, \dots, i_k] + \sum_{j < l} \text{score}(S_j[i_j, n_j], S_l[i_l, n_l]) < L$

Then $T[i_1, i_2, \dots, i_k]$ cannot be on an optimal path.

Multiple Alignment to a Phylogenetic Tree

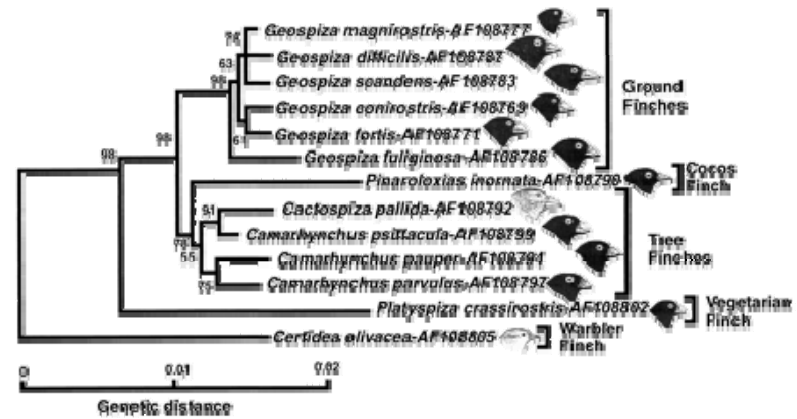
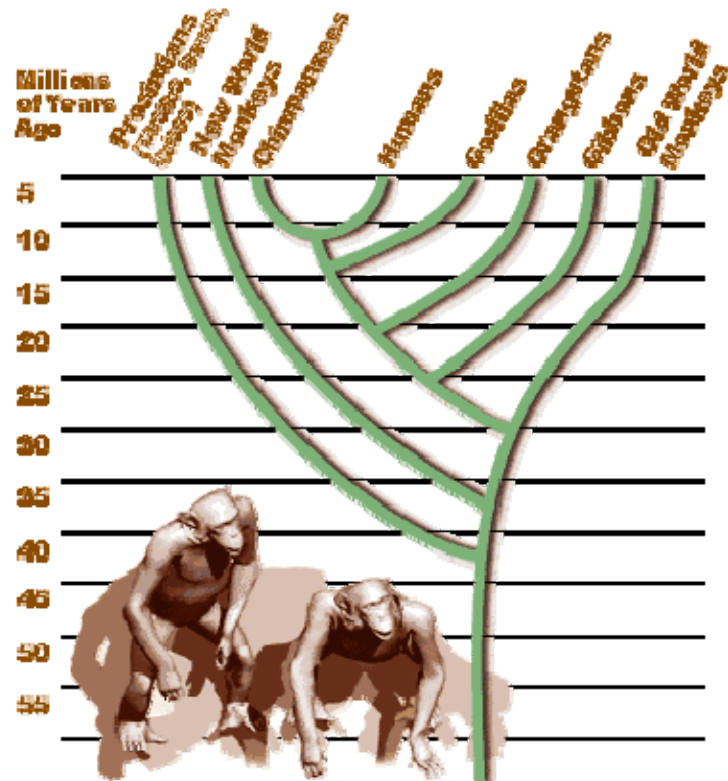
- A tree showing the evolutionary relationship between sequences is available.
- Compute multiple alignment such that for each edge (i,j) in the tree

Induced alignment between S_i and S_j .

= Optimal alignment between S_i and S_j .

Examples

Primates



Darwin's Finches

<http://members.aol.com/darwinpage/trees.htm>

Multiple Alignment to a Tree

- Build the multiple alignment incrementally.
- To add a new sequence, an edge should connect it in the tree to a sequence already incorporated in the multiple alignment.
- Insert the new sequence according to its optimal alignment with the other sequence connected by the edge.
- Adjust other sequences in the multiple alignment.
- Run-time – time for k pairwise alignments.

Multiple Alignment Software

- Clustalw (<http://www.ebi.ac.uk/clusaw>)
- MSA (<http://softlib.rice.edu/softlib/msa.html>)
- HMMER (<http://hmmerr.wustl.edu/>)
- SAM (<http://www.cse.ucsc.edu/research/compbio/sam.html>)