February 5, 2020. **Molecular data II: Sequence alignment**

Required reading: *Tree Thinking*: pp 86-89, 195-200.

1. Similarity, Identity & Homology:

Sequence similarity can simply be a mathematical distance between two sequences given events such as insertions, deletions, and substitutions. Sometimes the percent of matching nucleotides or amino acids is referred to as the percent similarity, sometimes, incorrectly called percent homology, and sometimes as referred to as identity. In some cases similarity is used to refer to shared properties (e.g. charge or hydrophobicity) even when nucleotides are different. Even NCBI is inconsistent in the use of these terms and you will need to consider the context when you see them. When identity or similarity refers to matching nucleotides or amino acids, and the amount is 25% or more, this suggests possible shared function.

Generally alignment is when two or more sequences (bases, amino acids, proteins, etc.) are matched pairwise, 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).
In the simplest model this is the "Edit distance" or the minimal number of events/edits required to transform one sequence into another.

Example: to go from **acctga** to **agcta**:

accgta <<[substitution]>> **a**g**ctga** <<[deletion]>> **agct-a**

This edit distance = 2. Of course there are many possible ways to go from one sequence to another. We want the best justified, but how do you tell?

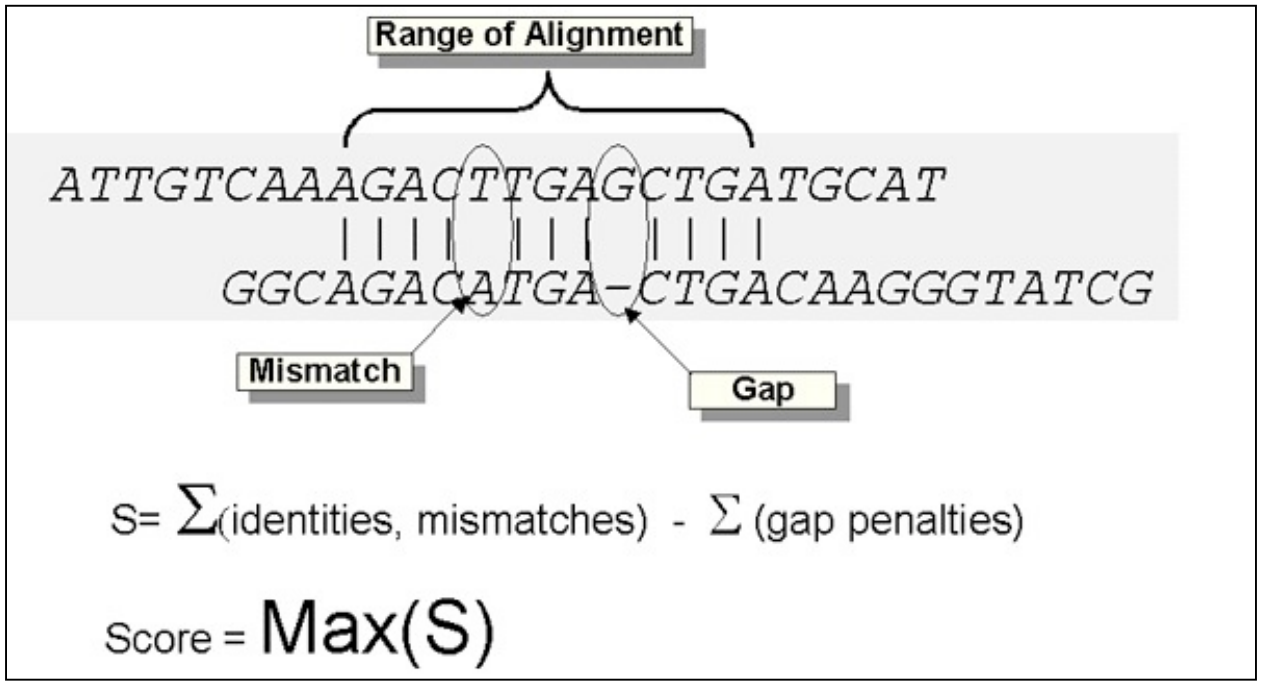


Image from: https://www.ncbi.nlm.nih.gov/books/NBK62051/

```
Taxon1:  ATTCCGAATTTGGCT
Taxon2:  ACTCGATTGCCT
```

What's the best alignment?

Minimize substitutions:

```
     A-TTCCGAATTTGG-CT
     | |  |||  |||  ||     = 7 gaps; 0 substitutions
     ACT--CGA--TTG-CCT
```

Minimize ind/dels:

```
     ATTCCGAATTTGGCT
     |*||********          = 3 gaps; 9 substitutions
     ACTCGATTGCCT---
```

Somewhere in between:

```
     ATTCCGAATTTGGCT
     |*|| || || |*||       = 3 gaps; 2 substitutions
     ACTC-GA-TT-GCCT
```

One approach to quantify quality of alignment: D = y + wz (where D = distance; y = mismatches, w = gap penalty, z = total length of gaps)

Here is an example to try yourself:

```
     Taxon1:  TCAGACGATTG
     Taxon2:  TCGGAGCTG
```

Some possibilities

	(I)	(II)	(III)
	TCAG-ACG-ATTG	TCAGACGATTG	TCAG-ACGATTG
	TC-GGA-GC-T-G	TC**G**GA**GCTG**--	TC-GGA-G**C**T**G**-
If y=1; w= 2:	*D=12*	*D=9*	*D=10*

Can you find a better one? Try in class

---

2. Alignment for phylogenetic analysis.

For phylogenetic analysis we need to establish character homology (aligned columns in the matrix) and primary homology hypotheses of character states. The very simplicity of molecular characters (i.e., no ontogeny, few possible character states) leads to problems with determining homology (despite our intuitions, which might suggest the opposite). Unlike morphology we have essentially a one-dimensional string, although we may also have some additional dimensions added by structural constraints.

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 1029 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.

Alignment attempts to balance the amount of gaps (inferred 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).

Gaps are usually needed for alignment, but are not real when considering a single genome. Evolutionary gaps might result from several processes:

- point mutation
- unequal crossing over during meiosis
- DNA slippage during replication
- retrovirus insertions
- movement of sequences in the genome (e.g. translocation or transposition)

BLAST (Altschul et al. 1990). Basic local alignment search tool. 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

3. Brief overview of some methods used for alignment

Dynamic Programming and global alignment: (Needleman-Wunsch) underlies or is part of most alignment methods.

Simultaneous alignment- Simultaneous multiple alignments synchronise the information of all input sequences in a hyperspace lattice, e.g. so-called exact alignment algorithms using the divide-and-conquer (DCA) strategy (Tönges et al. 1996). It cuts down the input sequences at carefully chosen positions to align in segments. Current algorithms cannot handle large/complex data sets.

Progressive alignment- As used 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.

Progressive, consistency-based alignment- These strategies are incorporating "local signals" into global alignment construction.

Iterative, segment-based alignment- One example is DIALIGN, which iteratively collects local similar segments, which can be merged into a common multiple alignment. Iteration continues until no more local signals can be found or until all positions are aligned. Recent benchmarks have shown that this strategy can even handle long sequences of a low overall similarity. No explicit gap cost or input trees.

Direct optimization - POY (Varón et el. 2010)- The correspondences among homologues are determined and evaluated simultaneously with transformations. Tree space 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.

Adjustments: 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.

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

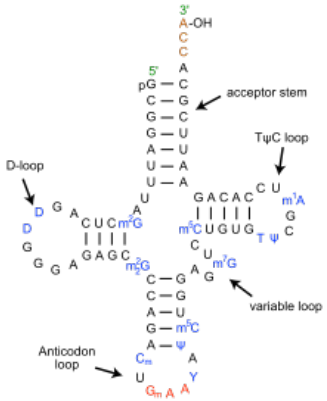
Alternatively, the variable region can be converted into a character in each taxon and scored. This has all the problems above and adds another layer of difficulty in determining how to code the states.

Nucleotide alignments of coding sequences informed by amino acids. This works well and can easily be implemented in commonly used software such as Mesquite.

Coding larger gaps that appear homologous - this is particularly applicable when some OTUs appear to be missing one or more triplets in a protein coding gene. But sometimes you see large gaps in some OTUs relative to others, that seem quite aligned, even in non-coding regions. E.g.:

```
ATTCCGAATTT--------GAATTTGGCT
ATTCCGAATTTGGCTTTCCGAATTTGGCT
ATTCCGAATTT--------GAATTTGGCT   Best to code this as one binary character?
ATTCCGAATTTGGCTTTCCGAATTTGGCT
ATTCCGAATTTGGCTTTCCGAATTTGGCT
```

Alignments informed by consideration of secondary structure in RNA or protein-



1. Does not solve the problem of nucleotide homology.

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.

Detail your methods!

However you derive your alignment it is necessary to explain clearly how it was done. Cite the software, its version number and settings used. A good idea in any case, but especially if you make manual edits, is to deposit the aligned matrix online, e.g. at *TreeBASE* <https://treebase.org/treebase-web/home.html>.

References cited:

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Tönges, U., Perrey, S.W., Stoye, J. and Dress, A.W.M. 1996. A general method for fast multiple sequence alignment. Gene 172GC33-GC41

Varón, A., L. S. Vinh, W. C. Wheeler. 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics, 26:72-85.

---

answer from pg. 2:

TCAGACGATTG

TC**G**GA**GC**--TG

D=7