Integrative Biology 200 “PRINCIPLES OF PHYLOGENETICS” Spring 2014
University of California, Berkeley K.W. Will

Alignment

Similarity: Two or more sequences (bases, amino acids, proteins, etc.) are matched in a pairwise alignment either globally (two sequences matched over their whole length) or locally (some subset of the sequences matched while other regions are not expected to match). Sequence similarity can simply be a mathematical distance between two sequences given events such as insertions, deletions and substitutions.

In the simplest model this is the “Edit distance” or the minimal number of events required to transform one sequence into another.

Example to go from acctga to agcta:
accgta <<[substitution]>> a**g**ctga <<[deletion]>> agcta
The edit distance = 2.

BLAST (Altschul, SF, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. J Mol Biol 215(3):403-10, 1990). For example, a gene is newly identified and function understood in Drosophila, a researcher can BLAST the database of the human genome to look for similar gene sequences.

Very basic description of BLAST
1. Uses short segments of sequence to find other sequences that contain the same set.
2. Does “ungapped” alignment extending from the matched subsequence regions to find high-scoring matches
3. Does a rapid gapped alignment to select and rank close matches

Practical issues:

Two problems- how to find alignments and how to choose.

Alignment really attempts to balance the amount of indels with the amount of base substitution, normally based on some cost differential. Of course it is possible to account for all differences by inserting enough gaps (trivial alignment).

```
Taxon1 ACTTCCGAATTTGGCT
Taxon2 ACTCGATTGCCT
```

Minimize substitutions

```
ACTTCCGAATTTGG-CT
|||  |||  |||  ||
ACT--CGA--TTG-CCT
```

Minimize ind/dels

```
ACTTCCGGAATTTGGCT
|||*     **|||*||
ACTC-----GATTGCCT
```

Dynamic Programming and global alignment: (Needleman-Wunsch) underlies or is part of most alignment methods. See tutorial at http://www.sbc.su.se/~pjk/molbioinfo2001/dynprog/dynamic.html

For the attached example: a diagonal move that is a match= 1 or mismatch = 0, vertical move is a gap in top sequence = 0, horizontal move is a gap in side sequence = 0.

For two sequences, i.e. pairwise alignment, of length **n**, if no gaps are allowed then there is one or few optimal alignments. If gaps are allowed, i.e. there is sequence length variation, then... $(2n)!/(n!)^2$ e.g. **n**=50 then 10^{29} alignments. Enumeration is not an option! We need heuristic searches based on Optimality and scoring. For phylogenies, pairwise comparison is not sufficient. What must be done is multiple sequence alignment, a global solution for the whole data matrix or primary homology for the characters (columns) in the matrix.

Many methods have been used to do this. Here are some.....

Manual or by eye- For very simple data this may be sufficient, however, it violates any criterion of repeatability as there is no obvious costs matrix. The counter argument is that the aligned matrix can be made available. However, what if I want to add or subtract OTUs? This would influence the alignment, but how? This is subject to individual pattern recognition abilities for thousands of bases and hundreds of sequences. It is also likely to increase the number of editing errors because of additional "handling" of sequences.

>>Manual alignments informed by consideration of secondary structure-

1. Does not solve the problem of nucleotide homology. At best it places constraints on changes by establishing putative limits between loop and stem regions. Nucleotides within each of those units must still be homologized and all the problems still apply.

2. Determination of secondary structure is not simple and not unambiguous. Generally the actual pattern of bonding is probabilistic and depends on the minimization of free energy and the thermodynamic stability of the resulting structure. Programs explicitly designed to model secondary structure are not very realistic (yet) in terms of the actual cell environment and might find multiple, equally probable models. In phylogenetic studies, secondary structure is typically inferred by aligning with a sequence of “known” secondary structure, although the basis of that knowledge remains uncertain and applicability to the study taxa is unclear in many cases, but this is heading in the right direction.

3. There might be reasonable to expect selective pressures to apply to secondary structure interactions (that is, requirements of compensatory changes), it is unclear just how relevant those interactions are compared to selective pressures applied at other structural levels.

>>Purging "bad" data or scoring variable regions as single characters.

Another method frequently used get around problems in hard to align sections is the elimination of gap heavy regions in alignments. Exactly which columns should be eliminated (left-right boundaries) is subjective and obviously they may have an impact on the results (otherwise why bother).

Alternatively, the variable region can be converted into a character in each taxon and scored. This has all the problems above and adds another layer of difficulty in determining how to code the states.

Approximate Progressive alignment- As in Clustal W(X) the most prominent program for progressive alignment strategies.

1. All sequences are compared to each other (pairwise alignments)
2. A dendrogram is constructed, describing the approximate groupings of the sequences by similarity.
3. Final multiple alignment uses the guide tree

Basically, the multiple alignment is created by iteratively aligning sequences from the input to an already partially constructed solution. Obviously, the order is a crucial point in this method as uses a sort of UPGMA tree-based alignment order and requires sequence weighting.

It could be argued that it doesn't make sense to determine alignment order with one optimality criterion (e.g., phenetics) and then analyze the alignment later with another (e.g., parsimony, ml) but to re-align on a parsimony tree derived from the first alignment to get an "improved" alignment may be circular.

See also, Progressive, consistency-based alignment, **iterative**, segment-based alignment, simultaneous multiple alignments in a hyperspace lattice etc.

Direct optimization - POY (W.Wheeler)- The correspondences among homologues are determined and evaluated simultaneously with transformations. Treespace is explored without a static alignment.

-Single process of alignment and tree construction.
-Insertion and deletion events are counted are real events (transformations) as opposed to being implied by the pattern as in multiple sequence alignments.

Eliminates inconsistent treatment of data between alignment and tree construction steps. Tree-alignment methods like this are ones that simultaneously deals with base changes and insertion/deletion events on the tree, with simultaneous estimation of the parameters of change (including rates in insertions and deletions, which is what in the parsimony world is referred to as gap costs, or rate matrices for nucleotide change (e.g., TV/TS ratios)).

Direct optimization and iterative-pass optimization strive to construct HTU sequences such that the overall cladogram is of minimal length. This is done through modified two and three dimensional string-matching, respectively.

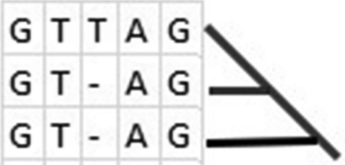
Fixed-states optimization and search-based optimization draw optimal HTU sequences from a pool of predetermined sequences. This can be a small or large collection of possible sequences. Dynamic programming is used to identify the best HTU sequences and determine cladogram length.

		G	A	A	T	T	C
	0	0	0	0	0	0	0
G	0						
G	0						
A	0						
T	0						
C	0						

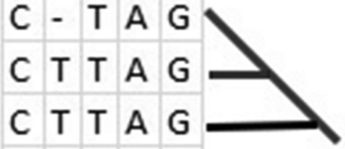
		G	A	A	T	T	C
	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1
G	0	1	1	1	1	1	1
A	0	1	2	2	2	2	2
T	0	1	2	2	3	3	3
C	0	1	2	2	3	3	4

		G	A	A	T	T	C
	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1
G	0	1	1	1	1	1	1
A	0	1	2	2	2	2	2
T	0	1	2	2	3	3	3
C	0	1	2	2	3	3	4

```
-  1  1 -  1 -  1 =4
-  G  A  A  T  T  C
G  G  A  -  T  -  C
```



G	T	T	A	G
G	T	-	A	G
G	T	-	A	G



C	-	T	A	G
C	T	T	A	G
C	T	T	A	G

G	T	T	A	G
G	T	-	A	G
G	T	-	A	G
C	-	T	A	G
C	T	T	A	G
C	T	T	A	G