BCBio 444: Bioinformatics Analysis

Karin S. Dorman

Department of Statistics
Department of Genetics, Development & Cell Biology
Program in Bioinformatics & Computational Biology
Iowa State University

October 14, 2017

Multiple Sequence Alignment

Definition (multiple sequence alignment)

A multiple sequence alignment is an alignment involving three or more sequences that are completely or partially aligned.

- You know why *multiple sequence alignments* (MSAs) are useful [Friedberg], but now we need to know how to obtain them.
- Many ways! Wikipedia's List of Sequence Alignment Software.
- With so many aligners, it can be hard to choose one for a project. To choose among them, search for recent publications, such as Thompson et al. (2011) and Le, Sievers, and Higgins (2017), that compare multiple alignment software on data with *known* alignments.

"Known" Alignments

- Because we did not observe evolution as it happened, there are no known alignments.
- **Benchmarking.** However, because amino acids change slower than nucleotides and structure and function change slower than sequences, we can use the slow evolving feature, when they are available, to assess the quality of an alignment of the fast evolving feature (the sequence).

See Le, Sievers, and Higgins (2017)

Types of MSAligners

- **Exact.** Too slow for all but a few sequences, so not used.
- **Progressive.** The original MSA, first proposed (Fitch and Yasunobu, 1975) and first applied (Hogeweg and Hesper, 1984) decades ago. Most famous variant is ClustalW (Thompson, Higgins, and Gibson, 1994), which is now retired.
- **Iterative.** Includes MAFFT (Katoh et al., 2002), PRALINE (Heringa, 1999), InterAlign (Pible, Imbert, and Pellequer, 2005), and MUSCLE (Edgar, 2004).
- **Consistency-based.** Includes MAFFT, ProbCons (Do et al., 2005), T-Coffee (Notredame, Higgins, and Heringa, 2000).
- **Structure-based.** Includes Expresso (Armougom et al., 2006).

[Pevsner (2015)]

Exact MSA

Extend dynamic programming to > 2 sequences by filling out a n-dimensional matrix for n sequences. Unfortunately, the algorithm complexity is $O(2^n l^n)$, where l is the average length of the sequences.

[Pevsner (2015)]

Progressive – ClustalW

While ClustalW may be the best-known MSA, it is no longer recommended (see benchmarks for alternatives).

1. **Pairwise alignments.** Perform Needleman-Wunsch pairwise global alignment on all pairs of sequences: $\frac{n(n-1)}{2}$ alignments. Compute some kind of distance metric on each alignment, *e.g.* one minus the fractional matches or $D = -\ln S$, where

$$S = 100 \times \frac{S_o - S_r}{S_s - S_r}, \tag{1}$$

   S_o is the observed PWA score, S_r is the average PWA score from multiple random reshufflings, and S_s is the average of self-alignment scores.

2. **Construct guide tree.** Same algorithms used to estimate phylogenetic trees, which we will learn next week, but here done using distances computed from unrefined PWAs. [try ClustalW alignment at Japanese site]

3. **Alignment.** Align two closest sequences, then align next two closest, *etc.* At some point, you need to align an alignment to a sequence or an alignment to an alignment (we will discuss).

Pevsner (2015)

Interpretting Eq. (1)

Here is an motivation for Eq. (1).

- You subtract S_r so that pairs of random (nonhomologous) sequences have a score of 0, and therefore an infinite distance.
- You scale by $S_s - S_r$ to remove the length and content effects. Pairs of longer proteins will have larger alignment scores, but it does not necessarily indicate more homology. Pairs of proteins filled with rarer, less easily substituted amino acids will have larger alignment scores, but once again, it may not indicate greater homology. This scaling normalizes for these effects.

Progressive – ClustalW (Cons)

- Once a mistake is made, it is impossible to back out of it, and it will cause downstream mistakes too.
- One may think that deletion events are rare in evolution, thus one may prefer to align deletions across the whole alignment even tolerating a few mismatches at the ends of the deletion. This is accomplished by changing the gap penalties to be position specific based on the presence of gaps already there in the alignment.
- A single insertion while aligning a sequence to a profile costs a gap open (and extension) penalty in *all* sequences of the profile. That is an imbalance, solved much later in different software.
- Alignment depends on order from guide tree, but guide tree is estimated with poor PWAs. Circular reasoning.

Iterative – MAFFT (FFT-NS-i)

1. **Initialize.** Use some kind of progressive alignment to get an initial alignment A and guide tree T.
2. **Optimize.** Given MSA A of n sequences and a guide tree T with branch lengths, define

$$WSP(A) = \sum_{j=2}^{n} \sum_{k=1}^{j-1} w_{j,k} S_{j,k}, \tag{2}$$

where $S_{j,k}$ is the PWA score of the jth and kth sequences and $w_{j,k}$ is a weight of the pair (j, k) (to be discussed). Iteratively improve the alignment to increase $WSP(A)$.

   1. Split T on each of the branches, producing two alignments A_1 and A_2.
   2. Realign A_1 and A_2 producing alignment A' using a group-to-group algorithm and replace A with A' if $WSP(A') > WSP(A)$. Otherwise, end.

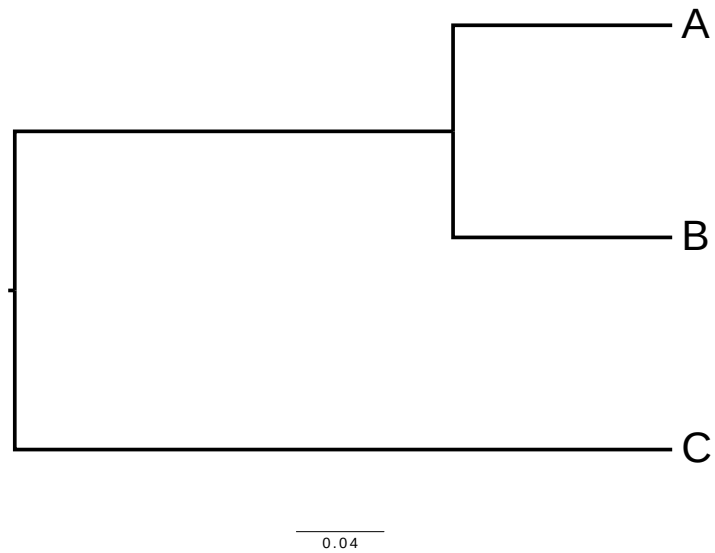
Katoh and Toh (2008)

Iterative – MAFFT Variations

- **FFT-NS-i** uses FFT-based group-to-group alignment algorithm to maximize $WSP(A)$.
- **G-INS-i** uses a weighted consistency-based score.
- **L-INS-i** is G-INS-i with local PWA and affine gap penalty.

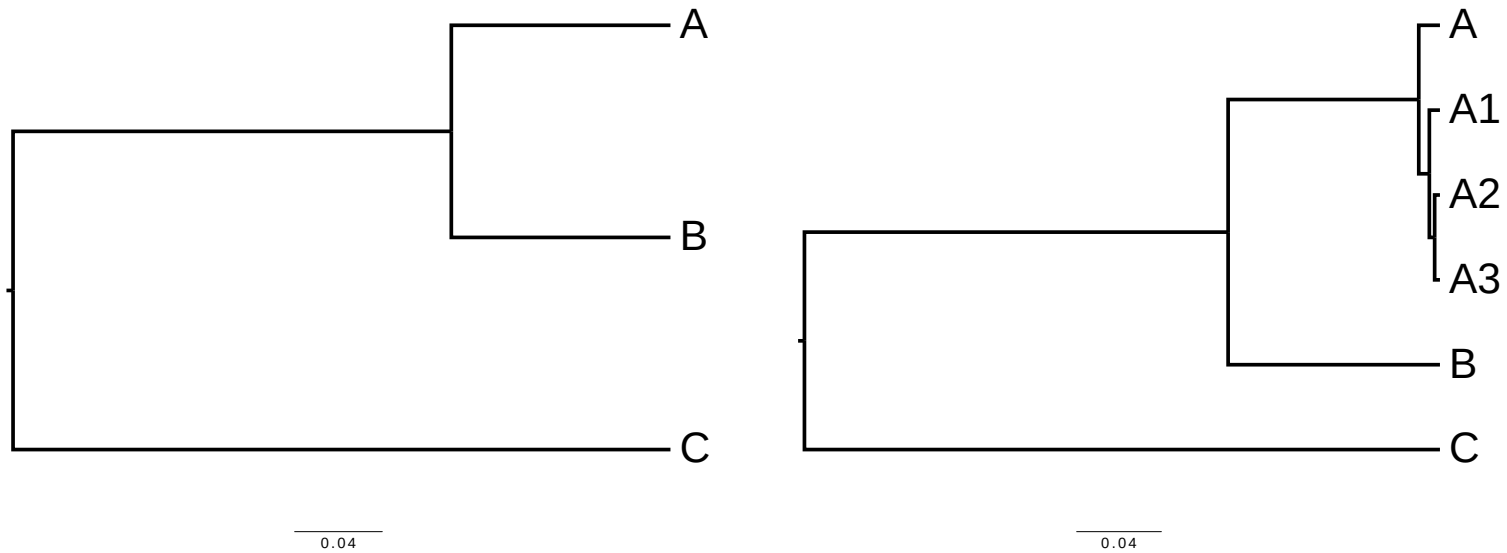
Katoh and Toh (2008)

Weighting Pairwise Alignments



Tree drawn with FigTree (Rambaut, 2016).

Weighting Pairwise Alignments



Trees drawn with FigTree (Rambaut, 2016).

Choosing weights

- Assume you know the true tree $T = (V, E, L)$, where V are the nodes (internal branch points and leaves), E are the edges (branches between nodes) of the tree and L are the lengths of the edges (we use the guide tree).
- Consider a mutation on edge $e \in E$ of the tree. It contributes to the distance between any pair of sequences separated by edge e.
- PRINCIPLE: One mutation on edge e should not count once for every traversing sequence pair: it should count once for the one mutation that it is.
- IMPLEMENTATION: The sum of the weights for all pairs (j, k) that use edge e should sum to 1.
- Thus, for every $e \in E$, we want

$$\sum_{(j,k) \in \mathcal{P}_e} w_{j,k} = 1,$$

  where $\mathcal{P}_e$ is the set of sequence pairs whose connecting path goes through edge e.
- Can we solve this system of equations? Yes...many ways; it is an underdetermined system.

Altschul, Carroll, and Lipman (1989)

Altschul, Carroll, and Lipman (1989)/1

- **Builing a model**: We believe that each pair of sequences carries the same information, we'll call it μ, about the MSA, but the information x_p held by (j_1, k_1) with path p and the information x_q held by (j_2, k_2) with path q is correlated if they share some edges. (Here, x_p and x_q are hypothetical, *unobserved* measures of information.)
- Let l_p be the length of path p and l_{pq} be the shared length of paths p and q. It seems reasonable that the correlation of information in path p and q is

$$\text{Cor}(x_p, x_q) := M_{pq} = \frac{l_{pq}^2}{l_p l_q}.$$

- Thus, we have (unobserved) information vector $\boldsymbol{x}$ with constant mean vector $\boldsymbol{\mu}$ and covariance

$$\text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{M},$$

with the further assumption of constant variance $\text{Var}(x_{j,k}) = \sigma^2$ for all $j \neq k$. (This latter assumption is valid if the *molecular clock* assumption is satisfied – to be discussed next week, I think.)

- What probability model can we use for $\boldsymbol{x}$?

Altschul, Carroll, and Lipman (1989)/2

- **The model**: The authors assumed the multivariate Normal model,

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{M}\right).$$

- Use statistics to get an estimate of $\boldsymbol{\mu}$ (the maximum likelihood estimate: covered in Stat 342 and 510, so not required for this course):

$$\hat{\boldsymbol{\mu}} = \frac{(1,1,\ldots,1)\boldsymbol{M}^{-1}\boldsymbol{x}}{(1,1,\ldots,1)\boldsymbol{M}^{-1}(1,1,\ldots,1)^T},$$

  which is a weighted average of the observations $\boldsymbol{x}$, with weights

$$\boldsymbol{w}^T = \frac{(1,1,\ldots,1)\boldsymbol{M}^{-1}}{(1,1,\ldots,1)\boldsymbol{M}^{-1}(1,1,\ldots,1)^T},$$

  which are the normalized row sums of $\boldsymbol{M}^{-1}$ (compute $\boldsymbol{M}^{-1}$ [Math 207], sum rows, and divide by the sum of elements in $\boldsymbol{M}^{-1}$).

- Now, we do not observe x_p for pair (j,k), but we do observe scores $S_{j,k}$. Suppose $S_{j,k}$ estimates x_p. We combine all the scores into a single metric $WSP(A)$ (2), after weighting each score by $w_{j,k}$, such that each single mutation counts only once in the combined metric.

What's important

It is all important, but obviously you haven't all taken Math 207, Stat 342, and Stat 510, so parts of it will be beyond your current knowledge (but not far!). So, I want you to remember these main ideas that once again demonstrate the beautiful combination of algorithms, math/stats, and biology that is bioinformatics:

- We **assume** the PWAs are good quality (from an exhaustive search via dynamic programming, aka "exact") and thus are trustworthy.
- Each PWA contains information about the multiple sequence alignment: we **assume** equal information. (Think about it the other way around: If you had a multiple sequence alignment, you could certainly get the pairwise alignments from it, so it is sensible that PWAs inform on the MSA).
- We **assume** the PWA score $S_{j,k}$ is a measure of that information. (FYI, statistical theory supports the use of the score as *the* numerical summary of information in alignment because the score is the log likelihood.)
- However, the information from the PWAs are correlated because they represented evolutionary history that is shared. To the extent that history is shared between two pairs of sequences, they will be more or less correlated. We **assume** the correlation is summarized by matrix $\boldsymbol{M}$ and the distances in the tree T.
- To account for the correlation and to downweight duplicate information from highly correlated features, we weight the scores by $w_{j,k}$.
- The weights $w_{j,k}$ are estimated from the estimate of the evolutionary tree T, which indicates how correlated PWA scores are and a statistical model of reality: the multivariate normal.

Critique of assumptions

When we list a bunch of assumptions, we should also consider how they might be wrong:

- The PWAs are good quality and the scores summarize the information they contain as long as the parameters (the match, mismatch, and gap penalities) are correct. These are only *roughly* estimated from data under another bunch of assumptions (see PAM, BLOSUM). Most importantly, they assume a knokwn level of relatedness between the pairs of sequences. Don't use scores for distantly related proteins, when your proteins are closely related!
- I'm not sure how well we can confirm or refute the assumption that sequence pairs provide equal information about the alignment.
- Since T is estimated with error, the correlation matrix $\boldsymbol{M}$ is estimated with error. Further, the theory that motivates the entries of $\boldsymbol{M}$ is constant Brownian motion in time (Altschul, Carroll, and Lipman, 1989), which also motivates the multivariate Normal model. This model is only valid if the molecular clock is true and the sequences are infinite in length.

Alignments of alignments (MUSCLE)

MAFFT must align alignments to alignments, but we will not study how MAFFT does it. Instead, we will study how MUSCLE does it (see Iteration step).

- **Initialize.** Obtain a draft alignment using progressive alignment by fraction identify from NW PWA or k-mer counting.
- **Guide tree.** Reassess similarity of pairs as fractional identity and estimate guide tree.
- **Iteration.** For each edge in the guide tree, partition into alignment A_1 and A_2 and refine the alignment, assessing improvement as:
  - Let $Q_{i,j} = \log\left(\frac{p_{ij}}{p_i p_j}\right)$ be the score for aligning character i and j, where p_i, p_j are the relative frequencies of i and j and $p_{i,j}$ the probability of aligning i and j.
  - Let f_i^x be observed count of character i in column x.
  - Score of aligning column x in A_1 to column y in A_2 is

$$PSP^{xy} = \sum_i \sum_j f_i^x f_j^y Q_{i,j}.$$

Example

Can you imagine filling out a scoring matrix for two alignments using match score of 4, mismatch of -1 and gap open and extension of -5?

Here is a tiny example.

		A	C
		G	C
A	A	$2 \times 1 \times 4$	$2 \times (-5)$
C	T	$2 \times (-5)$	$\nwarrow 8 + 1 \times 2 \times 4$

References I

S F Altschul, R J Carroll, and D J Lipman. "Weights for data related by a tree." In: *Journal of molecular biology* 207 (4 1989), pp. 647–653.

Fabrice Armougom et al. "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee." In: *Nucleic acids research* 34 (2006), W604–W608.

Chuong B Do et al. "ProbCons: Probabilistic consistency-based multiple sequence alignment." In: *Genome research* 15 (2 2005), pp. 330–340.

Robert C Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Research* 32.5 (2004), pp. 1792–1797.

References II

W M Fitch and K T Yasunobu. "Phylogenies from amino acid sequences aligned with gaps: the problem of gap weighting." In: *Journal of molecular evolution* 5 (1 1975), pp. 1–24.

J Heringa. "Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment." In: *Computers & chemistry* 23 (3-4 1999), pp. 341–364.

P Hogeweg and B Hesper. "The alignment of sets of sequences and the construction of phyletic trees: an integrated method." In: *Journal of molecular evolution* 20 (2 1984), pp. 175–186.

References III

Kazutaka Katoh and Hiroyuki Toh. “Recent developments in the MAFFT multiple sequence alignment program.” In: *Briefings in bioinformatics* 9 (4 2008), pp. 286–298.

Kazutaka Katoh et al. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.” In: *Nucleic acids research* 30 (14 2002), pp. 3059–3066.

Quan Le, Fabian Sievers, and Desmond G Higgins. “Protein multiple sequence alignment benchmarking through secondary structure prediction.” In: *Bioinformatics (Oxford, England)* 33 (9 2017), pp. 1331–1337.

References IV


C Notredame, D G Higgins, and J Heringa. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment". In: *J Mol Biol* 302.1 (2000), pp. 205–217.


Jonathan Pevsner. *Bioinformatics and Functional Genomics*. 3rd. Wiley Blackwell, 2015.


Olivier Pible, Gilles Imbert, and Jean-Luc Pellequer. "INTERALIGN: interactive alignment editor for distantly related protein sequences." In: *Bioinformatics (Oxford, England)* 21 (14 2005), pp. 3166–3167.


Andrew Rambaut. *FigTree – Tree Figure Drawing Tool*. Version 1.4.3. 2016. URL: `http://tree.bio.ed.ac.uk/software/figtree/`.

References V

J D Thompson, D G Higgins, and T J Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." In: *Nucleic acids research* 22 (22 1994), pp. 4673–4680.

Julie D Thompson et al. "A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives." In: *PloS one* 6 (3 2011), e18093.