CS273B: Deep learning for Genomics and Biomedicine

Lecture 2: Convolutional neural networks and applications to functional genomics

09/28/2016

Anshul Kundaje, James Zou, Serafim Batzoglou

Outline

- Anatomy of the human genome
- Introduction to next-gen sequence and protein-DNA binding maps
- Convolutional neural networks for predicting protein-DNA binding maps from DNA sequence
- Multi-modal convolutional neural networks for predicting protein-DNA binding maps
- Convolutional neural networks on images

Anatomy of the human genome

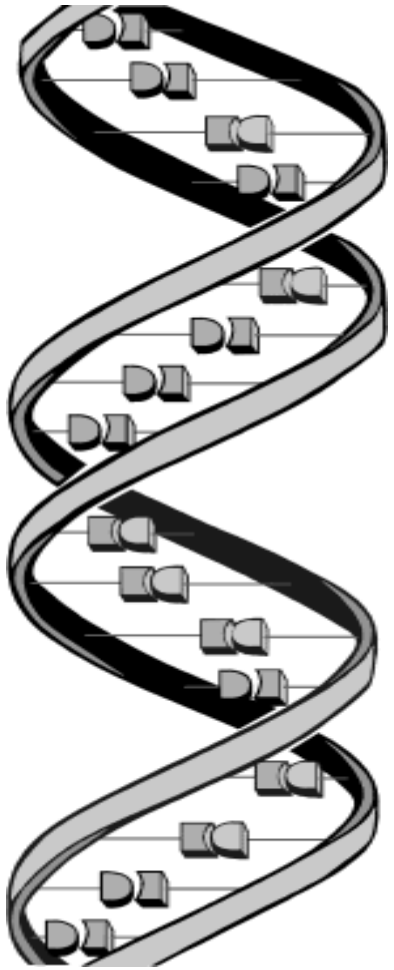
TGCCAAGCAGCAAAGTTTTGCTGCTGTTTATTTTTGTAGCTCTTACTATATTCT
ACTTTTACCATTGAAAATATTGAGGAAGTTATTTATATTTCTATTTTTTATATAT
TATATATTTTATGTATTTTAATATTACTATTACACATAATTATTTTTTATATATATGA
AGTACCAATGACTTCCTTTTCCAGAGCAATAATGAAATTTCACAGTATGAAA
ATGGAAGAAATCAATAAAATTATACGTGACCTGTGGCGAAGTACCTATCGTG
GACAAGGTGAGTACCATGGTGTATCACAAATGCTCTTTCCAAAGCCCTCTCC
GCAGCTCTTCCCCTTATGACCTCTCATCATGCCAGCATTACCTCCCTGGACCC
CTTTCTAAGCATGTCTTTGAGATTTTCTAAGAATTCTTATCTTGGCAACATCTT
GTAGCAAGAAAATGTAAAGTTTTCTGTTCCAGAGCCTAACAGGACTTACATA
TTTGACTGCAGTAGGCATTATATTTAGCTGATGACATAATAGGTTCTGTCATA
GTGTAGATAGGGATAAGCCAAAATGCAATAAGAAAAACCATCCAGAGGAA
ACTCTTTTTTTTTTCTTTTTCTTTTTTTTTTTTCCAGATGGAGTCTCGCACTTC
TCTGTCACCCGGGCTGGAGCGCAGTGGTGCAATCTTGGCTCACTGCAACCT
CCACCTCCTGGGTTCAGGTGATTCTCCCACCTCAGCCTCCCGAGTAGTAGCT
GGAATTACAGGTGCGCGCTCCCACACCTGGCTAATTTTTTGTATTCTTAGTA
GAGATGGGGTTTCACCATGTTGGCCAGGCTGGTCTCAAACTCCTGCCCTCA
GGTGATCTGCCCACCTTGGCCTCCCAGTGTTGGGTTTACAGGCGTGAGCCA
CCGCGCCTGGCCTGGAGGAAACTCTTAACAGGGAAACTAAGAAAGAGTTG
AGGCTGAGGAACTGGGGCATCTGGGTTGCTTCTGGCCAGACCACCAGGCT
CTTGAATCCTCCCAGCCAGAGAAAGAGTTTCCACACCAGCCATTGTTTTCCT
CTGGTAATGTCAGCCTCATCTGTTGTTCCTAGGCTTACTTGATATGTTTGTAA
ATGACAAAAGGCTACAGAGCATAGGTTCCTCTAAAATATTCTTCTTCCTGTGT
CAGATATTGAATACATAGAAATACGGTCTGATGCCGATGAAAATGTATCAGCT
TCTGATAAAAGGCGGAATTATAACTACCGAGTGGTGATGCTGAAGGGAGAC
ACAGCCTTGGATATGCGAGGACGATGCAGTGCTGGACAAAAGGCAGGTAT
CTCAAAAGCCTGGGGAGCCAACTCACCCAAGTAACTGAAAGAGAGAAACA
AACATCAGTGCAGTGGAAGCACCCAAGGCTACACCTGAATGGTGGGAAGC
TCTTTGCTGCTATATAAAATGAATCAGGCTCAGCTACTATTATT ……

The Human Genome

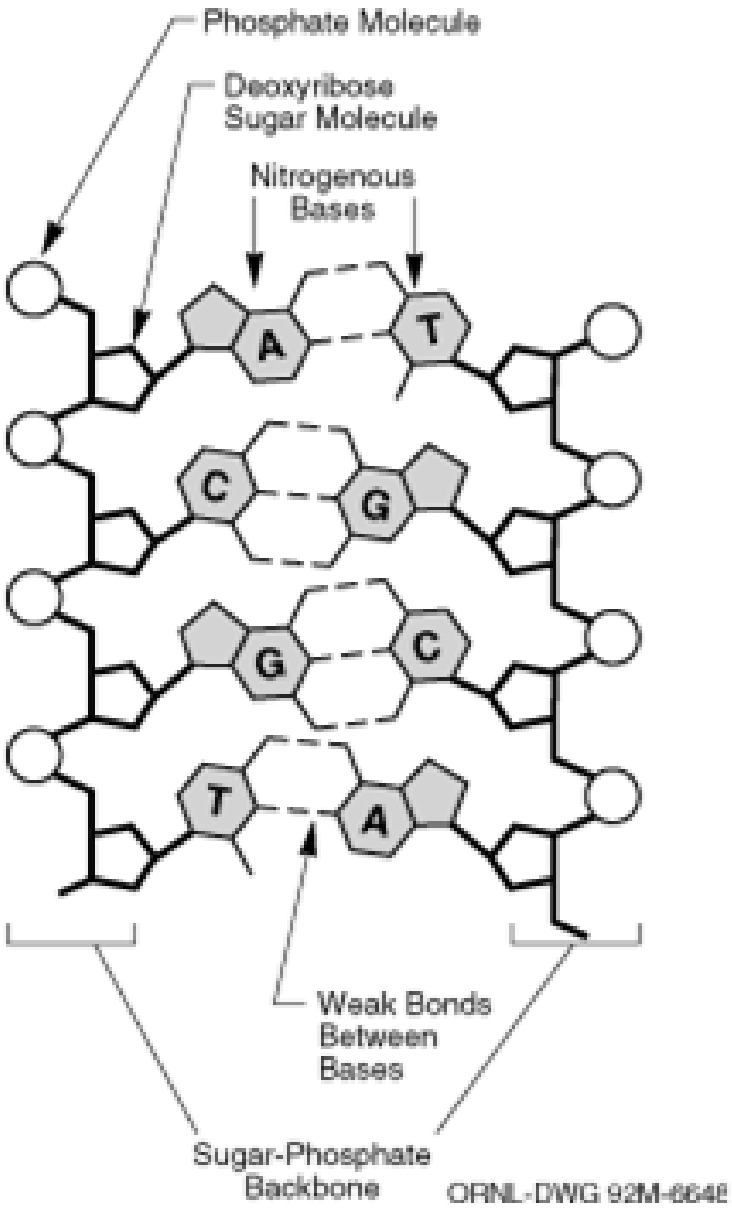
2003

~ 3 billion nucleotides

DNA: the molecule of heredity



Double helix
(double stranded)



Forward strand

5’ → 3’

TGCCAAGCA

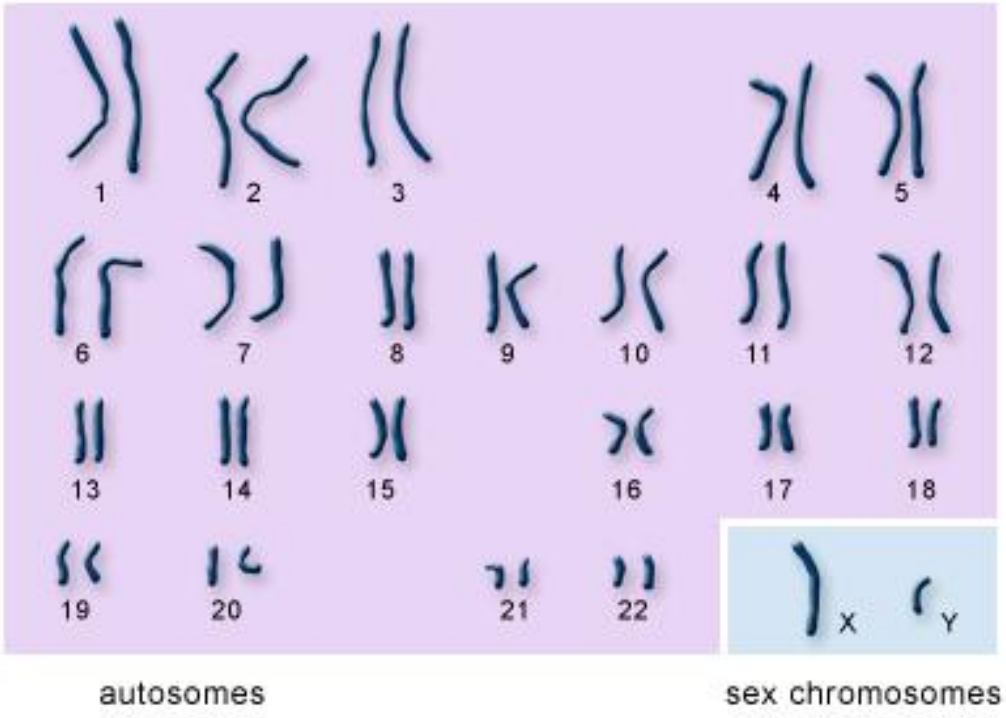
||||||||||

ACGGTTCGT

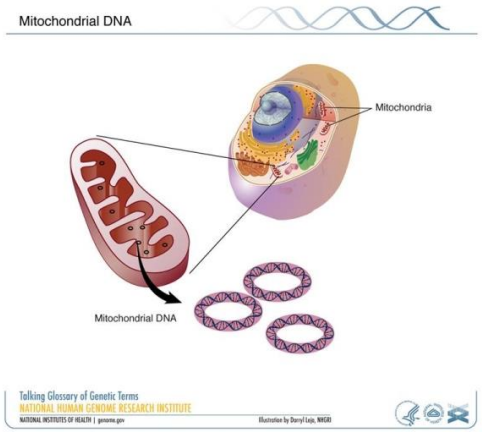
3’ ← 5’

Reverse complement

Chromosomes in humans



U.S. National Library of Medicine

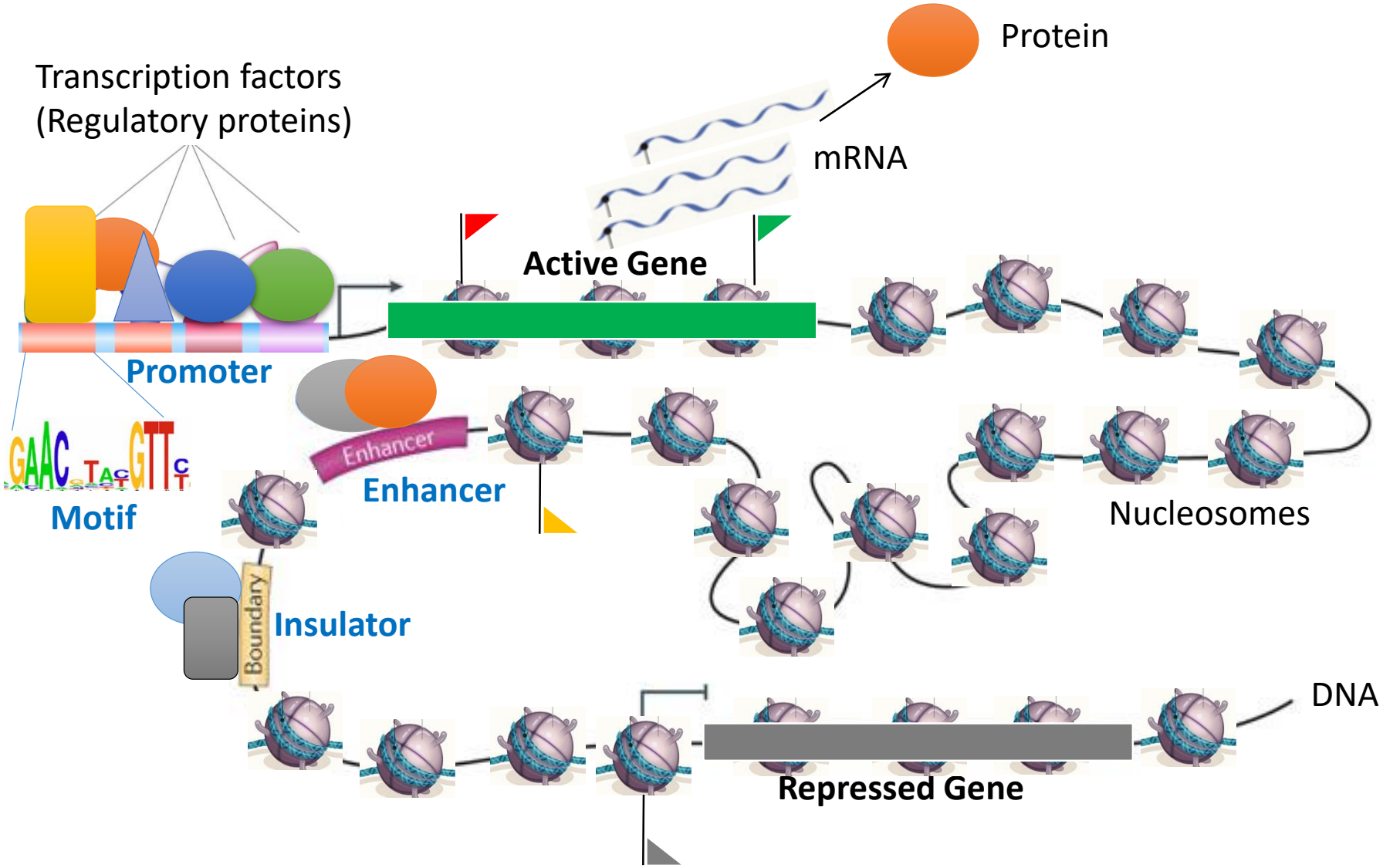


TGCCAAGCA
||||||||||
ACGGTTCGT

TGCCAAGCA
||||||||||
ACGGTTCGT

- Humans are diploid (2 copies of each chromosome)
- 22 pairs of autosomes
- Sex chromosomes: female (X,X) , male (X,Y)
- Mitochondrial DNA (circular, many copies per cell)
- Diploid Human genome = ~3 billion bp X 2

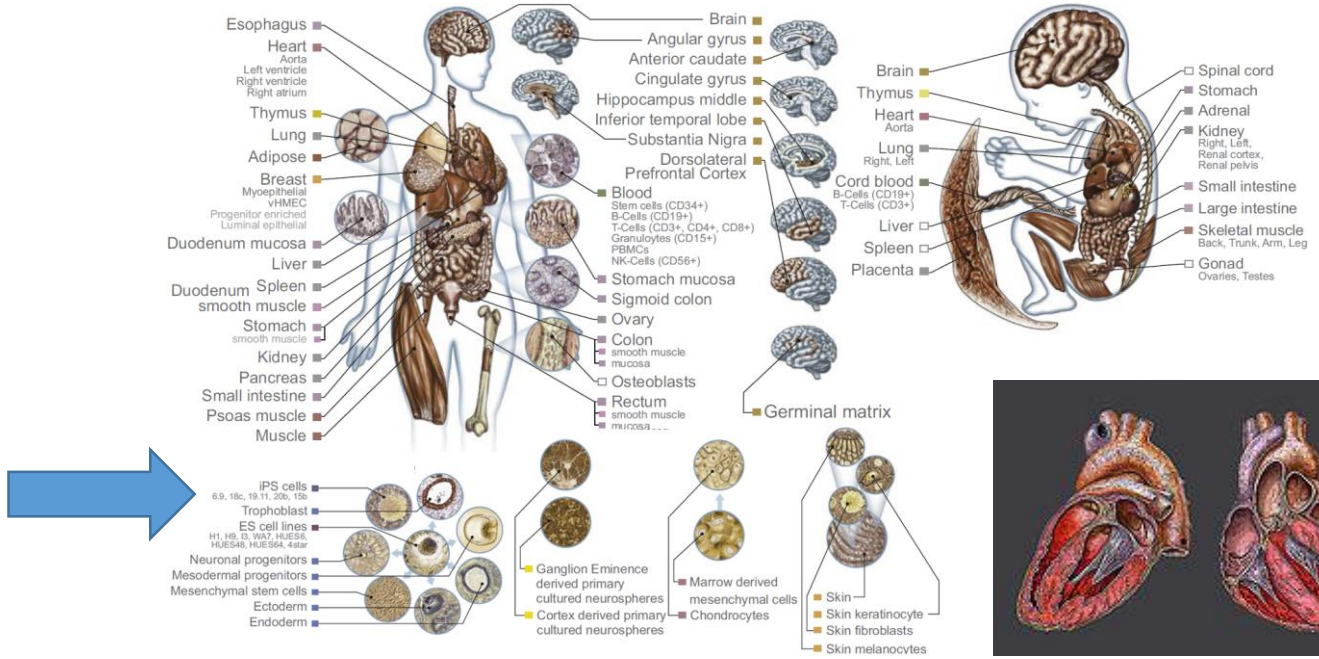
Functional elements in the genome

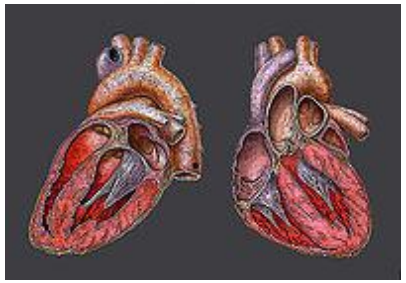


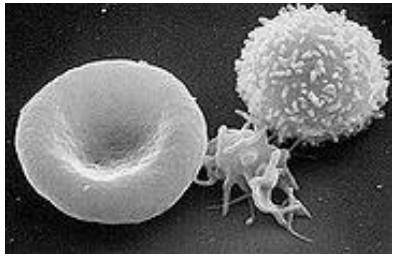
Chromatin (epigenetic) modifications

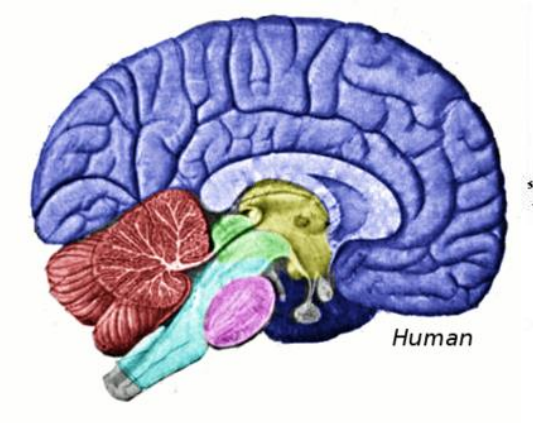
One genome ⇔ Many cell types

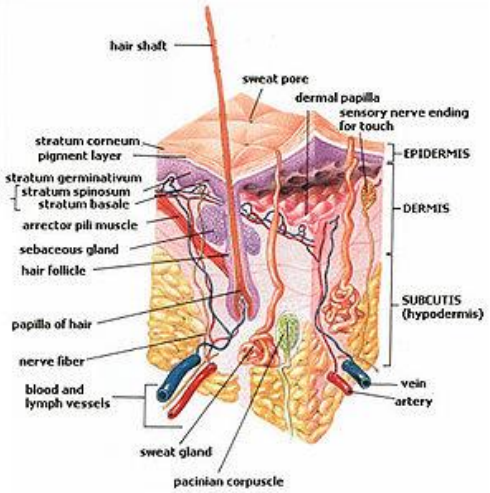
ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA







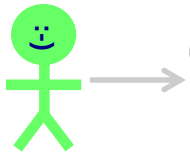




Introduction to functional genomics & next-gen sequencing

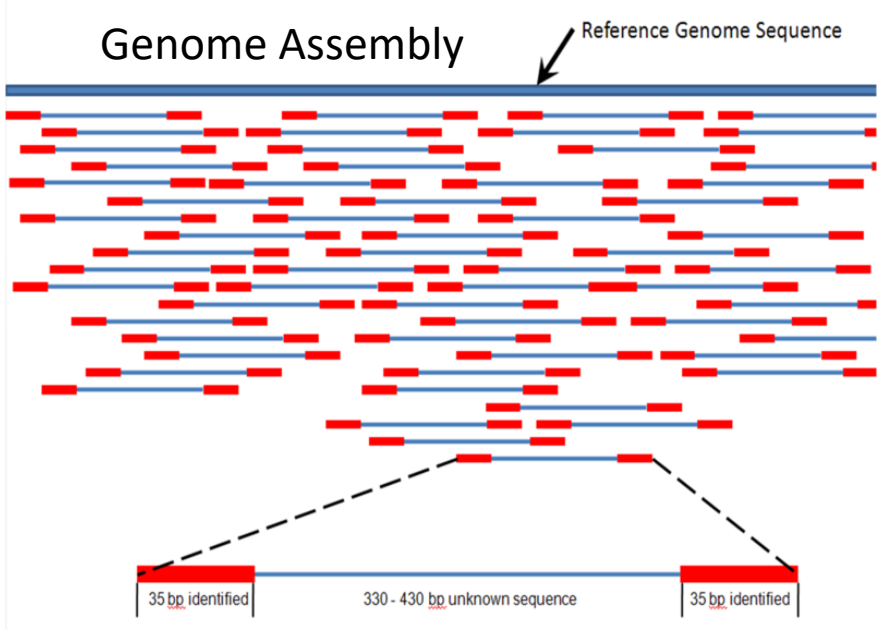
What is Functional Genomics?

Function?



GGCAATACGATATTAGCAAATAAACGATAGTATACAAATCGTATTAC...

~ 3 billion bases



2003

Genomic sequence => Static

What is the context-specific function of different regions (bases) of the genome?

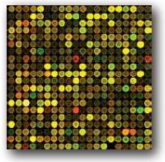
How to explain diversity of cell-types?

How to explain dynamic cellular repsonse?

Sequencing technologies

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

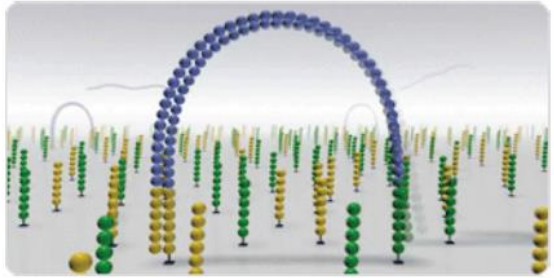
2nd-generation DNA sequencing

Since ~2007

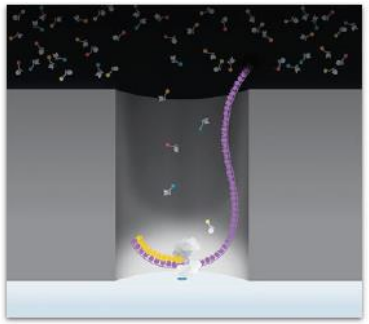
3rd-generation & single-molecule DNA sequencing

Since ~2010

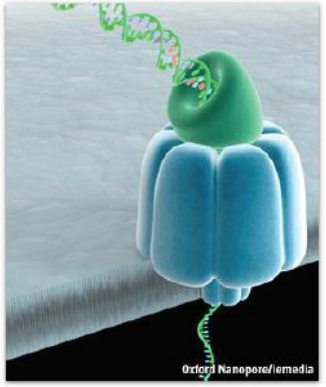
Since 2005, many DNA sequencing instruments have been described and released. They are based on a few different principles



Synthesis / ligation



SMRT cell

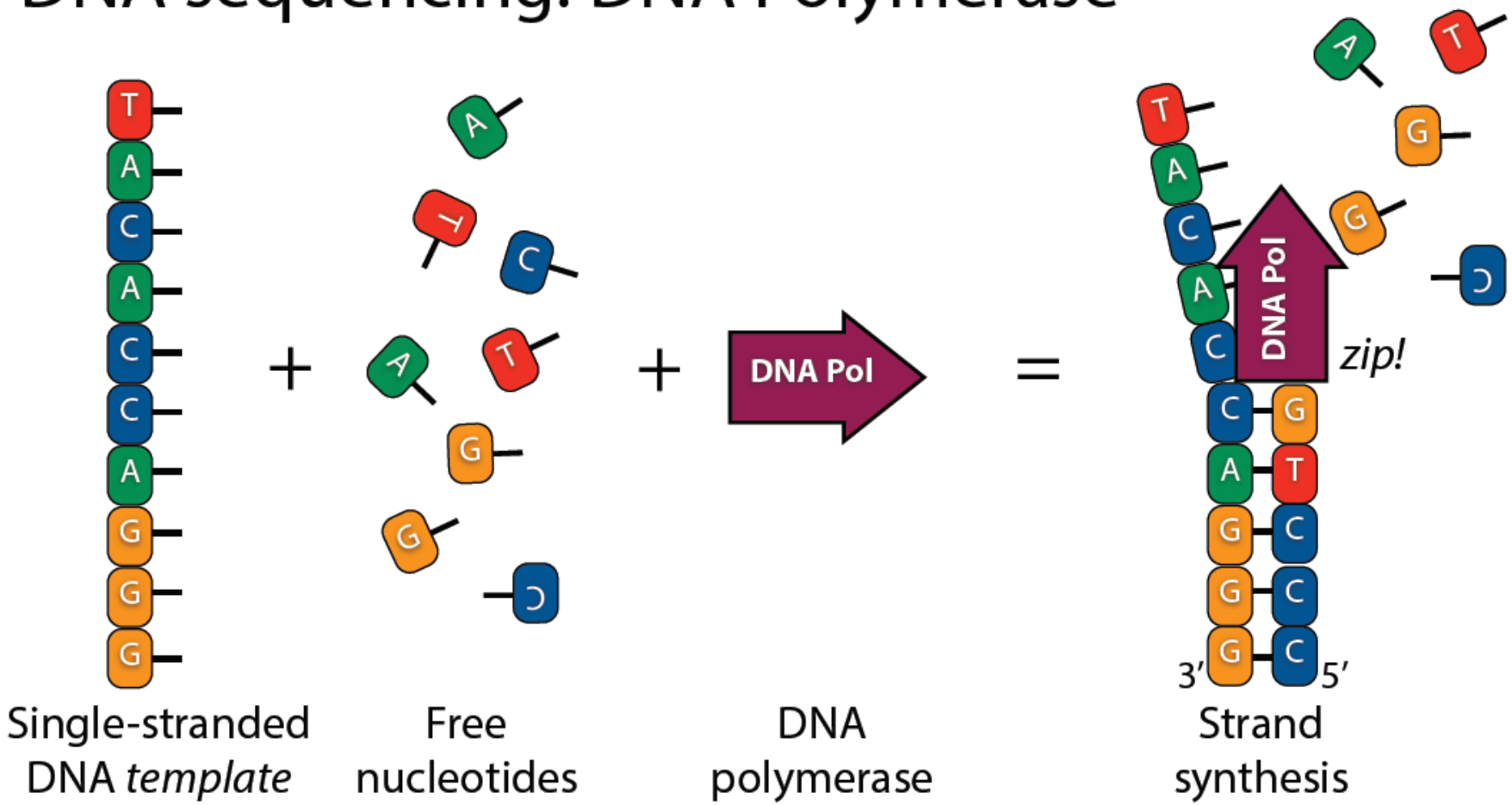


Nanopore

Sequencing by synthesis ("massively parallel sequencing") provides greatest throughput, and is the most prevalent today

Pictures: http://www.illumina.com/systems/miseq/technology.ilmn, http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

DNA sequencing: DNA Polymerase



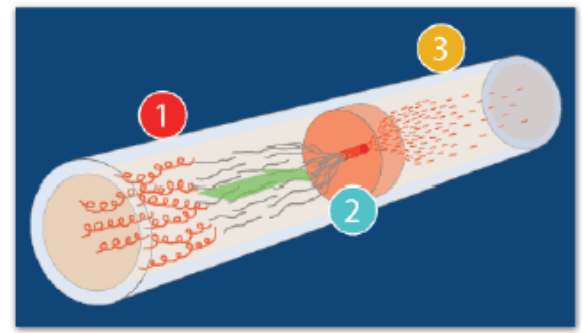
DNA polymerase moves along the template in one direction, integrating complementary nucleotides as it goes

Slides from Ben Langmead

Sequencing by synthesis

1. Take DNA sample, which includes many copies of the genome, and chop it into single-stranded fragments ("templates")

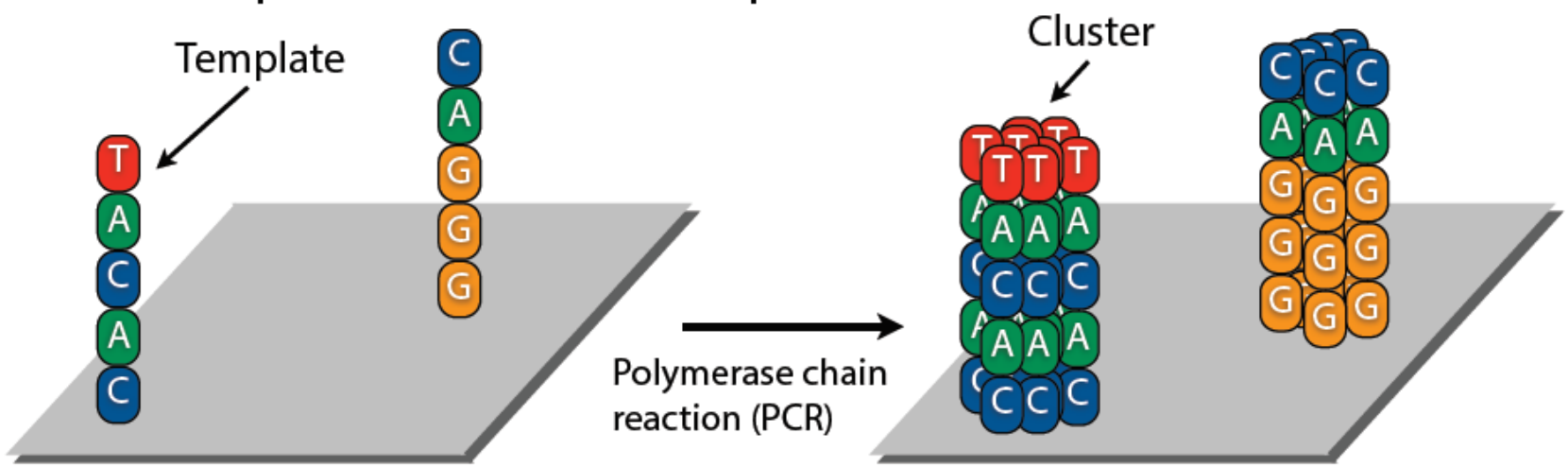
   E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



Picture: http://www.jgi.doe.gov/sequencing/education/how/how_1.html

2. Attach templates to a surface

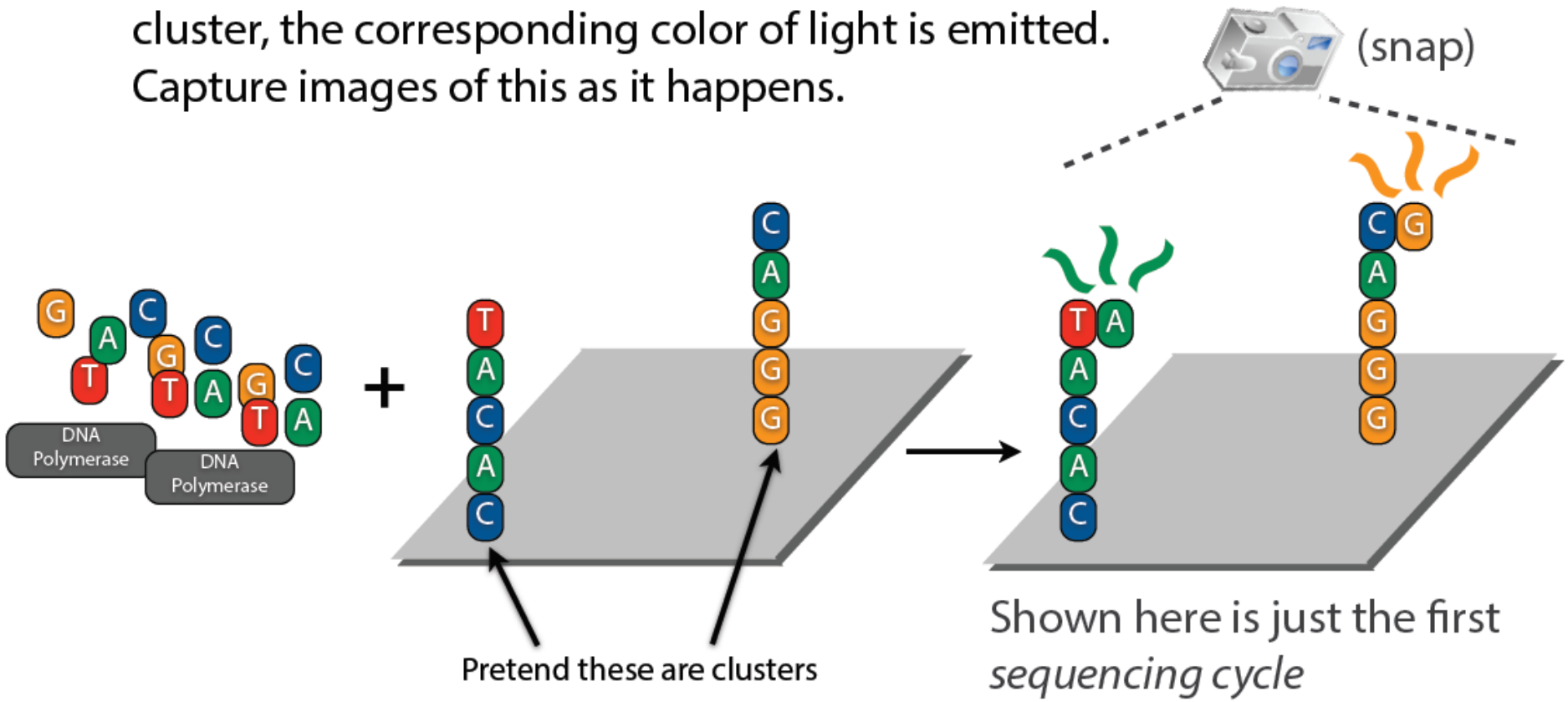
3. Make copies so that each template becomes a "cluster" of clones



Slides from Ben Langmead

Sequencing by synthesis

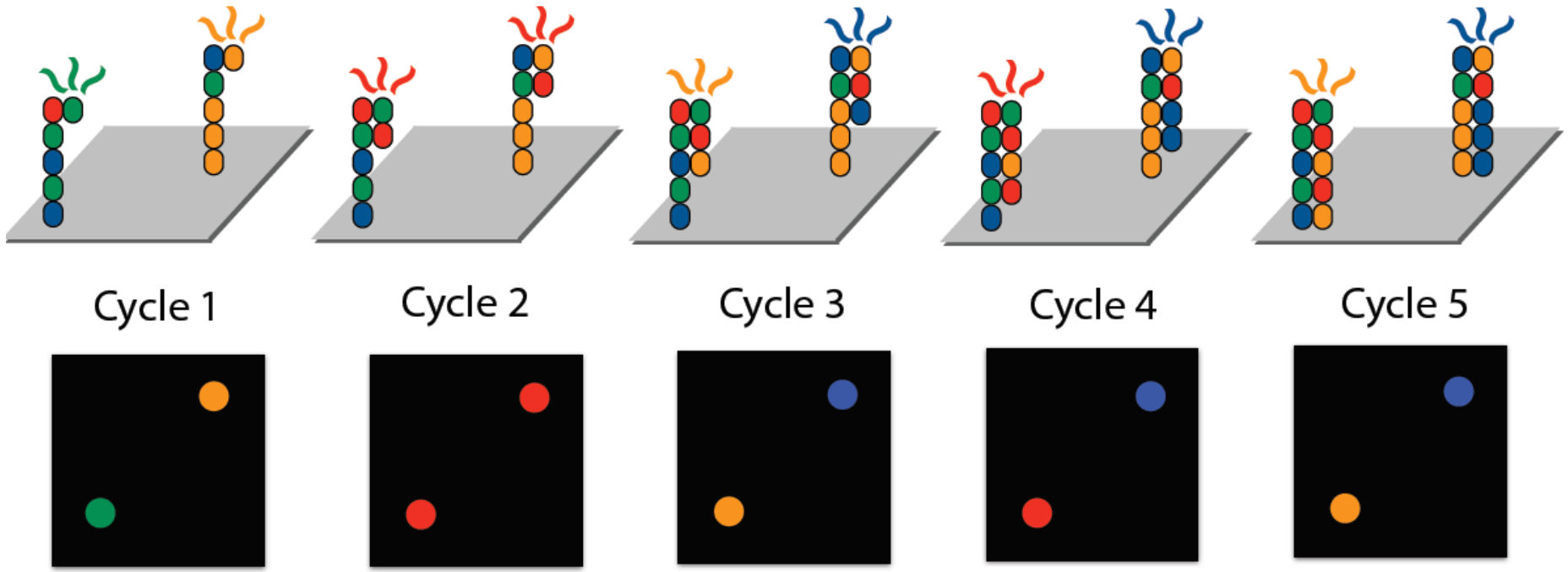
4. Repeatedly inject mixture of *color-labeled* nucleotides (A, C, G and T) and DNA polymerase. When a complementary nucleotide is added to a cluster, the corresponding color of light is emitted. Capture images of this as it happens.



Slides from Ben Langmead

Sequencing by synthesis

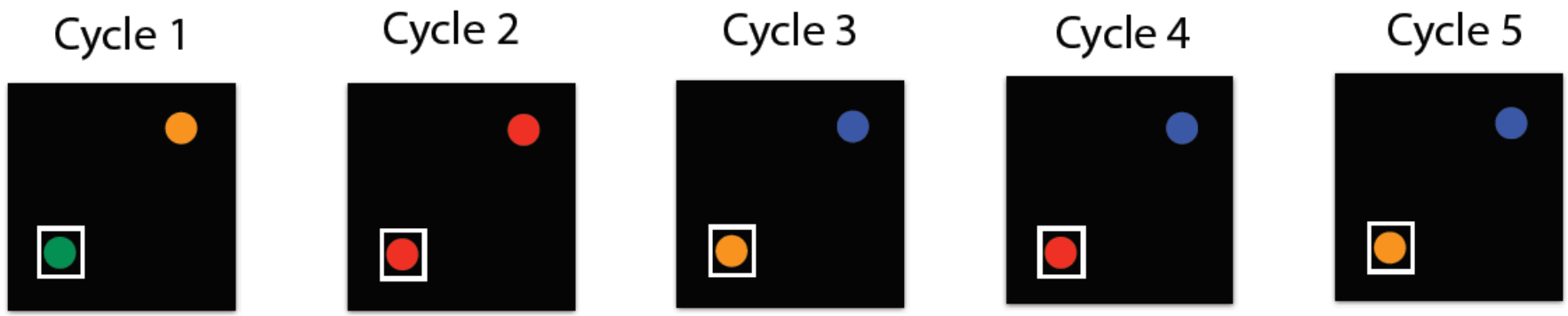
5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



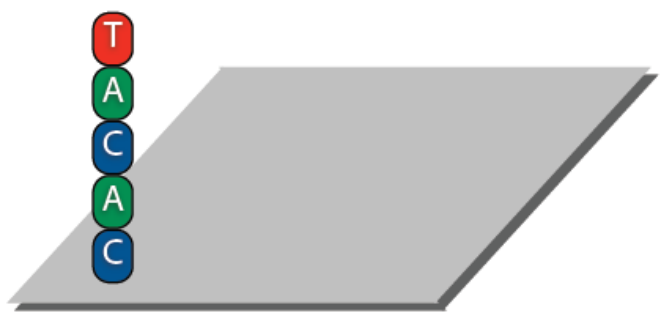
Slides from Ben Langmead

Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: TACAC, corresponding to the template sequence



TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

Slides from Ben Langmead

Mapping short-reads to reference genome

Naïve method

- Scan whole genome with every read
- Problem: Too slow

Indexing + Alignment approach

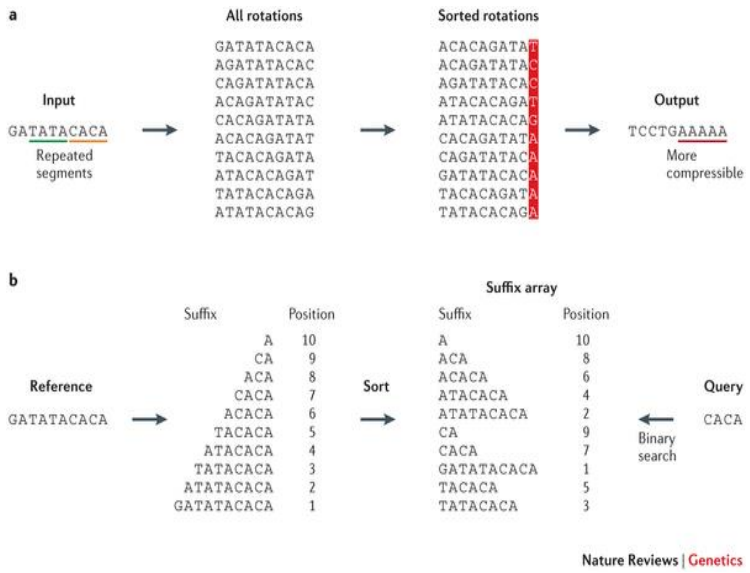
- Create a compressed reference 'genome index'
  - a map of where each short subsequence of length 'k' hits the genome
- Map reads using index via smart alignment algorithms and data structures (e.g suffix array)
- Allow for errors: insertions, deletions, mismatches in alignments

Run times for indexing alignment

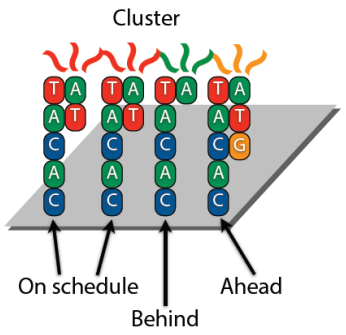
- Indexing human genome ~ 3 hours
- Alignment speed: 2 million 35 bp reads on 1 processor ~20 mins
- Alignment speed depends on error rate

ACGTTACCGAATCGATCAAGTCGA

TAC

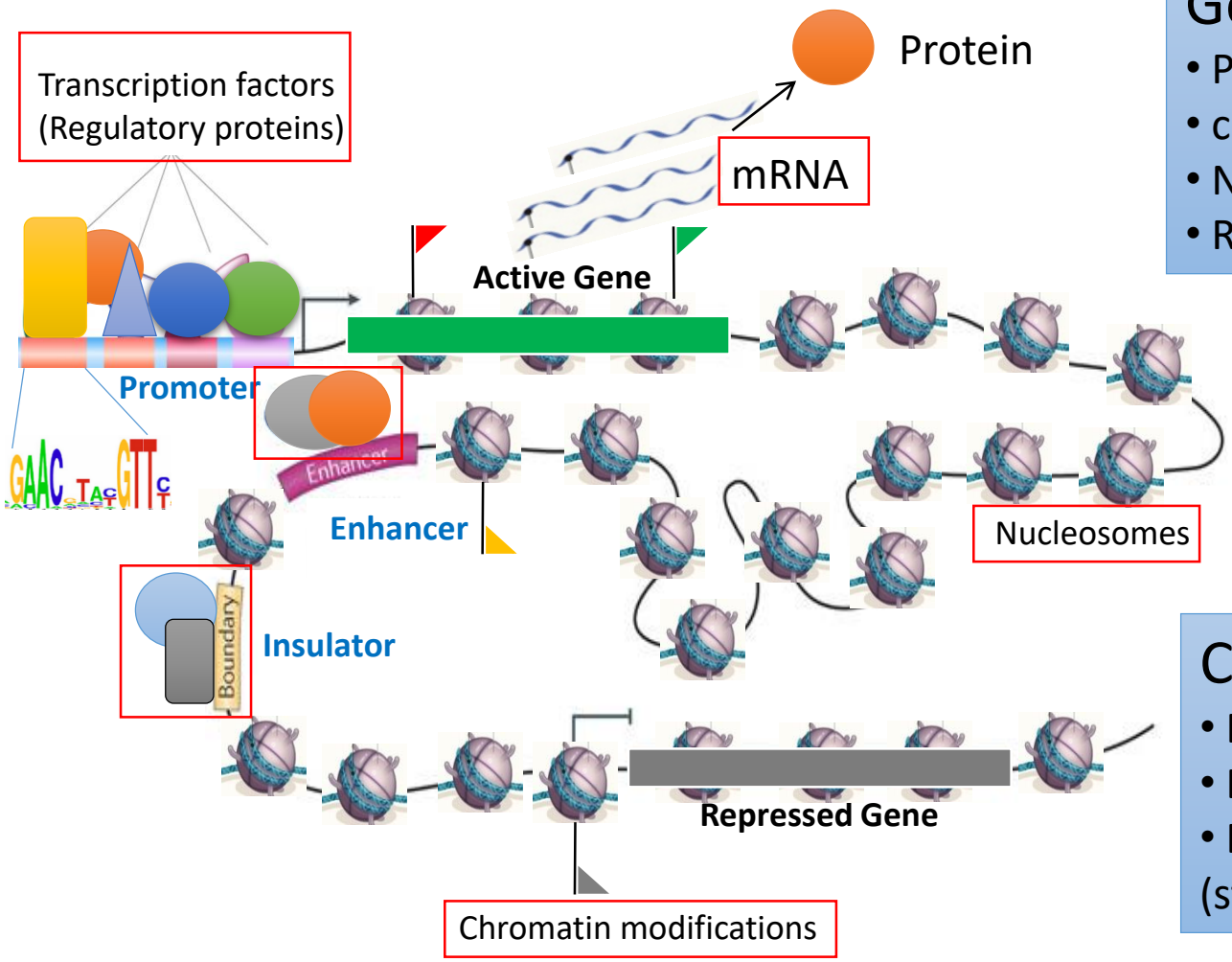


http://www.nature.com/nrg/journal/v14/n5/box/nrg3433_BX2.html



Using sequencing for functional genomics

Genome-wide maps of biochemical activity



Genome-wide expts.

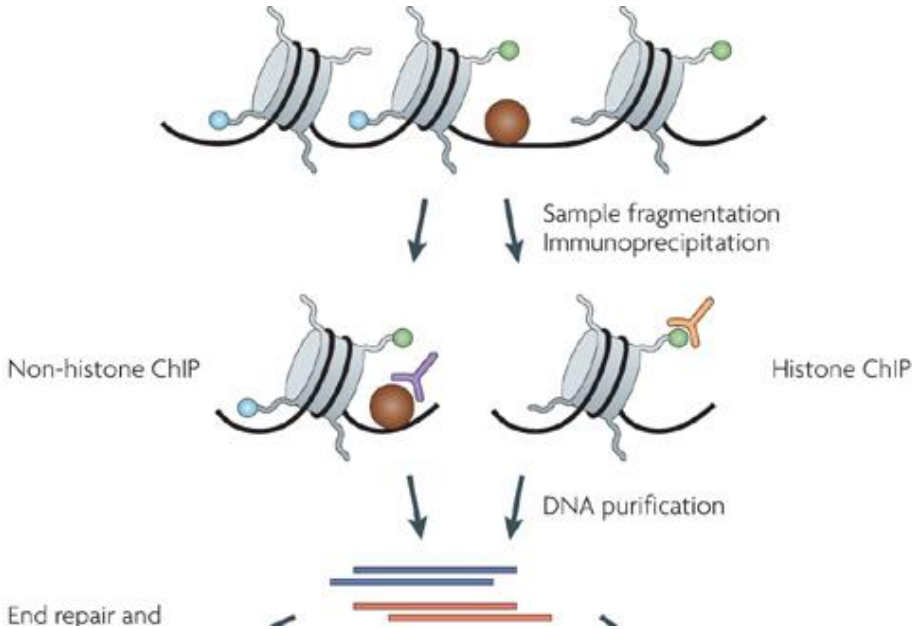
- Protein-DNA binding maps
- chromatin modification maps
- Nucleosome positioning maps
- RNA expression

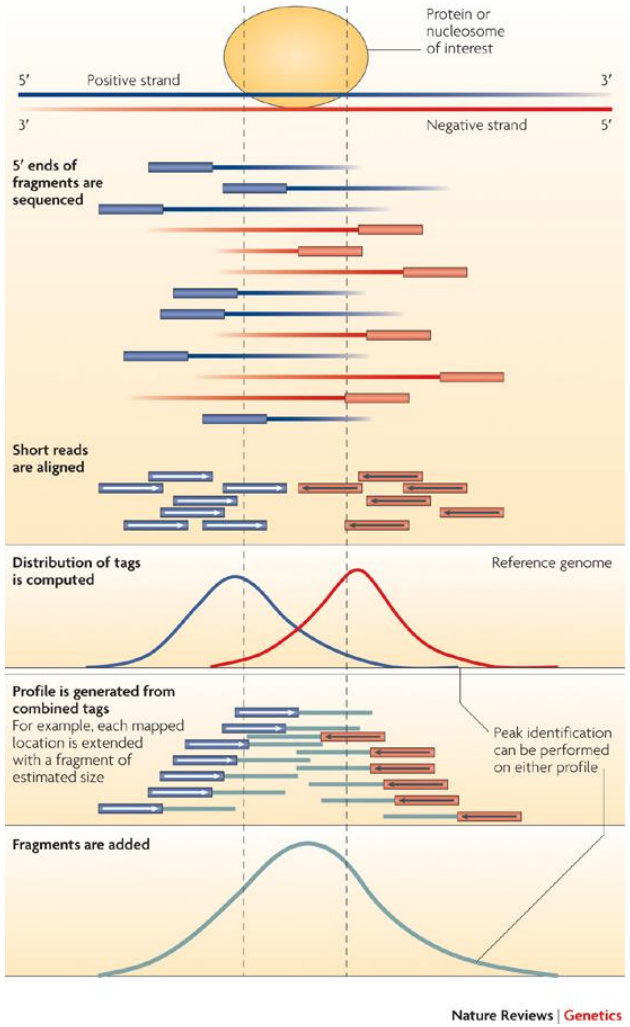
Cellular Dynamics

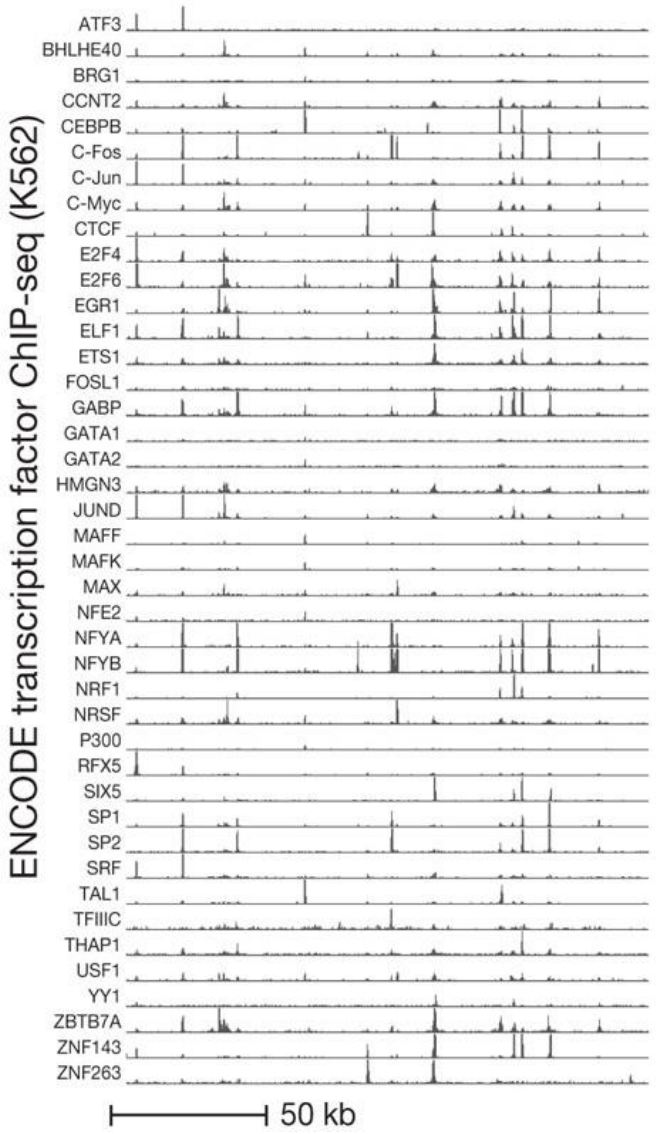
- Different cell-types/tissues
- Diseased states (e.g. cancer)
- Different perturbations (stimuli)

Protein-DNA binding maps

Chromatin immunoprecipitation (ChIP-seq)

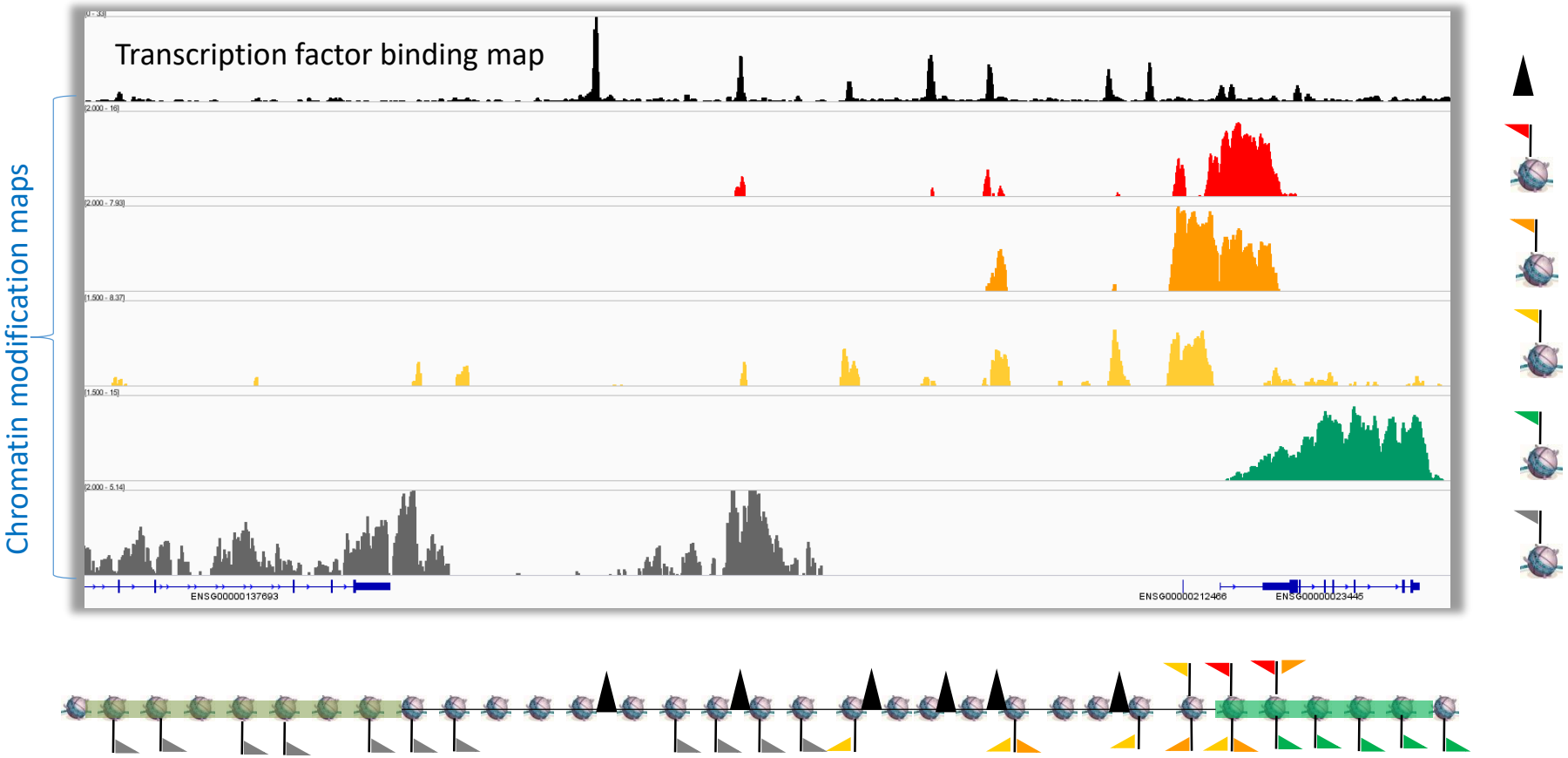






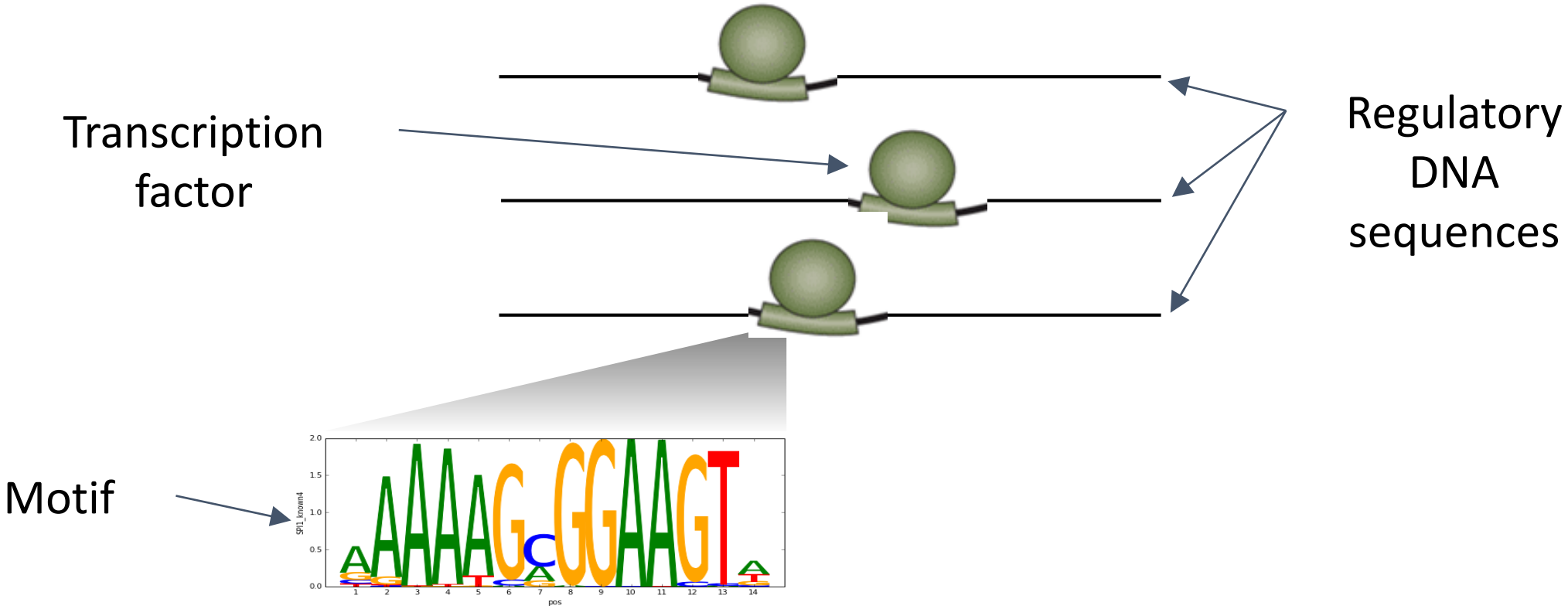
Protein-DNA binding maps
Maps of histone modifications
Maps of histone variants

Genome-wide ChIP-seq signal maps



DNA sequence determinants of protein-DNA interactions

Key properties of regulatory sequence



TRANSCRIPTION FACTOR BINDING

Regulatory proteins called **transcription factors (TFs)** bind to high affinity sequence patterns (**motifs**) in regulatory DNA

Sequence motifs

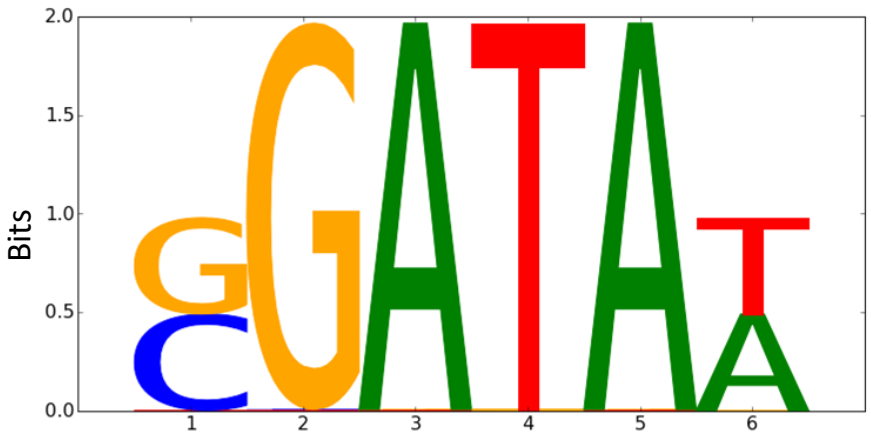
GGATAA
CGATAA
CGATAT
GGATAT

Set of aligned sequences
Bound by TF

$$p_i(x_i = a_i)$$

A	0	0	1	0	1	0.5
C	0.5	0	0	0	0	0
G	0.5	1	0	0	0	0
T	0	0	0	1	0	0.5

Position weight matrix
(PWM)



PWM
logo

https://en.wikipedia.org/wiki/Sequence_logo

The information content (y-axis) of position i is given by:[2]

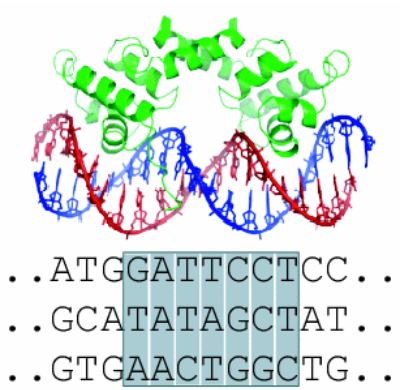
$$R_i = \log_2(4) - (H_i + e_n)$$

where H_i is the uncertainty (sometimes called the Shannon entropy) of position i

$$H_i = -\sum f_{a,i} \times \log_2 f_{a,i}$$

The height of letter a in column i is given by

$$\text{height} = f_{a,i} \times R_i$$



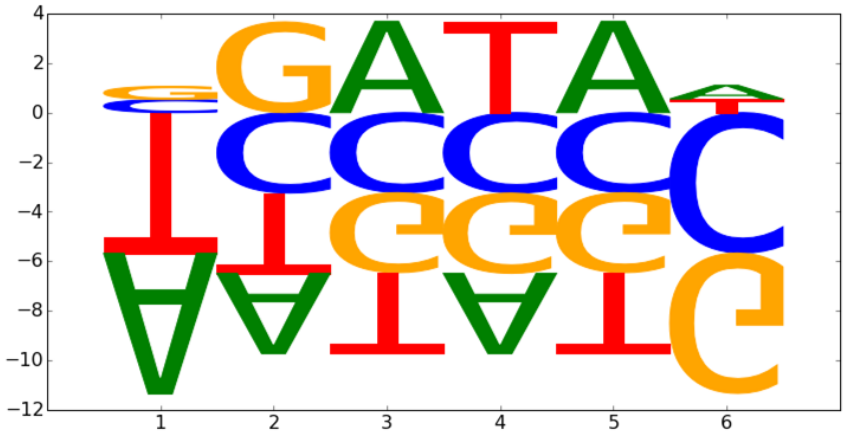
Sequence motifs

Accounting for genomic background nucleotide distribution

Position-specific scoring matrix (PSSM)

$$\log_2\left(\frac{p_i(x_i = a_i)}{p_{bg}(x_i = a_i)}\right)$$

A	-5.7	-3.2	3.7	-3.2	3.7	0.6
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5



PSSM logo

Scoring a sequence with a motif PSSM

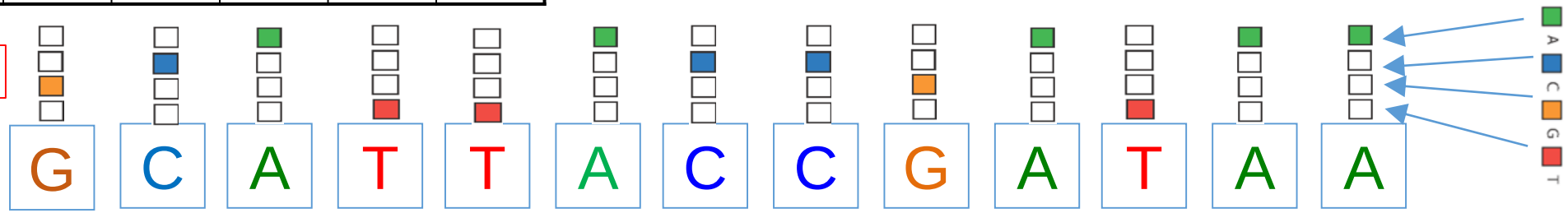
PSSM parameters

Scoring weights **W**

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding **(X)**

Input sequence



Convolution:
Scoring a sequence with a PSSM

Motif match Scores
sum(W * x)

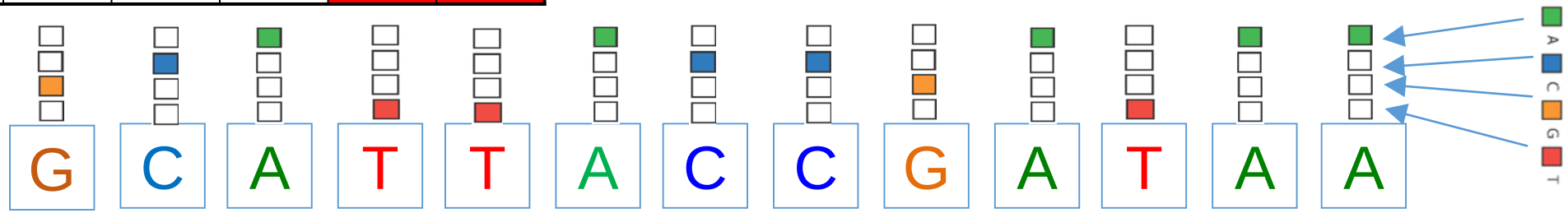
	-5.4											

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding (X)

Input sequence



Convolution

Motif match Scores
sum(W * x)

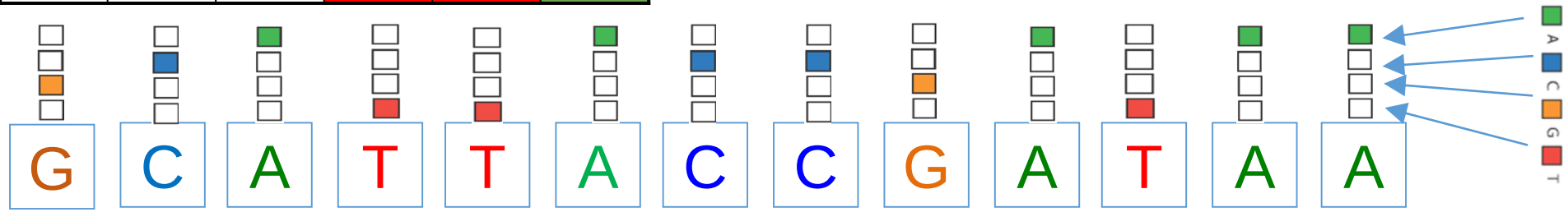
	-5.4	2.0										

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding (X)

Input sequence



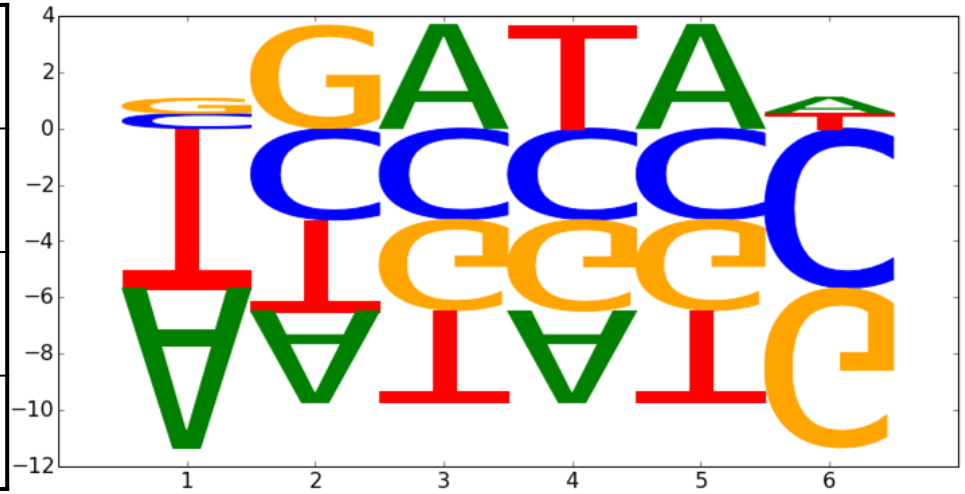
Convolution

Motif match Scores
sum(W * x)

-2.2	-5.4	2.0	-4.3	-24	-17	-18	-11	-12	16	-5.5	-8.5	-5.2

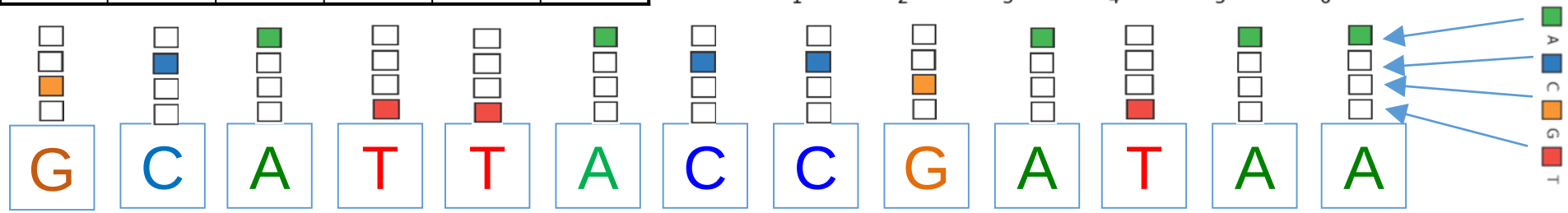
Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**



One-hot encoding (X)

Input sequence

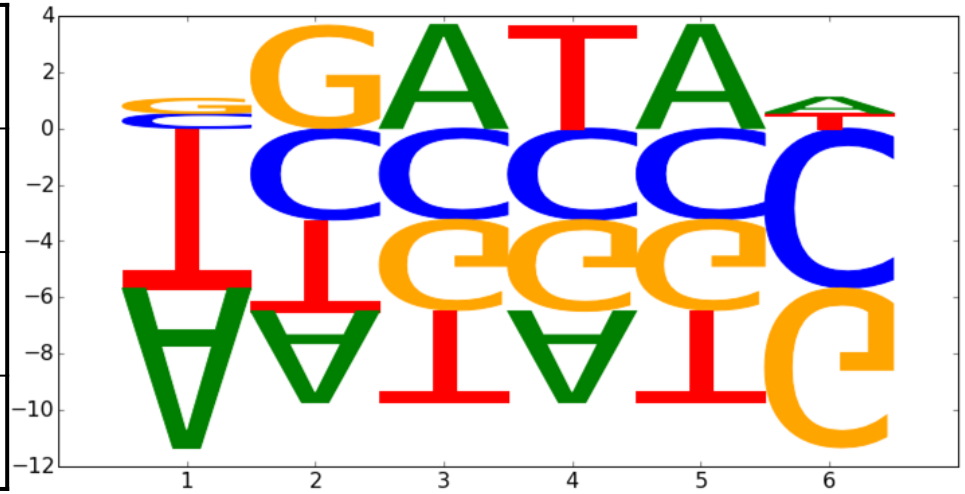


Thresholding scores

Thresholded Motif Scores **max(0, W*x)**	0	0	2.0	0	0	0	0	0	0	16	0	0	0
Motif match Scores W*x	-2.2	-5.4	2.0	-4.3	-24	-17	-18	-11	-12	16	-5.5	-8.5	-5.2

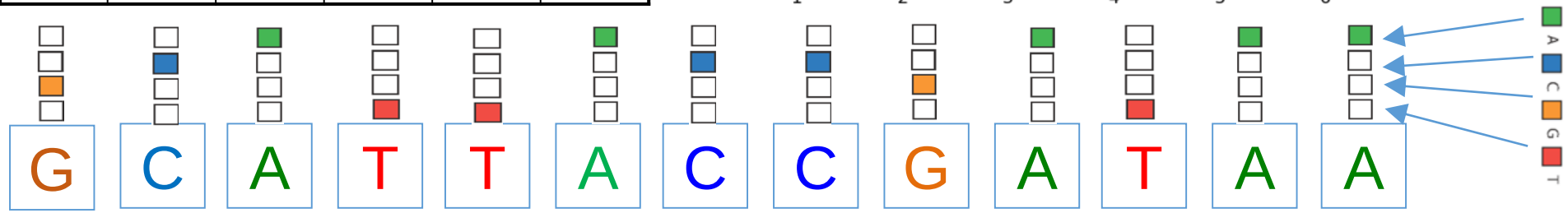
Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**



One-hot encoding (X)

Input sequence



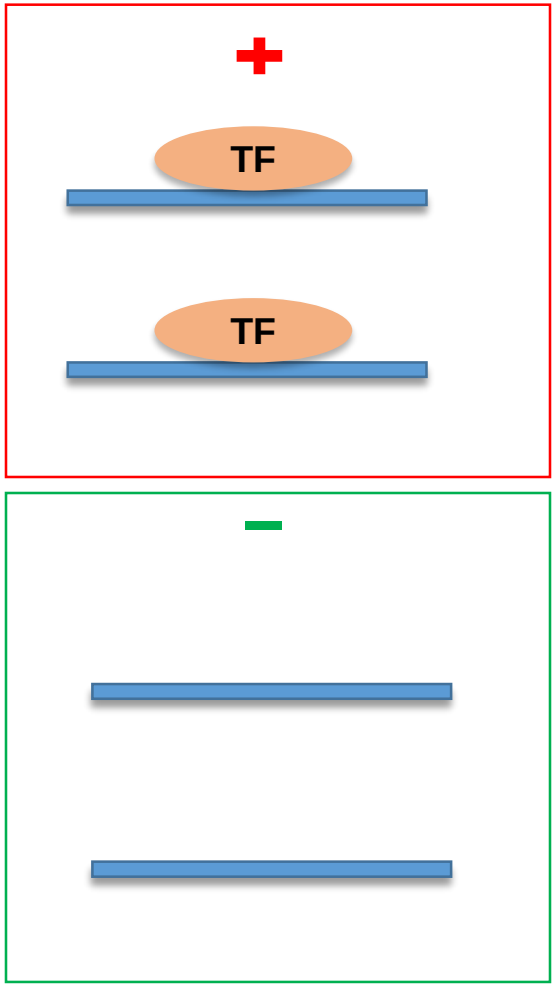
Convolutional neural networks for learning from DNA sequence

Learning patterns in regulatory DNA sequence

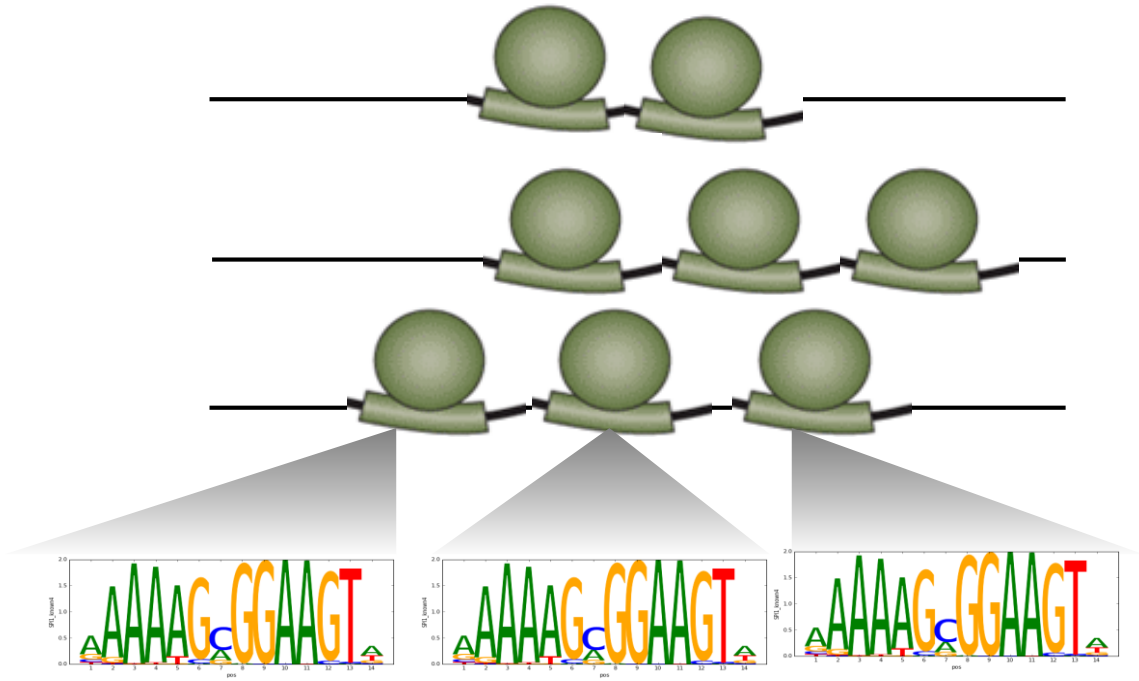
- Positive class of genomic sequences bound a transcription factor of interest

Can we learn patterns in the DNA sequence that distinguish these 2 classes of genomic sequences?

- Negative class of genomic sequences not bound by a transcription factor of interest



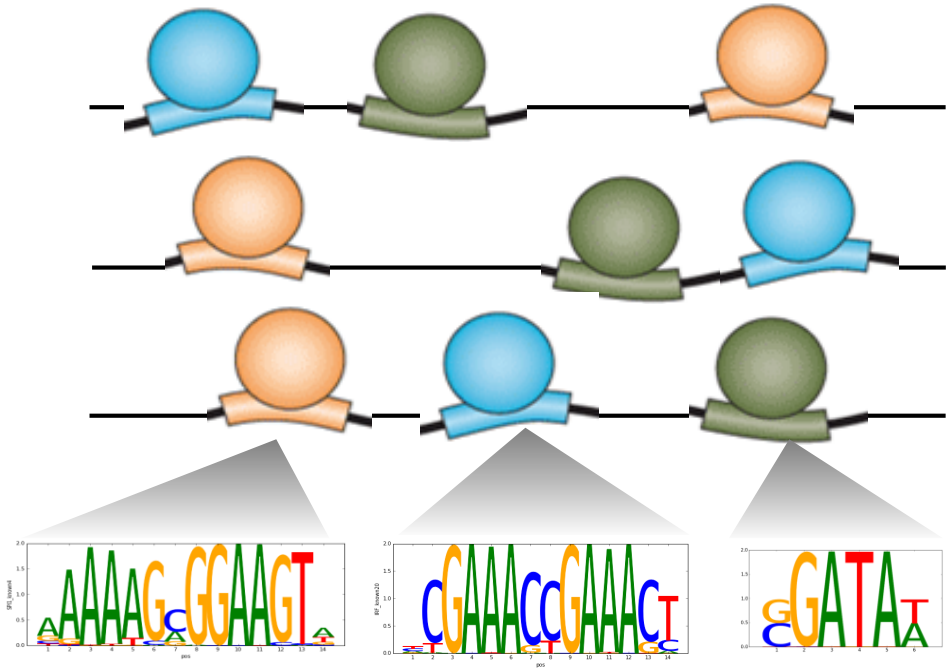
Key properties of regulatory sequence



HOMOTYPIC MOTIF DENSITY

Regulatory sequences often contain **more than one binding instance** of a TF resulting in **homotypic clusters of motifs of the same TF**

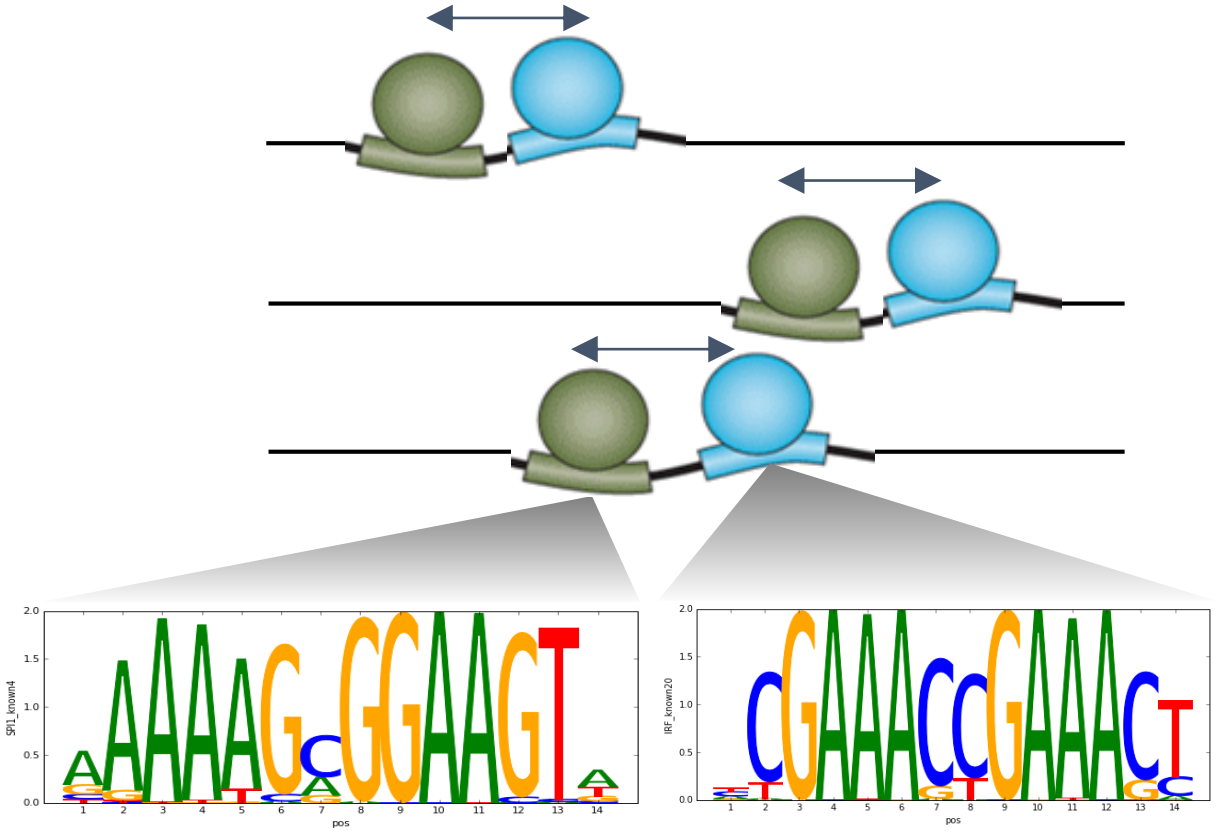
Key properties of regulatory sequence



HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences often bound by **combinations of TFs** resulting in **heterotypic clusters of motifs of different TFs**

Key properties of regulatory sequence



SPATIAL GRAMMARS OF HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences are often bound by **combinations of TFs** with specific **spatial and positional constraints** resulting in distinct **motif grammars**

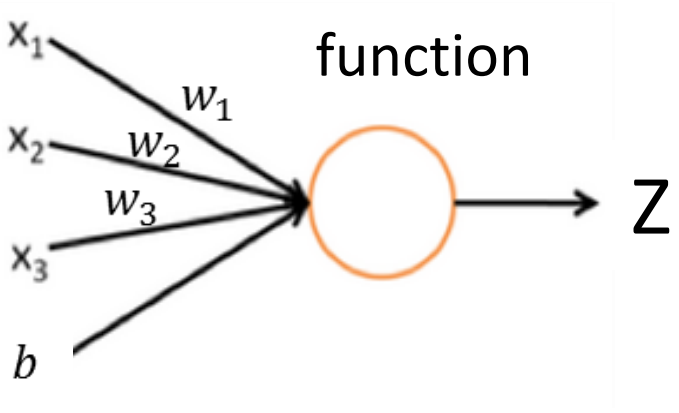
A simple classifier (An artificial neuron)

$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1 . x_1 + w_2 . x_2 + w_3 . x_3 + b$$

Linear function



Training the neuron means learning the optimal w's and b

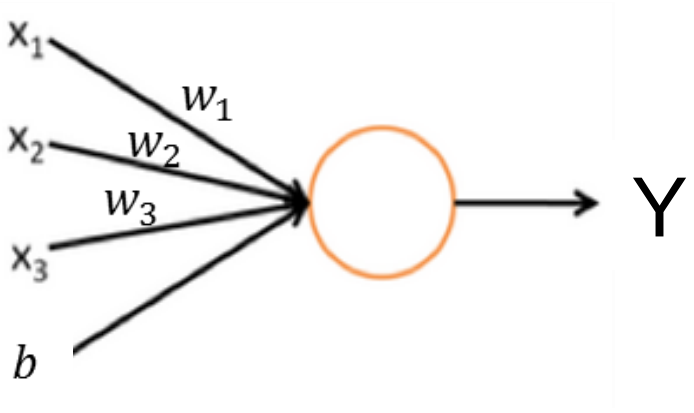
A simple classifier
(An artificial neuron)

$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1.x_1 + w_2.x_2 + w_3.x_3 + b$$

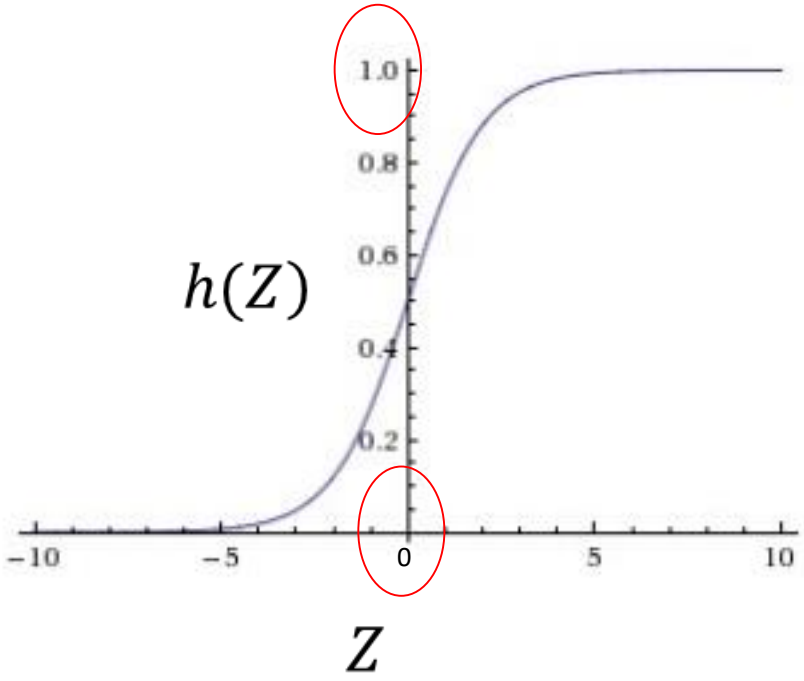
$$Y = h(Z)$$



Logistic / Sigmoid

Useful for predicting probabilities

Non-linear function



Training the neuron means learning the optimal w's and b

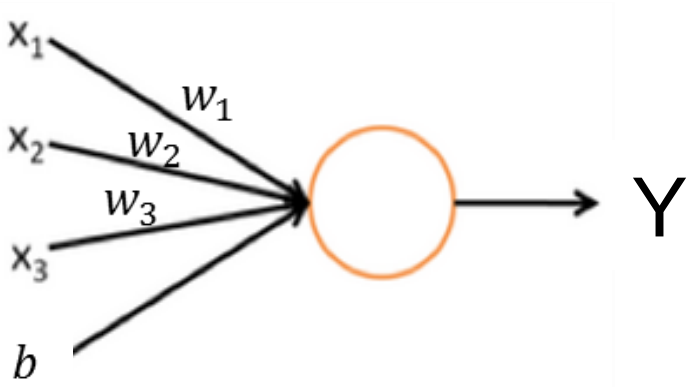
A simple classifier (An artificial neuron)

$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1.x_1 + w_2.x_2 + w_3.x_3 + b$$

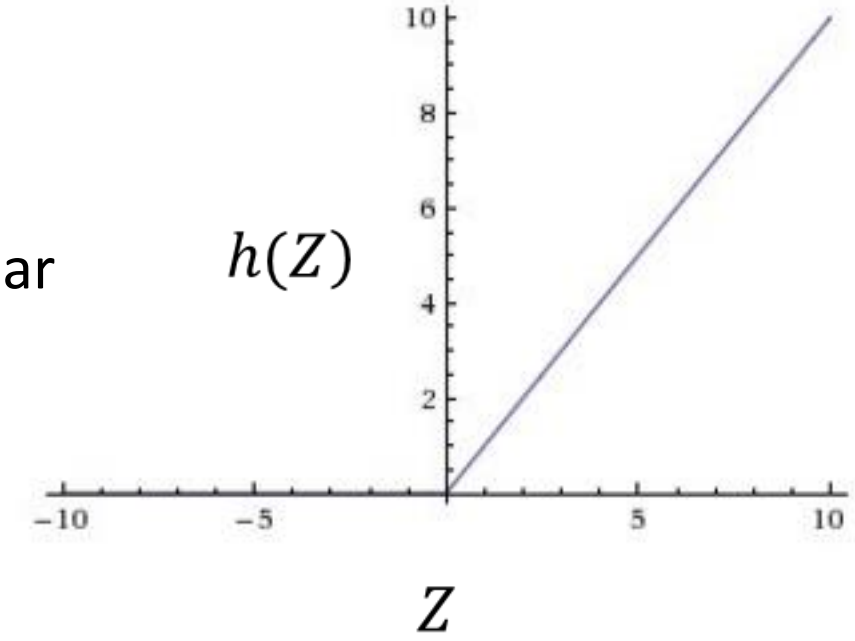
$$Y = h(Z)$$



ReLu (Rectified Linear Unit)

Useful for thresholding

Non-linear function



Training the neuron means learning the optimal w's and b

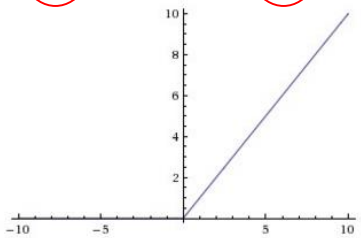
Artificial neuron can represent a motif

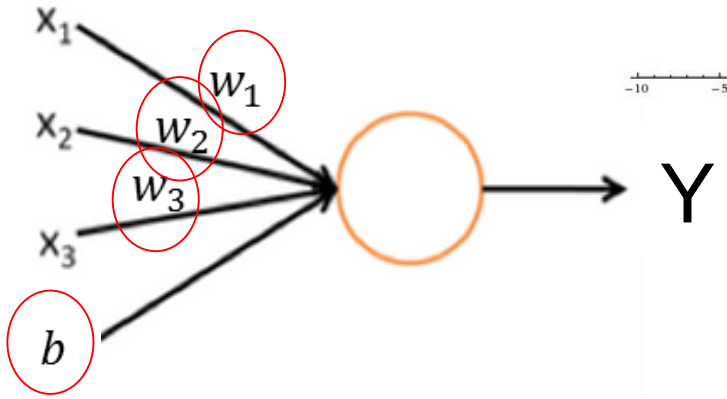
$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1.x_1 + w_2.x_2 + w_3.x_3 + b$$

$$Y = h(Z)$$

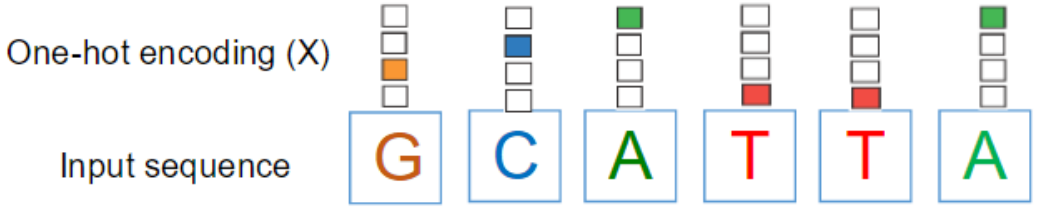




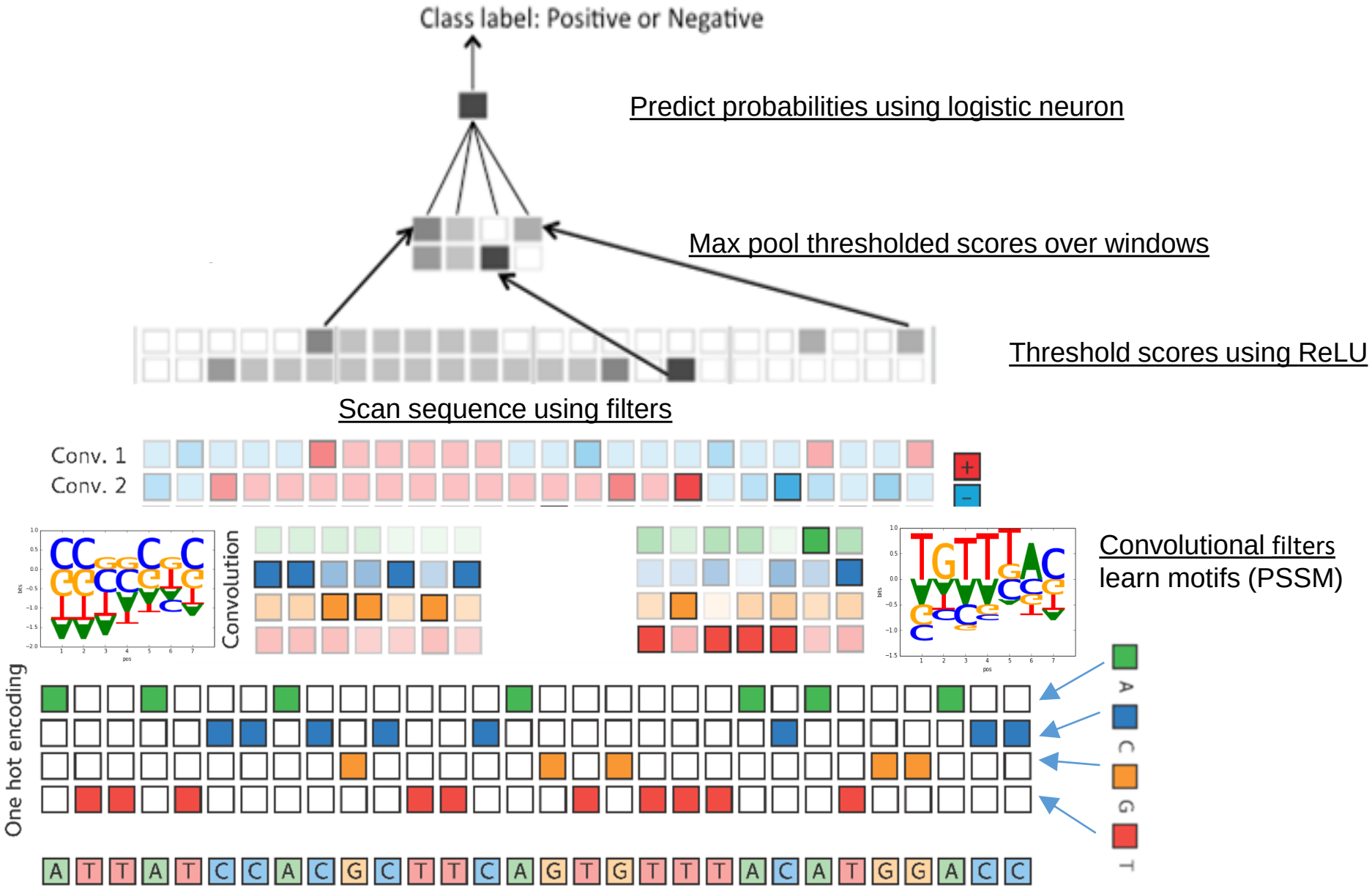
Thresholded Motif Scores max(0, W*x)	0	0	2.0	0	0	0
Motif match Scores sum(W * x)	-2.2	-5.4	2.0	-4.3	-24	-17

Scoring weights W

A	-5.7	-3.2	3.7	-3.2	3.7	0.6
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5



Biological motivation of DCNN



Deep convolutional neural network

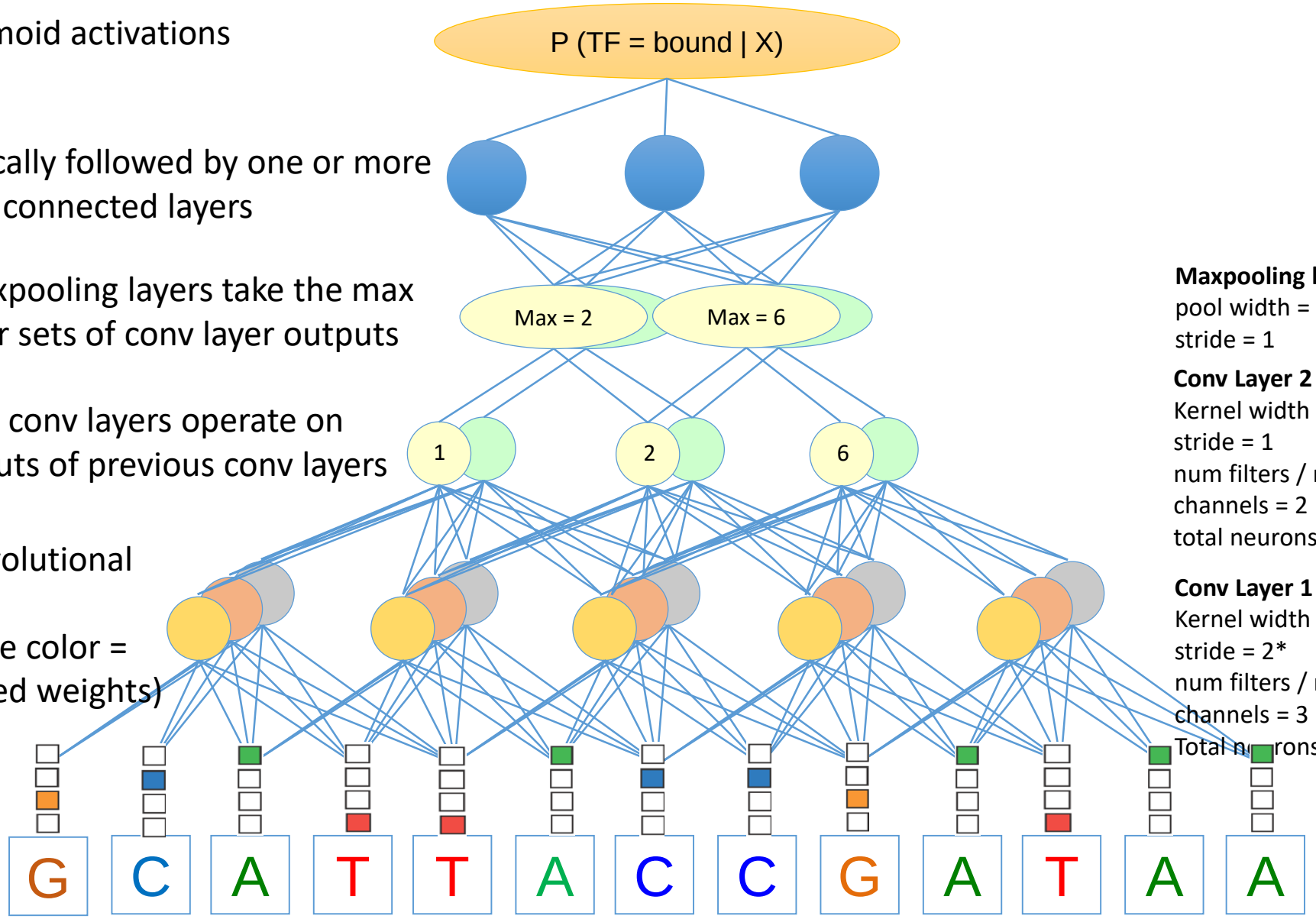
Sigmoid activations

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer (same color = shared weights)



Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num channels = 2
total neurons = 6

Conv Layer 1
Kernel width = 4
stride = 2*
num filters / num channels = 3
Total neurons = 15

***for genomics, a stride of 1 for conv layers is recommended**

Multi-task CNN

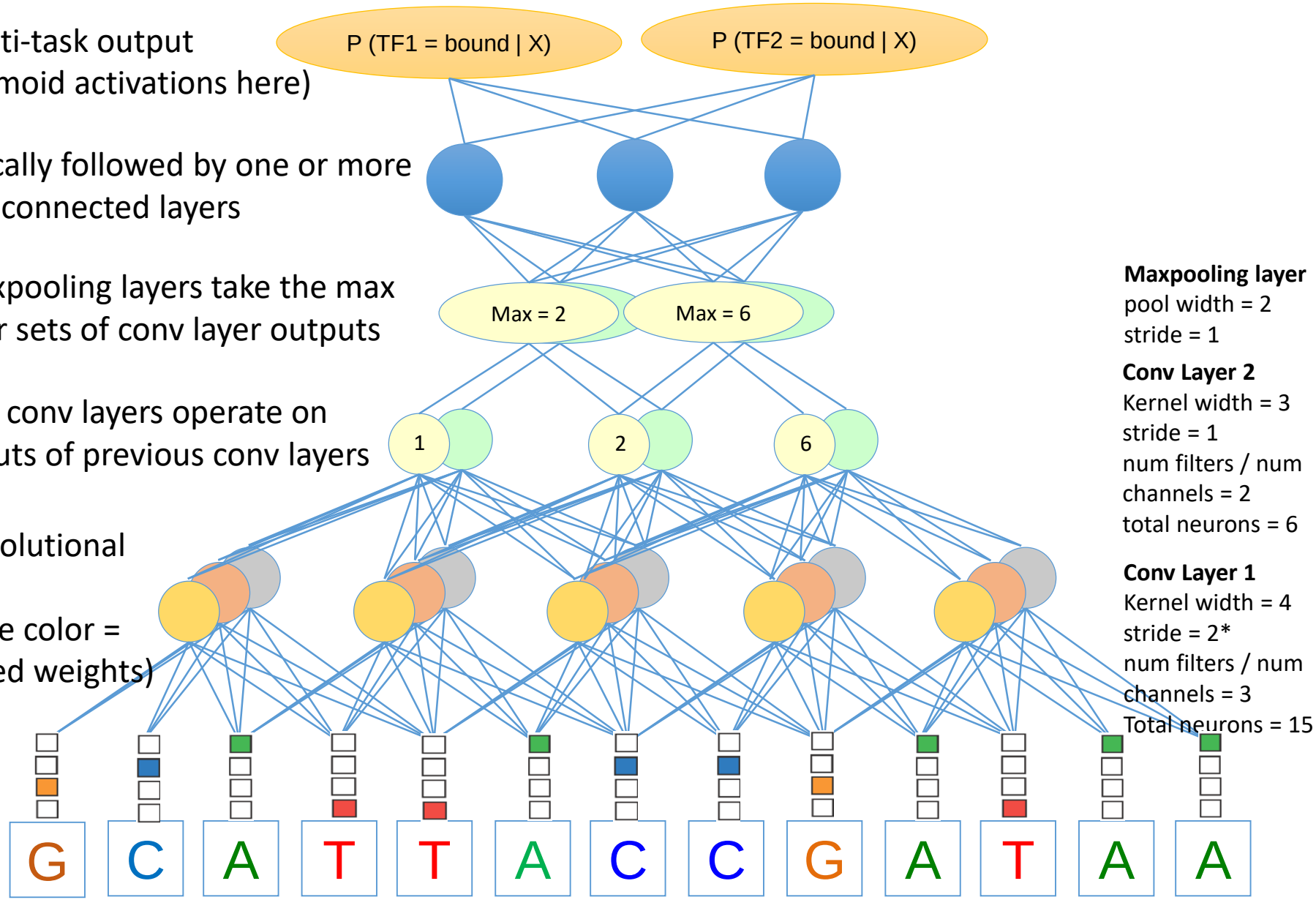
Multi-task output
(sigmoid activations here)

Typically followed by one or more
fully connected layers

Maxpooling layers take the max
over sets of conv layer outputs

Later conv layers operate on
outputs of previous conv layers

Convolutional
layer
(same color =
shared weights)



Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num
channels = 2
total neurons = 6

Conv Layer 1
Kernel width = 4
stride = 2*
num filters / num
channels = 3
Total neurons = 15

NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

日本語要約

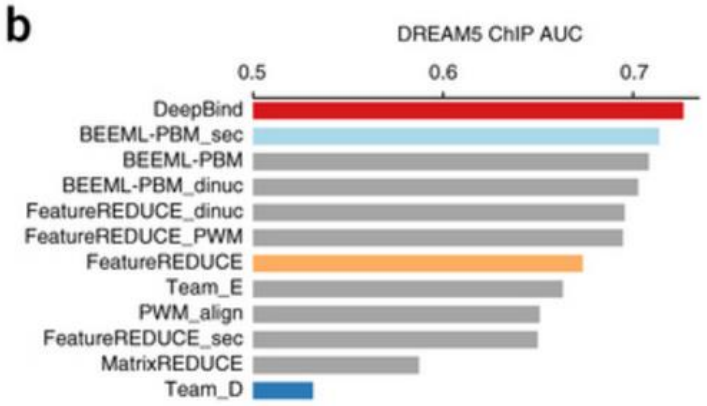
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

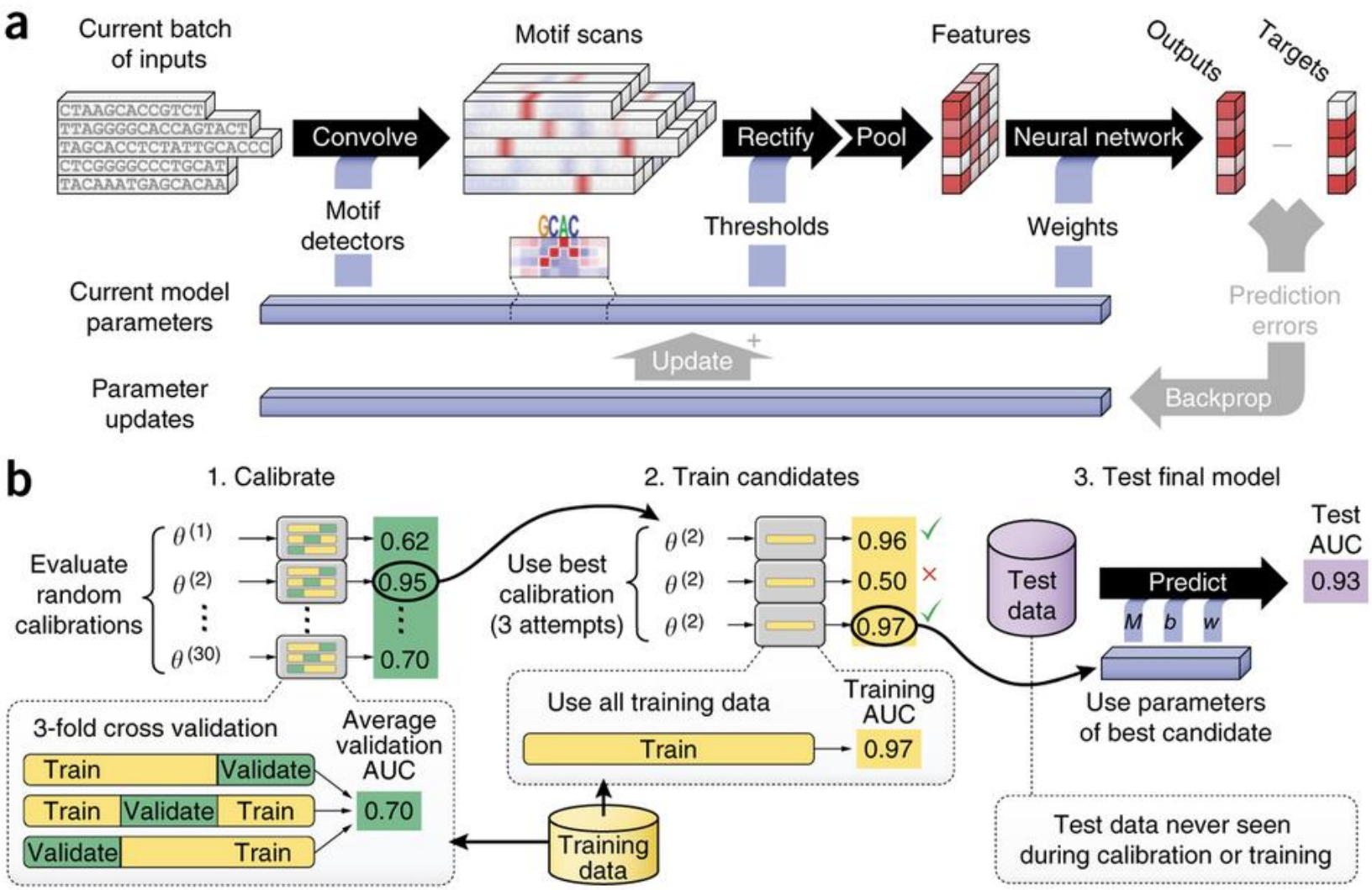
Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Affiliations | Contributions | Corresponding author

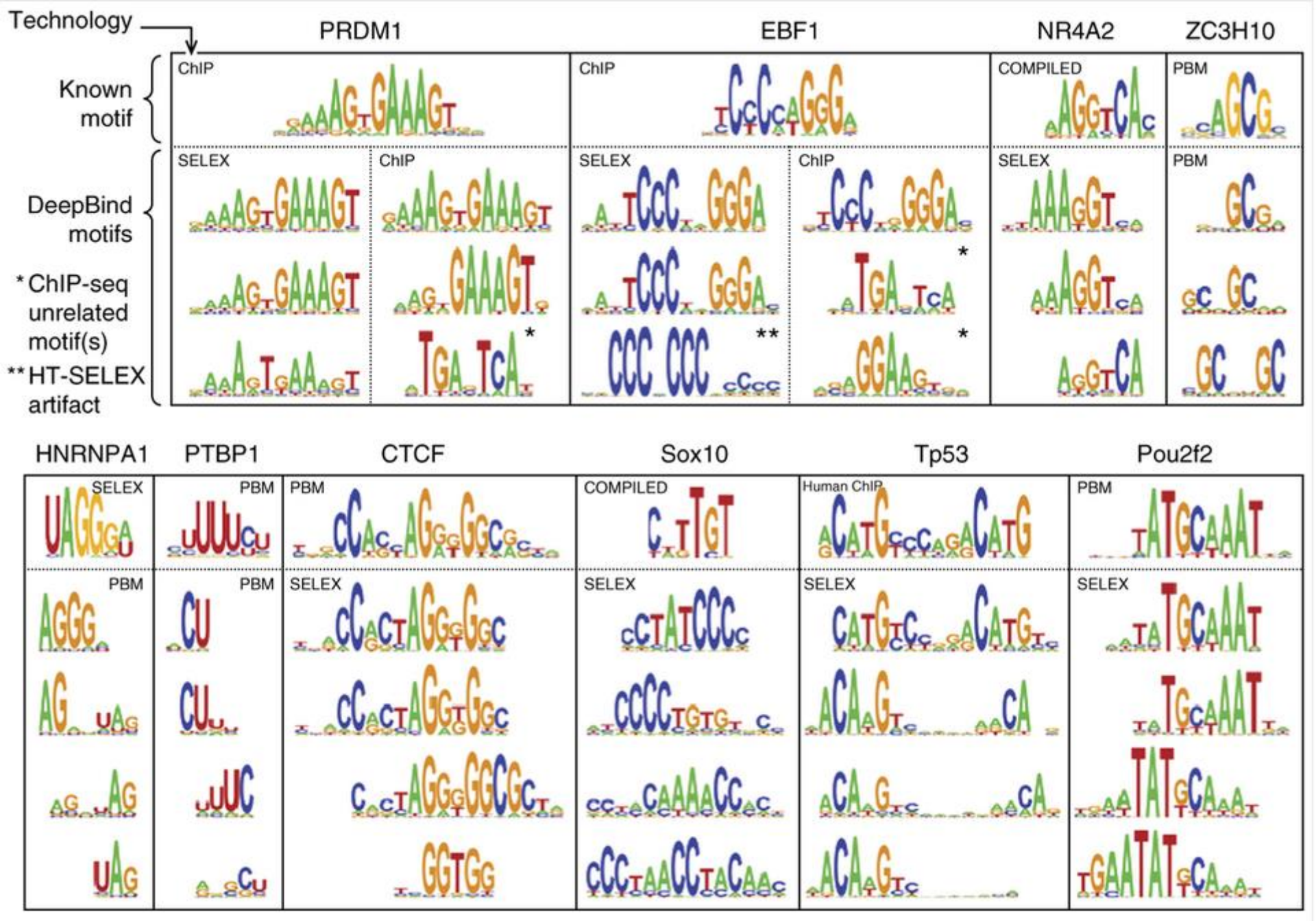
Nature Biotechnology **33**, 831–838 (2015) | doi:10.1038/nbt.3300

Received 28 November 2014 | Accepted 25 June 2015 | Published online 27 July 2015





http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html

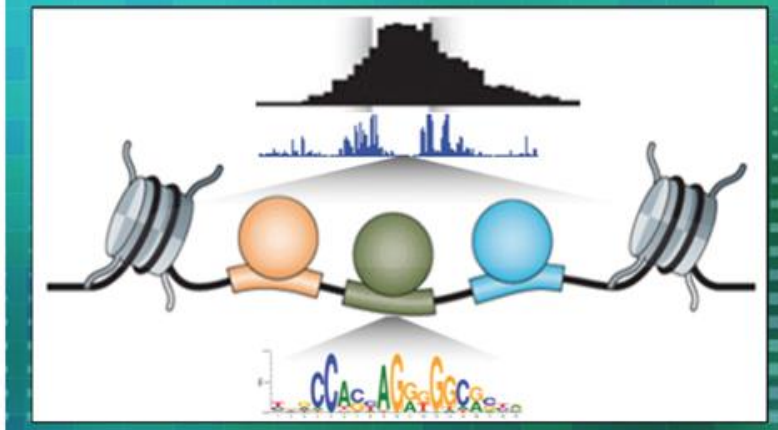

Technology
PRDM1
EBF1
NR4A2
ZC3H10
Known motif
DeepBind motifs
* ChIP-seq unrelated motif(s)
** HT-SELEX artifact
ChIP
SELEX
COMPILED
PBM
HNRNPA1
PTBP1
CTCF
Sox10
Tp53
Pou2f2
Human ChIP

Regulatory DNA sequence simulator + simple CNN models + hands tutorial

http://kundajelab.github.io/dragonn/

Many open questions on what are optimal CNN (or other deep learning) architectures for learning from DNA sequence data

ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge



DREAM CHALLENGES powered by Sage Bionetworks

Sage BIONETWORKS

Stanford University

OHSU

IBM Research

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

SCHOOL OF MEDICINE
Department of Pharmacology
UNIVERSITY OF COLORADO ANSCHUTZ MEDICAL CAMPUS

To receive email updates about this Challenge including a launch announcement, please pre-register.

Pre-register

Pre-registration open
Launch: Late June 2016
Close: September 30, 2016

http://dreamchallenges.org/

Additional optional readings

In Canvas

Name	Date Created	Date Modified	Modified By	Size
2004-LifeAndItsMolecules.pdf	11:23am	11:23am	Anshul Kundaje	637 KB
2010-Review-Genomics.pdf	11:23am	11:23am	Anshul Kundaje	549 KB
Backpropagation In Convolutional Neural Networks - DeepG...	11:19am	11:19am	Anshul Kundaje	675 KB
Guide2ConvArithmetic.pdf	11:19am	11:19am	Anshul Kundaje	879 KB
Understanding Convolutions - colah's blog.pdf	11:19am	11:19am	Anshul Kundaje	2.2 MB

https://canvas.stanford.edu/courses/51037/files/folder/LectureMaterial/Lecture2