“Central Dogma” of Biology

**DNA =**

- double-stranded, linear molecule
- each strand is string over \{A,C,G,T\}
- strands are complements of each other \(A \leftrightarrow T; \ C \leftrightarrow G\)
- substrings encode for genes, most of which encode for proteins
Finding transcription factor binding sites can tell us about the cell's regulatory network.
RNA Polymerase

- b/c it makes RNA into a polymer
- is an enzyme

Discovered in 1960; Nobel prize for its discovery in 1959... oops

1959 Nobel awarded to Severo Ochoa and Arthur Kornberg for discovering what was mistakenly believed to be RNA pol.

1960 Sam Weiss and Jared Hurwitz discover the real RNA pol.

2006 Nobel awarded to Roger Kornberg (son of Arthur) for detailed structure of RNA pol.

Image of transcription occurring.
Each “hair” is a piece of RNA that RNA pol is growing off of the DNA.
Transcription Factor Binding Sites

RegulonDB (Feb 27, 2010)
Transcription Factor Binding Sites

RegulonDB (Feb 27, 2010)
Transcription Factor Binding Sites

Activation DNA Binding Sites Distribution

Position -44

Repression DNA Binding Sites Distribution

Position -10

X-axes of plots are not to scale.
Transcription Network

169 transcription factors (excluding sigmas)

3322 edges
- 1753 activation,
- 1369 repression,
- 185 both,
- 3 unknown
CCT domain, often found near one end of plant proteins.

Suppose we want to search for other examples of this domain.

How can we represent the pattern implied by these sequences?

One way is a Sequence Profile
Sequence Profiles (PSSM)

Motif Position

Amino Acid

Color \approx \text{Probability that the } i^{\text{th}} \text{ position has the given amino acid } = e_i(x).

\sum = 1
Sequence Logos

Height of letter \(\approx\) fraction of time that letter is observed at that position.

(Height of all the letters in a column \(\approx\) to how conserved the column is)
Scoring a Sequence

\[ \text{Score}(x) = \Pr(x \mid M) = \prod_{i=1}^{L} e_i(x_i) \]

Score of a string according to profile \( M \) = Product of the probabilities you would observe the given letters.

Color \( \approx \) Probability that the \( i^{\text{th}} \) position has the given amino acid = \( e_i(x) \).
Background Frequencies

Interested in how different this motif position is from what we expect by chance.

Correct for “expect by chance” by dividing by the probability of observing $x$ in a random string:

$$
\text{ScoreCorrected}(x) = \frac{\Pr(x \mid M)}{\Pr(x \mid \text{background})} = \prod_{i=1}^{L} \frac{e_i(x_i)}{b(x_i)}
$$

$b(x_i)$ := probability of observing character $x_i$ at random.
Usually computed as $(\# x_i \text{ in entire string}) / (\text{length of string})$

Often, to avoid multiplying lots of terms, we take the log and then sum:

$$
\text{ScoreCorrectedLog}(x) = \log \prod_{i=1}^{L} \frac{e_i(x_i)}{b(x_i)} = \sum_{i=1}^{L} \log \left( \frac{e_i(x_i)}{b(x_i)} \right)
$$
Problem: What about gaps?

- The PSSM doesn’t handle either:
  - **insertions** of characters in the string that are not in the profile.
  - **deletions** of positions in the profile (that don’t have a match in the string).

- A solution: use an HMM to model the profile!
A Simple HMM

- A profile is equivalent to a simple HMM:

\[ e_i(x) = \]

No choice about which state to visit.

Emission probabilities given by Sequence Profile
Handling Insertions
characters in the string that are not in the profile

The “I” state allows any number of non-profile characters to be output.

The emission probabilities for “I” states = random probability of observing each character.
Handling Deletions

positions in the profile that are not matched in the string

We could add $O(n^2)$ edges that allow us to skip any number of match states.

But this is too many edges.

Instead we add some delete states that don’t emit any characters:
Combining Insertions & Deletions
Every alignment corresponds to some path in this HMM.

Every path in this HMM corresponds to some alignment.
PROSITE

Database of protein domains

Patterns specified by these HMMs
The exact way the parameters are encoded is not important for this class.
Recap

• Short sequence patterns can be used to model protein domains (functional units of proteins)

• They also can match transcription factor binding sites.

• We can encode these patterns as Sequence Profiles (often called Position-Specific Scoring Matrices or PSSMs).

• To handle insertions and deletions, we can model the patterns as an HMM.

• Next: how do we find these motifs...