

Multiple Sequence Alignment — Lecture Notes

1. Why multiple sequence alignment?

A single pairwise comparison gives limited information. If two protein sequences differ at one amino acid, we can't tell whether that difference is functionally important or neutral variation. Aligning many related sequences (e.g., the same protein across species) shows which positions are **conserved** (likely important for structure or function) and which vary freely. More sequences → more confident conclusions.

2. Cost of naive MSA (dynamic programming)

Pairwise alignment fills a 2D table; aligning 3 sequences fills a 3D (cube) table. For k sequences of length n :

- **Number of cells:** about $(n + 1)^k$
- **Predecessors per cell:** $2^k - 1$ (each sequence either contributes a residue or a gap, minus the all-gap case)
- **Total time:** cells $\times$ predecessors $\approx O(2^k \cdot n^k)$

The cost is exponential in k , the number of sequences, so the naive approach is impractical beyond a handful of sequences. Finding the optimal MSA is NP-hard under standard scoring, so heuristics are necessary.

Heuristics used in practice

- **Progressive alignment:** build a guide tree from pairwise distances, align the most similar sequences first, then add the rest (e.g., Clustal Omega). Fast, but early mistakes are never corrected ("once a gap, always a gap").
- **Iterative refinement:** repeatedly realign subsets to improve the overall score (e.g., MUSCLE, MAFFT).
- **Profile-based alignment:** align new sequences to a statistical summary of an existing alignment (see profile HMMs, section 5).

3. Application: protein folding and AlphaFold2

AlphaFold2 predicts how a protein folds into its native **conformation** (its 3D structure). An MSA is one of its main inputs.

Forces involved in folding

- Hydrophobic effect (usually considered the main driving force: hydrophobic residues are buried, polar ones exposed)
- Hydrogen bonds, ionic interactions, van der Waals interactions, and disulfide bonds
- Side-chain properties: size, charge, flexibility
- Backbone geometry: the peptide bond is planar and rigid; rotation happens around the bonds on either side of each α -carbon (ϕ and ψ angles) and in the side chains

Why computational prediction is hard: there are many interactions and possible angles to account for, so the number of possible conformations is astronomical (Levinthal's paradox). Yet real proteins fold in milliseconds to seconds.

Why prediction is possible at all: Anfinsen's experiment showed that the amino acid sequence alone determines the fold.

Historical approaches

Type	Approach
Experimental (determines structure)	X-ray crystallography
Experimental	NMR / mass spectrometry to identify folding intermediates
Experimental	Protein engineering to find stabilizing residues
Computational (predicts structure)	Lattice models
Computational	Molecular dynamics
Computational	Monte Carlo methods
Computational	MSA-based methods (this lecture)

4. Coevolution: how an MSA reveals structure

Analogy: a hummingbird's beak is shaped to fit the flower, and the flower is shaped to fit the beak. Each one's evolution constrains the other.

In proteins: if two amino acids sit close together in the folded structure, a mutation in one often forces a matching (compensatory) mutation in the other to keep the structure stable. For example, if a positive charge pairs with a negative charge and one flips, the partner tends to flip too.

What this means for an MSA: the residues at two positions may differ across sequences, but they change *together* so the interaction is preserved. Columns that covary like this are likely in contact, even if far apart in the 1D sequence, and the overall fold stays the same. Many predicted contacts together constrain the fold.

Caveat: correlations can be indirect. If A contacts B and B contacts C, then A and C may look correlated without touching. Methods like direct coupling analysis (DCA), and AlphaFold2's network, aim to separate direct from indirect signals.

5. Profile HMMs

A profile HMM is a **sequence-generating model** for a protein family, trained on an MSA of hundreds of family members.

- **States:** Match, Insert, and Delete states for each column of the alignment. Delete states are silent (emit nothing).
- **Transition probabilities** between states.
- **Emission probabilities:** Match and Insert states emit amino acids, with probabilities learned from the MSA columns (conserved columns have peaked distributions).

Using it on a new sequence: ask which path of hidden states most likely generated the sequence (e.g., Match → Match → Insert → Match → Delete → Match). The probability of a path is the product of the transition and emission probabilities along it. This works because of

the Markov property: each step depends only on the current state, not the full history.

- **Most likely path:** Viterbi algorithm
- **Total probability of the sequence (summed over all paths):** forward algorithm

Note: in common designs (e.g., HMMER), Insert and Delete states don't connect directly.

Why this matters for structure prediction: the path shows how the query lines up with the family (which query positions correspond to which MSA columns). AlphaFold2's pipeline uses profile-HMM tools (e.g., HHblits, jackhmmer) to search sequence databases for homologs of the query and build its MSA. Its network (the Evoformer) then reads the MSA directly and learns the coevolution signal implicitly, rather than computing contacts as a separate step. It can also use known structures of similar proteins as templates.