Lecture Title and Date

Multiple Sequences, 02/10

Objectives of the Lecture

By the end of this lecture, students should be able to:

1. Understand the role and applications of Multiple Sequence Alignment (MSA) in molecular biology.
2. Learn the basics of progressive alignment and clustering methods (e.g., UPGMA, agglomerative clustering).
3. Describe the role of motifs and profiles in characterizing protein families (e.g., PROSITE, EGF-like patterns).
4. Recognize the challenges in MSA, such as the domain problem and local minimum issues.
5. Learn how Position Weight Matrices (PWMs) are used for sequence scanning and pattern recognition.
6. Explore how probabilistic models, such as Expectation-Maximization (EM) and Gibbs Sampling, are used to determine PWMs.
7. Understand the basic principles of Hidden Markov Models (HMMs) in sequence analysis.

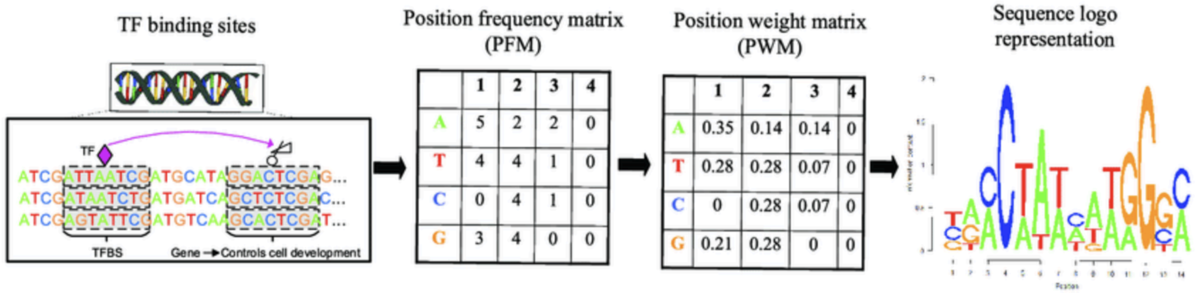
Key Concepts and Definitions

- **Multiple Sequence Alignments**: a method for aligning multiple biological sequences to identify regions of similarity, which can be used to infer homology, analyze functional conservation, and study evolutionary relationships.
- **Progressive Alignment**: a heuristic MSA method that builds alignments iteratively based on pairwise alignments and phylogenetic relationships.
- **Agglomerative Clustering**: a bottom-up hierarchical clustering method that iteratively merges the closest clusters until a single hierarchy is formed.
- **UPGMA (Unweighted Pair Group Method with Arithmetic Mean)**: a hierarchical clustering algorithm that iteratively merges the closest clusters based on pairwise distances, using the arithmetic mean to update cluster distances.
- **Domain Problem**: the difficulty of accurately aligning protein sequences when they contain different domain architectures.
- **Local Minimum Problem**: a situation where an alignment algorithm gets stuck in a suboptimal alignment solution, unable to find a better alignment due to the greedy nature of alignment.
- **Motifs**: a short, conserved sequence pattern in DNA or proteins that is functionally significant and specific enough to identify related family members in a DNA/protein database

- **Prosite Pattern**: a small collection of motifs used to identify protein families and domains.
- **Profile**: a position-specific scoring matrix that captures amino acid or nucleotide frequencies at each position in a multiple sequence alignment, enabling sensitive sequence comparison and motif detection.
- **Position Weight Matrix (PWM)**: a scoring matrix for motif scanning, which represents the log-likelihood of nucleotide or amino acid occurrence at each position in a motif.
- **PSI-BLAST**: a method used to generate multiple sequence alignments by iteratively searching a protein database with a position-specific scoring matrix (PSSM) derived from a preliminary alignment.
- **Expectation-Maximization (EM) Algorithm**: an iterative probabilistic method for estimating parameters, such as motif models, by alternating between the Expectation step (estimating motif locations) and the Maximization step (optimizing parameters) to refine sequence alignments.
- **Gibbs Sampling**: a stochastic method for motif discovery that iteratively refines motif predictions by sampling and updating sequence alignments.
- **Hidden Markov Models (HMMs)**: a probabilistic model for sequence alignment, where states represent sequence positions, emit symbols based on defined probabilities, and transition between states to generate a sequence, enabling pattern recognition in biological data.
- **Forward Algorithm**: a probability-based method in HMMs that computes the probability of an observed sequence by summing over all possible state paths that could generate it.
- **Viterbi Algorithm**: a dynamic programming method for finding the most probable path through a HMM for a given sequence.

Main Content/Topics

Motifs. In molecular biology, motifs are commonly repeated patterns in nucleic or amino acid sequences that usually correlate with some conserved function. Mark describes them as "basic molecular units." Examples include transcription factor binding motifs (TFBMs) in DNA sequences, or helix-turn-helix domains in DNA-binding proteins. Motifs have been used to advance sequence alignment algorithms, often through the use of motif profiles. Below is an example of a TFBM and its corresponding profiles.


TF binding sites
TF
ATCGATTAATCGATGCATAGGACTCGAG...
ATCGATAATCTGATGATCAGCTCTCGAC...
ATCGAGTATTCGATGTCAAGCACTCGAT...
TFBS
Gene → Controls cell development
Position frequency matrix (PFM)
Position weight matrix (PWM)
Sequence logo representation

Position frequency matrix (PFM)

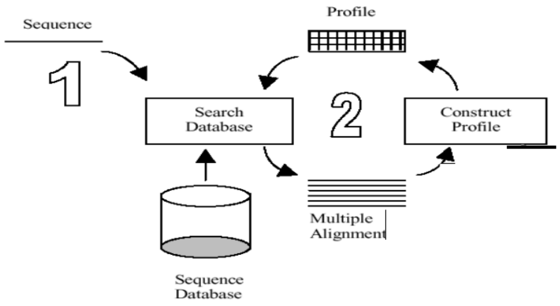
	1	2	3	4
A	5	2	2	0
T	4	4	1	0
C	0	4	1	0
G	3	4	0	0

Position weight matrix (PWM)

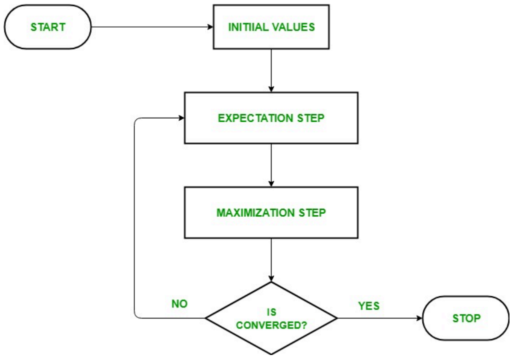
	1	2	3	4
A	0.35	0.14	0.14	0
T	0.28	0.28	0.07	0
C	0	0.28	0.07	0
G	0.21	0.28	0	0

Profiles. In simplest terms, a profile is a matrix denoting the probability of encountering a given nucleotide or amino acid at a certain position in a set of aligned sequences. A commonly used profile is a Position Weight Matrix (PWM), which represents probabilities as log odds. Some methods of generating motif profiles are described below.

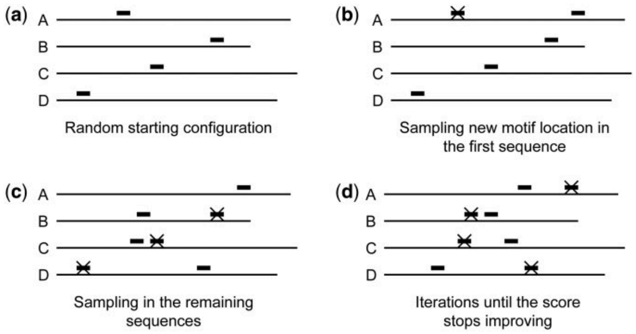
- **PSI-BLAST** is an iterative method of generating Position-Specific Scoring Matrices (PSSMs) using the BLAST sequence alignment algorithm. Multiple sequences are aligned to an input sequence, from which an initial PSSM is constructed. This PSSM is fed back into the database to identify sequences that match the motif patterns in the PSSM. These sequences are then added to the alignment to produce a new PSSM, and the process repeats. Eventually, the algorithm will converge on a PSSM to which no new significant matches can be added–this is the result of your query. A good explanation of PSI-BLAST can be found here.



- **Expectation-Maximization (EM)** is a statistical algorithm adapted for constructing PWMs. The basic steps are to 1.) construct a set of initial values, 2.) calculate the expected values of desired (usually latent) variables, 3.) maximize the expectation in 2.) and 4.) repeat until convergence is reached. In the context of generating a PWM, the steps would be to 1.) initialize a starting PWM, 2.) identify sequence alignments that best match the PWM, 3.) re-estimate the PWM using these sequences, and 4.) repeat until the algorithm converges on a PWM. EM is greedy in that it only incorporates sequences that best match the PWM–as such, it may not return the best possible PWM.



- **Gibbs Sampling** is another probabilistic method of generating PWMs. In the first step, a random k-mer (sequence of length k) is taken from each input sequence. One k-mer is discarded and the remaining are used to define an initial PWM. A new k-mer is chosen at random from the probability distribution of matching k-mers to update the PWM as an old one is discarded. Because this k-mer is



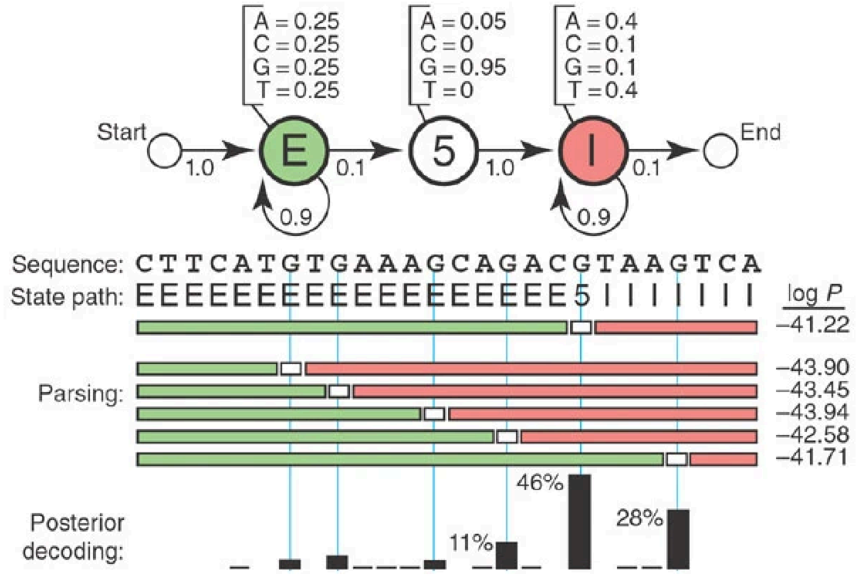
Random starting configuration

Sampling new motif location in the first sequence

Sampling in the remaining sequences

Iterations until the score stops improving chosen at random, it may not necessarily be the best match. This process repeats for as long as desired and returns the highest-scoring PWM observed.

- **Hidden Markov Models (HMMs)** are generally used to model observations determined by latent processes. When adapted to PWMs, each state in the model represents a position in the sequence, and emitted symbols correspond to observed nucleotides. The probability of observing a sequence is the product of the probabilities of all paths taken to produce the sequence. **Forward** and **Viterbi** algorithms are two methods of traversing HMMs, with the former returning the probability that a given sequence is emitted, and the latter returning the most probable path through the model for a given sequence. (Figure from *Eddy, Nature Biotechnology 2004*).



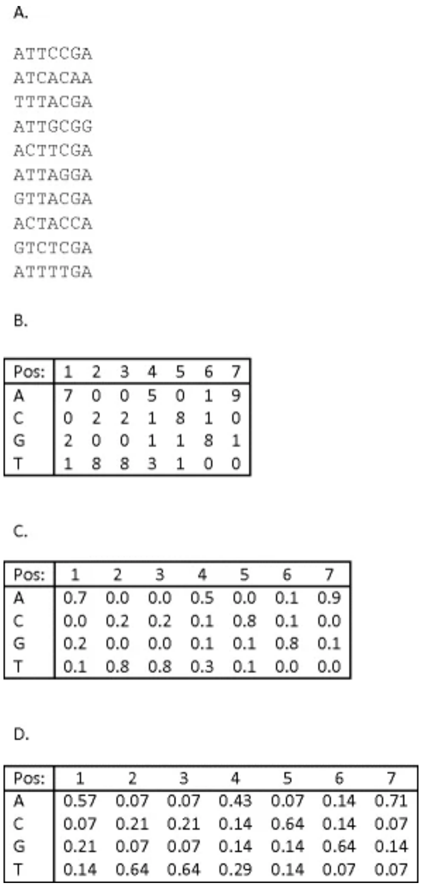
Discussion/Comments

- We have huge amounts of sequences and want to look for "patterns of conservation."
- Going through and individually checking each sequence would skyrocket computational complexity → instead do it heuristically
- In regards to the domain problem, we can chop the sequences into domains because we do not know where domain boundaries exactly are.
- Motifs are similar to regular expressions in computational terms.
- EGF Profile Generated for SEARCHWISE: matrix of sequence by sequence, tabulate how often we observe an amino acid relative to baseline → identify enriched amino acids
- In PSI-BLAST, instead of taking a sequence and running it against the database, we take an abstracted pattern and matrix and run it against the database.
- Convergence indicates that we sensitively found the "best" match for the sequence.
- Explosion is when our "best" matches are the whole database, indicating that nothing of value was discovered.

Required Readings

- Stormo, G. D. (2010). Methods in Molecular Biology, 85–95. Motif discovery using expectation maximization and Gibbs' sampling. https://doi.org/10.1007/978-1-60761-854-6_6

Motif discovery using expectation maximization and Gibbs' sampling neatly presents the motif "problem" (know these binding sites exist but not exactly sure where) and the need for discovery algorithms such as Expectation-Maximization and Gibbs' Sampling. It also introduces the idea of a PWM (or a Position Weight Matrix) and provides a digestible example of one. For EM, the authors provide four scenarios with varying amounts of known information and explain how to leverage EM accordingly. They also bring up an important point that EM does not guarantee convergence to the correct solution and recommend starting with multiple initial PFMs over several runs. Their explanation of Gibbs' sampling was not as clear to me, but still provides helpful background on the differences between EM and Gibbs'.



A.

```
ATTCCGA
ATCACAA
TTTACGA
ATTGCGG
ACTTCGA
ATTAGGA
GTTACGA
ACTACCA
GTCTCGA
ATTTTGA
```

B.

Pos:	1	2	3	4	5	6	7
A	7	0	0	5	0	1	9
C	0	2	2	1	8	1	0
G	2	0	0	1	1	8	1
T	1	8	8	3	1	0	0

C.

Pos:	1	2	3	4	5	6	7
A	0.7	0.0	0.0	0.5	0.0	0.1	0.9
C	0.0	0.2	0.2	0.1	0.8	0.1	0.0
G	0.2	0.0	0.0	0.1	0.1	0.8	0.1
T	0.1	0.8	0.8	0.3	0.1	0.0	0.0

D.

Pos:	1	2	3	4	5	6	7
A	0.57	0.07	0.07	0.43	0.07	0.14	0.71
C	0.07	0.21	0.21	0.14	0.64	0.14	0.07
G	0.21	0.07	0.07	0.14	0.14	0.64	0.14
T	0.14	0.64	0.64	0.29	0.14	0.07	0.07

- R Durbin, S Eddy, A Krogh, G Mitchison Published by Cambridge University Press 1st (first) edition (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids https://www.amazon.com/Biological-Sequence-Analysis-ProbabilisticUniversity/dp/B00HQ0TD64 (http://www.mcb111.org/w06/durbin_book.pdf)

Chapter 5 in *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* adequately presents an in-depth explanation of Hidden Markov Models, some of which may be a little outside the scope of the class (i.e. Dirichlet mixtures). It is also quite reliant on mathematical equations, which may not be the best presentation of information for the diverse backgrounds of students in this course.

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert An Introduction to Statistical Learning: with Applications in R [ISLR (2nd edition)] https://www.amazon.com/Introduction-Statistical-Learning-ApplicationsStatistics/dp/1071614177/ + https://www.statlearning.com

ISL Chapter 12.4.2 describes hierarchical clustering, and more specifically agglomerative clustering. It provides an explanation of how to interpret similarity on a dendrogram properly and how the formation of clusters is dependent on the height chosen. It also describes the steps in a hierarchical clustering algorithm, and definitions for complete, single, and average linkage.

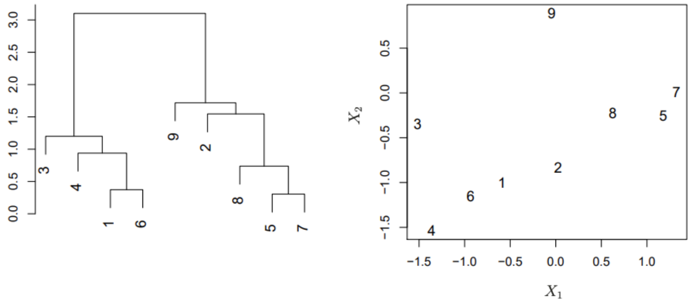
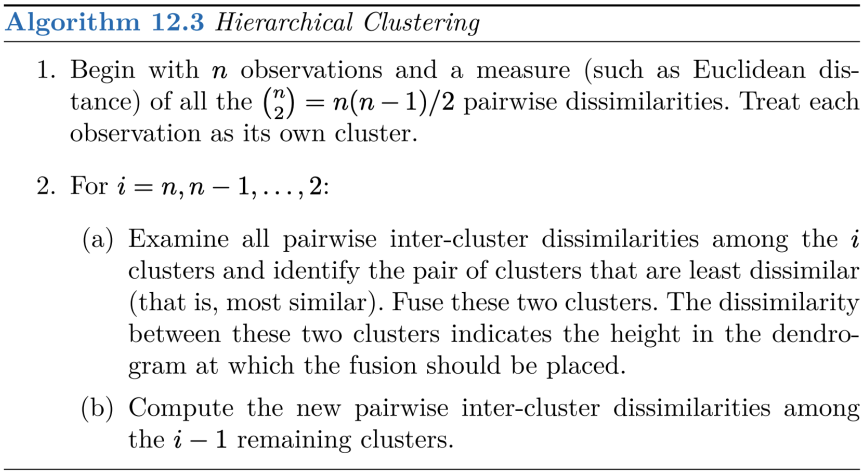


FIGURE 12.12. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6*. However, observation* 9 is no more similar to *observation* 2 *than it is to observations* $8, 5,$ *and* 7*, even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* $2, 8, 5,$ *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* $8, 5,$ *and* 7.



Algorithm 12.3 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

References ISL/ESL (if any)

Progressive Multiple Alignment & Clustering Approaches:

- **ESLII**: Chapter 14 (*Unsupervised Learning*) introduces clustering methods such as hierarchical clustering, which is a fundamental approach in progressive sequence alignment.

Motif Discovery (Expectation-Maximization, Gibbs Sampling)

- **ESLII:** Chapter 8 (*Model Inference and Averaging*) explains the Expectation-Maximization (EM) algorithm, which is widely used for motif discovery in biological sequences.
- **ISLP:** While ISLP does not explicitly cover EM, Chapter 5 (*Resampling Methods*) discusses probabilistic methods, including techniques used in motif searching.

Other suggested references for many of the key concepts

Multiple Sequence Alignment (MSA)

- **Durbin, R., Eddy, S., Krogh, A., & Mitchison, G.** (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.
- **Sievers, F., & Higgins, D. G.** (2014). "Clustal Omega." Current Protocols in Bioinformatics, 48(1), 3.13.1-3.13.16.
  - Overview of Clustal Omega, a widely used progressive alignment tool.
- **Nalbantoğlu Ö. U.** (2014). Dynamic programming. *Methods in molecular biology (Clifton, N.J.)*, *1079*, 3–27. https://doi.org/10.1007/978-1-62703-646-7_1
  - Introduction to various methods and techniques for dynamic programming in multiple sequence alignment.

Progressive Alignment & Clustering Approaches

- **Felsenstein, J.** (2004). *Inferring Phylogenies.* Sinauer Associates.

  - Explores phylogenetic tree construction, which is fundamental to progressive alignment.
- **Katoh, K., & Standley, D. M.** (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." *Molecular Biology and Evolution*, 30(4), 772-780.
- **Gotoh O.** (2014). Heuristic alignment methods. *Methods in molecular biology (Clifton, N.J.)*, *1079*, 29–43. https://doi.org/10.1007/978-1-62703-646-7_2
  - Outlines steps of progressive alignment.

Hidden Markov Models (HMMs)

- **Durbin, R. et al.** (1998). *Biological Sequence Analysis.* Cambridge University Press.
  - Covers HMMs in the context of sequence analysis.
- **Sjolander, K.** (1997). "Hidden Markov models and other probabilistic models for genomic sequence analysis." *Current Opinion in Structural Biology*, 7(3), 333-340.
  - Discusses different probabilistic models for sequence analysis.
- **L. Smith, L. Yeganova, W.J. Wilbur.** Hidden Markov models and optimized sequence alignments, Computational Biology and Chemistry, Volume 27, Issue 1, 2003, Pages 77-84, ISSN 1476-9271, https://doi.org/10.1016/S1476-9271(02)00096-8. (https://www.sciencedirect.com/science/article/pii/S1476927102000968)
  - Application of Hidden Markov to sequence alignment.