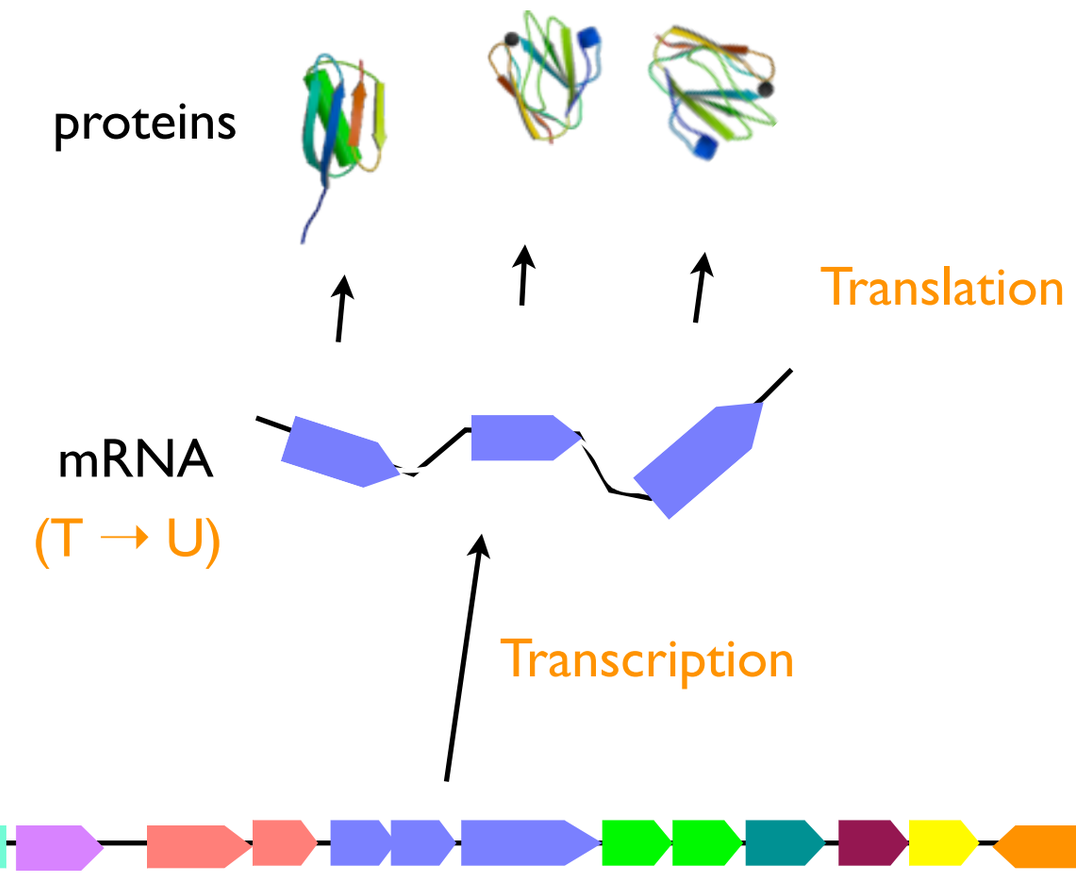
Motif Search

CMSC 423

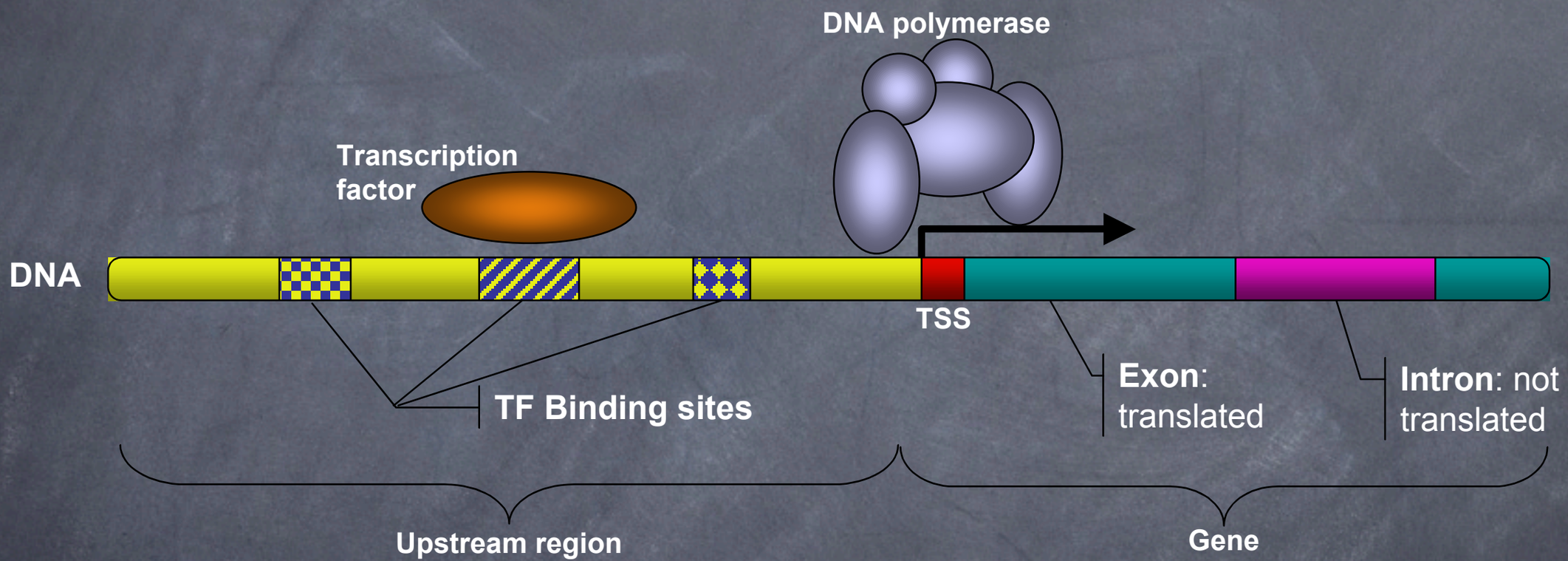
"Central Dogma" of Biology



DNA =

- double-stranded, linear molecule
- each strand is string over {A,C,G,T}
- strands are complements of each other (A ↔ T; C ↔ G)
- substrings encode for genes  most of which encode for proteins

DNA -> mRNA -> Protein



- 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.

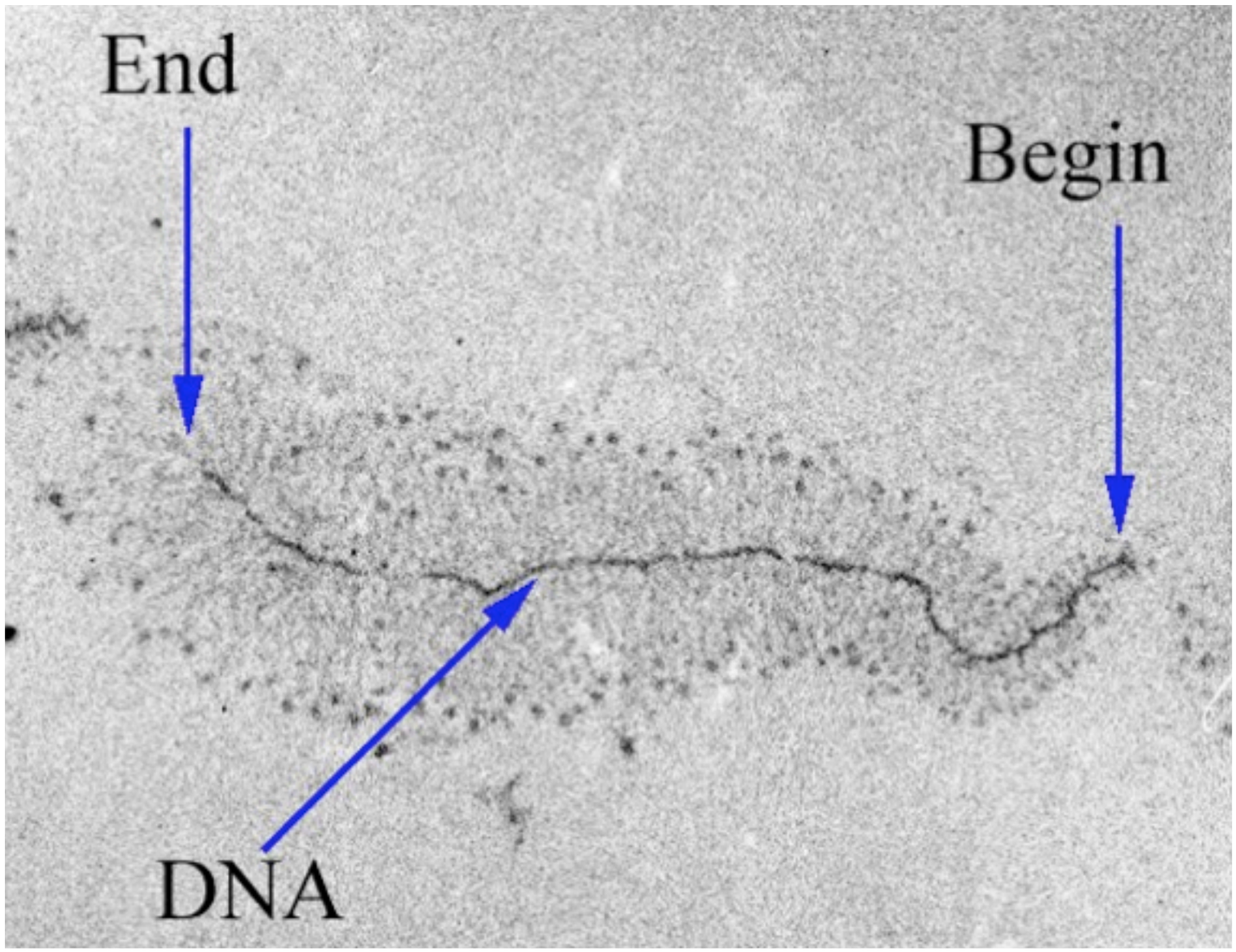
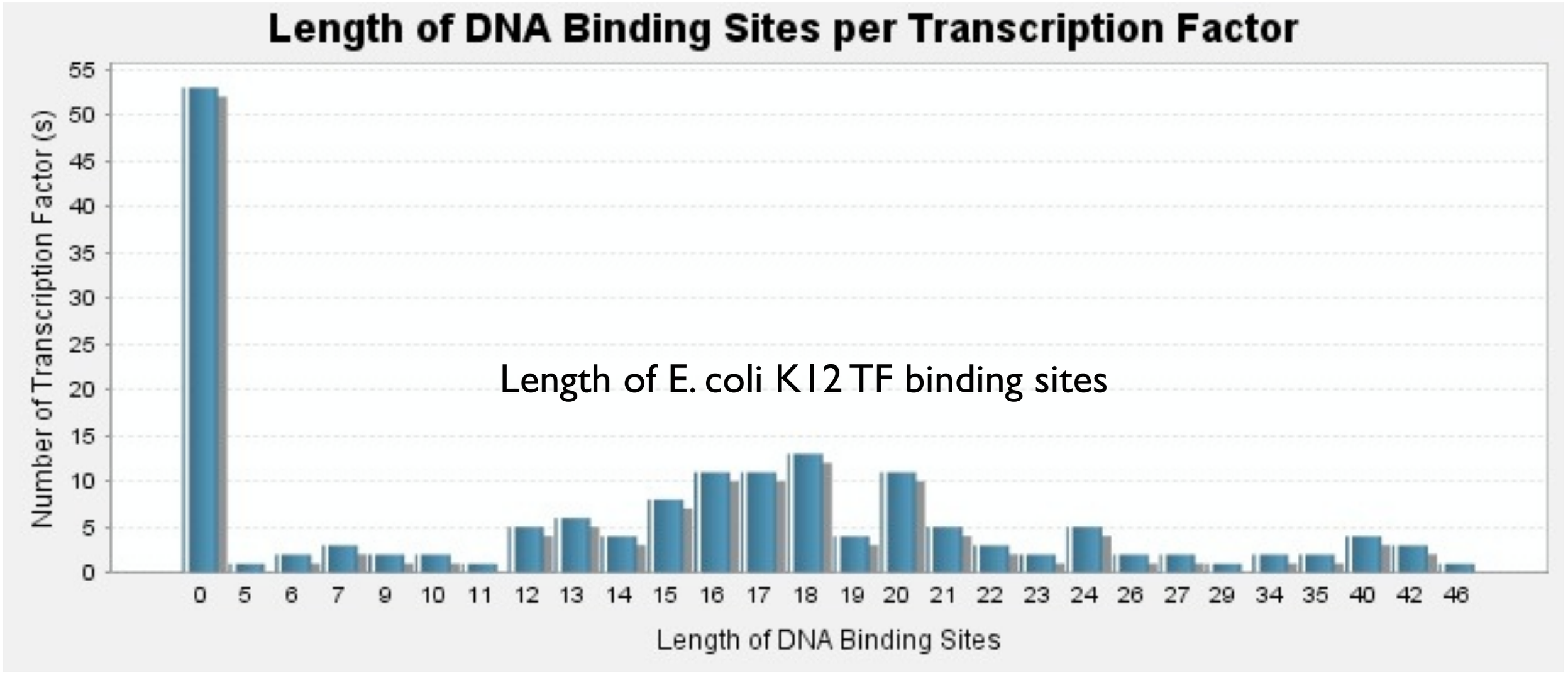


Image of transcription occurring.

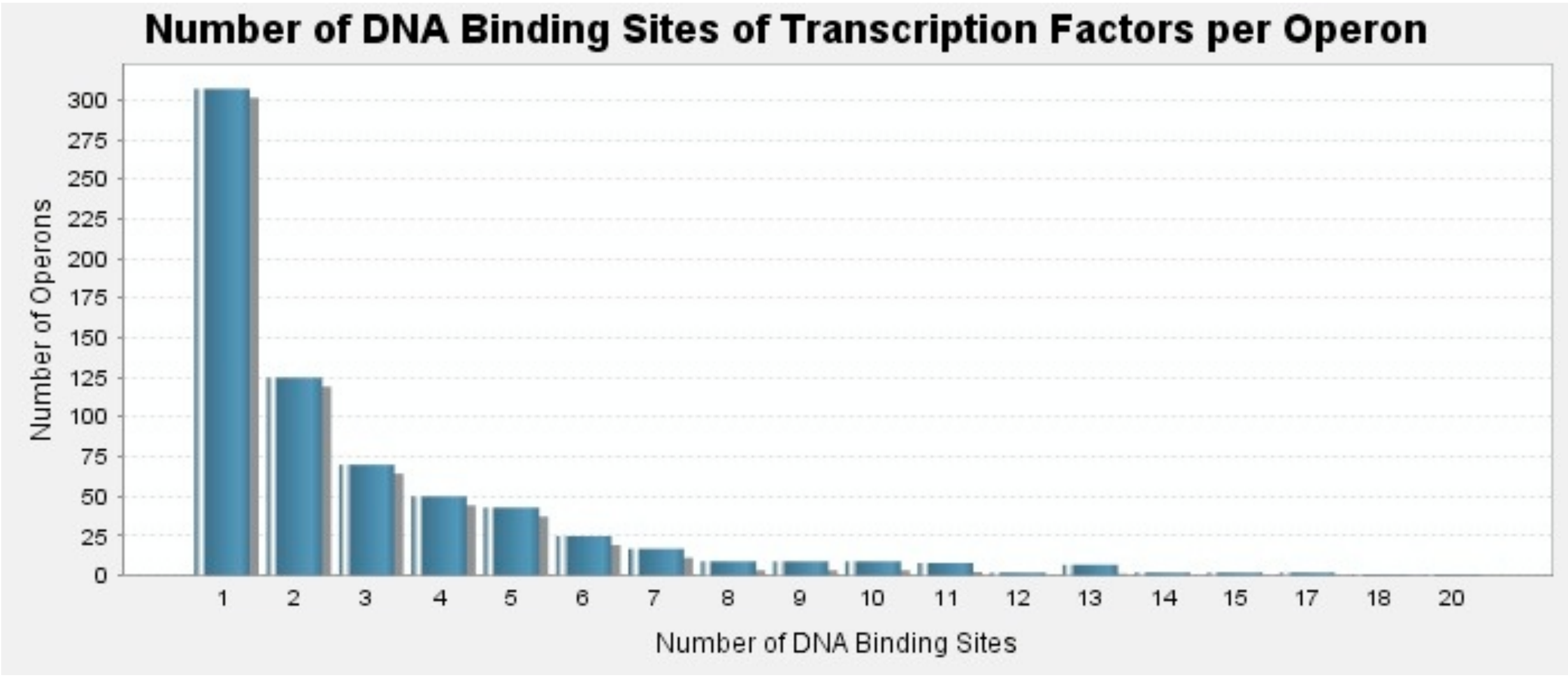
Each "hair" is a piece of RNA that RNA pol is growing off of the DNA.

Transcription Factor Binding Sites



RegulonDB (Feb 27, 2010)

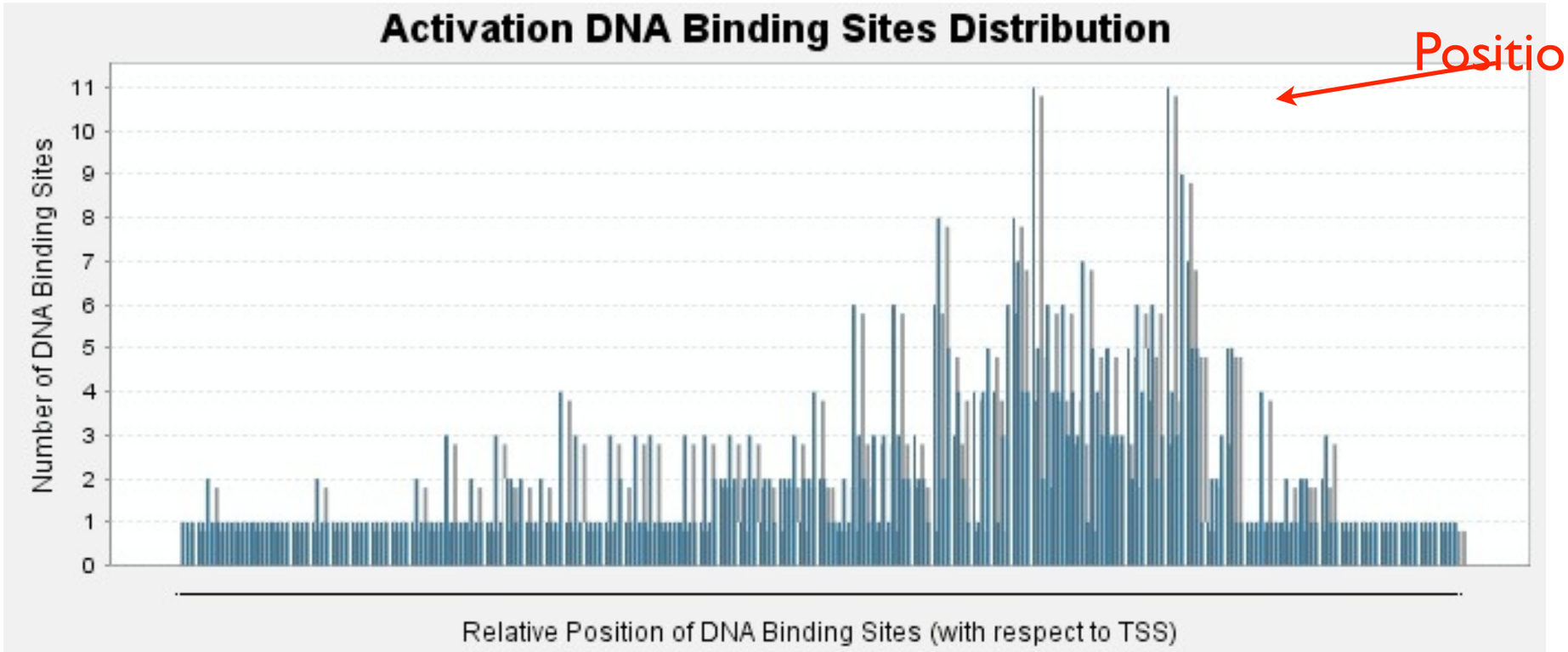
Transcription Factor Binding Sites



RegulonDB (Feb 27, 2010)

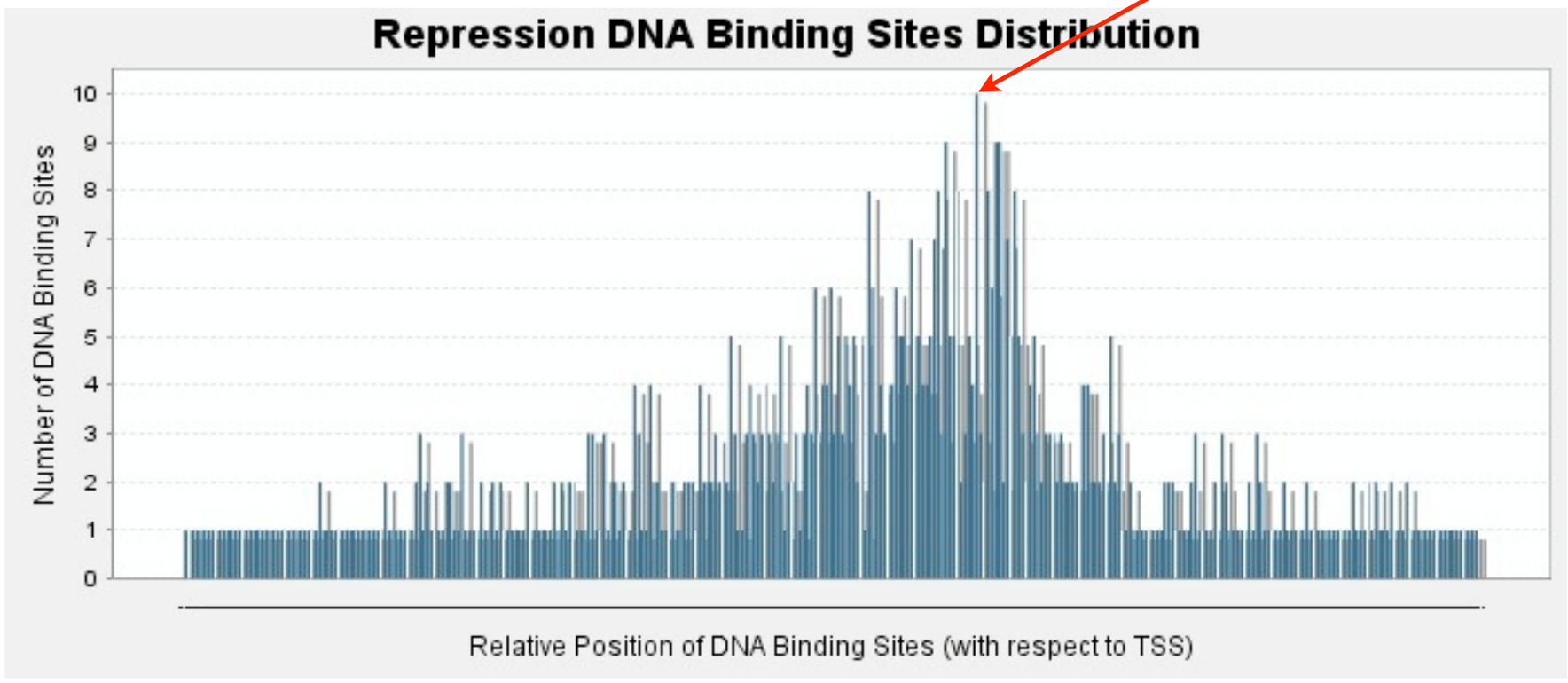
Transcription Factor Binding Sites

RegulonDB (Feb 27, 2010)



Position -44

X-axes of plots are not to scale.

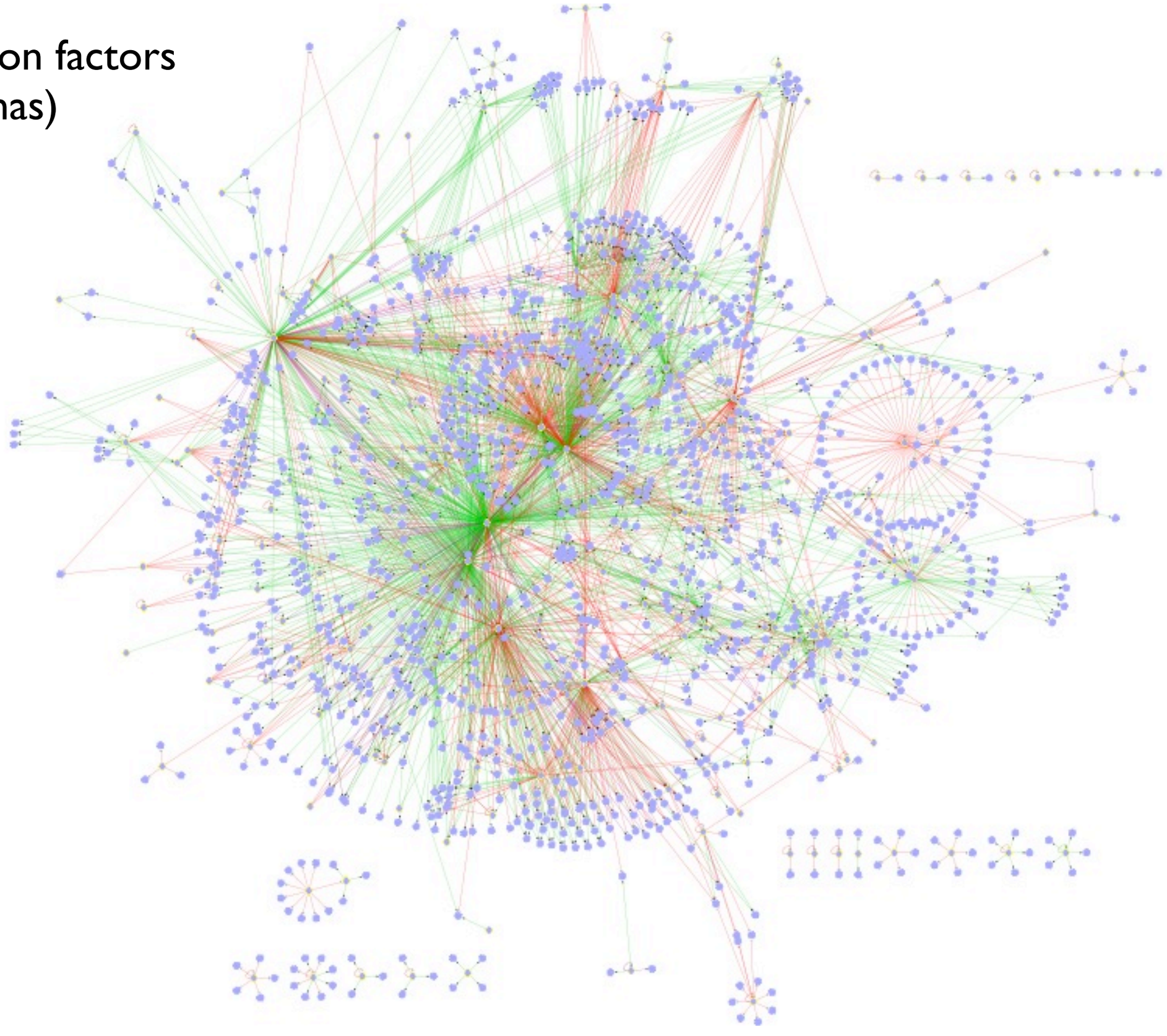


Position -10

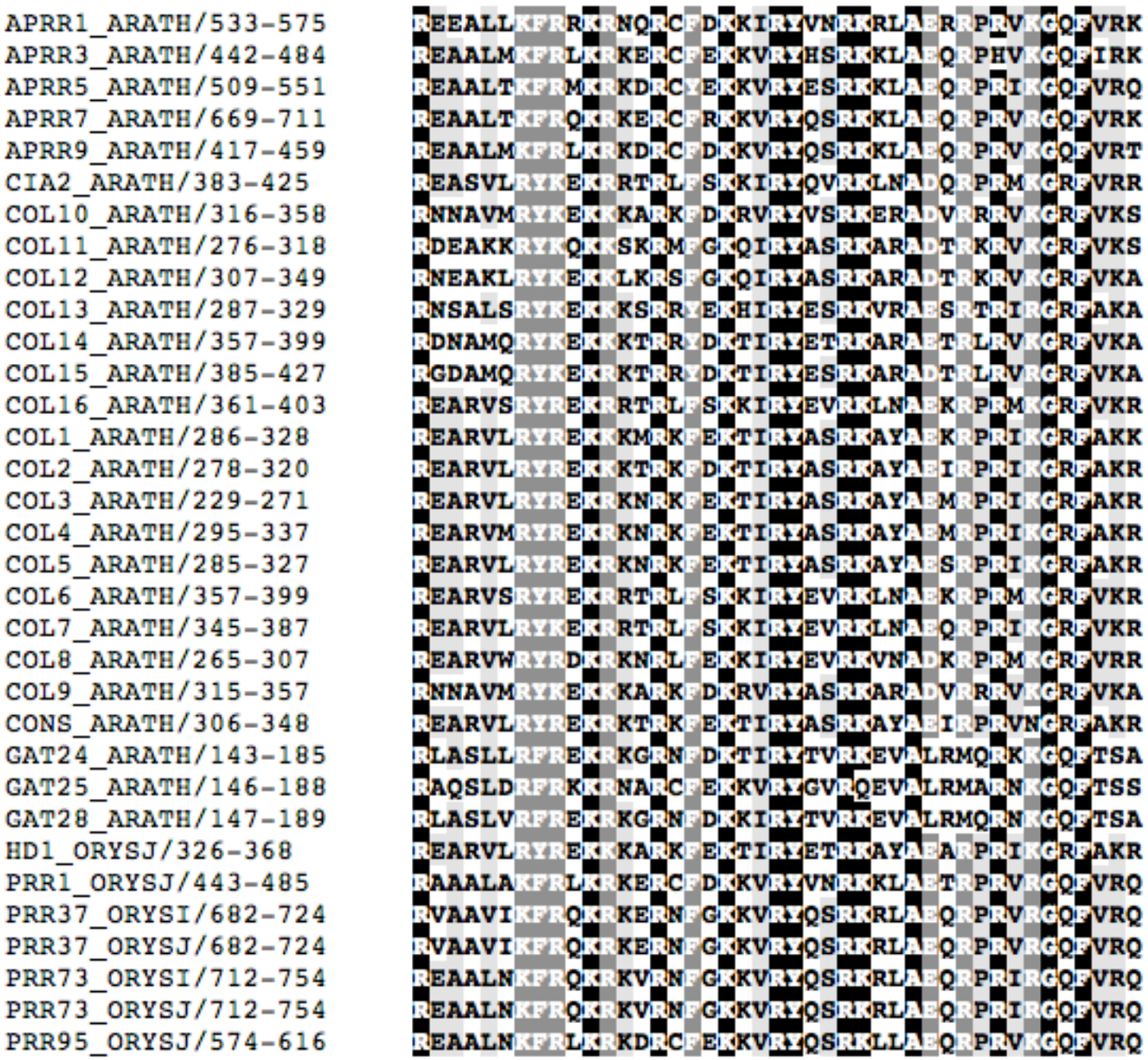
Transcription Network

169 transcription factors
(excluding sigmas)

3322 edges
- 1753 activation,
- 1369 repression,
- 185 both,
- 3 unknown



Sequence Profiles



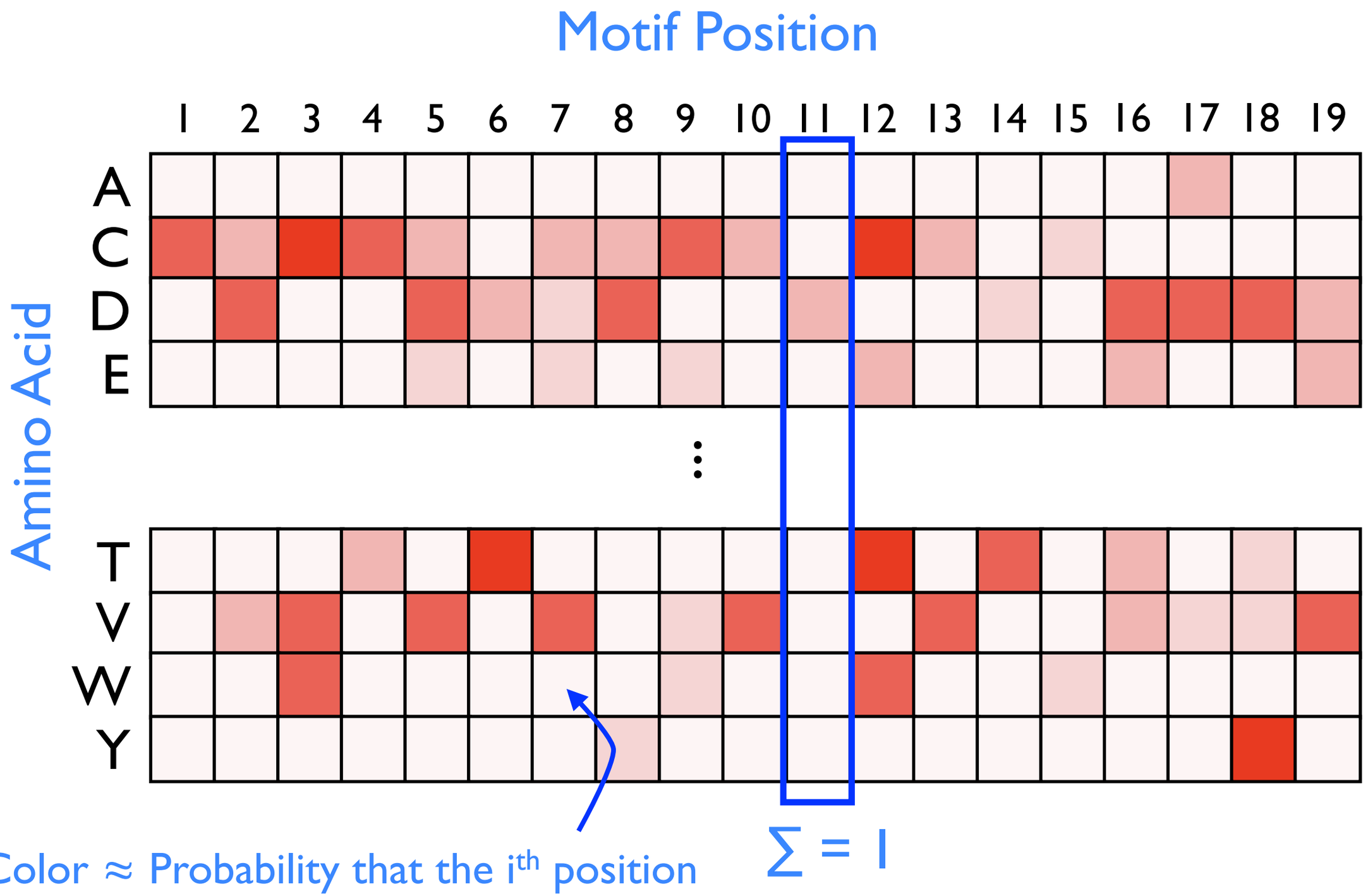
←CCT domain, often found near one end of plant proteins.

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

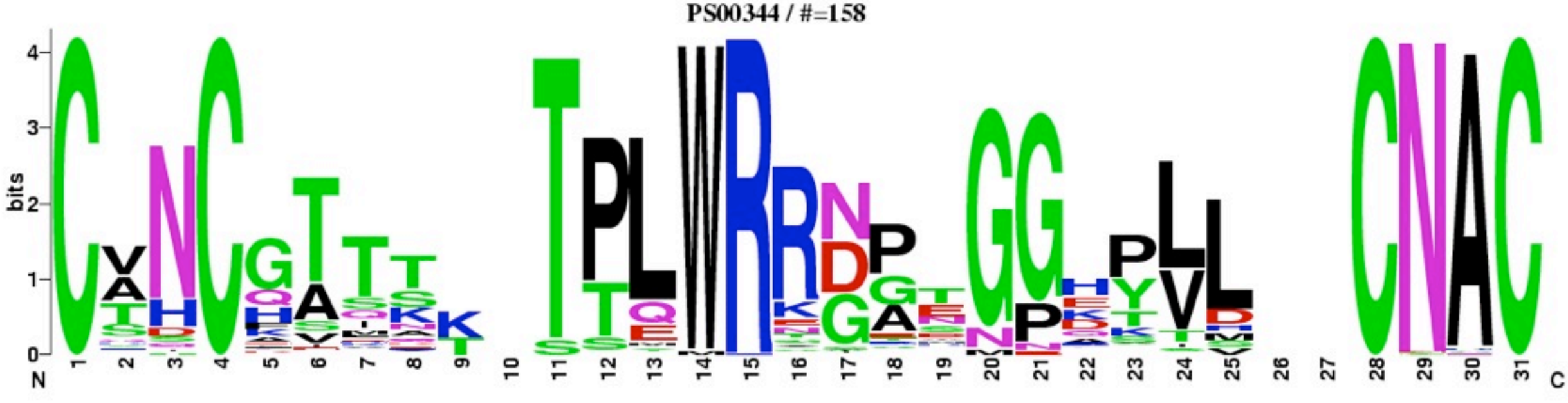
Sequence Profiles (PSSM)



Sequence Logos

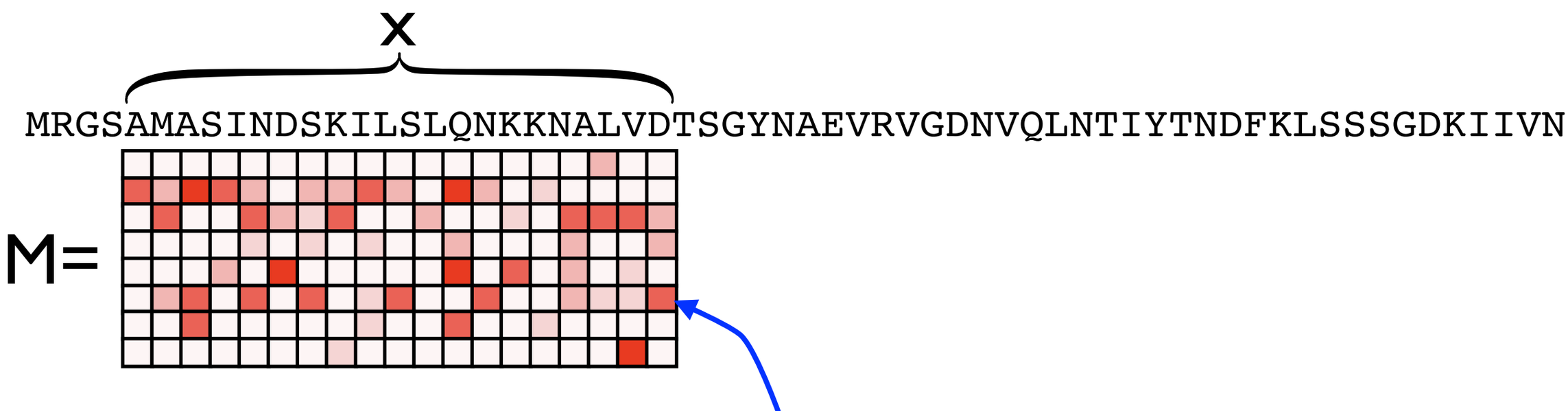
Height of letter ≈ fraction of time that letter is observed at that position.

(Height of all the letters in a column ≈ to how conserved the column is)



Motif Position

Scoring a Sequence



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

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

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

Background Frequencies

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

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

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

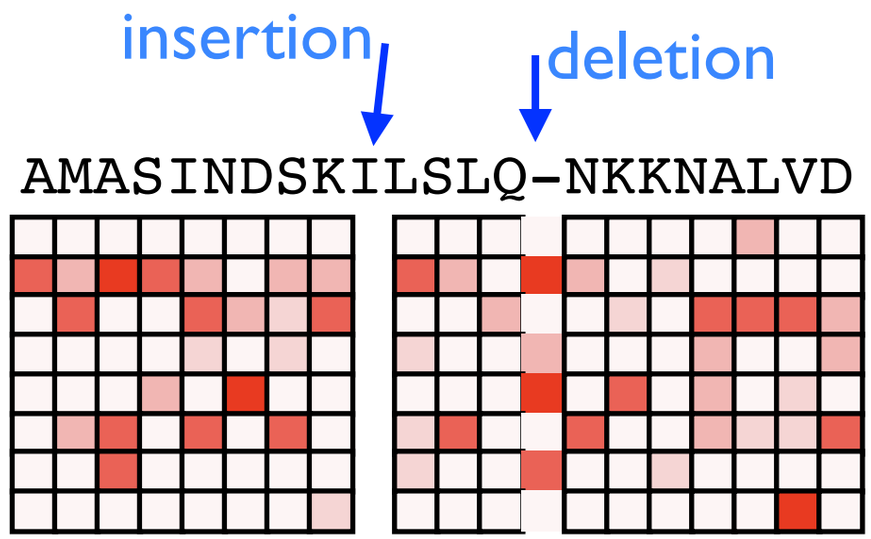
$b(x_i)$:= probability of observing character x_i at random.
Usually computed as (# x_i in entire string) / (length of string)

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).

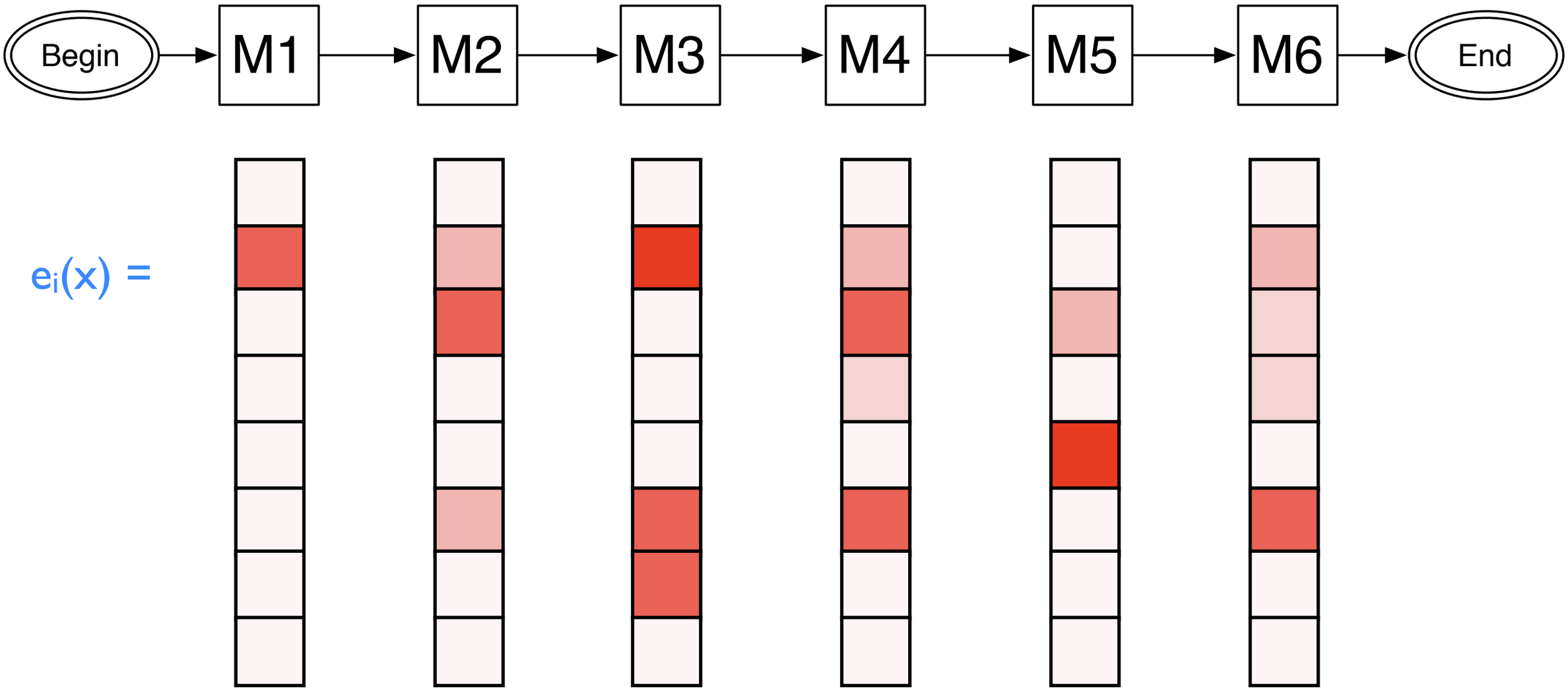


- A solution: use an HMM to model the profile!

A Simple HMM

- A profile is equivalent to a simple HMM:

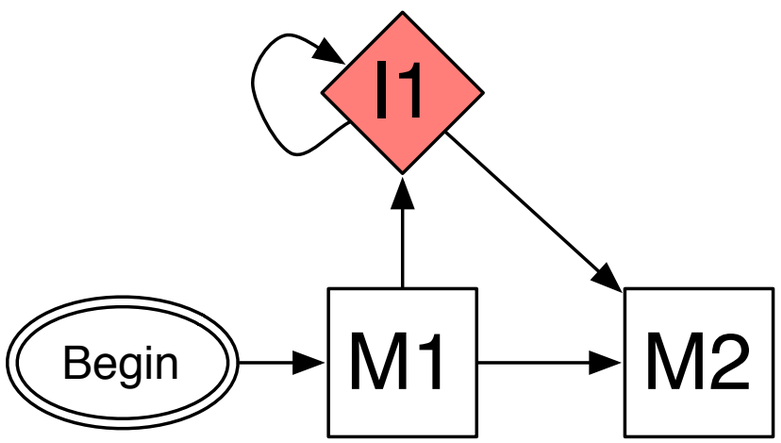
No choice about which state to visit.



Emission probabilities given by Sequence Profile

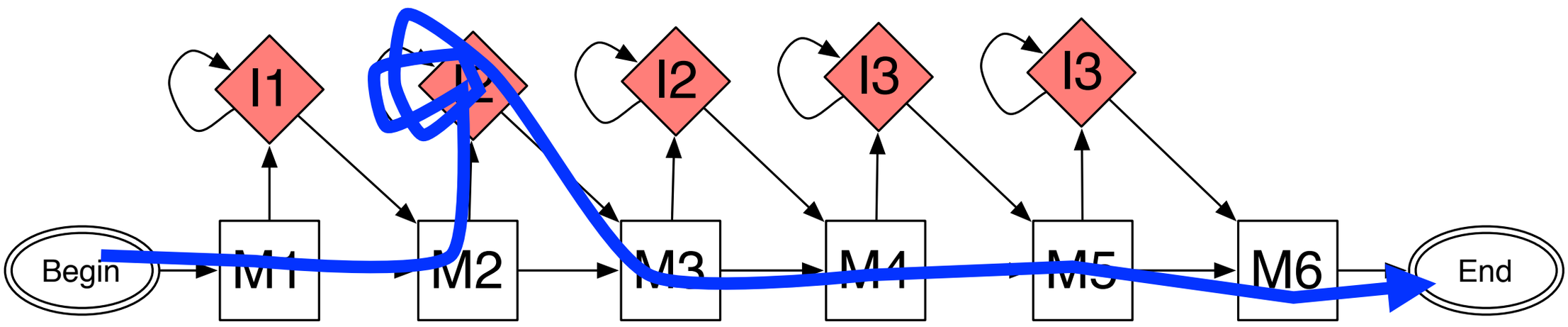
Handling Insertions

characters in the string that are not in the profile



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

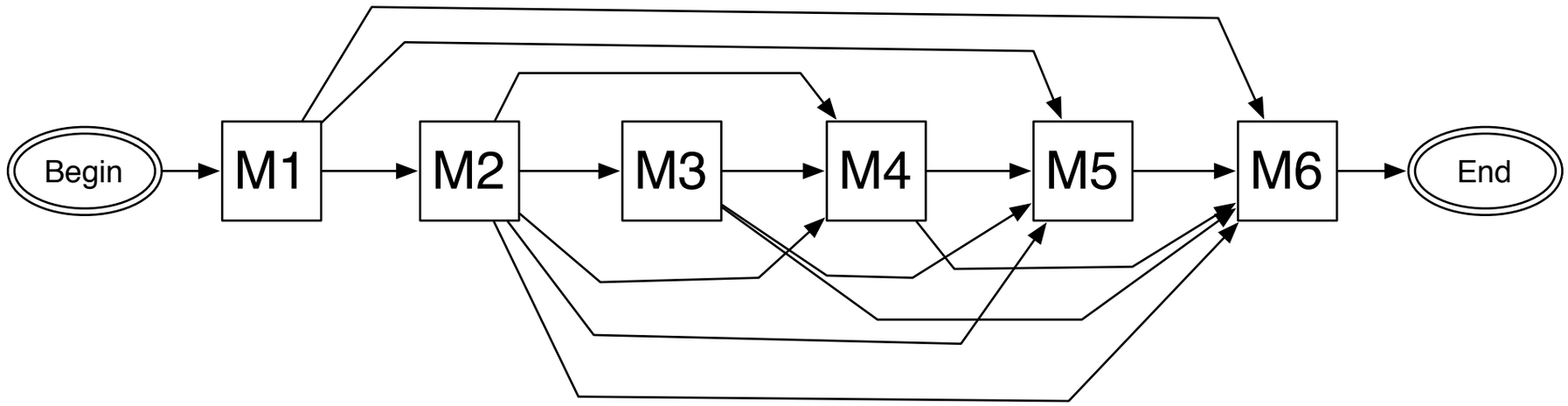
The emission probabilities for "I" states = random probability of observing each character.



Handling Deletions

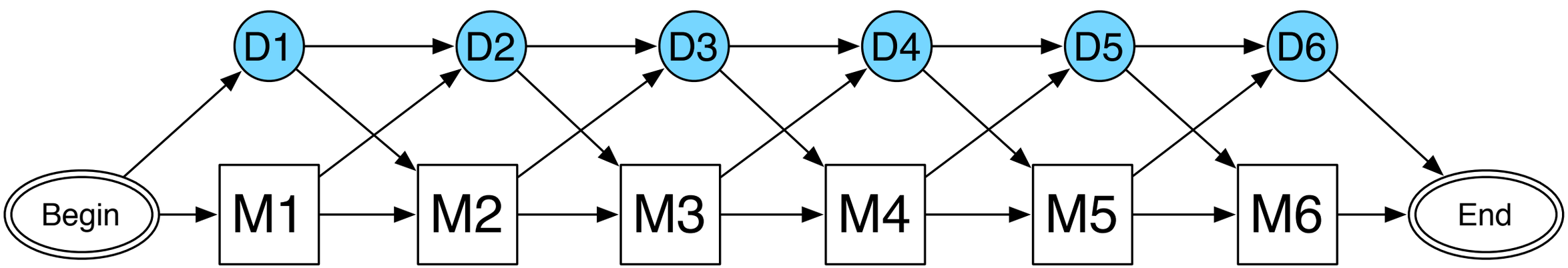
positions in the profile that are not matched in the string

We *could* add O(n2) edges that allow us to skip any number of match states.

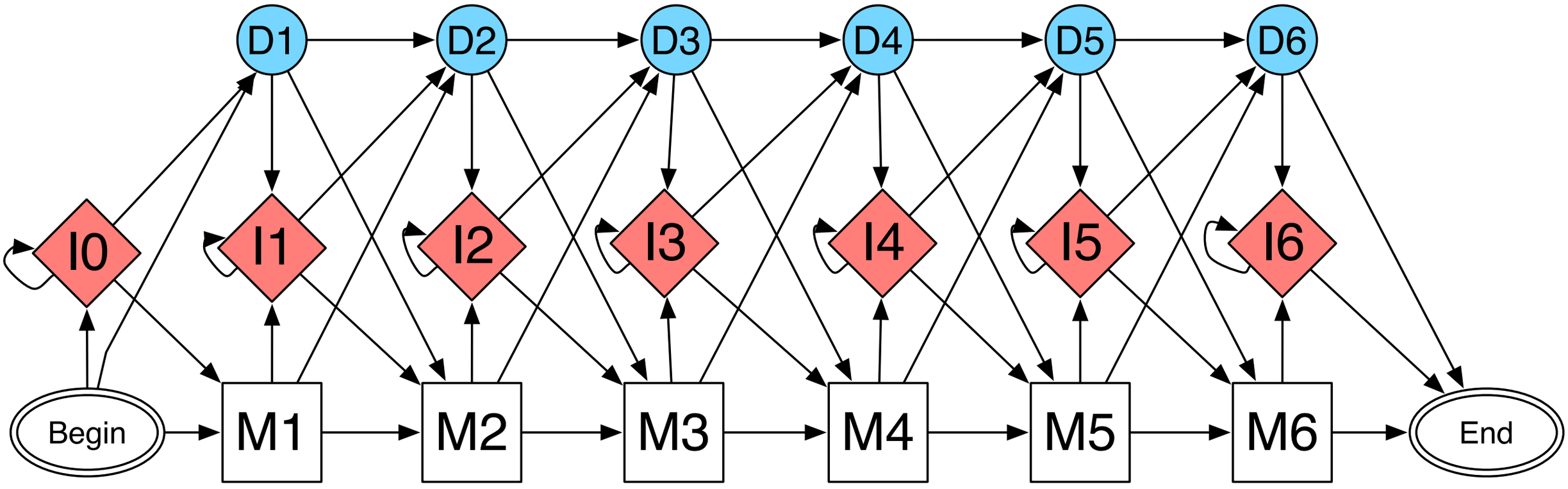


But this is too many edges.

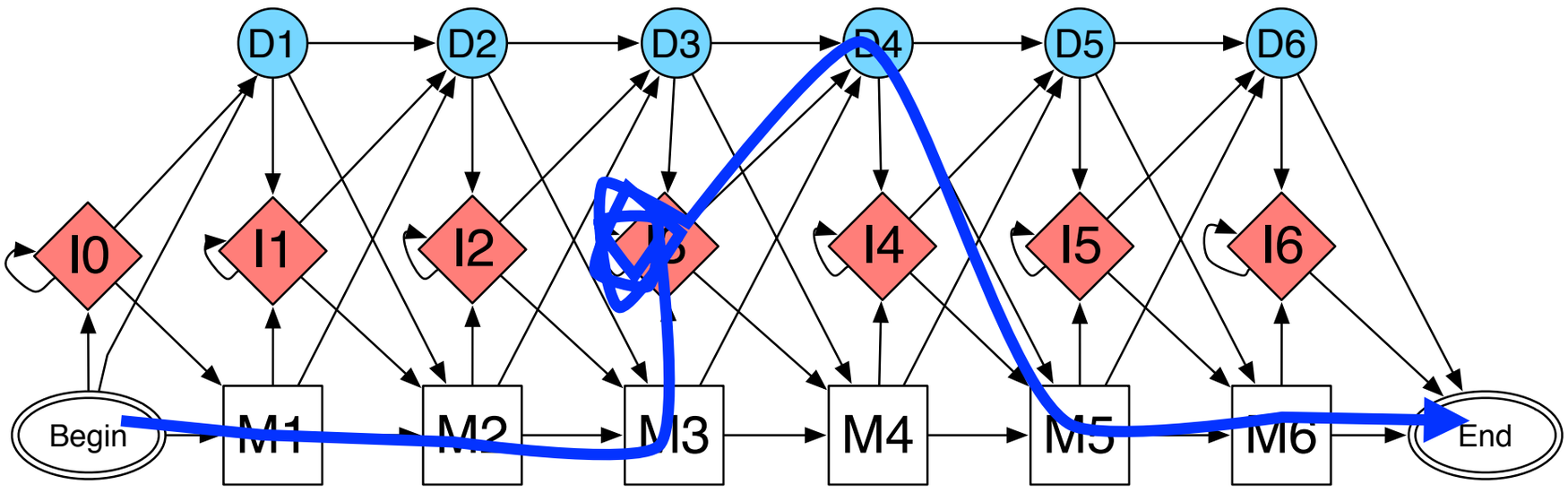
Instead we add some delete states that don't emit any characters:



Combining Insertions & Deletions

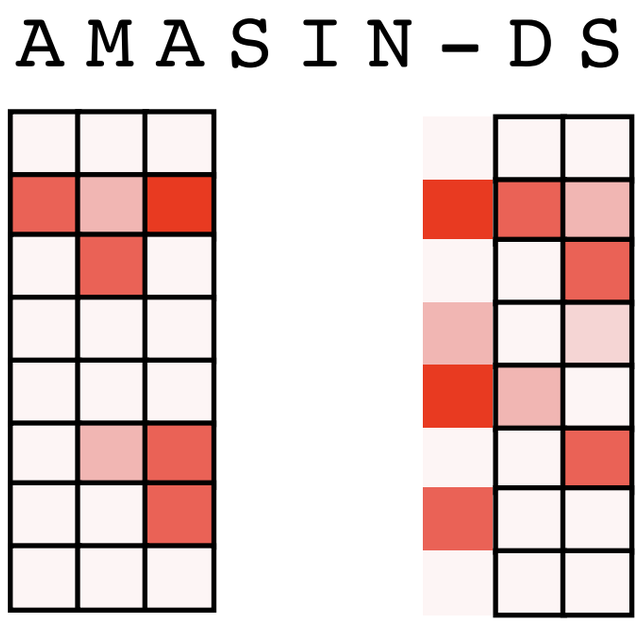


Example



Every alignment corresponds to some path in this HMM.

Every path in this HMM corresponds to some alignment.



PROSITE documentation PDOC51017

http://expasy.org/prosite/PDOC51017

Swiss Institute of Bioinformatics

ExPASy Proteomics Server

Search ExPASy web site for zinc finger

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Databases > PROSITE

Home ScanProsite ProRule Documents Downloads Links Funding

PDOC51017

CCT domain profile

Description:

The CCT (CONSTANS, CO-like, and TOC1) domain is a highly conserved basic module of ~43 amino acids, which is found near the C-terminus of plant proteins often involved in light signal transduction. The CCT domain is found in association with other domains, such as the B-box zinc finger (see <PDOC50119>), the GATA-type zinc finger (see <PDOC00300>), the ZIM motif or the response regulatory domain (see <PDOC50110>). The CCT domain has been shown to be involved in nuclear localization and probably also has a role in protein-protein interaction [1,2].

Some proteins known to contain a CCT domain are listed below:

- Plant CONSTANS family of transcription factors.
- Plant GATA factor subfamily III [3].
- Arabidopsis thaliana timing of CAB expression 1 (TOC1) or ABI3-interacting protein 1 (AIP1).
- Arabidopsis thaliana TOC1-Like (TL).

The profile we developed covers the entire CCT domain.

Last update:
September 2004 / First entry.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

CCT, PS51017; **CCT domain profile** (MATRIX)

Sequences known to belong to this [illegible] detected by the profile:	ALL
Other sequence(s) detected in Swiss-Prot:	NONE

- Domain architecture view of Swiss-Prot proteins matching PS51017

- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS51017
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS51017
- Scan Swiss-Prot/TrEMBL entries against PS51017
- view ligand binding statistics

References:

1
Authors Strayer C., Oyama T., Schultz T.F., Raman R., Somers D.E., Mas P., Panda S., Kreps J.A., Kay S.A.
Title *Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog.*
Source Science 289:768-771(2000).
PubMed ID 10926537

2
Authors Robson F., Costa M.M.R., Hepworth S.R., Vizir I., Pineiro M., Reeves P.H., Putterill J., Coupland G.
Title *Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants.*
Source Plant J. 28:619-631(2001).
PubMed ID 11851908

PROSITE

Database of protein domains

Patterns specified by these HMMs

```
/DEFAULT: M0=-7; D=-50; I=-50; B1=-500; E1=-500; MI=-105; MD=-105; IM=-105; DM=-105;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:         B1=0; BI=-105; BD=-105;
/M: SY='R'; M=-19, -8,-30, -7,  3,-21,-19, 10,-30, 25,-20,-10,  1,-19, 10, 58, -9,-11,-22,-22, -7,  2;
/M: SY='E'; M=  1,  0,-24,  1, 16,-22,-15, -8,-19,  8,-13,-10, -2,-11,  4,  4, -4, -8,-15,-25,-15,  9;
/M: SY='A'; M= 18,  2,-18, -2, 10,-24, -8, -9,-18,  0,-17,-13,  1, -9,  4, -7,  7, -2,-13,-26,-18,  7;
/M: SY='R'; M=  5, -8,-20,-11, -2,-16,-13,-11,-18, 10,-16, -9, -4,-14,  1, 13,  3, -2,-10,-22,-12, -1;
/M: SY='I'; M= -6,-26,-18,-28,-22,  1,-29,-23, 23,-20, 22, 14,-24,-26,-20,-18,-17, -5, 23,-24, -5,-22;
/M: SY='M'; M= -8, -9,-22,-12, -5,-13,-20,  0, -3, -5,  1, 12, -8,-17,  5, -3, -6, -2, -6,-24, -5, -1;
/M: SY='R'; M=-16, -4,-29, -3,  3,-23,-18, -3,-30, 31,-23,-12,  2,-17,  9, 53, -7, -9,-20,-22,-11,  3;
/M: SY='Y'; M=-19,-18,-28,-19,-18, 31,-29,  8, -1,-13,  0, -1,-17,-27,-12,-12,-18,-10, -8, 18, 57,-18;
/M: SY='R'; M=-15, -9,-28,-10,  0,-16,-21, -7,-23, 27,-16, -7, -3,-17,  4, 42,-10, -7,-16,-19, -7,  0;
/M: SY='E'; M= -9,  5,-28, 11, 33,-26,-19, -4,-23,  8,-17,-14, -1, -2, 12,  2, -3, -8,-21,-29,-17, 22;
/M: SY='K'; M= -9, -3,-28, -4,  5,-22,-16,-11,-26, 36,-24, -9, -2,-12,  4, 21, -8, -9,-17,-20, -9,  5;
/M: SY='R'; M=-15, -9,-28,-10, -1,-16,-21, -6,-23, 29,-18, -5, -4,-18,  5, 44,-11,-10,-14,-19, -7,  0;
/M: SY='K'; M= -9,  0,-26, -2,  6,-24,-14, -7,-22, 23,-21, -9,  3,-14,  8, 18, -5, -7,-17,-24,-12,  6;
/M: SY='T'; M= -3,  1,-20, -3,  1,-17,-10, -8,-14, -2,-11, -6,  4,-15,  1, -1,  2,  5,-13,-28,-13,  1;
/M: SY='R'; M=-19,-10,-29,-10, -2,-17,-21,  5,-27, 24,-18, -9, -1,-20,  8, 58, -9, -7,-19,-18, -4, -2;
/M: SY='R'; M=-11, -2, -5, -7, -5,-17,-18, -4,-18,  7,-12, -7,  4,-21, -3,  8, -7, -7,-15,-29,-12, -5;
/M: SY='F'; M=-19,-22,-23,-31,-25, 57,-26,-13, -2,-24,  5, -2,-13,-29,-30,-17,-18,-10, -5, 13, 29,-25;
/M: SY='D'; M= -7,  9,-25, 14, 12,-26,  1, -6,-26, -4,-20,-17,  5,-13,  0, -5,  2, -7,-21,-30,-19,  5;
/M: SY='K'; M=-10, -5,-22, -6,  4,-25,-22,-12,-23, 37,-22, -7, -4,-14,  5, 23,-12,-10,-15,-21,-10,  4;
/M: SY='K'; M= -8, -3,-25, -6,  3,-23,-20, -6,-21, 21,-18, -7, -2,-13, 11, 19, -2,  3,-15,-22, -9,  7;
/M: SY='I'; M= -8,-27,-20,-33,-27, -5,-30,-25, 32,-26, 10, 11,-20,-17,-21,-26,-16, -9, 24,-24, -6,-27;
/M: SY='R'; M=-14,-11,-27,-12, -2,-18,-20, -5,-23, 24,-15, -7, -4,-19,  5, 51,-10, -8,-14,-21,-10, -2;
/M: SY='Y'; M=-20,-20,-30,-20,-20, 30,-30, 20,  0,-10,  0,  0,-20,-30,-10,-10,-20,-10,-10, 30, 80,-20;
/M: SY='A'; M= 15, -3,-20, -5, 13,-25,-10, -8,-16, -2,-14,-10, -6, -9,  9, -9,  4, -3,-12,-24,-17, 11;
/M: SY='C'; M=  2,-11, 17,-14,-14,-14,-15,-19, -9,-17,-15,-11, -6,-22,-14,-17, 14,  8,  4,-39,-19,-14;
/M: SY='R'; M=-20,-10,-30,-10,  0,-20,-20,  0,-30, 30,-20,-10,  0,-20, 10, 70,-10,-10,-20,-20,-10,  0;
/M: SY='K'; M= -9, -1,-29, -1,  9,-29,-19, -9,-29, 45,-28,-10,  0,-11, 11, 31, -9,-10,-20,-20,-10,  9;
/M: SY='A'; M=  7, -7,-19,-10,  2,-17,-16,-14,-12,  2, -7, -6, -7,-12, -1, -2,  1,  5, -6,-24,-13,  0;
/M: SY='L'; M=-11,-15,-21,-18,-16,  2,-24,-10,  6,-13, 14,  5,-10,-26,-13, -5,-15, -4,  3,-18,  5,-16;
/M: SY='A'; M= 50,-10,-10,-20,-10,-20,  0,-20,-10,-10,-10,-10,-10,-10,-10,-20, 10,  0,  0,-20,-20,-10;
/M: SY='D'; M=-12, 20,-21, 31, 26,-29,-17, -5,-27, -1,-17,-19,  4,-10,  4, -9, -3, -9,-22,-34,-18, 15;
/M: SY='Q'; M= -4, -2,-18, -3,  1,-23,-14, -7,-19,  6,-19, -9,  1,-14,  9,  9,  7,  4,-13,-28,-13,  4;
/M: SY='R'; M=-18,-10,-29,-11, -1,-19,-20,  0,-24, 26,-16, -2, -2,-19, 11, 58,-11,-10,-17,-20, -9,  1;
/M: SY='P'; M= -8,-17,-35,-10,  0,-27,-20,-16,-18, -3,-23,-14,-16, 57, -3,-10, -9, -9,-24,-27,-24, -5;
/M: SY='R'; M=-20,-10,-30,-10,  0,-20,-20,  1,-30, 29,-20,-10,  0,-20, 10, 68,-10,-10,-20,-19, -9,  0;
/M: SY='V'; M= -2,-22,-16,-26,-23, -3,-26,-22, 23,-16,  8, 12,-18,-24,-19,-16,-10, -3, 29,-27, -8,-22;
/M: SY='K'; M=-13, -2,-30, -3,  5,-27,-15, -6,-30, 38,-27,-11,  3,-14,  9, 38, -9,-10,-21,-21,-11,  6;
/M: SY='G'; M=  1,-12,-28,-12,-20,-27, 61,-21,-34,-20,-26,-17, -3,-21,-20,-20, -1,-18,-24,-21,-28,-20;
/M: SY='R'; M=-15, -7,-27, -7,  5,-25,-19,  1,-26, 21,-20, -8,  0,-17, 22, 47, -5, -7,-22,-21,-11, 10;
/M: SY='F'; M=-19,-26,-20,-36,-28, 72,-28,-19, -1,-28,  8, -1,-16,-29,-37,-19,-18, -8, -1,  6, 26,-28;
/M: SY='V'; M= 18,-18,-13,-23,-18,-10,-17,-23,  9,-14,  0,  0,-16,-17,-18,-17,  0,  3, 19,-25,-14,-18;
/M: SY='R'; M=-11, -4,-27, -4,  4,-25,-16, -7,-28, 32,-26,-12,  1,-14,  8, 38, -2, -5,-18,-23,-12,  4;
/M: SY='N'; M=  0,  6,-19, -1,  0,-21,-11, -5,-17,  3,-20,-11, 13,-15,  6,  5,  8,  4,-14,-29,-15,  2;
/I:         E1=0; IE=-105; DE=-105;
```

Probabilities for leaving insertion states.

Emission probabilities for each match state

The exact way the parameters are encoded 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...