Lecture 8: Patterns, Profiles, and Motifs

- Finding patterns in protein and DNA sequences
- Calculating profiles of DNA sequences

Some slides adapted from slides from Dr. Keith Dunker

Some slides adapted from slides created by Dr. Zhiping Weng (Boston University)

Definitions and Resources

§ Motif: A region of a protein or DNA sequence that may be functionally or structurally significant and/or conserved in other sequences

- Motifs usually contain biologically important sequences

§ Pattern: Describes a motif using a qualitative consensus sequence (e.g., IUPAC or regular expression)

§ Profile: Describes a motif using quantitative information captured in a position specific scoring matrix (weight matrix)

§ PROSITE is a protein sequence pattern and profile database

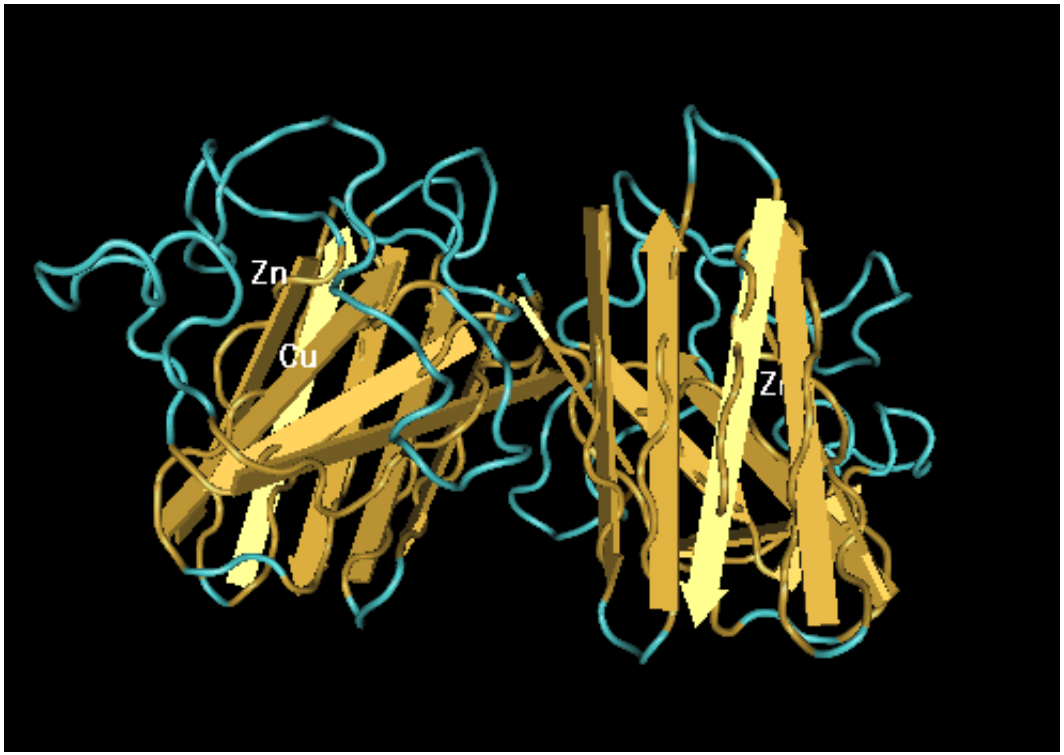
- http://www.expasy.ch/prosite
- Contains >1100 entries describing >1600 patterns and profiles

§ DNA pattern and profile databases are more fragmented

- JASPAR (http://jaspar.genereg.net/) and *S. cerevisiae* Promoter Database (SCPD) (http://rulai.cshl.edu/SCPD/)

Importance of Sequence Patterns in Proteins

- Conserved patterns in protein sequences usually have important biological functions
- Conserved sequence patterns may be indicative of e.g., a protein structural domain, enzyme active site, or a binding site for another protein or metal ion



Cu,Zn Superoxide Dismutase

Image created using Cn3D from pdb id: 1XSO

Steps in the Development of a New PROSITE Pattern

(1) Construct a multiple sequence alignment of a protein family

(2) Use the alignment to identify conserved or biologically significant residues (e.g., residues in catalytic/active site, binding domain, structural features)

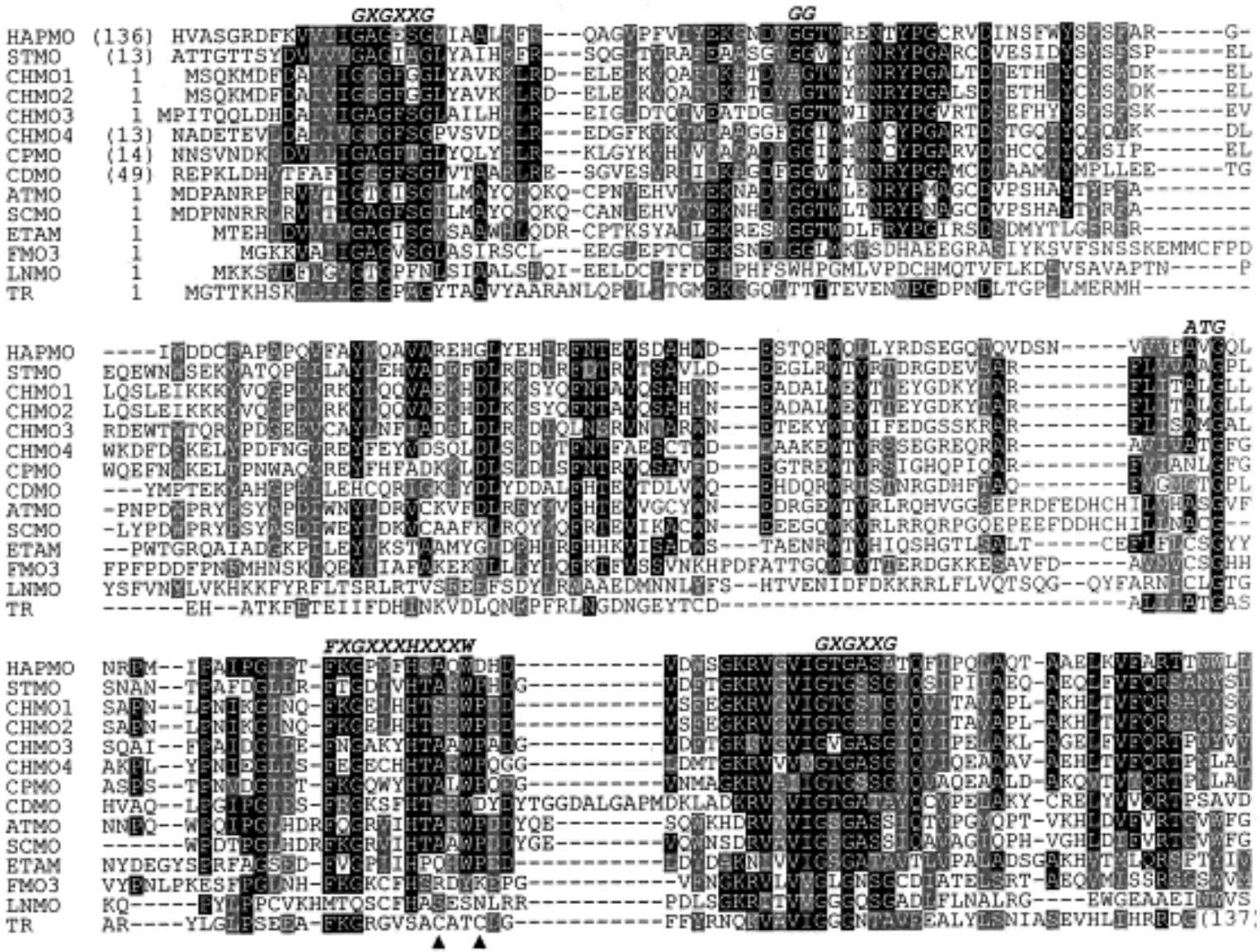
(3) Start by creating a core sequence pattern (approximately 4-5 contiguous amino acids in length)

(4) Expand the pattern to improve its sensitivity and specificity for detecting known protein family members

- Sensitivity: Test the trial pattern against known positive sequences
- Specificity: Test the trial pattern against known negative sequences

Information from http://ca.expasy.org/prosite/prosuser.html

Finding Patterns in Multiple Sequence Alignments



Baeyer-Villiger monooxygenases (BVMOs) - taken from Fraaije, et al (2002) *FEBS Letters* **518**:43-47

Pattern of Cu/Zn Superoxide Dismutase

§ Example of a PROSITE pattern (**PS00087; SOD_CU_ZN_1**):

[GA]-[IMFAT]-H-[LIVF]-H-{S}-x-[GP]-[SDG]-x-[STAGDE]

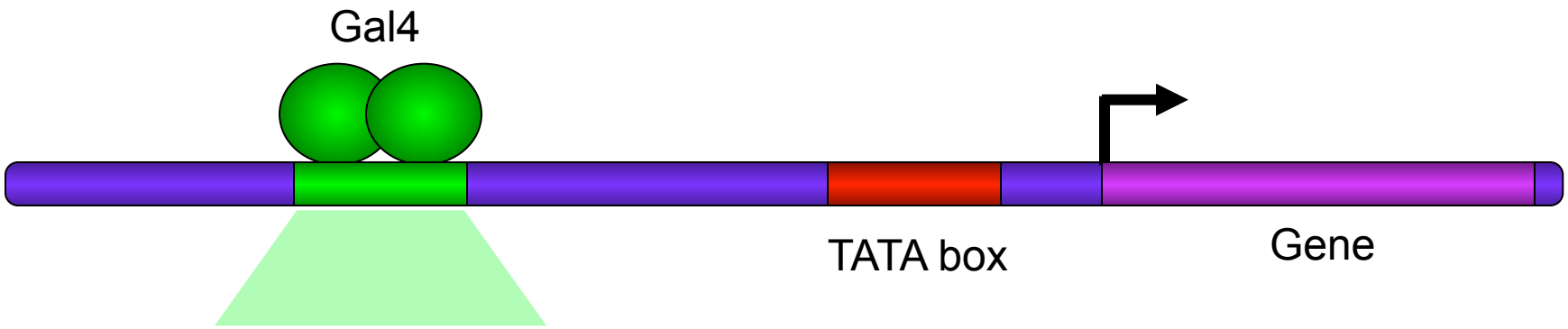
§ The two histidines (H) are copper ligands

§ Pattern Nomenclature for PROSITE database:

- Each position in pattern is separated with a hyphen
- **x** can match any residue
- **[ ]** are used indicate ambiguous positions in the pattern
  e.g., [SDG] means the pattern can match S, D, or G at this position
- **{ }** are used to indicate residues that are not allowed at this position
  e.g., {S} means NOT S (not Serine)
- **( )** surround repeated residues, e.g., A(3) means AAA
- **<** or **>** indicate the pattern is restricted to the N- or C-terminus of the sequence

Information from http://ca.expasy.org/prosite/prosuser.html

Many DNA patterns are binding sites for Transcription Factors



GAL3	CGGTCCACTGTGTGCCG
GAL7	CGGAGCACTGTTGAGCG
GAL80	CGGCGCACTCTCGCCCG
GCY1	CGGGGCAGACTATTCCG
GAL1	CGGATTAGAAGCCGCCG
	CGGGCGACAGCCCTCCG
	CGGAAGACTCTCCTCCG
GAL10	CGGAGGAGAGTCTTCCG
	CGGAGGGCTGTCGCCCG
	CGGCGGCTTCTAATCCG
GAL2	CGGAAAGCTTCCTTCCG
	CGGCGGTCTTTCGTCCG
	CGGAGATATCTGCGCCG
	CGGGGCGGATCACTCCG
	CGGATCACTCCGAACCG
PCL10	CGGAGTATATTGCACCG
MTH1	CGGGGAAATGGAGTCCG

Gal4 binding sequence:

C-G-G-N(11)-C-C-G

TATA Box:

T-A-T-A-A-[AT](3)

IUPAC DNA Codes for DNA Patterns

Symbol	Meaning	Rationale
A	A	Adenine
B	C or G or T	Not-A (B follows A in alphabet)
C	C	Cytosine
D	A or G or T	Not-C (D follows C in alphabet)
G	G	Guanine
H	A or C or T	Not-G (H follows G in alphabet)
K	G or T	Keto
M	A or C	aMino
N	A or C or G or T	aNy
R	A or G	puRine
S	C or G	Strong interaction (3 H-bonds)
T	T	Thymine
V	A or C or G	Not-T (or Not-U)
W	A or T	Weak interaction (2 H-bonds)
Y	T or C	pYrimidine

Pattern of TF binding site consensus sequences

- Alignment of transcription factor binding sites

```
Motif

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
TCTATTTTGGAAA
CCAATTTTCAAAA
CCAAATTGGCAAA
```

Consensus: YCHAWTWNSNAWA or CCHAWTTNGNAWA

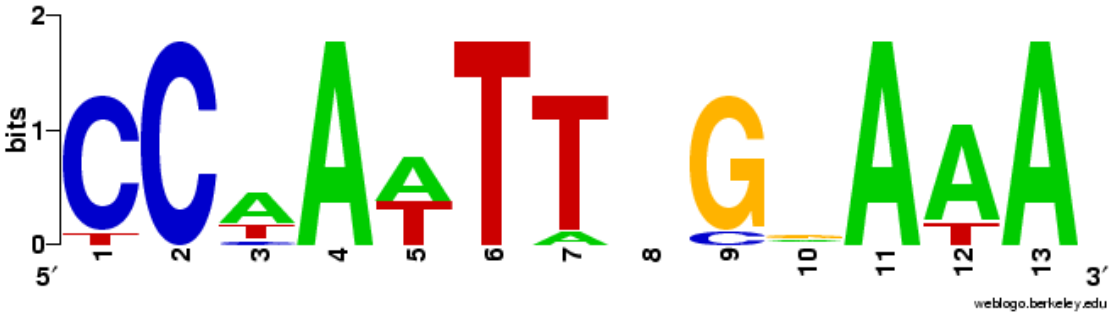


Image created by: http://weblogo.berkeley.edu/logo.cgi

Patterns: Summary

§ Advantages:

- Relatively straightforward to identify => exact pattern matching is fast
- Patterns are relatively intuitive to read and understand
- Databases with large numbers of protein and DNA sequence patterns are available (e.g., PROSITE, etc)

§ Disadvantages:

- Patterns are a qualitative description of motif => lose information about relative frequency of each residue or nucleotide at an ambiguous position, e.g. [GAC] versus 0.6 G, 0.28 A, and 0.12 C
- Can be difficult to write complex motifs using regular expression notation
- Cannot represent subtle sequence motifs

DNA Sequence Profiles

§ A profile is a position-specific scoring matrix that gives a quantitative description of a sequence motif

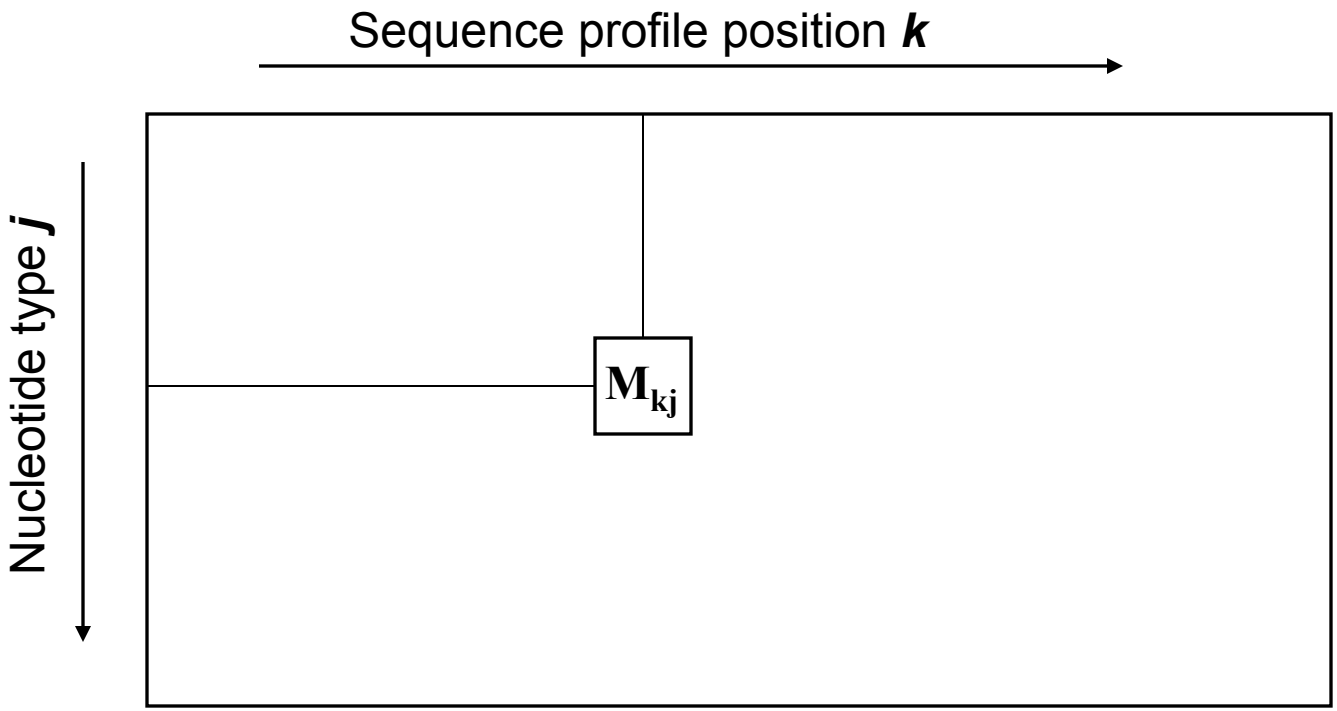
§ For DNA sequences, the profile scoring matrix has **N** columns and 4+ rows, **N** being the length of the profile (# of sequence positions)

§ The first 4 rows of each column specify the score (log odds ratio) for finding, at that position in the target sequence, each of the 4 nucleotides (A, C, G, T)

§ The rows after the first 4 rows contain penalties for insertions/deletions at that position in the target sequence

§ $\boldsymbol{M_{kj}}$ = score for the j^{th} nucleotide (or gap) at the k^{th} position in the sequence

DNA Profile: Position Specific Scoring Matrix



$$M_{kj} = \log\left(\frac{p_{kj}}{p_j}\right)$$

- p_{kj} = probability of nucleotide j at position k in the profile
- p_j = "background" probability of nucleotide j in genome sequence

Simple Method for Calculating DNA Sequence Profiles

Adapted from Hertz and Stormo, *Bioinformatics* **15**:563-577

§ Recall: $M_{kj} = \log_e\left(\frac{p_{kj}}{p_j}\right)$

§ As the number of aligned sequences grow (for large Z): $p_{kj} = \frac{C_{kj}}{Z}$

- C_{kj} = Number of j^{th} type nucleotide at position k
- Z = Total number of aligned sequences

§ For small numbers of aligned sequences, better to use the following method of calculating p_{kj}:

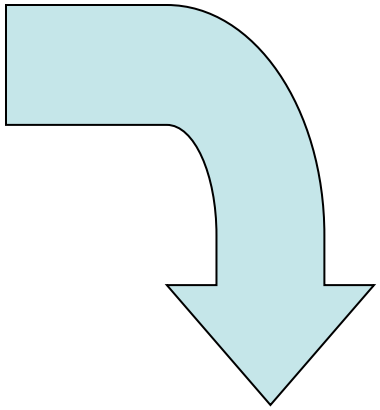
$$p_{kj} = \frac{C_{kj} + p_j}{Z + 1}$$

- Where p_j = background probability of that nucleotide type in the genome (based on GC content of genome)

Example of calculating a DNA sequence profile (PSSM)

Alignment of Transcription factor consensus binding sequence:

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA



Alignment Matrix:

Position k =	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0
Consensus:	**C**	**C**	**[ACT]**	**A**	**[AT]**	**T**	**T**	**N**	**G**	**N**	**A**	**[AT]**	**A**

Computing the DNA Sequence Profile (PSSM)

Alignment Matrix:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0

Recall: $$M_{kj} = \log_e\left(\frac{p_{kj}}{p_j}\right) = \log_e\left(\frac{(C_{kj} + p_j)/(Z+1)}{p_j}\right)$$

Profile matrix values for k = 1 (assume $p_j = 0.25$ for all nucleotides):

$$M_{1A} = \log_e\left(\frac{(C_{1A} + p_A)/(Z+1)}{p_A}\right) = \log_e\left(\frac{(0+0.25)/(10+1)}{0.25}\right) = -2.4$$

$$M_{1C} = \log_e\left(\frac{(C_{1C} + p_C)/(Z+1)}{p_C}\right) = \log_e\left(\frac{(9+0.25)/(10+1)}{0.25}\right) = 1.2$$

$$M_{1G} = \log_e\left(\frac{(C_{1G} + p_G)/(Z+1)}{p_G}\right) = \log_e\left(\frac{(0+0.25)/(10+1)}{0.25}\right) = -2.4$$

$$M_{1T} = \log_e\left(\frac{(C_{1T} + p_T)/(Z+1)}{p_T}\right) = \log_e\left(\frac{(1+0.25)/(10+1)}{0.25}\right) = -0.8$$

Computing the DNA Sequence Profile (PSSM)

Alignment Matrix:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0

$$M_{kj} = \log_e\left(\frac{p_{kj}}{p_j}\right) = \log_e\left(\frac{(C_{kj} + p_j)/(Z+1)}{p_j}\right)$$

DNA Profile Matrix (PSSM):

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	-2.4	-2.4	0.8	1.3	0.6	-2.4	-0.8	0.6	-2.4	0.2	1.3	1.1	1.3
C:	1.2	1.3	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	0.6	1.3	1.2	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

Scoring a Test Sequence using the DNA profile (PSSM)

Test Sequence (potential binding site):

CCTATTTAGGATA

DNA sequence profile (PSSM) for Transcription Factor binding site:

Position k =	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
A:	-2.4	-2.4	0.8	**1.3**	0.6	-2.4	-0.8	**0.6**	-2.4	0.2	**1.3**	1.1	**1.3**
C:	**1.2**	**1.3**	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	**1.2**	**0.6**	-2.4	-2.4	-2.4
T:	-0.8	-2.4	**0.2**	-2.4	**0.6**	**1.3**	**1.2**	-0.2	-2.4	-0.8	-2.4	**-0.2**	-2.4
Test seq:	**C**	**C**	**T**	**A**	**T**	**T**	**T**	**A**	**G**	**G**	**A**	**T**	**A**

Total Score for test sequence:

Score = 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2 + 0.6 + 1.2 + 0.6 + 1.3 + -0.2 + 1.3

Score = 11.9

- Does the Test Sequence match the DNA sequence profile?

Simple Test for a Match to the DNA sequence profile

Score of Test Sequence (**CCTATTTAGGATA**): **11.9**

Maximum possible score (**CCAATTTAGGAAA**):

Position k =	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
A:	-2.4	-2.4	**0.8**	**1.3**	0.6	-2.4	-0.8	**0.6**	-2.4	0.2	**1.3**	**1.1**	**1.3**
C:	**1.2**	**1.3**	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	**1.2**	**0.6**	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	**0.6**	**1.3**	**1.2**	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

Max Score: **C C A A T T T A G G A A A**

<u>Total Score for Best Matching Sequence:</u>

Max Score = 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2 + 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3

Max Score = 13.8

Simple Test for a Match to the DNA sequence profile

Score of Test Sequence (**CCTATTTAGGATA**): **11.9**

Maximum possible score (**CCAATTTAGGAAA**): **13.8**

- Example threshold: if the score of the test sequence is >60% of the Maximum Score, we will designate it a match

Score Threshold for Match = 60% x Max Score = 0.6 x 13.8 = 8.28

For Match:

Score of test sequence > Score threshold

11.9 > 8.28

Hence, test sequence (**CCTATTTAGGATA**) matches the DNA sequence profile

Test sequence is a potential binding site of Transcription Factor

Following method in Harbison et al. (2004) *Nature* 431:99-104