

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by



This page is available in the following languages:

Afrikaans বাংলা Català Dansk Deutsch Ελληνικά English (GB) English (US) Esperanto
 Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
 Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
 Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik cncioi srpski (latinica) Sotho svenska
 中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

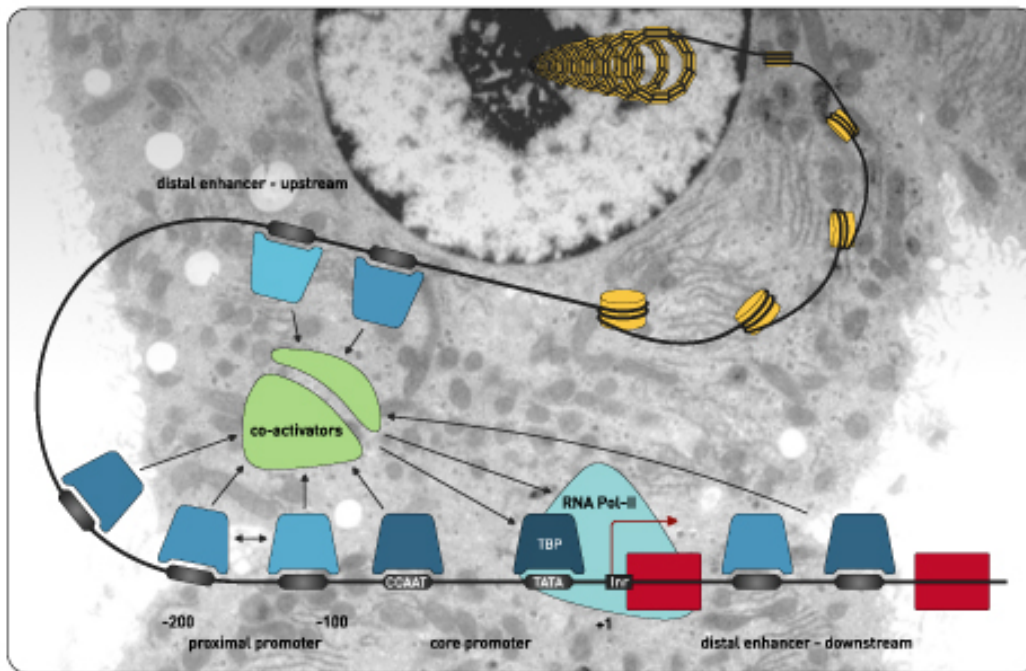
Contains material
 by Wyeth
 Wasserman,
 William Noble,
 Michael Hoffman,
 and Tim Bailey

Gene regulation and motif analysis

Michael M. Hoffman (@michaelhoffman)

Pathway and Network Analysis of -omics Data

July 27-29, 2020



<https://hoffmanlab.org/>

Learning Objectives

- By the end of this lecture, you will:
 - Understand challenges in predicting transcription factor (TF) binding
 - Be able to identify binding sites for known TFs
 - Be able to discover TF binding motifs in genomic regions like ChIP-seq peaks or promoters using iRegulon and Cytoscape

Creative Commons

This page is available in the following languages:

Afrikaans Afrikaans Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

[Learn how to distribute your work using this licence](#)

Overview

Part 1: Introduction to eukaryotic transcription

Part 2: Prediction of transcription factor binding sites

Part 3: Discovering novel motifs enriched in regulatory regions

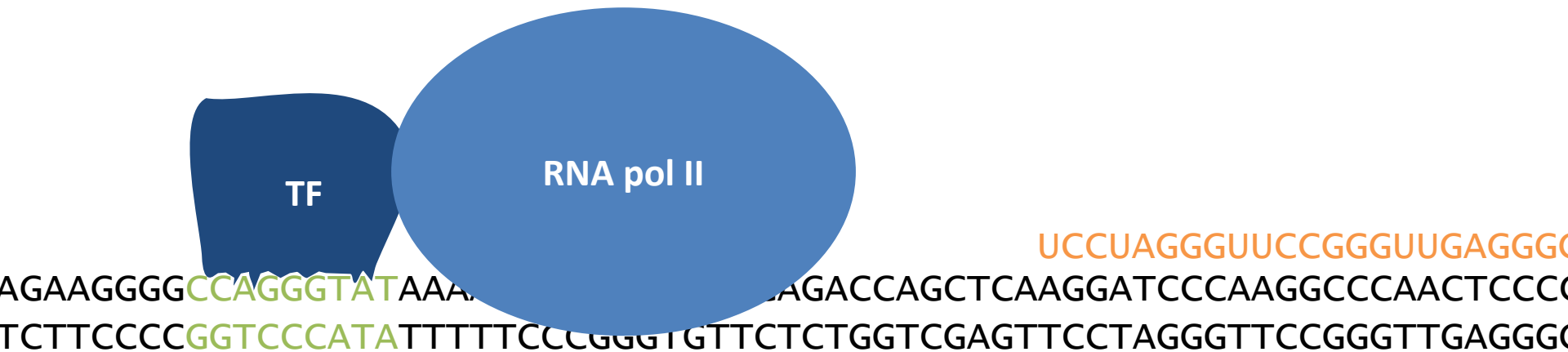
Part 4: Effectiveness of position weight matrix models

Part 5: Incorporating information about the biochemistry of gene regulation

Part 1
Introduction to
eukaryotic transcription

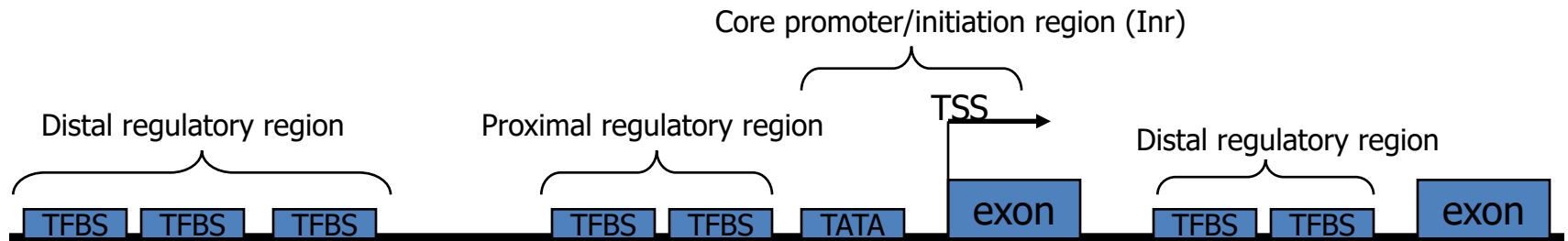
Transcription over-simplified

1. **TF** binds to DNA at **TF binding site**
2. **TF** recruits **RNA polymerase II**
3. **RNA polymerase II** produces **RNA**



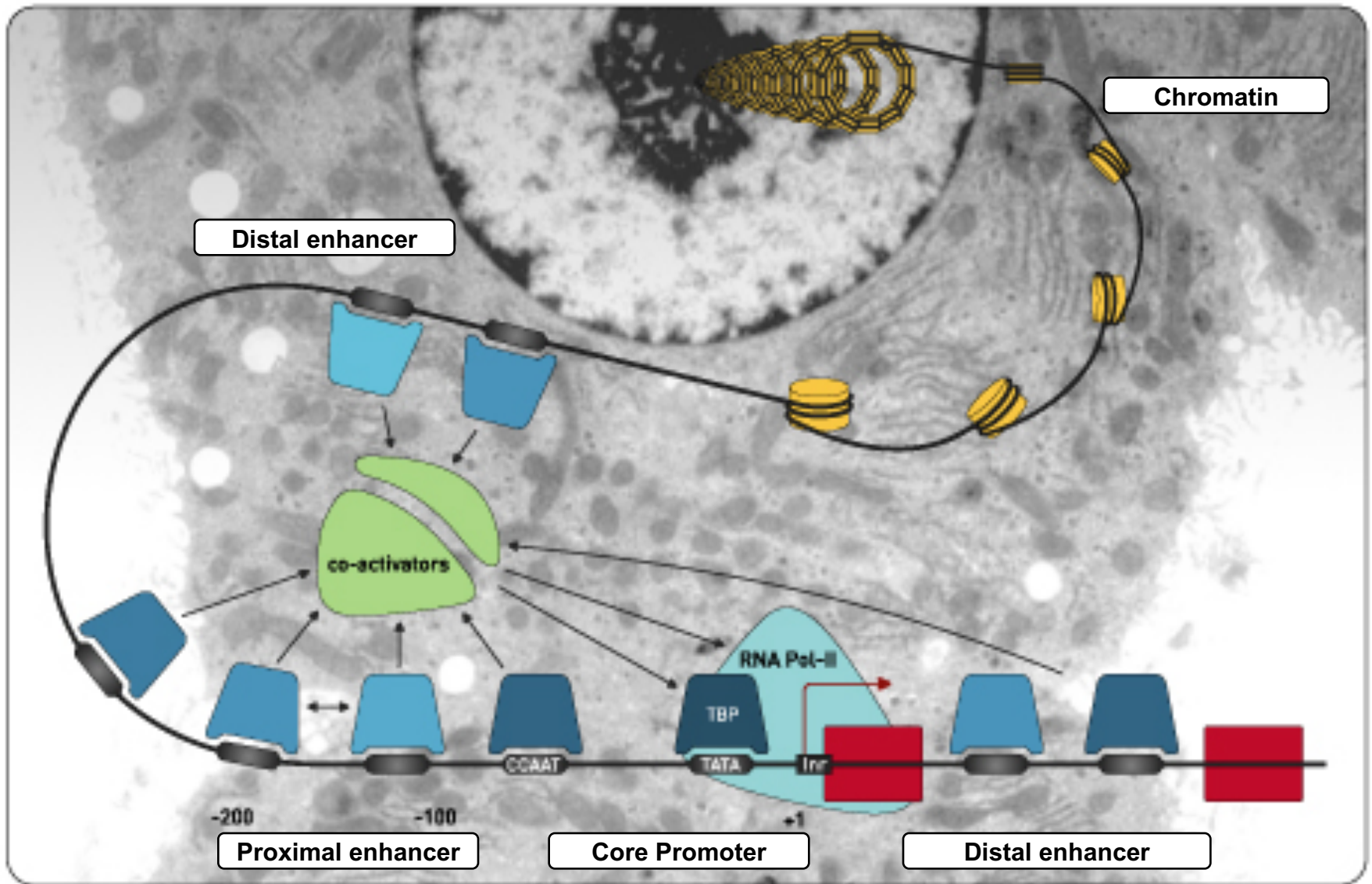
Anatomy of transcriptional regulation

WARNING: Terms vary widely in meaning between scientists

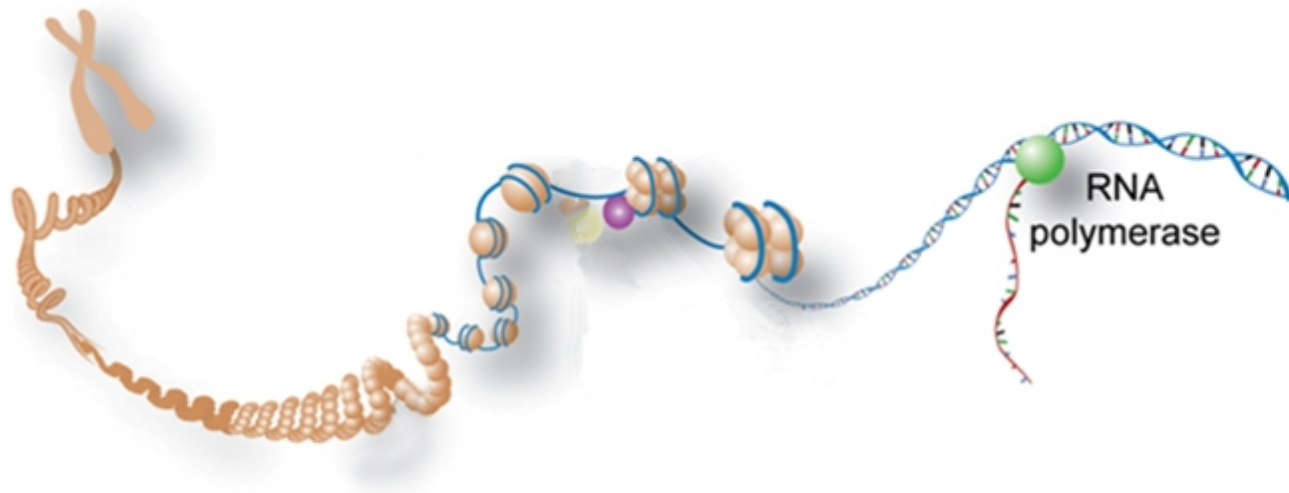


- Core promoter – Sufficient for initiation of transcription; orientation dependent
 - TSS – transcription start site
 - Often really a transcription start *region*
- TFBS – single transcription factor binding site
- Regulatory regions
 - Proximal/distal – vague reference to distance from TSS
 - May be positive (enhancing) or negative (repressing)
 - Orientation independent (generally)
 - Modules – Sets of TFBS within a region that function together
- Transcriptional unit
 - DNA sequence transcribed as a single polycistronic mRNA

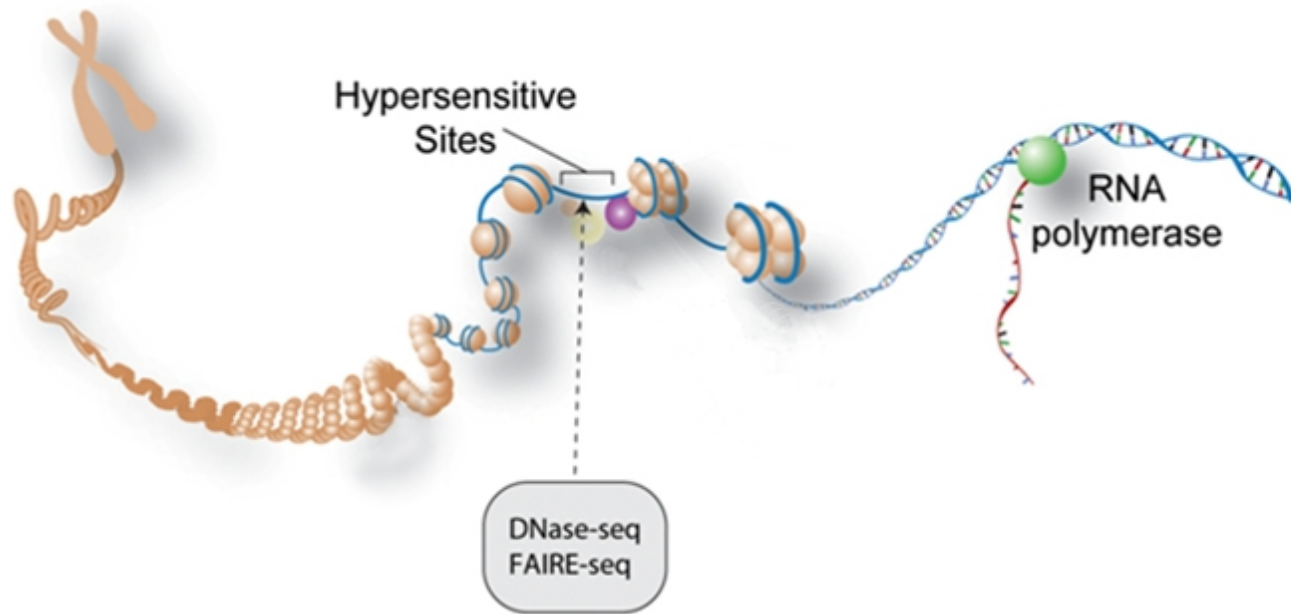
Complexity in transcription



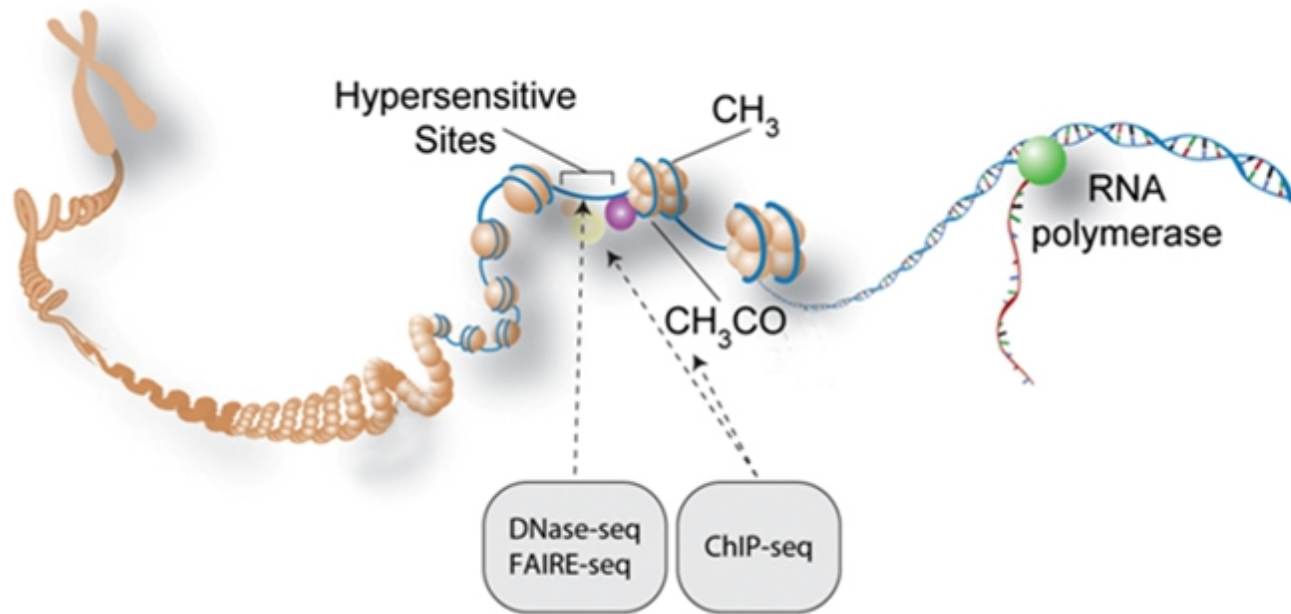
Functional genomics



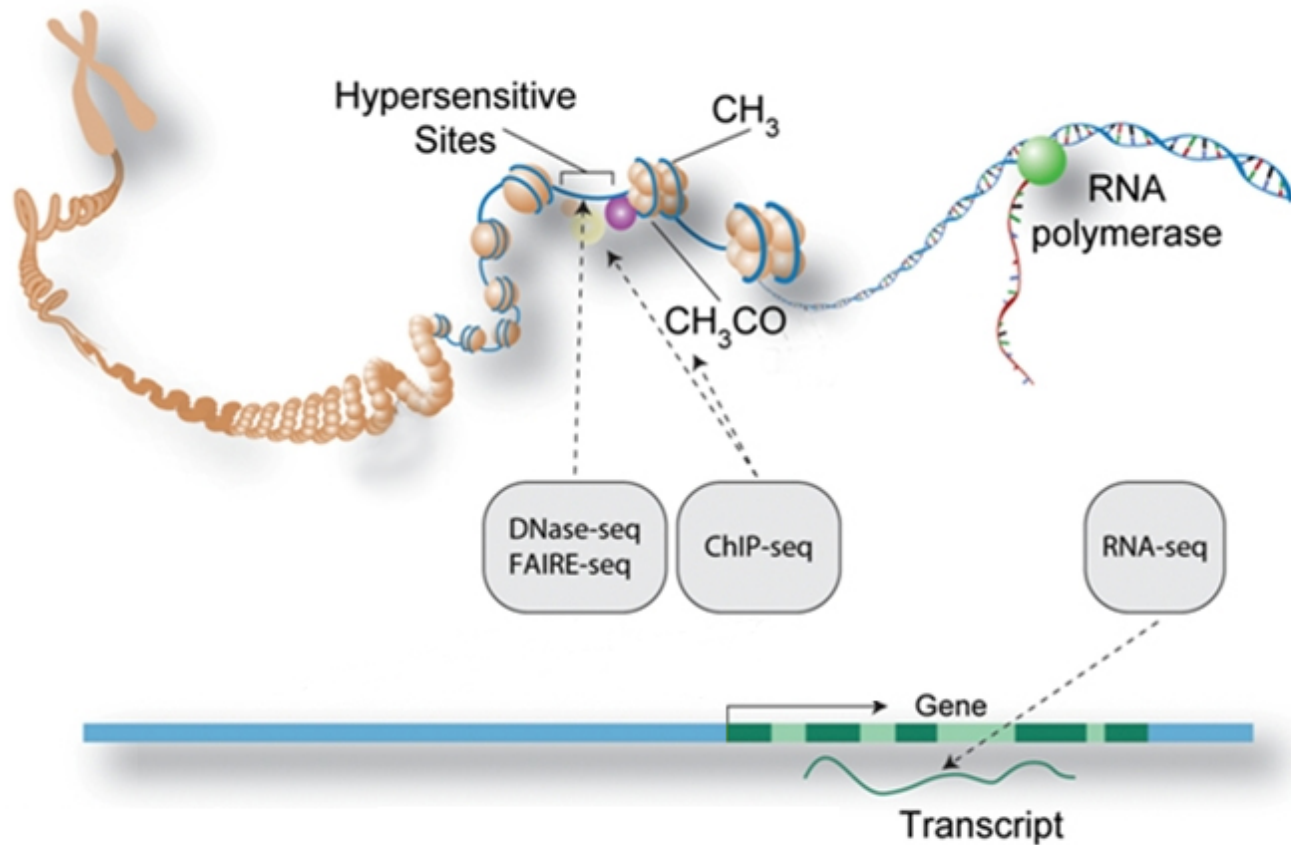
Functional genomics



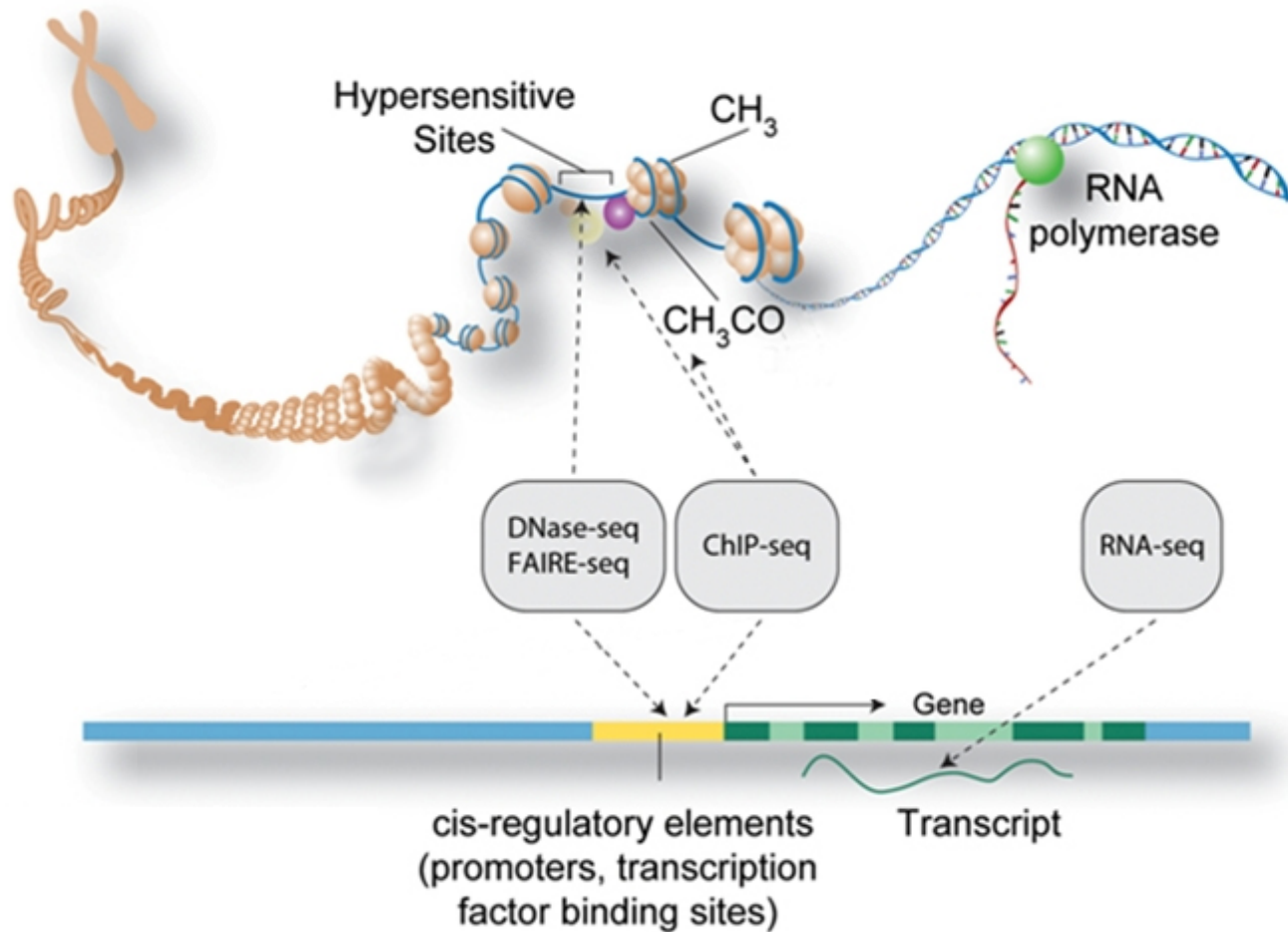
Functional genomics



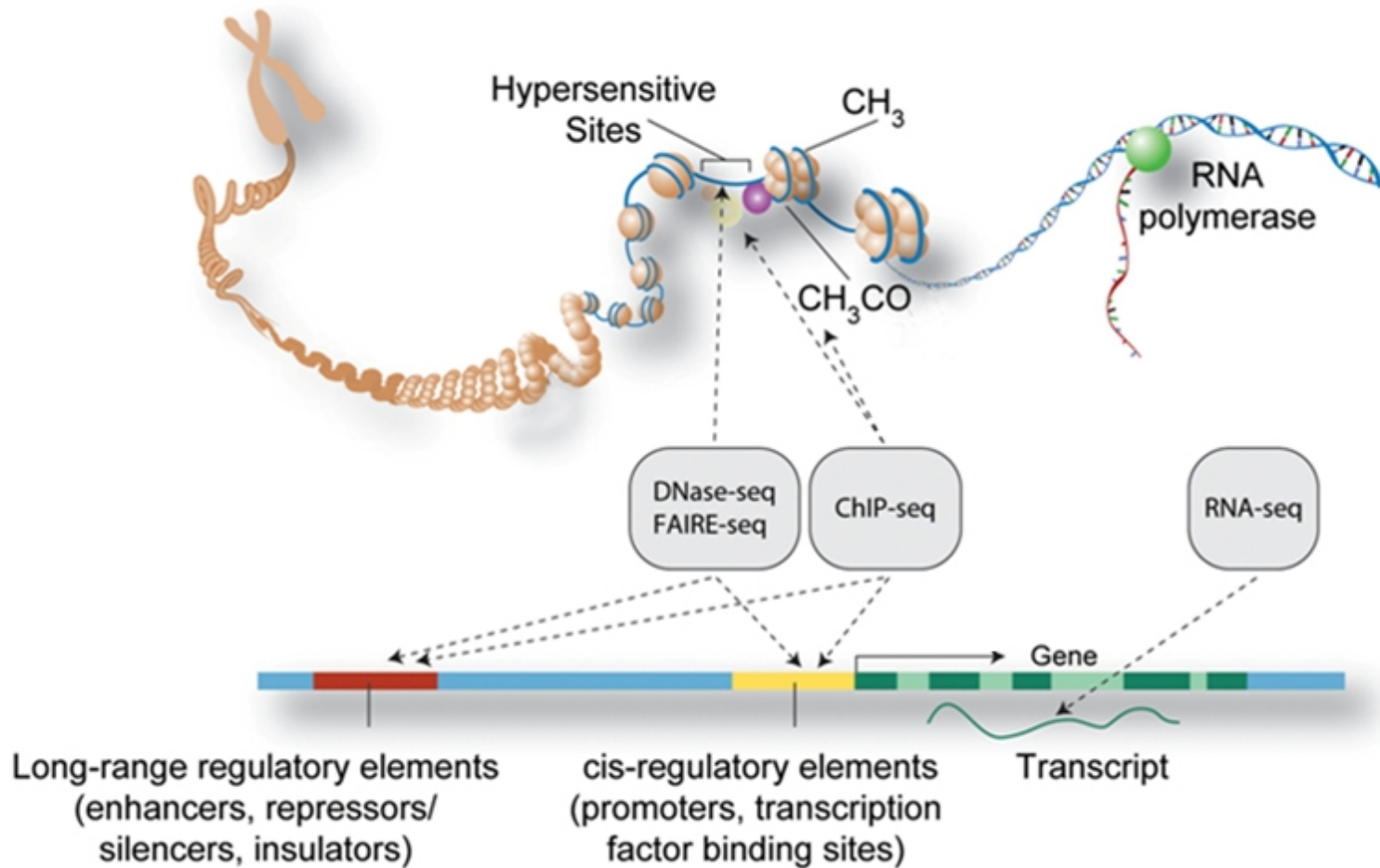
Functional genomics



Functional genomics

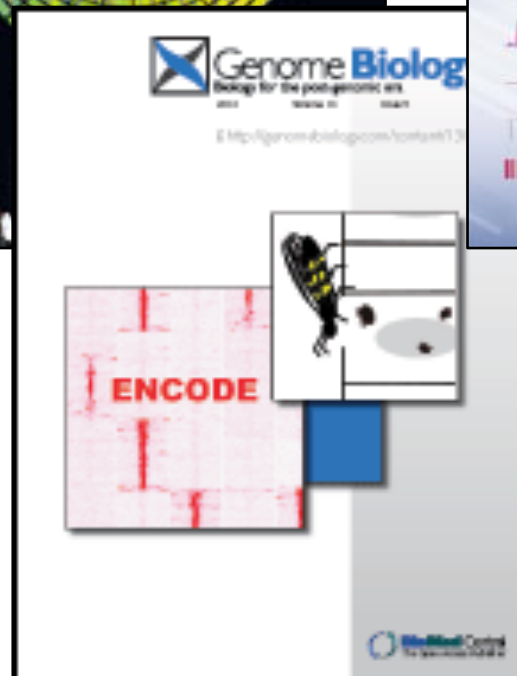


Functional genomics



ENCODE

The Encyclopedia of DNA Elements



Accessing regulatory data

- ENCODE Project
 - <http://encodeproject.org/>
- UCSC Genome Browser
 - <http://genome.ucsc.edu/>
- Ensembl
 - <http://ensembl.org/>
- Gene Expression Omnibus (GEO)
 - <http://www.ncbi.nlm.nih.gov/geo/>

Part 2

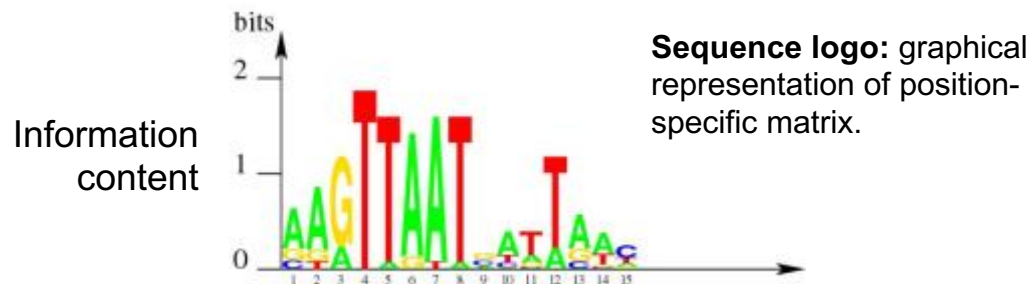
Prediction of TF binding sites

**Teaching a computer
to find transcription factor
binding sites**

Representing binding sites for a TF

- Single site
 - AAGTTAATGATTAAC
- Set of sites, represented as a consensus
 - VDRTWRWWSHDWVDH (IUPAC degenerate DNA)
- Set of sites, represented as a position frequency matrix (PFM)

A	14	16	4	0	1	19	20	1	4	13	4	4	13	12	3
C	3	0	0	0	0	0	0	0	7	3	1	0	3	1	12
G	4	3	17	0	0	2	0	0	9	1	3	0	5	2	2
T	0	2	0	21	20	0	1	20	1	4	13	17	0	6	4

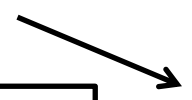


Set of binding sites

AAGTTAATGATTAAC
 CAGTTAATAAATAAC
 GAGTTAAACACTAAA
 CAGTTAATTAGTAAC
 GAGTTAATAAATAAC
 CAGTTATTCAGTAAC
 GAGTTAATAAATCAT
 CAGTTAATCAGTAAT
 AGATTAAAGAATAAT
 AAGTTAACGATTAAC
 AGGTTAACGATACAC
 ATGTTGATGATAAAC
 AAGTTAATGATAAAT
 AAGTTAACGATAAAC
 AAATTAATGATTCAC
 GAGTTAATGATTAAC
 AAGTTAATCATTGAC
 AAGTTGATGATTAAG
 AAATTAATGATTGAC
 ATGTTAATGATTAAC
 AAGTAAATGATTAAC
 AAGTTAATGATTGCC
 AAGTTAATGATTGAC
 AAATTAATGATTGAC
 AAGTTAATGATTAGG
 AAGTTAATGATTAAT
 AAGTTAATGATTAGC
 AAGTTAATGATTAAT

Position frequency matrix (PFM) → position weight matrix (PWM)

PFM f



base b

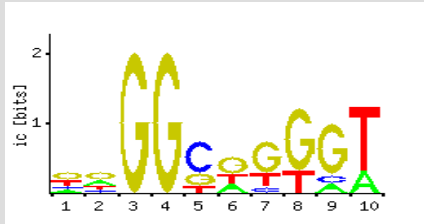
A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

column i

$f(b, i)$

Detecting binding sites in a single sequence

Sp1



Raw scores



ACCCTCCCCAGGGGCGGGGGCGGTGGCCAGGACGGTAGCTCC

A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

Abs_score = 13.4 (sum of column scores)

Relative scores

A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

Max_score = 15.2 (sum of highest column scores)

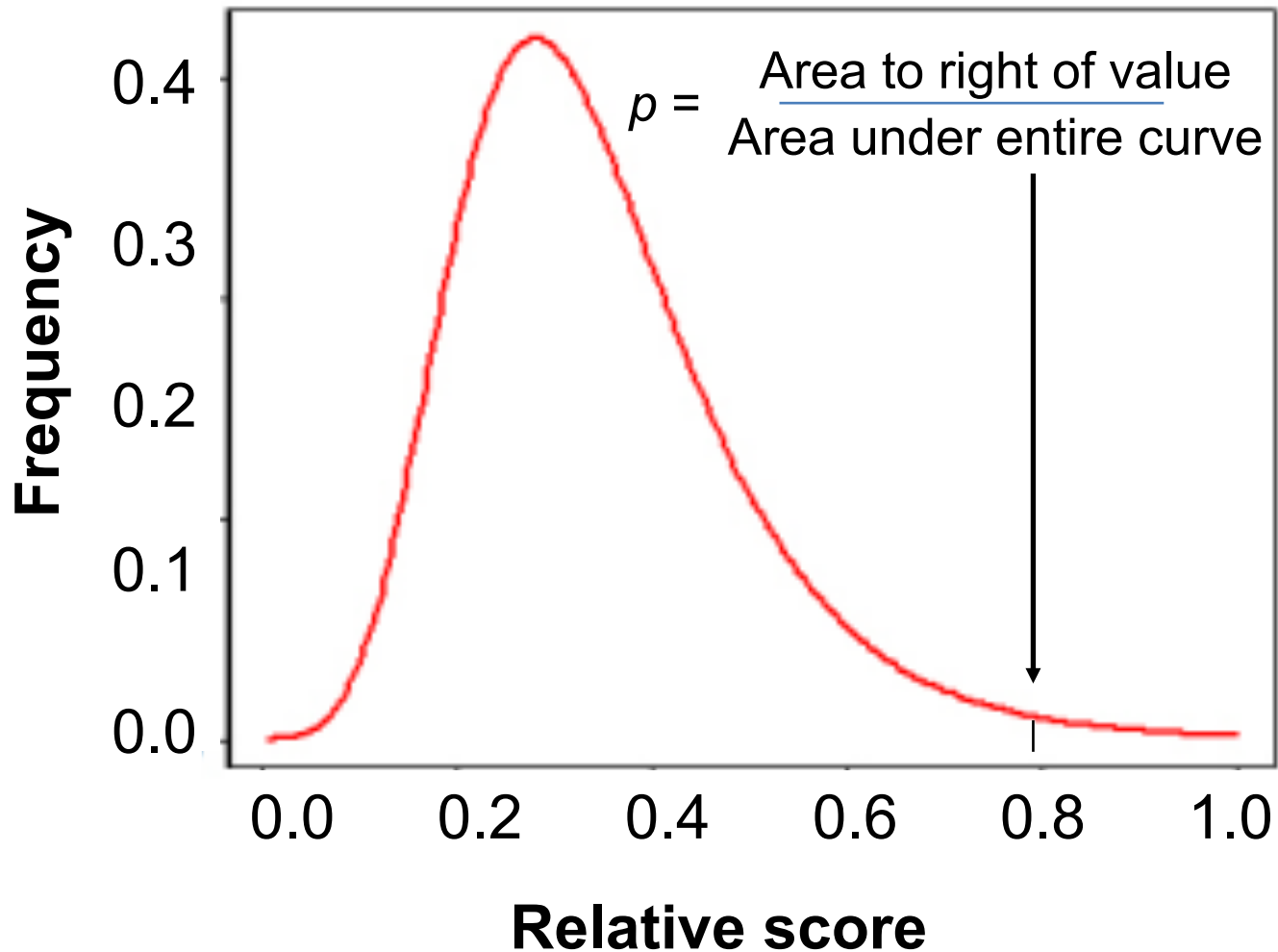
A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

Min_score = -10.3 (sum of lowest column scores)

$$\text{Rel_score} = \frac{\text{Abs_score} - \text{Min_score}}{\text{Max_score} - \text{Min_score}} \cdot 100\%$$

$$= \frac{13.4 - (-10.3)}{15.2 - (-10.3)} \cdot 100\% = \mathbf{93\%}$$

Empirical p-value score



The background of the slide is a 10x10 grid of 100 small plots, each representing a transcription factor (TF) binding profile. Each plot shows the relative frequency of nucleotides (A, C, G, T) at each position of a binding site, with the y-axis representing information content or probability. The profiles are diverse, showing various consensus sequences and binding site architectures.

JASPAR: **An open-access database** **of TF binding profiles**

<http://jaspar.genereg.net>

New Release Coming for 2018

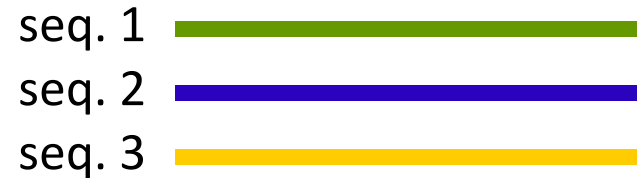
with Entirely New Interface

Part 3:

***De novo* discovery of transcription factor binding sites**

Motif discovery problem

- Given sequences

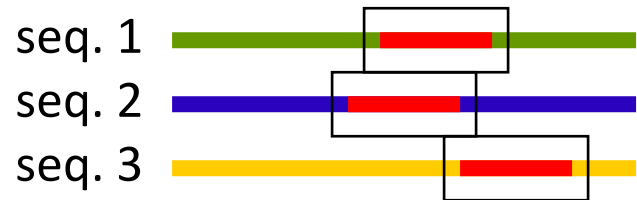


- Find motif

IGRGGFGEVY at position 515

LGEGCFGQVV at position 430

VGSGGFGQVY at position 682



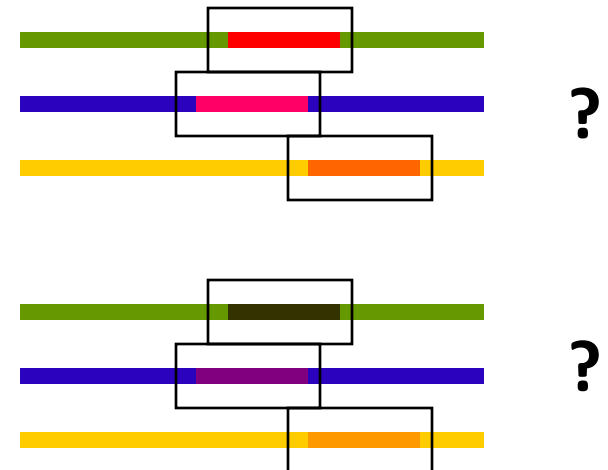
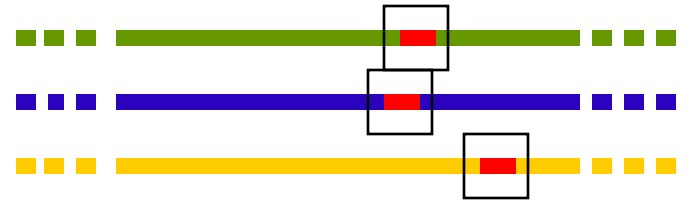
Motif discovery problem

- Given:
 - a sequence or family of sequences.
- Find:
 - the number of motifs
 - the width of each motif
 - the locations of motif occurrences



Why is this hard?

- Input sequences are long (thousands or millions of residues).
- Motif may be *subtle*
 - Instances are short.
 - Instances are only slightly similar.



TFBS motif discovery example

We are given a set of promoters from co-regulated genes.

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT
ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG
CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT
TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC
ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA
ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA
GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA

...*HIS7*
...*ARO4*
...*ILV6*
...*THR4*
...*ARO1*
...*HOM2*
...*PRO3*

TFBS motif discovery example

An unknown transcription factor binds to positions unknown to us, on either DNA strand.



TFBS motif discovery example

DNA binding motif of the transcription factor can be described by a position weight matrix (PWM).



5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAG	→	AAAAGAGTCA	←	GACATCGAAACATACAT	...	HIS7
5' - ATGGCAGAATCACTTTAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG	→	AAATGACTCA	←	ACG	...	AR04
5' - CACATCCAACGAATCACCTACCGTTATCGT	→	TGACTCACTT	←	TCTTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT	...	ILV6
5' - TGCGAAC	→	AAAAGAGTCA	←	ATTACAACGAGGAAATAGAAGAAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC	...	THR4
5' - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATAT	→	TGACTCATCC	←	GAAACATGAAA	...	AR01
5' - ATTGATT	→	TGACTCATT	←	TTTCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA	...	HOM2
5' - GGCGCCACAGTCCGCGTTTGGTTATCCGGC	→	TGACTCATTCTGACTCTTTT	←	TTGGAAAGTGTGGCATGTGCTTCACACA	...	PR03

TFBS motif discovery example

Sequence motif discovery problem is to discover the sites (or the motif) given just the sequences.

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT
5' - ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG
5' - CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT
5' - TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC
5' - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA
5' - ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA
5' - GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA

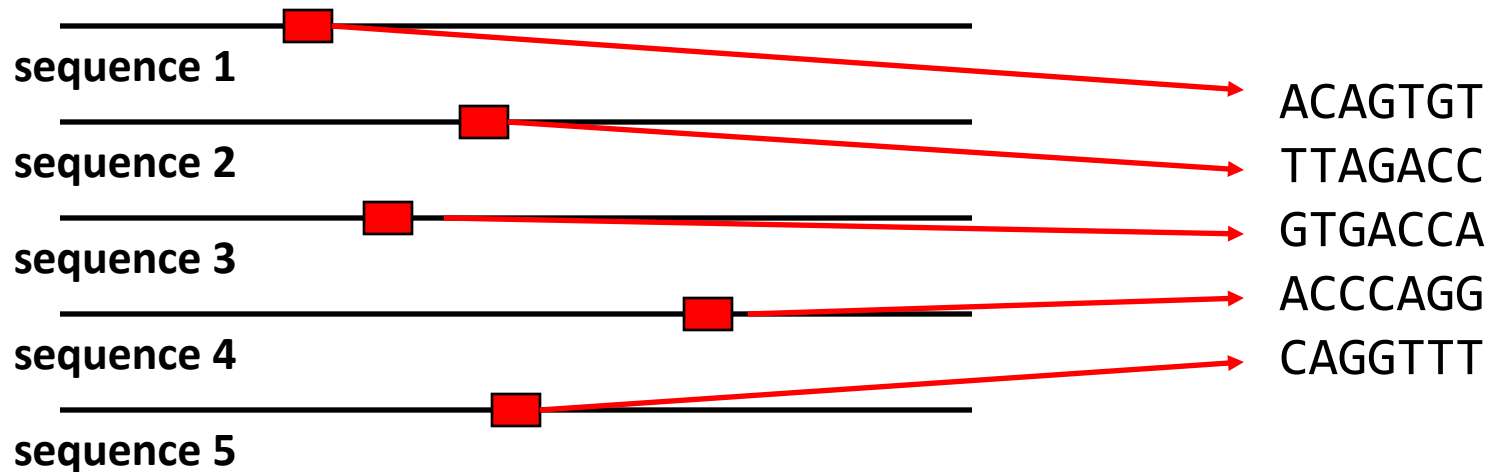
...HIS7
...AR04
...ILV6
...THR4
...AR01
...HOM2
...PRO3

Alternating approach

1. Guess an initial weight matrix
2. Use weight matrix to predict instances in the input sequences
3. Use instances to predict a weight matrix
4. Repeat 2 & 3 until satisfied.

Gibbs Sampler: 1. Initialization

- Randomly guess an instance s_i from each of t input sequences $\{S_1, \dots, S_t\}$.



Gibbs Sampler: 2a. Define PWM

ACAGTGT
TAGGCGT
ACACCGT
???????
CAGGTTT



A
C
G
T

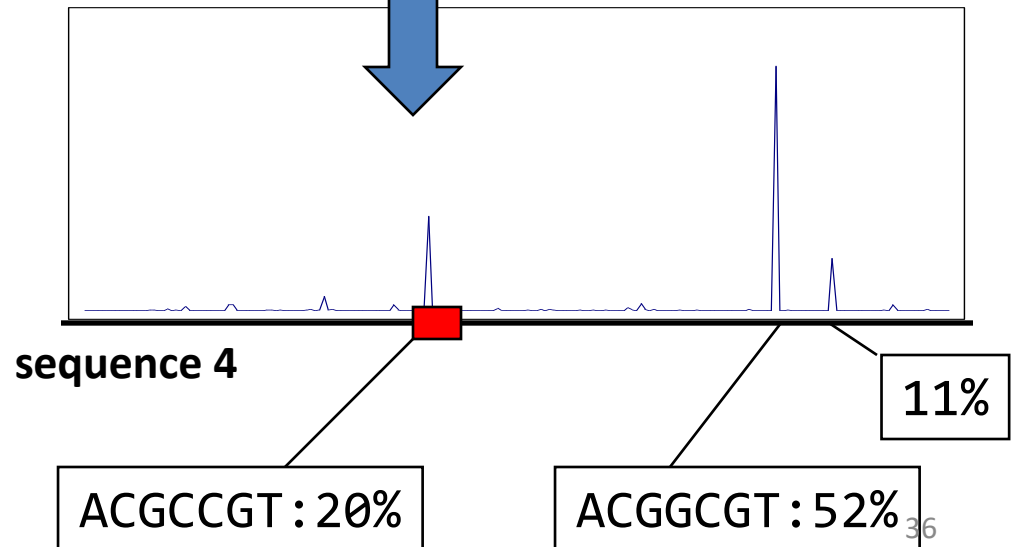
.45	.45	.45	.05	.05	.05	.05
.25	.45	.05	.25	.45	.05	.05
.05	.05	.45	.65	.05	.65	.05
.25	.05	.05	.05	.45	.25	.85

Gibbs Sampler: 2b. Predict instances

ACAGTGT
TAGGCGT
ACACCGT
???????
CAGGTTT



A	.45	.45	.45	.05	.05	.05	.05
C	.25	.45	.05	.25	.45	.05	.05
G	.05	.05	.45	.65	.05	.65	.05
T	.25	.05	.05	.05	.45	.25	.85



Gibbs Sampler: 3. Pick new instance

ACAGTGT
TAGGCGT
ACACCGT
???????
CAGGTTT



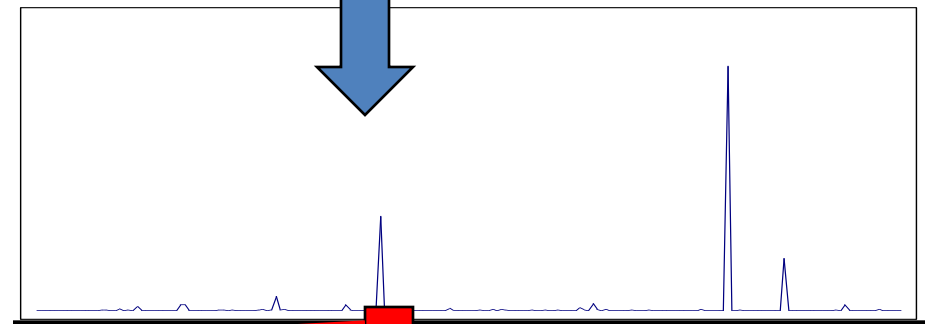
A	.45	.45	.45	.05	.05	.05	.05
C	.25	.45	.05	.25	.45	.05	.05
G	.05	.05	.45	.65	.05	.65	.05
T	.25	.05	.05	.05	.45	.25	.85



ACAGTGT
TAGGCGT
ACACCGT
ACGCCGT
CAGGTTT



sequence 4



11%

ACGCCGT : 20%

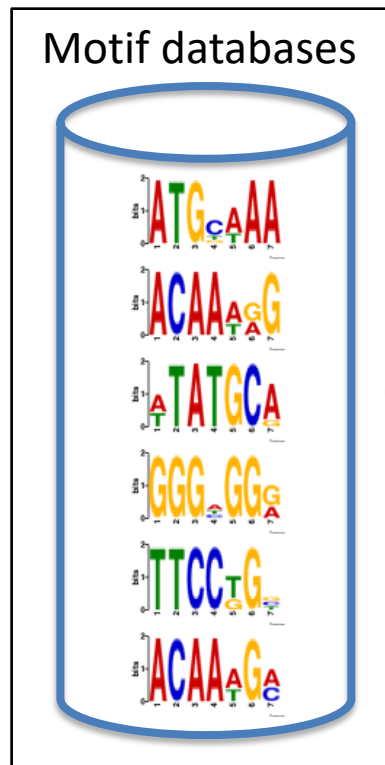
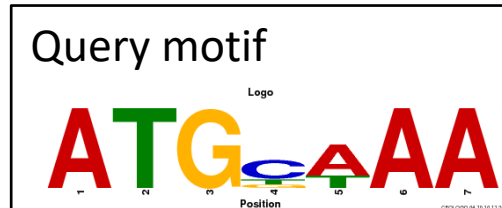
ACGGCGT : 52%

Gibbs sampler

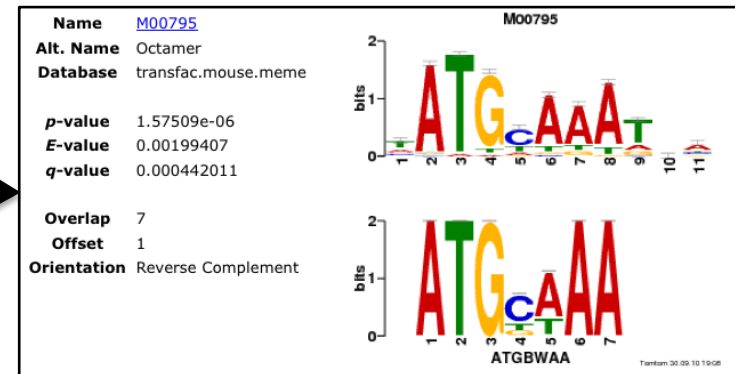
- Initially: randomly guess an instance s_i from each of t input sequences $\{S_1, \dots, S_t\}$.
- Steps 2 & 3 (search):
 - Throw away an instance s_i : remaining $(t - 1)$ instances define weight matrix.
 - Weight matrix defines instance probability at each position of input string S_i
 - Pick new s_i according to probability distribution
- Return highest-scoring motif seen

TOMTOM:

predict which proteins may bind a DNA motif



Alignment to
matching motif



- TOMTOM compares the query motif against all motifs in databases of known motifs (such as JASPAR).
- TOMTOM reports all statistically significant matches.

Part 4

Effectiveness of the position weight matrix model

The Good...

- Tronche (1997) tested 50 predicted HNF1 TFBS using an in vitro binding test and found that 96% of the predicted sites were bound!
- Stormo and Fields (1998) found in detailed biochemical studies that the best weight matrices produce scores highly correlated with in vitro binding energy



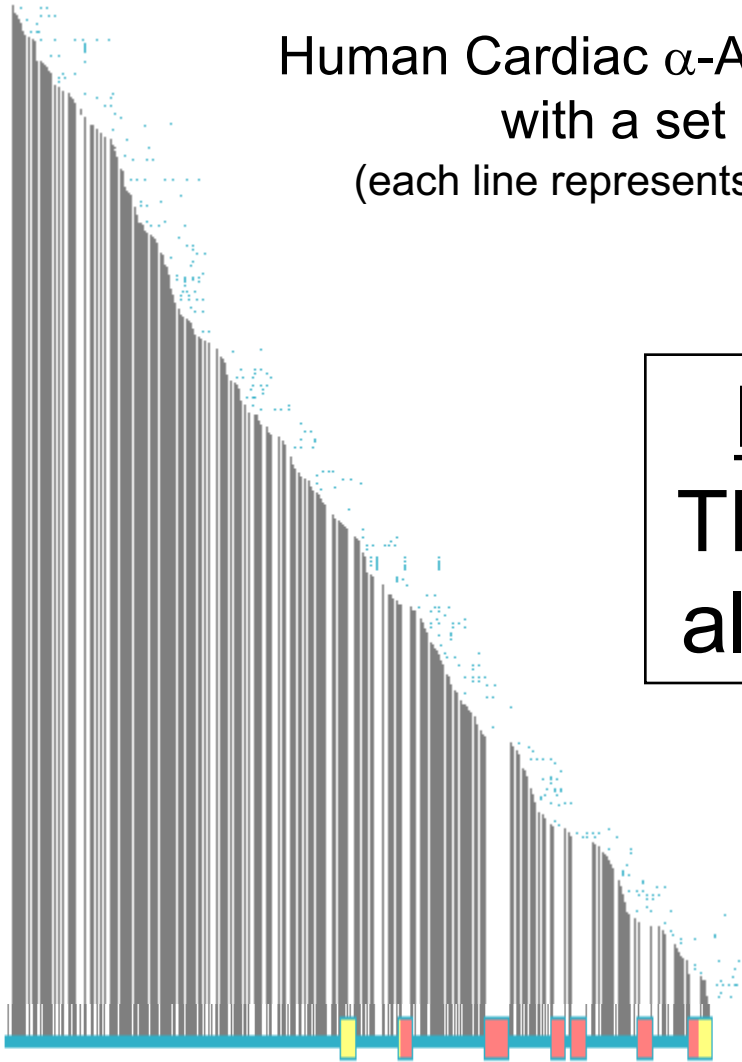
...the Bad...

- Fickett (1995) found that a profile for the MyoD TF made predictions at a rate of 1 per ~500 bp of human DNA sequence
 - This corresponds to an average of 20 sites / gene (assuming 10,000 bp as average gene size)

...and the Ugly!

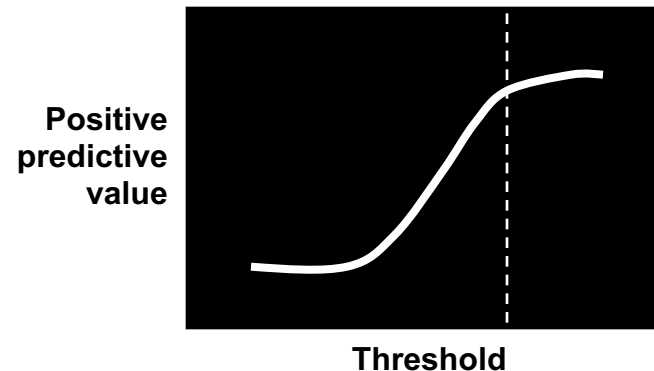
Human Cardiac α -Actin gene analyzed
with a set of profiles
(each line represents a TFBS prediction)

Futility conjecture:
TFBS predictions are
almost always wrong



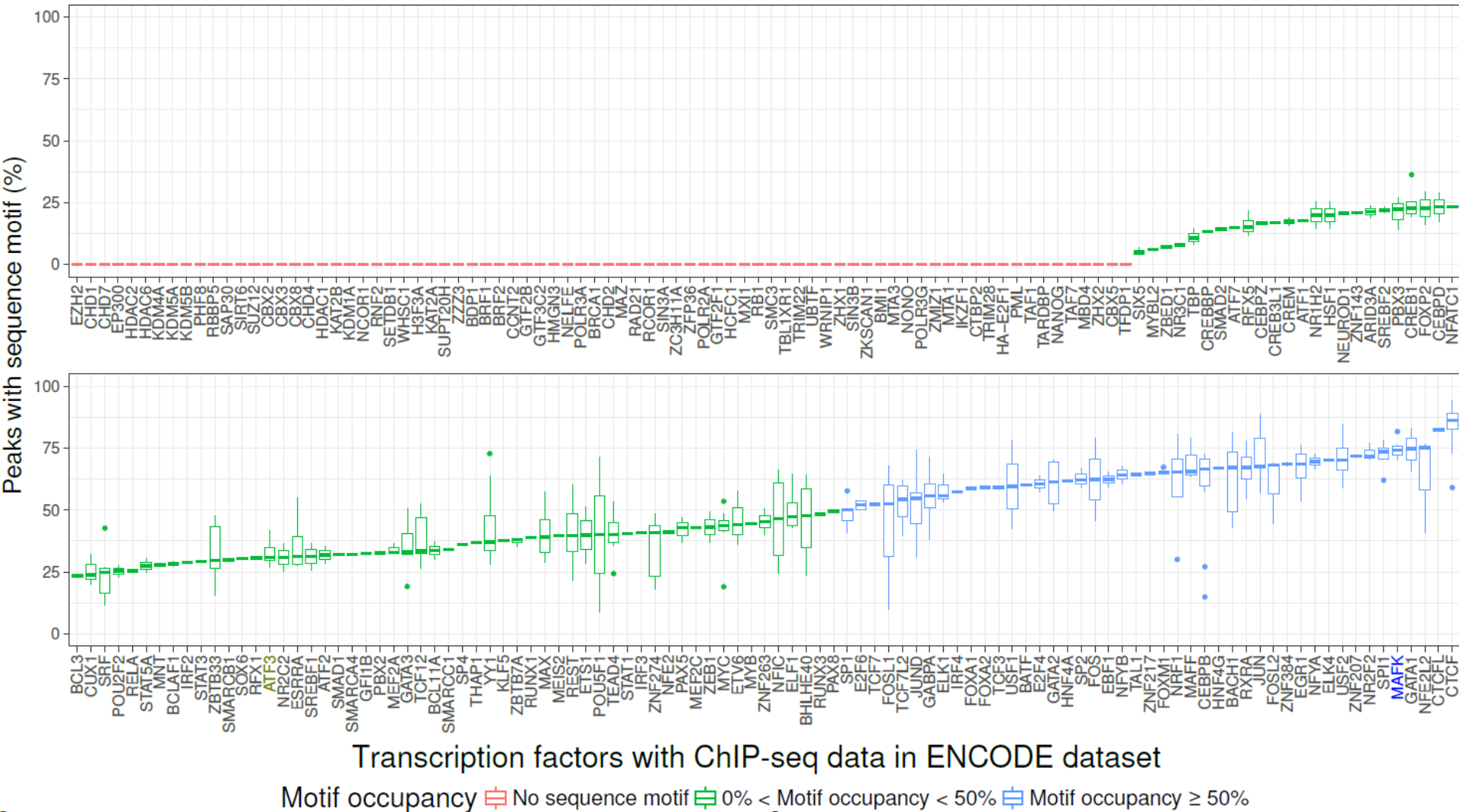
Red boxes are protein coding exons -
TFBS predictions excluded in this analysis

More stringency doesn't help

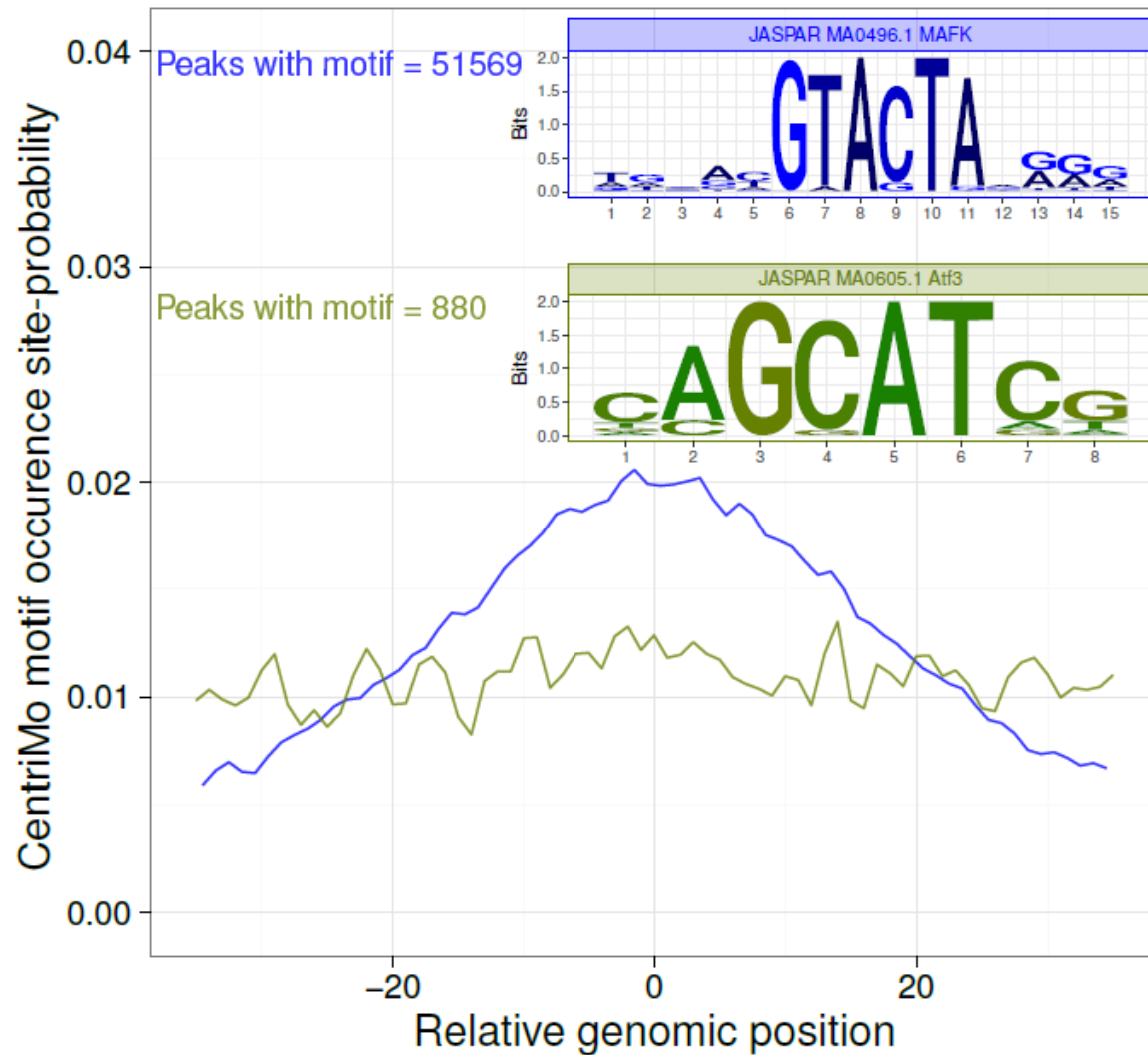


- Counter to intuition, the ratio of true positives to predictions fails to improve for “stringent” thresholds
 - For most predictive models this ratio would increase
- Why?
 - True binding sites are defined by properties not incorporated into the profile scores - above some threshold all sites *could* be bound if present in the right setting

It's even worse than we imagined



Please make it stop



Part 4

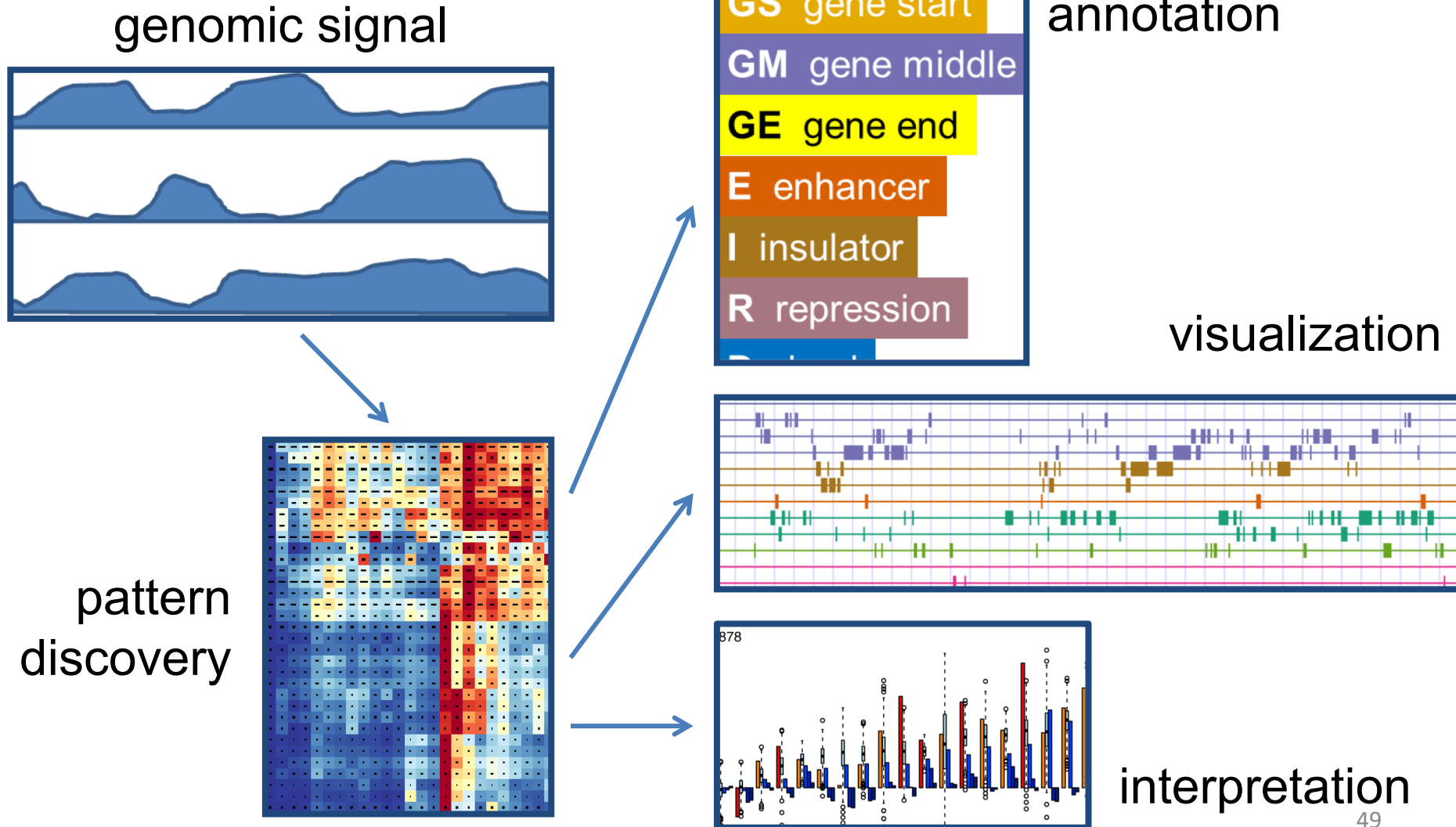
What have we learned?

- PWMs can accurately reflect *in vitro* binding properties of DNA-binding proteins
- Suitable binding sites occur at a rate far too frequent to reflect *in vivo* function
- *In vivo* presence of a DNA-binding protein often occurs without a strong motif
- Bioinformatics methods that use PWMs for binding site studies must incorporate additional information to enhance specificity
 - Unfiltered predictions are too noisy for most applications
 - Organisms with short regulatory sequences are less problematic (such as yeast and *E. coli*)

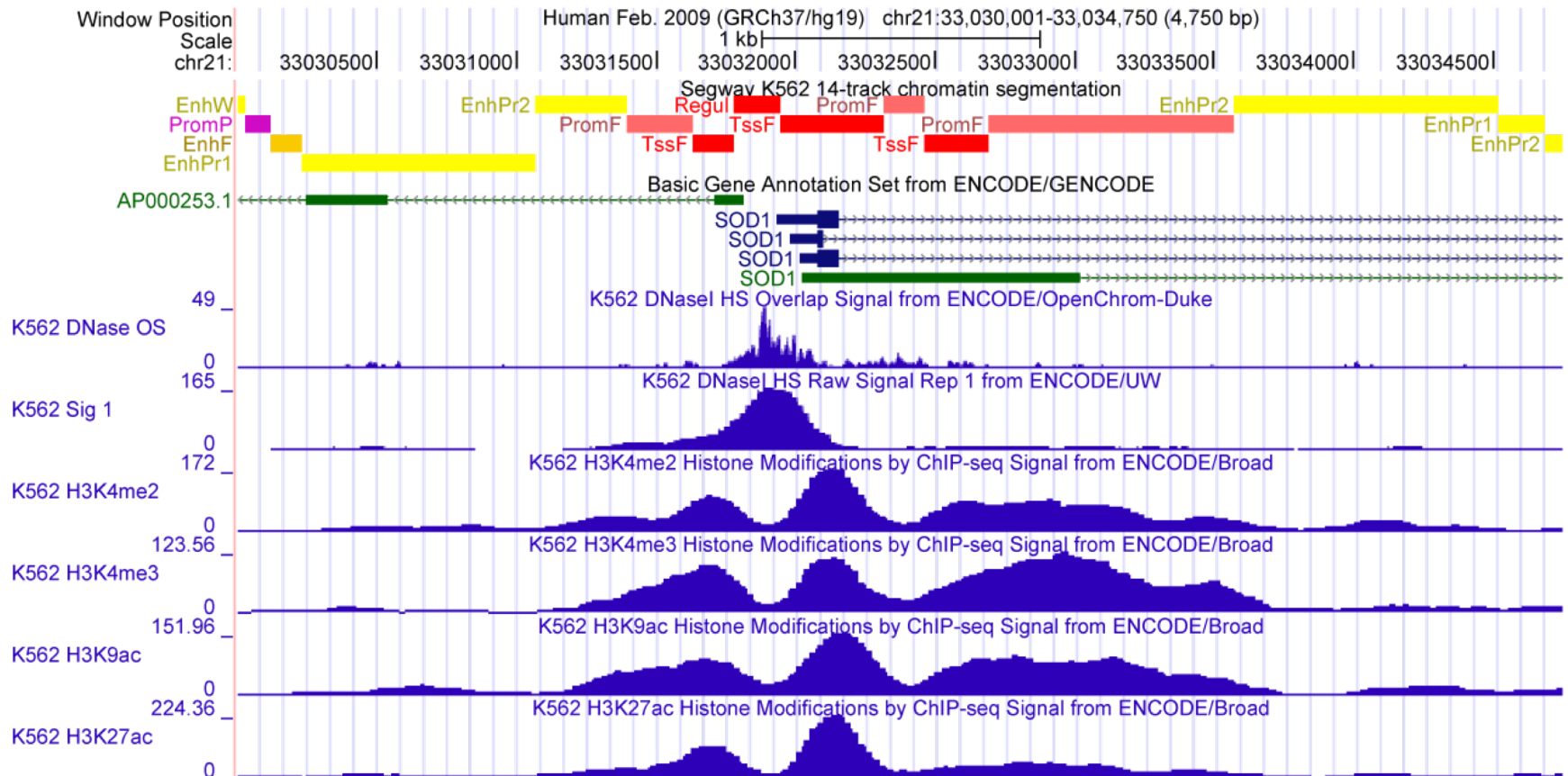
Part 5

Incorporating information about the biochemistry of gene regulation

Segway: semi-automated genome annotation



Transcription start site (TSS)



Segway semi-automated genomic annotation

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes J, Noble WS. 2012. [Unsupervised pattern discovery in human chromatin structure through genomic segmentation](#). *Nat Methods* 9:473–476. doi:10.1038/nmeth.1937. PubMed Central (free version): [PMC3340533](#) ([BibTeX](#))

Hoffman MM*, Ernst J*, Steven WP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, Hardison RC, Dunham I, Kellis M, Noble WS. 2012. [Integrative annotation of chromatin elements from ENCODE data](#). *Nucleic Acids Res* 41:827–841 doi: ([BibTeX](#))

The free Segway software package contains a novel method for analyzing multiple tracks of functional genomics data. Our method uses a dynamic Bayesian network (DBN) model, which enables it to analyze the entire genome at 1-bp resolution even in the face of heterogeneous patterns of missing data. This method is the first application of DBN techniques to genome-scale data and the first genomic segmentation method designed for use with the maximum resolution data available from ChIP-seq experiments without downsampling. Segway uses the [Graphical Models Toolkit \(GMTK\)](#) for efficient DBN inference. Our software has extensive documentation and was designed from the outset with external users in mind.

Segmentations

Human chromatin structure

There are two published segmentations of human chromatin structure available.

1. The regulatory segmentation from the [Ensembl Regulatory Build](#) viewable in [Ensembl](#)
2. The segmentation from our *Nature Methods* paper, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," viewable in the [UCSC Genome Browser](#)

Ensembl

The segmentation can be displayed by clicking the "Configure this page" option on the left navigation bar. The segmentations for each cell line can be selected under "Regulatory Features" and under the heading of "Enable/disable all Segmentation features". As an example you can try viewing the segmentations for [BRCA2](#) in [hg38](#).

For more details and instructions see the description of [Regulatory Segmentation](#).

UCSC Genome Browser

The [Ensembl Regulatory Build for GRCh38 \(hg38\)](#) can be viewed [here](#). It can also be loaded through the [Track Data Hub](#) interface. You can connect "Ensembl Regulatory Build" listed in the Public Hubs directory. After loading the track hub, you can show the "Cell Type Segmentations" supertrack which contains a Segway track for each of 18 cell types.

For older assemblies you can load, they can be browsed below:



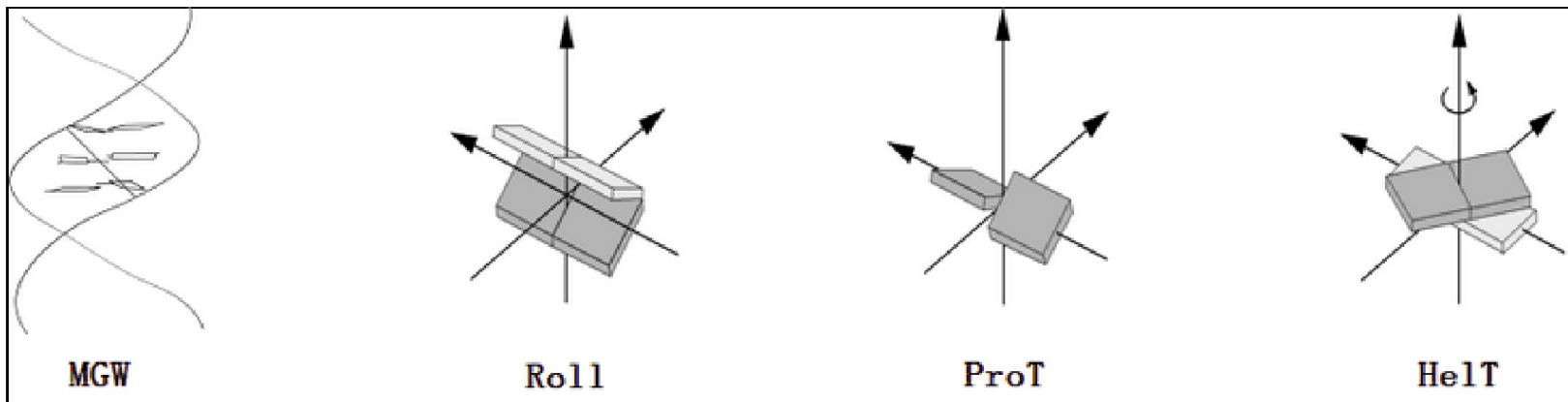
**GCAIACAGCAI CAI CAGAIACGACTACAGCA
TACATAGATATCAGCATAACAGCAGACTCATG
ACATCAGACAGCAGCGACGCAGACTCTCTC
ATCATAACATCAGACAGCAGCATAACCCACCA
AACGATAGA**CONTEXT**CATACTACTCATAGA
ACACACCATACTACGACTACAGACTCAGAC
CAAAGGGGTCCGCTCGACGCGCCTACTGCA
GCATCTCGGATCGCATCA**MATTERS**CGCAG
CTTCATCTCAGCGCAGCAGGGCCATTAGCG
AGCTACTCGAGCGATCAGCGACTCTCAGCG
ATCTACCGGGCTATTACAGAGCAGCTTACGC**

DNA shape features at Transcription Factor Binding Sites

Using data from JASPAR2014, the Rohs' lab developed the TFBSshape database storing DNA shape features of TFBSs.

Considered DNA shape features are :

- ▶ Minor Groove Width (MGW)
- ▶ Roll
- ▶ Propeller Twist (ProT)
- ▶ Helix Twist (HelT)



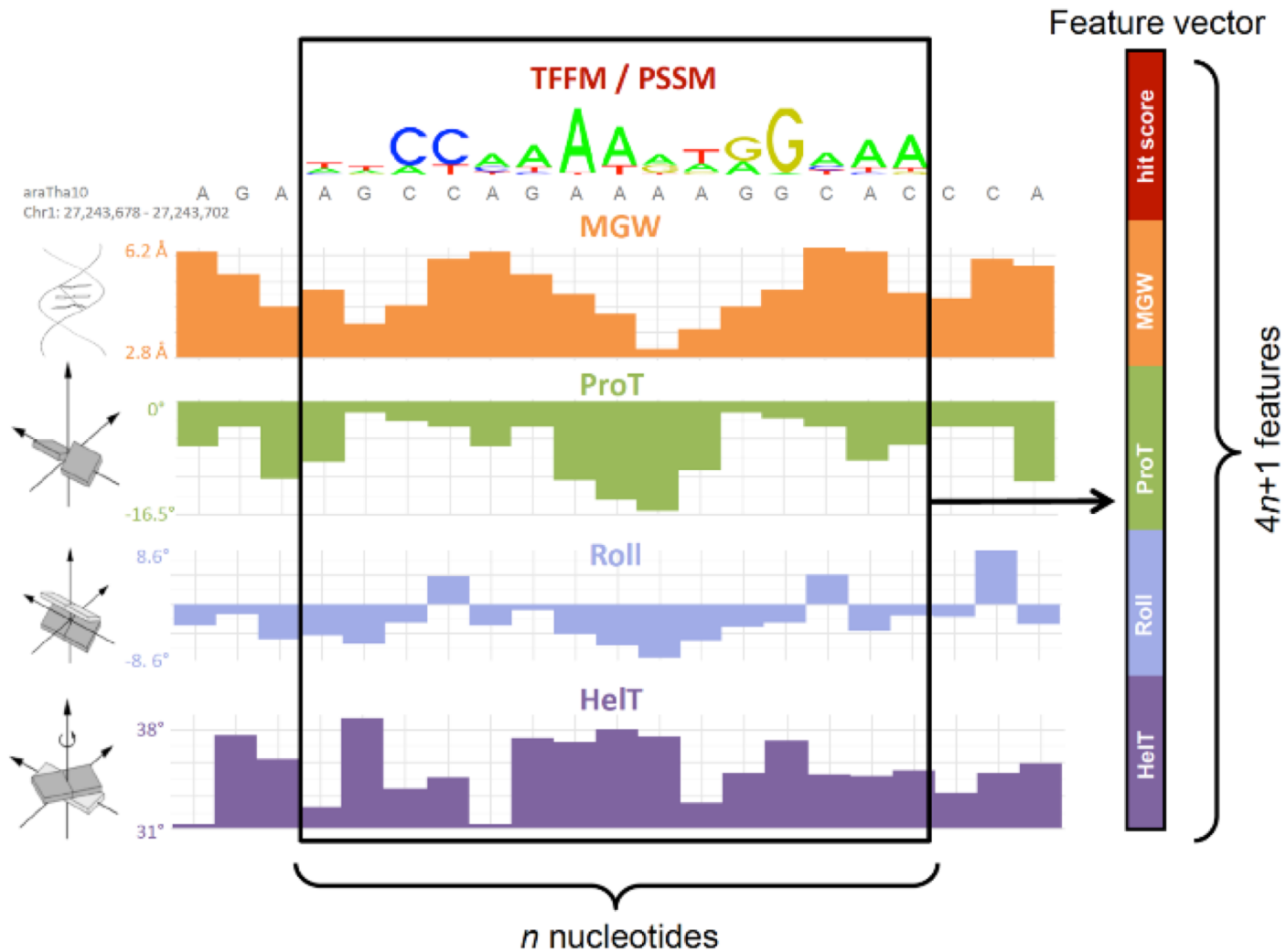
L. Yang, T. Zhou, I. Dror, A. Mathelier, W.W. Wasserman, R. Gordan, R. Rohs. *Nucl. Acids Res.*, 2014.



A. Mathelier and X. Zhao, *et al.*, B. Lenhard, A. Sandelin, W.W. Wasserman. *Nucl. Acids Res.*, 2014.

TFBSshape: <http://rohslab.cmb.usc.edu/TFBSshape/>

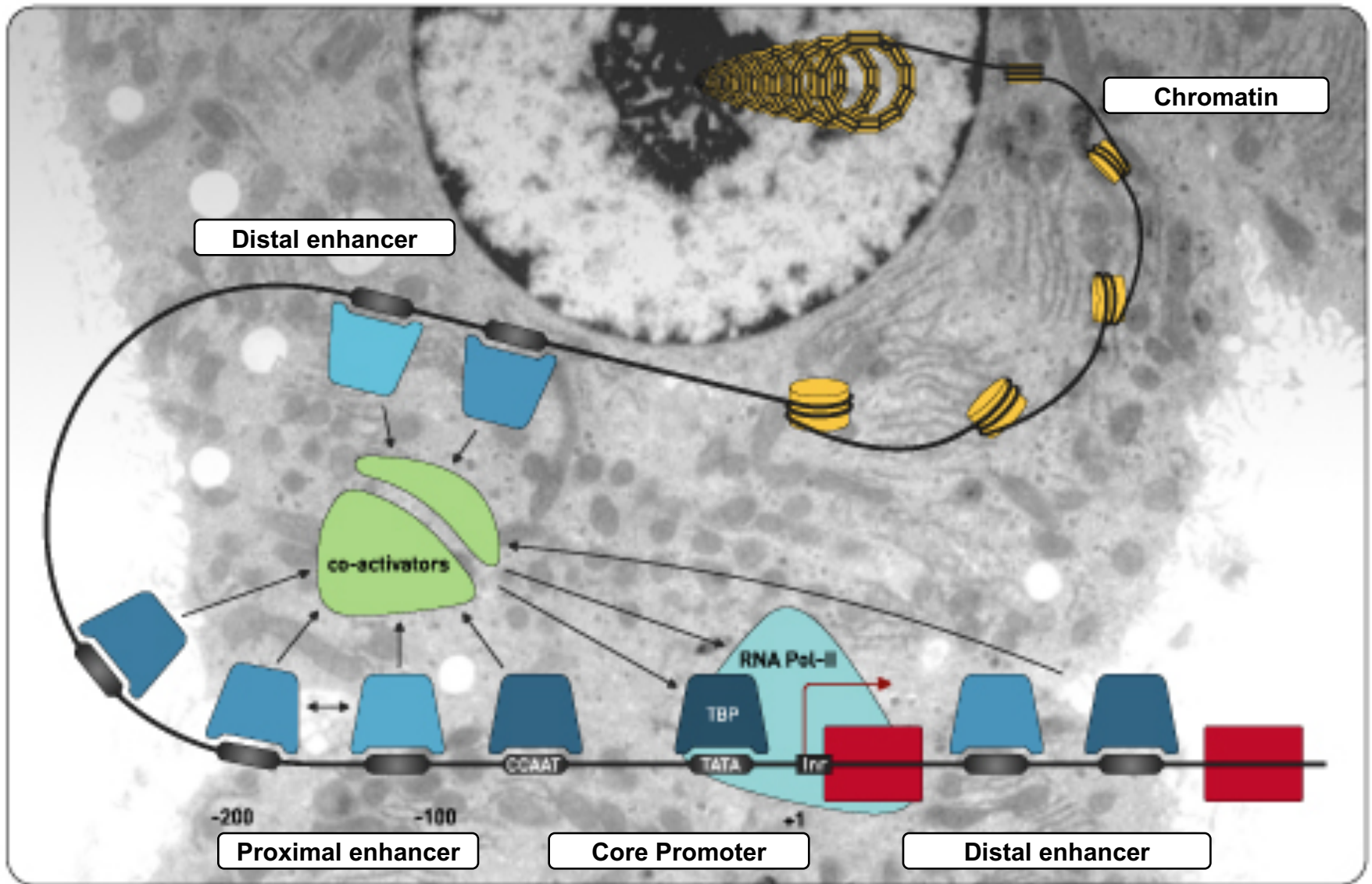
Shape Properties



Big challenges ahead

- Understanding all TFs across a developing organism
- Genetic variation in TFBS
- Integration of context and more complex predictive models
- Transition from matrices to hidden Markov models or energy models

Complexity in transcription



Reflections

- Futility conjecture – essentially predictions of individual TFBS have no relationship to an *in vivo* function
- Successful bioinformatics methods for site discrimination incorporate additional information (clusters, conservation)
- TFBS enrichment is a powerful means to identify TFs likely to contribute to observed patterns of co-expression
- Successful methods for pattern discovery will have to incorporate additional information (ChIP-seq, conservation, structural constraints on TFs, 3D genome organization)

We are on a Coffee Break & Networking Session

compute | calcul
canada | canada

Workshop Sponsors:



Canadian Centre for
Computational
Genomics

