Deep Learning Theory and Applications

Sequence-based Bioinformatics applications

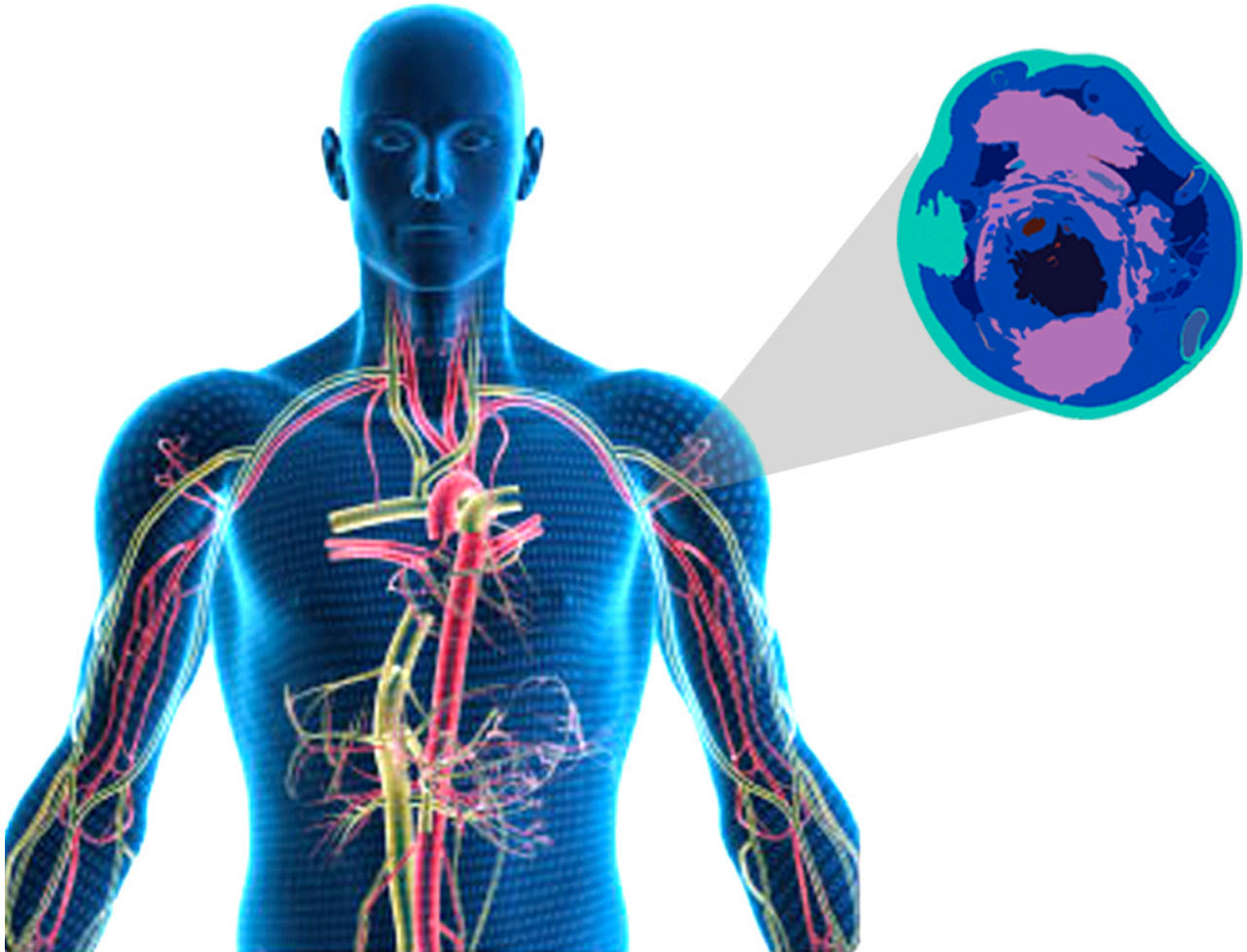
Yale

CPSC/AMTH 663

Outline

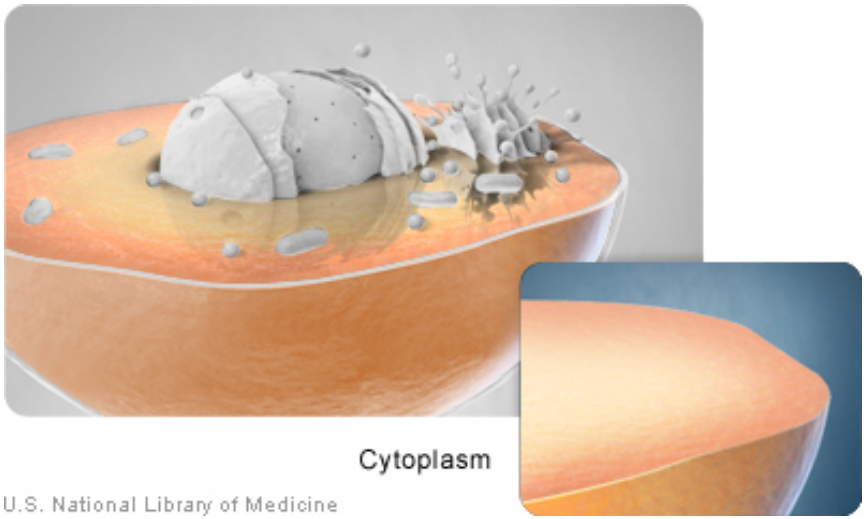
1. Autoencoders
2. Denoising Autoencoder
3. Contractive Autoencoder
4. Variational Autoencoder

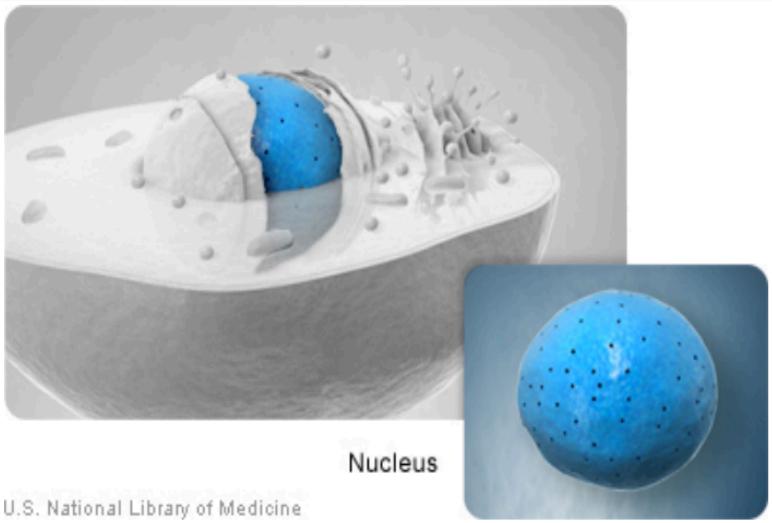
Biology Crash Course



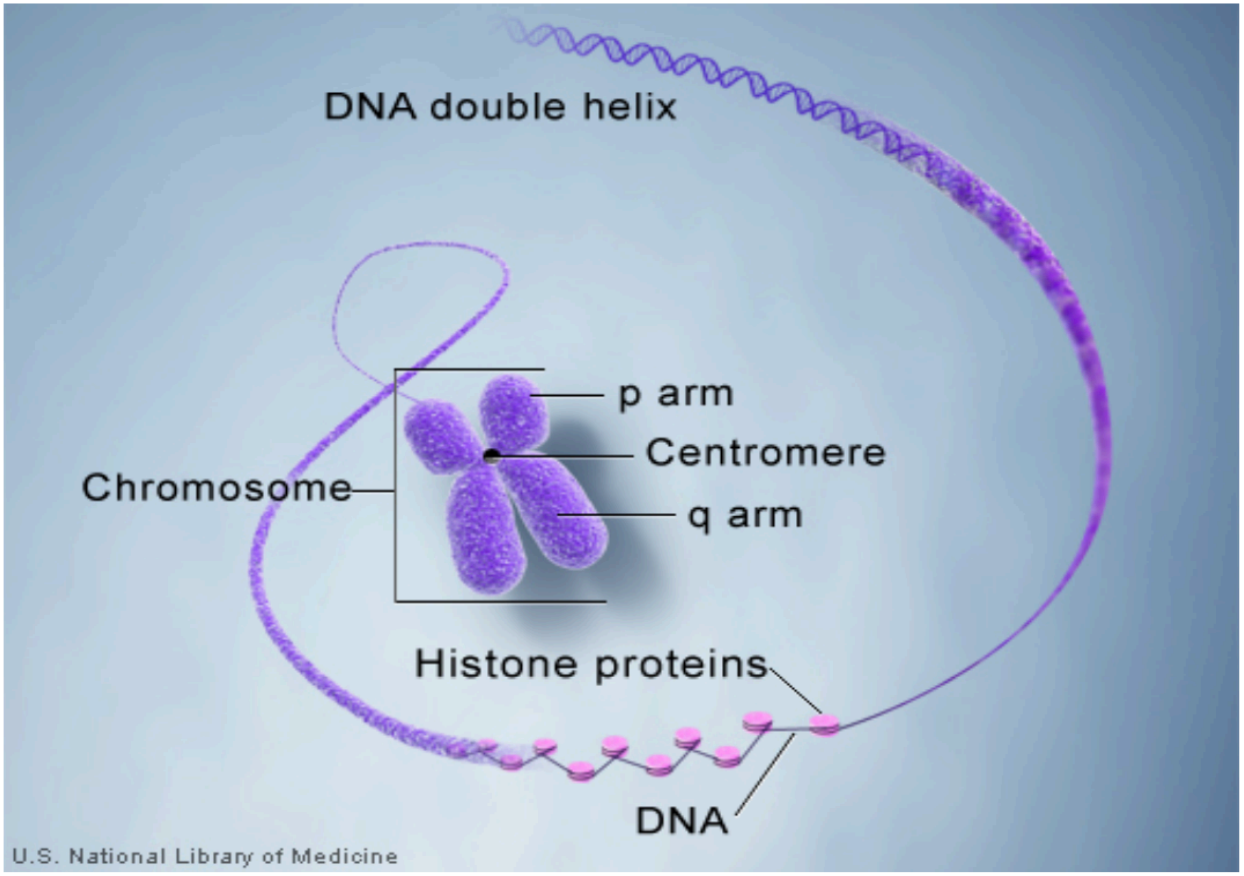
Parts of Cell

https://ghr.nlm.nih.gov/primer/basics.pdf



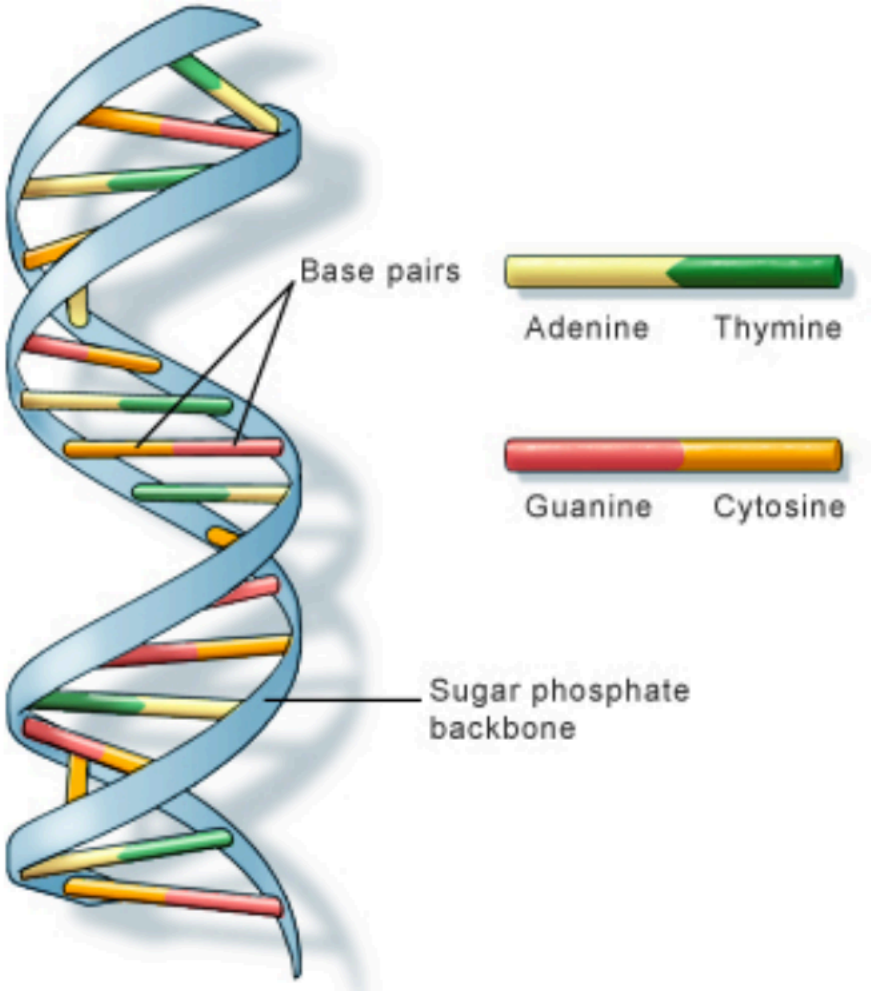


Chromosomes

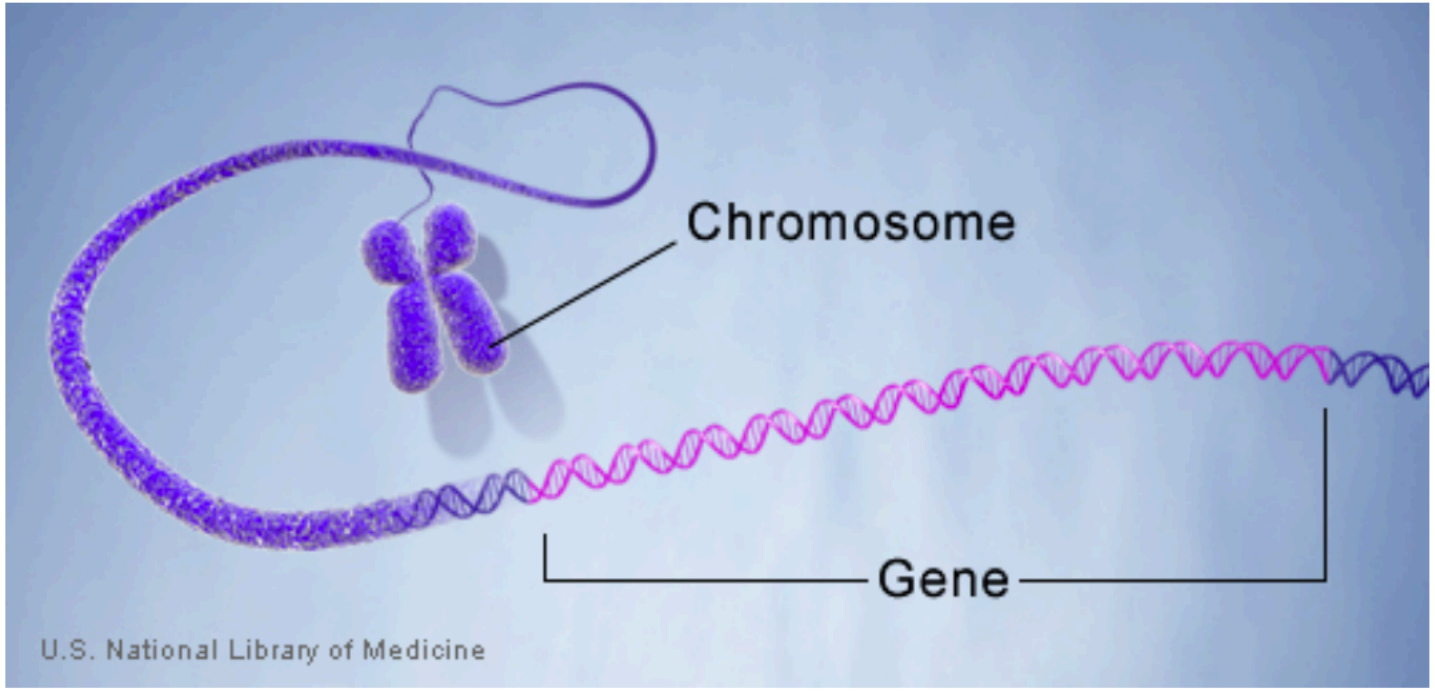


DNA and histone proteins are packaged into structures called chromosomes.

DNA

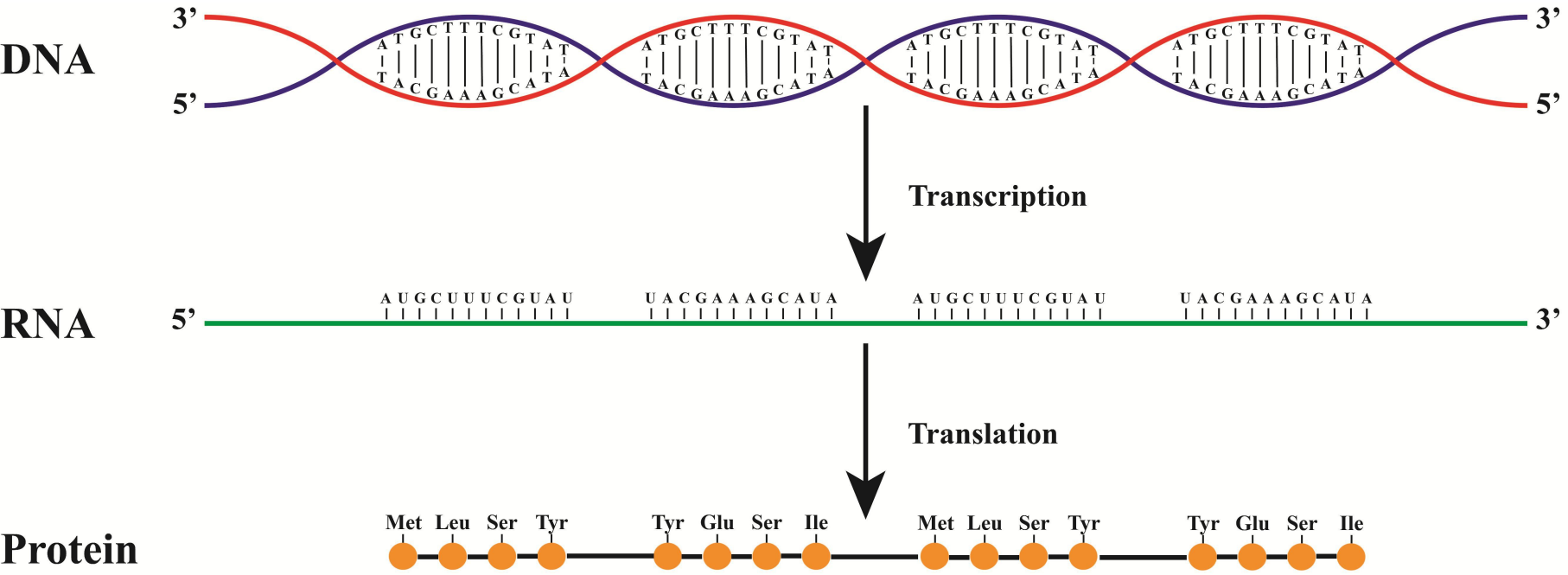

Base pairs
Adenine
Thymine
Guanine
Cytosine
Sugar phosphate backbone

Genes



Genes are made up of DNA. Each chromosome contains many genes.

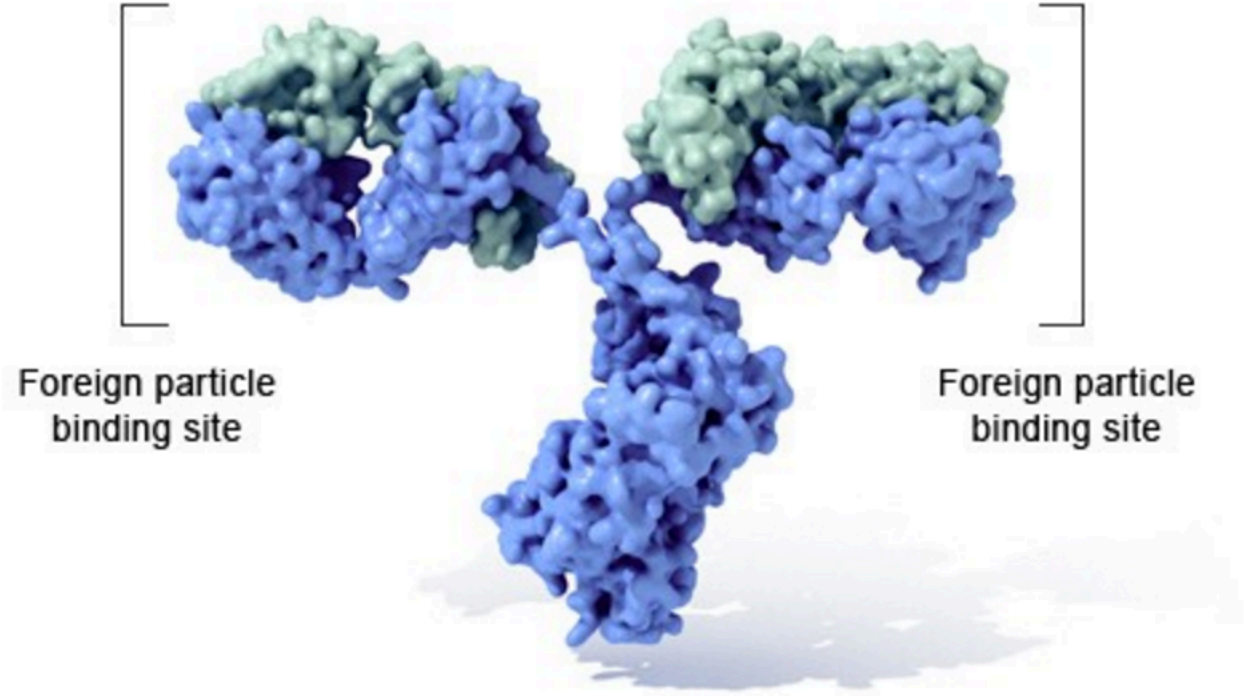
Central Dogma of Biology



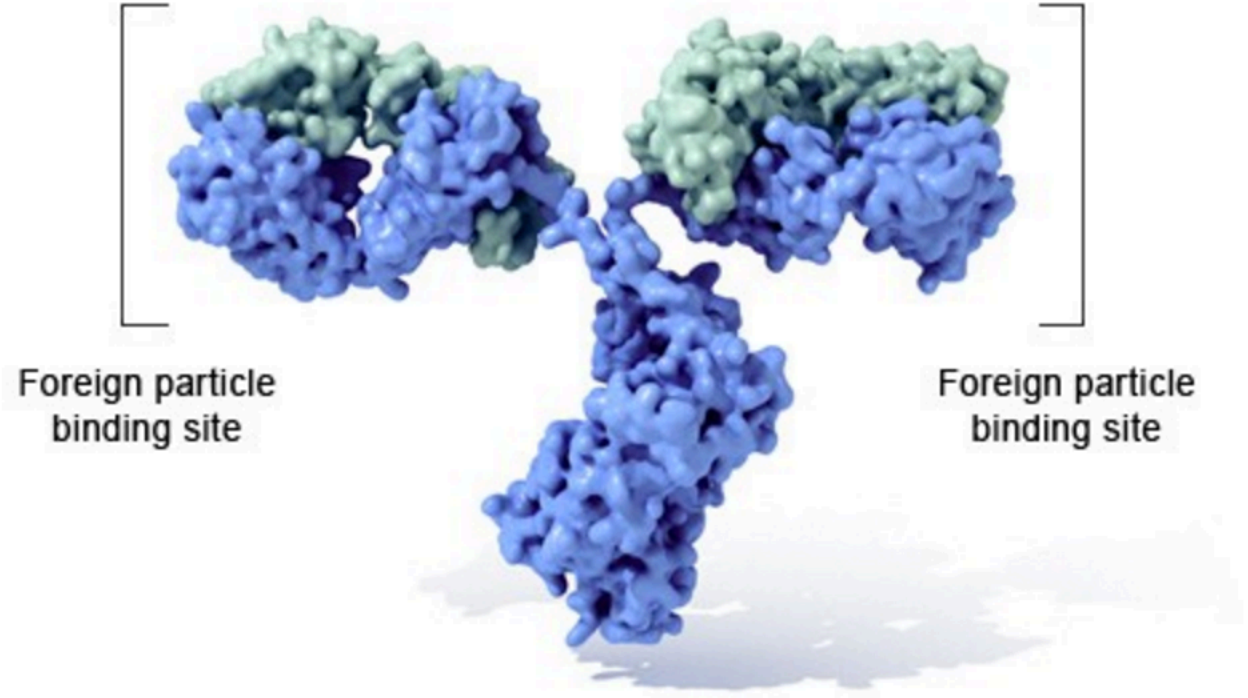
mRNA reveals what proteins the cell is making

What do proteins do?

Binds to surface receptors, help cells recognize invaders, viruses, chemical signals

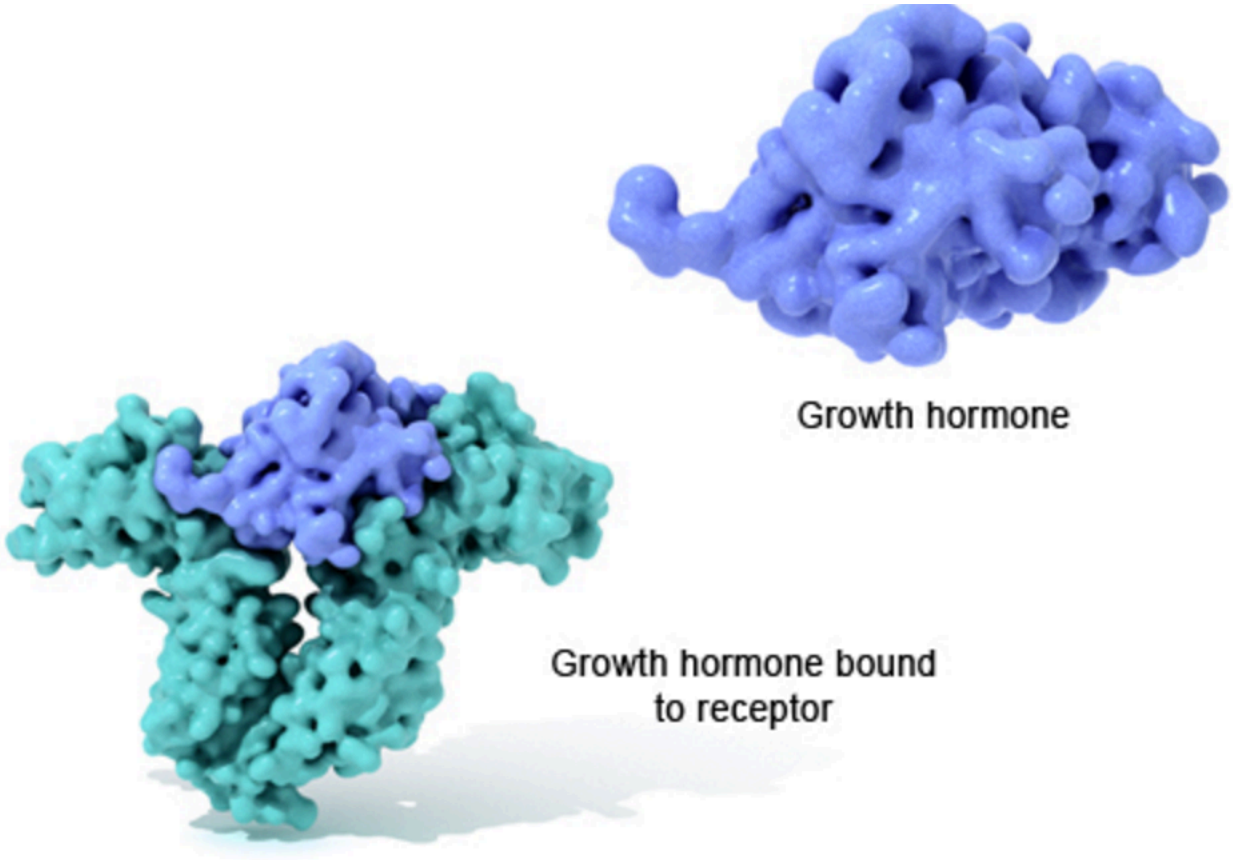


What do proteins do?



Antibodies: Binds foreign bodies, viruses, alerts immune system

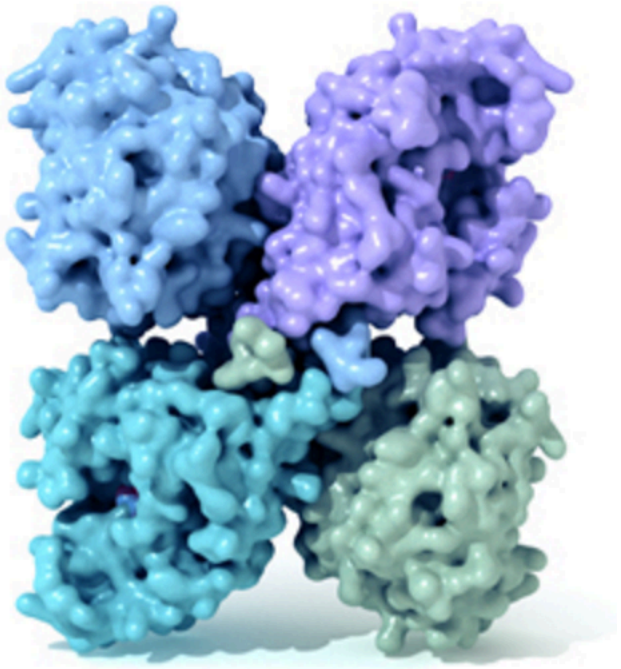
Chemical Messenger



Hormones, growth factors and other extracellular signals

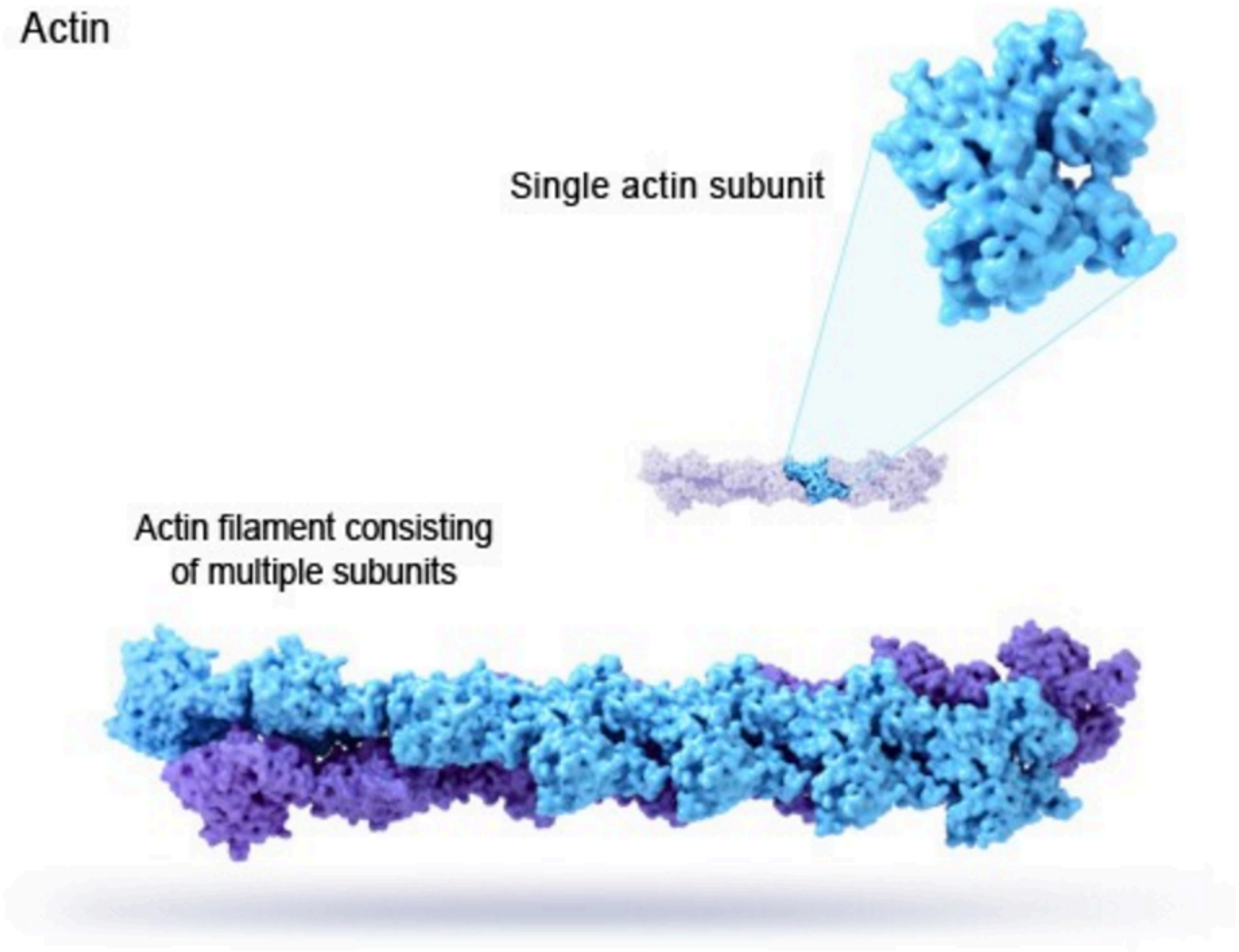
Enzyme

Carries out chemical reactions

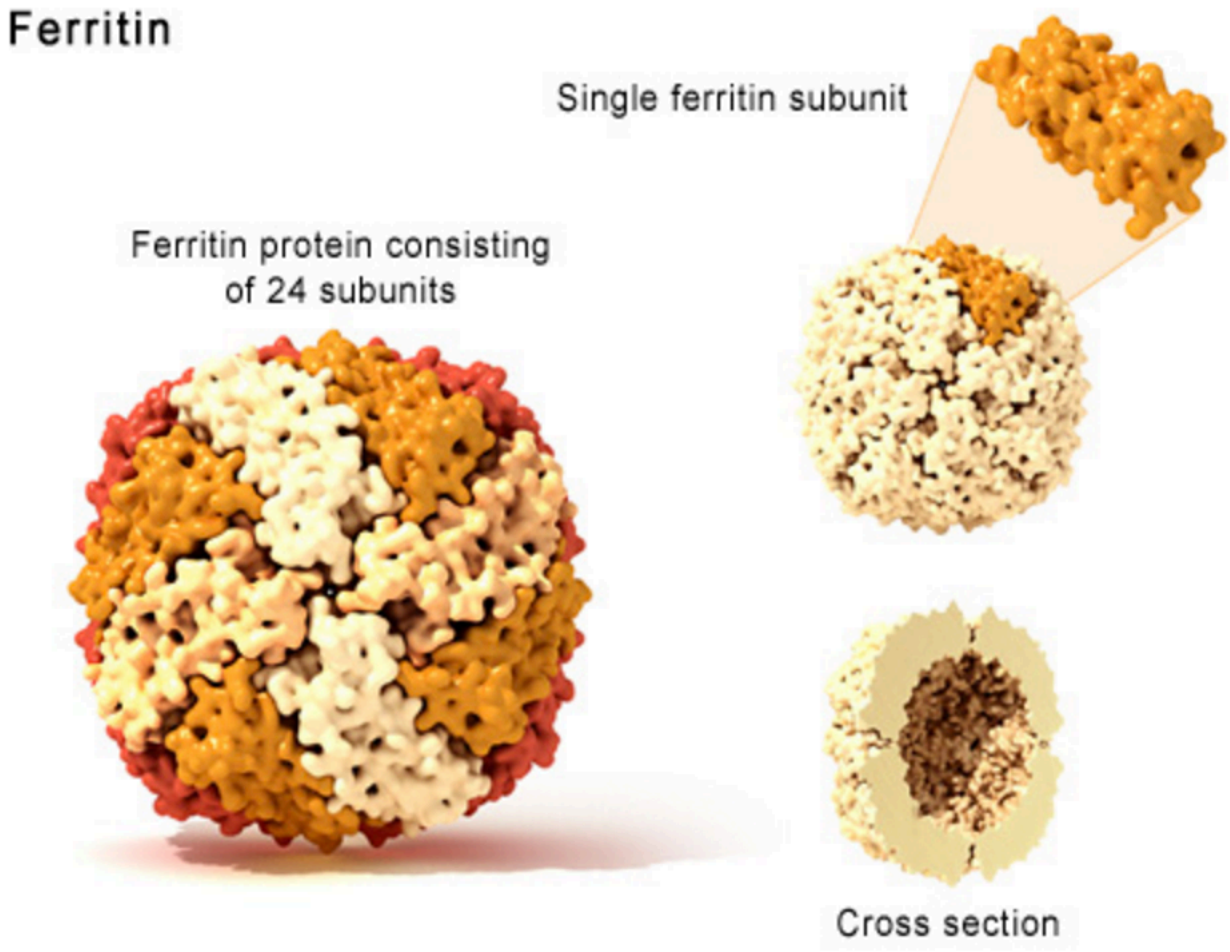


Phenylalanine hydroxylase
protein consisting of 4 subunits

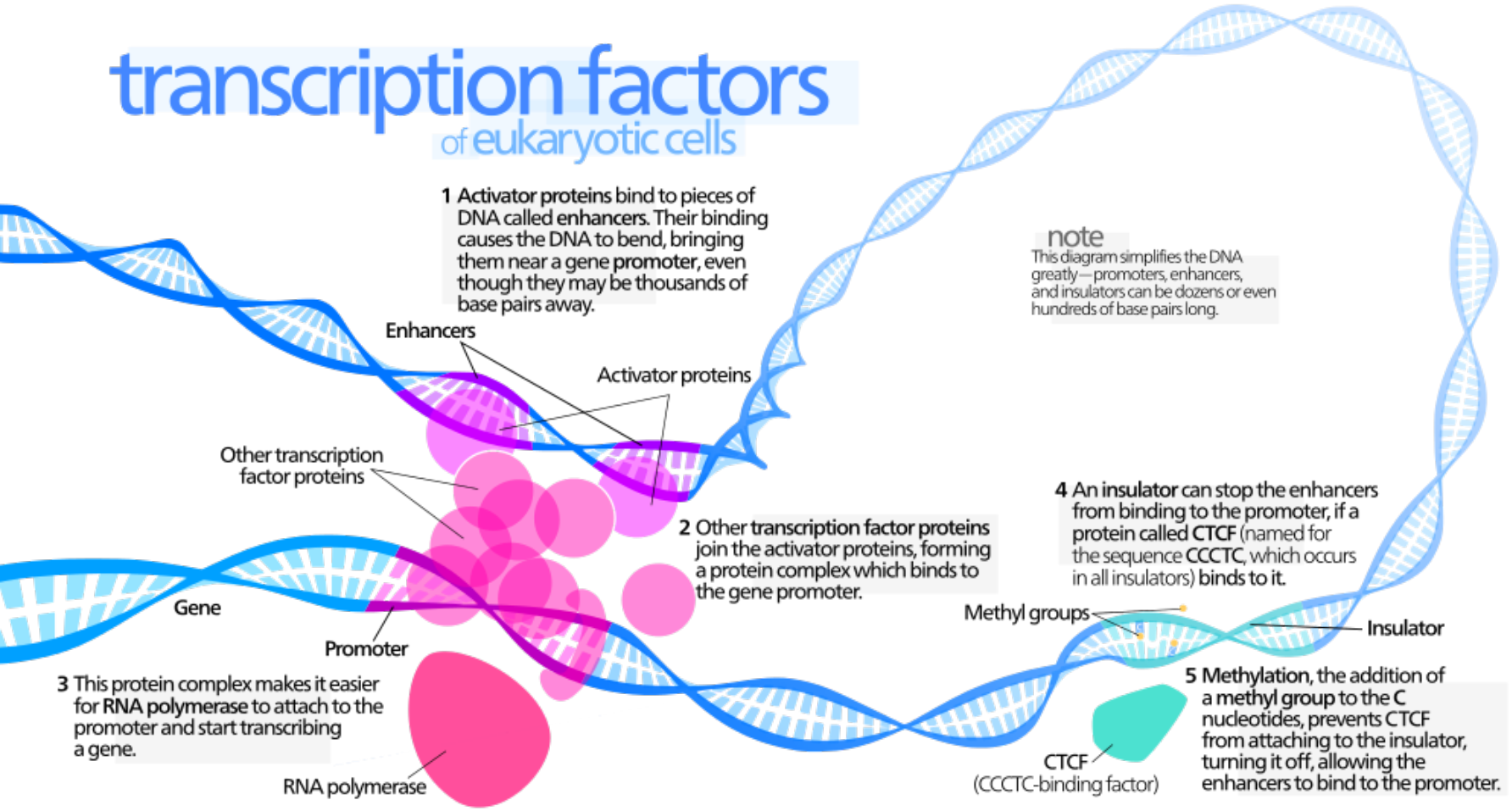
Structural Components



Transport/Storage



Proteins binding to DNA



DNA sequence understanding

- DNA and RNA-binding proteins can be ”approximately sequence specific”
- Knowing the sequence specificity can allow us to develop a model of the regulatory network in a biological system
- Mutations can disrupt the regulatory system due to this sequence specificity
- Can also identify disease-causing variants

Deep Bind

Analysis | Published: 27 July 2015

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

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

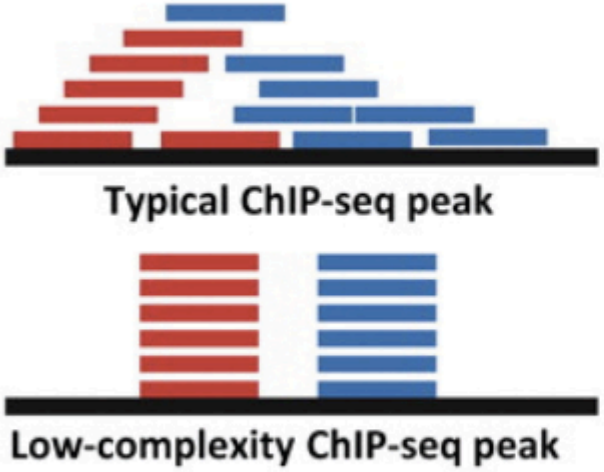
Nature Biotechnology **33**, 831–838 (2015) | Download Citation

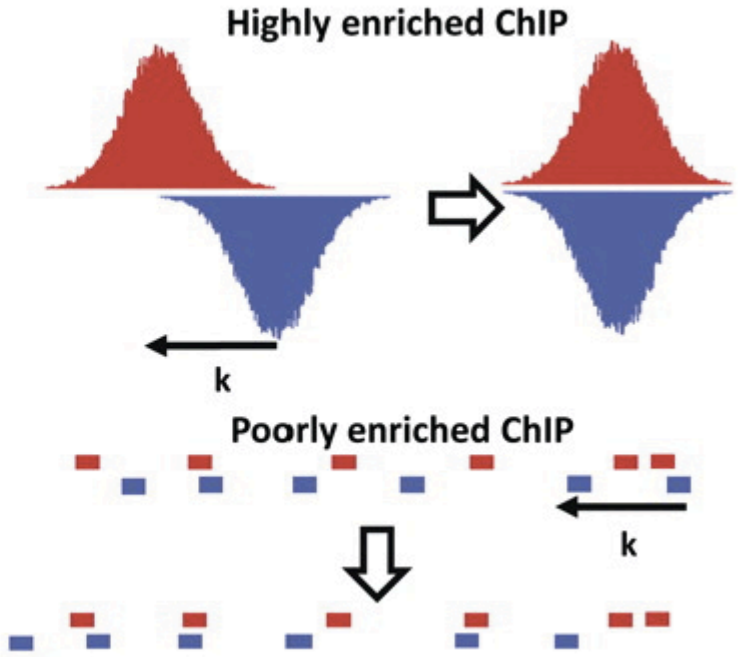
Where does binding data come from?

- Protein binding microarrays: contains a specificity coefficient for each probe sequence
- Chromatin immunoprecipitation (ChIP)-seq: provides a ranked list of putatively bound sequences of varying length
- HT-SELEX: generates a set of very high affinity sequences.
- LOTS of sequences:
  - A typical high-throughput experiment measures between 10,000 and 100,000 sequences

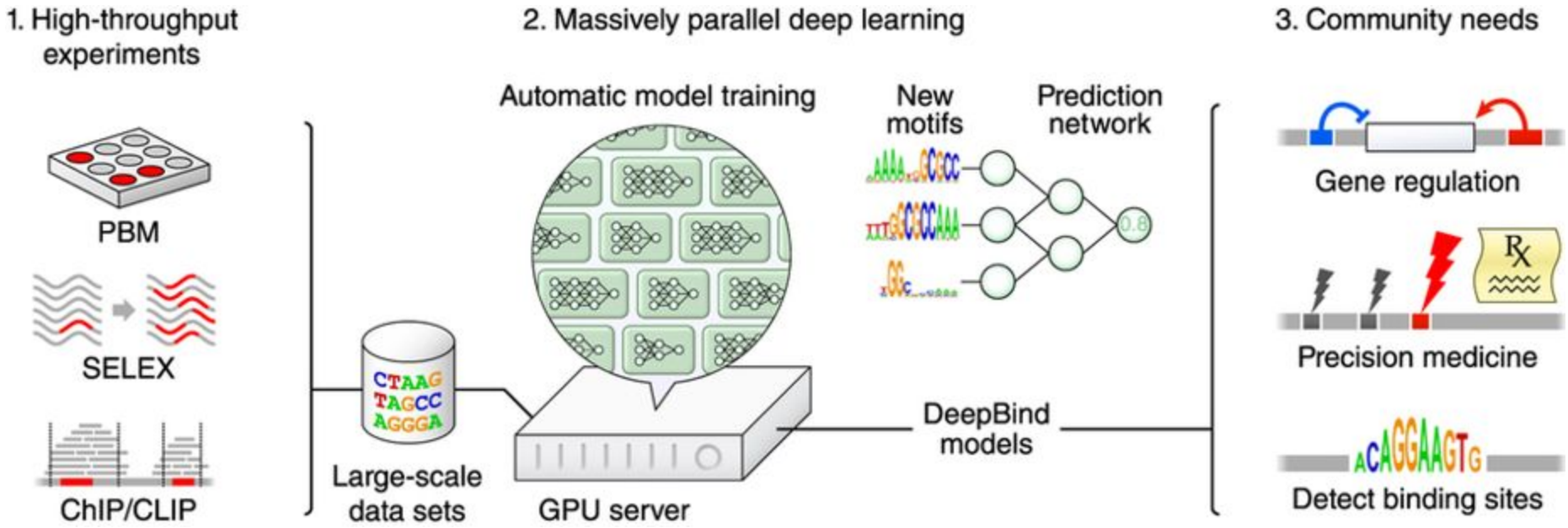
Artifacts, biases

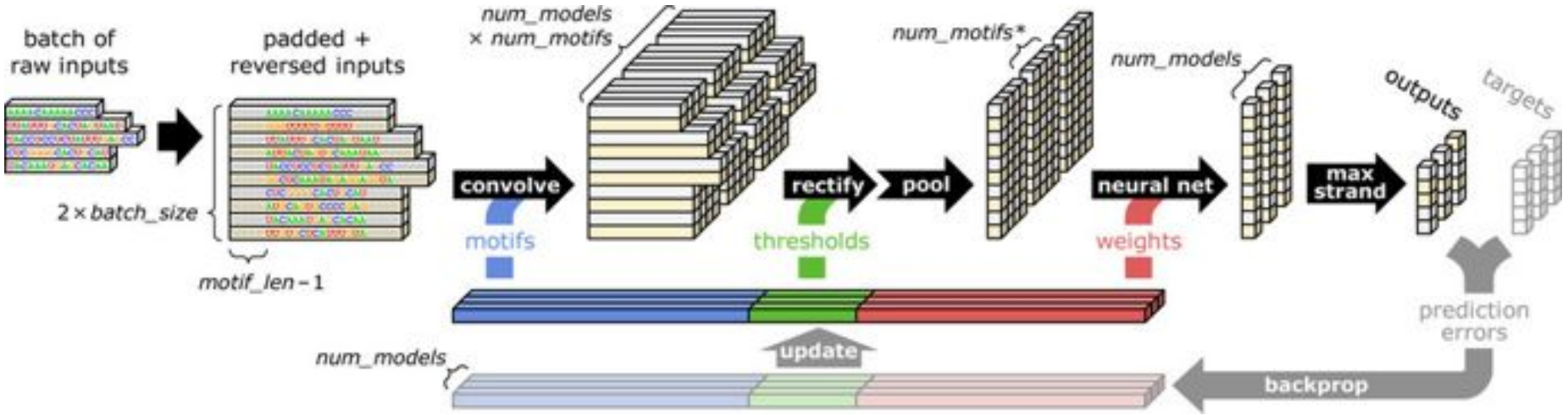
- Each data acquisition technology has its own artifacts, biases and limitations
- ChIP-seq reads often localize to “hyper-ChIPable” regions of the genome near highly expressed genes





DeepBind


1. High-throughput experiments
PBM
SELEX
ChIP/CLIP
2. Massively parallel deep learning
Automatic model training
Large-scale data sets
GPU server
New motifs
Prediction network
DeepBind models
3. Community needs
Gene regulation
Precision medicine
Detect binding sites


batch of
raw inputs
padded +
reversed inputs
2 × batch_size
motif_len – 1
num_models
× num_motifs
convolve
motifs
rectify
thresholds
pool
num_motifs*
neural net
weights
num_models
max
strand
outputs
targets
prediction
errors
update
num_models
backprop

Training

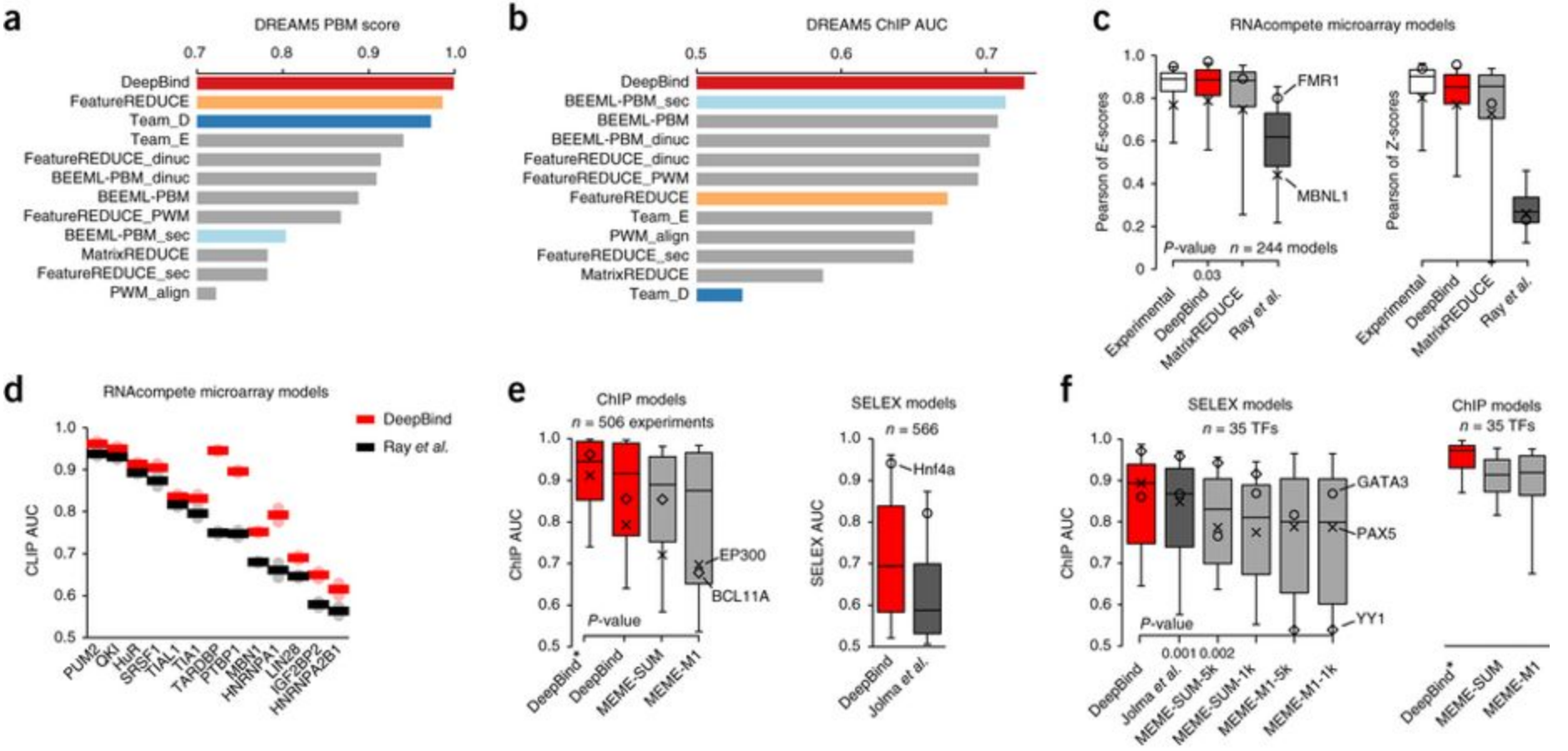
- Uses a set of sequences with experimentally determined “binding scores”
- Sequences are of varying length 14-101 nucleotides
- Binding scores can be real-valued or binary measurements

Architecture

- Convolutional stage: scans a set of ***motif detectors*** across the sequence space
  - Each motif detector is a 4xm matrix
- Rectification stage isolates positions with good pattern match by shifting the response detector of M_k by b_k and clamping negative values to 0
- Pooling stage computes maximum and average of each motif detector's rectified response across the sequence
  - Maximizing helps identify the presence of longer motifs
  - Averaging helps identify cumulative effect of shot motifs
- Values are fed into non-linear neural network which combines responses to produce score

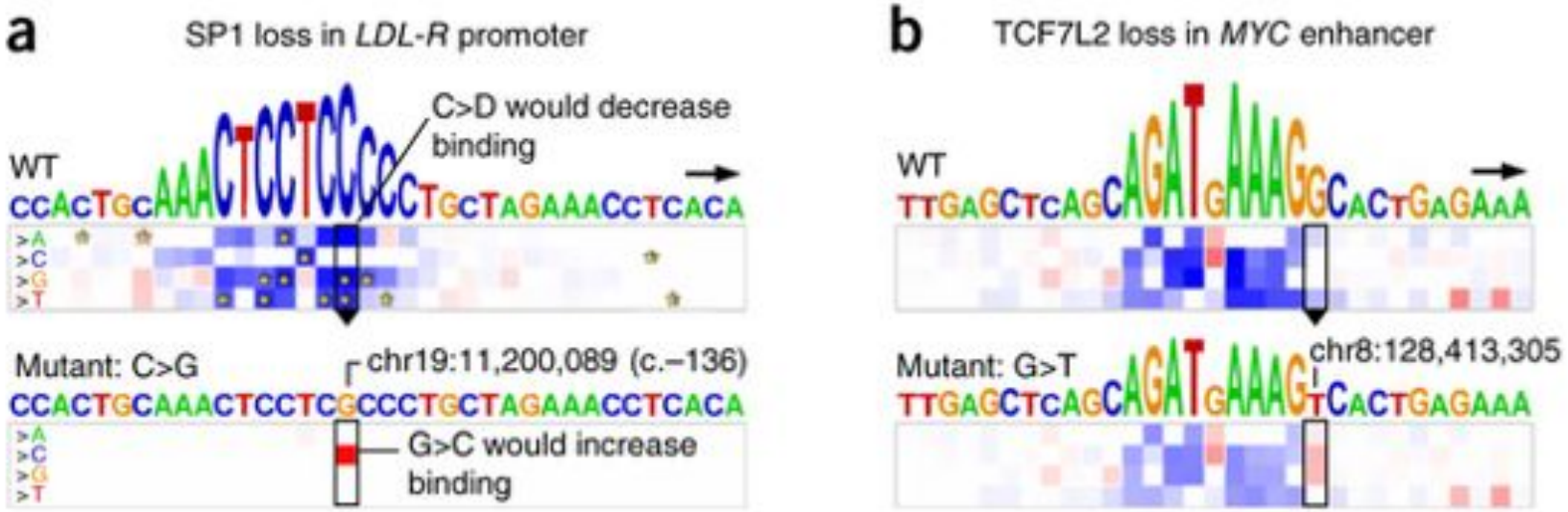
Repository

- (http://tools.genes.toronto.edu/deepbind/)
- 927 DeepBind models representing 538 distinct transcription factors and 194 distinct RBPs


a
DREAM5 PBM score
0.7
0.8
0.9
1.0
DeepBind
FeatureREDUCE
Team_D
Team_E
FeatureREDUCE_dinuc
BEEML-PBM_dinuc
BEEML-PBM
FeatureREDUCE_PWM
BEEML-PBM_sec
MatrixREDUCE
FeatureREDUCE_sec
PWM_align
b
DREAM5 ChIP AUC
0.5
0.6
0.7
DeepBind
BEEML-PBM_sec
BEEML-PBM
BEEML-PBM_dinuc
FeatureREDUCE_dinuc
FeatureREDUCE_PWM
FeatureREDUCE
Team_E
PWM_align
FeatureREDUCE_sec
MatrixREDUCE
Team_D
c
RNAcompete microarray models
Pearson of E-scores
FMR1
MBNL1
P-value
n = 244 models
0.03
Experimental
DeepBind
MatrixREDUCE
Ray et al.
Pearson of Z-scores
d
RNAcompete microarray models
DeepBind
Ray et al.
CLIP AUC
PUM2
QKI
HuR
SRSF1
TIAL1
TIA1
TARDBP
PTBP1
MBNL1
HNRNPA1
LIN28
IGF2BP2
HNRNPA2B1
e
ChIP models
n = 506 experiments
ChIP AUC
EP300
BCL11A
P-value
DeepBind*
DeepBind
MEME-SUM
MEME-M1
SELEX models
n = 566
SELEX AUC
Hnf4a
DeepBind
Jolma et al.
f
SELEX models
n = 35 TFs
ChIP AUC
GATA3
PAX5
YY1
P-value
0.001
0.002
DeepBind
Jolma et al.
MEME-SUM-5k
MEME-SUM-1k
MEME-M1-5k
MEME-M1-1k
ChIP models
n = 35 TFs
DeepBind*
MEME-SUM
MEME-M1

Identifying damaging genetic variants

- ‘mutation map’ illustrates the effect that every possible point mutation in a sequence may have on binding affinity



Visualization shows how important each base is for the analysis by the height of the Base letter

Heat map indicates how much mutation will increase or decrease binding score

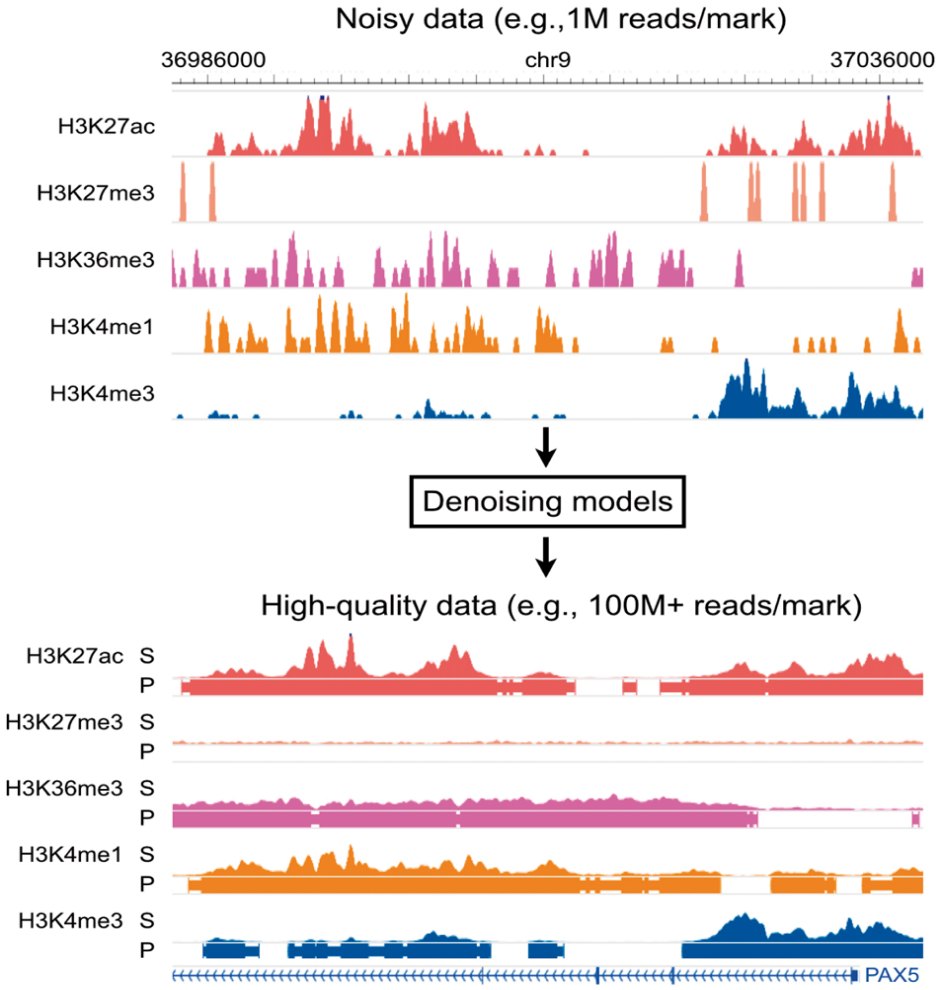
Coda

Denoising genome-wide histone ChIP-seq with convolutional neural networks

Pang Wei Koh, Emma Pierson, Anshul Kundaje ✉ Author Notes

Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages i225–i233,
https://doi.org/10.1093/bioinformatics/btx243
Published: 12 July 2017

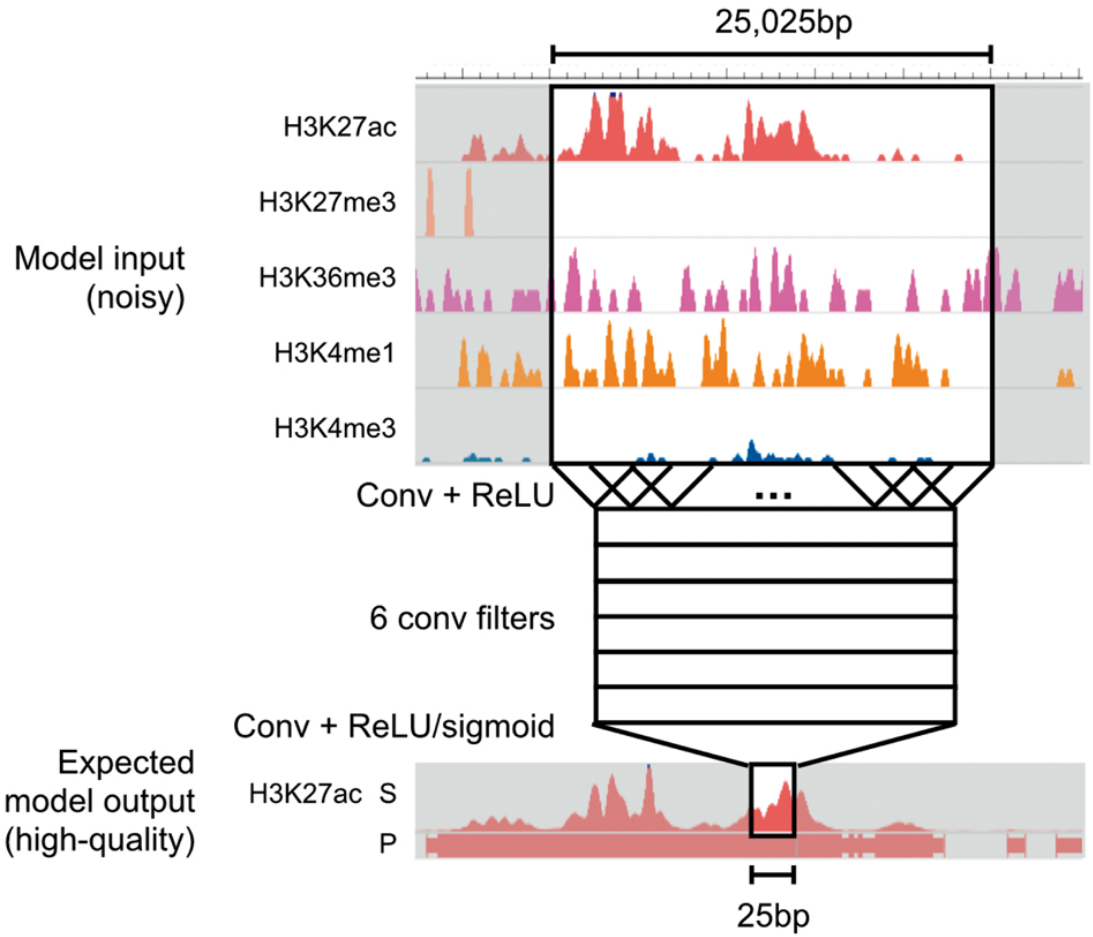
Denoising Chip-seq reads



Training data

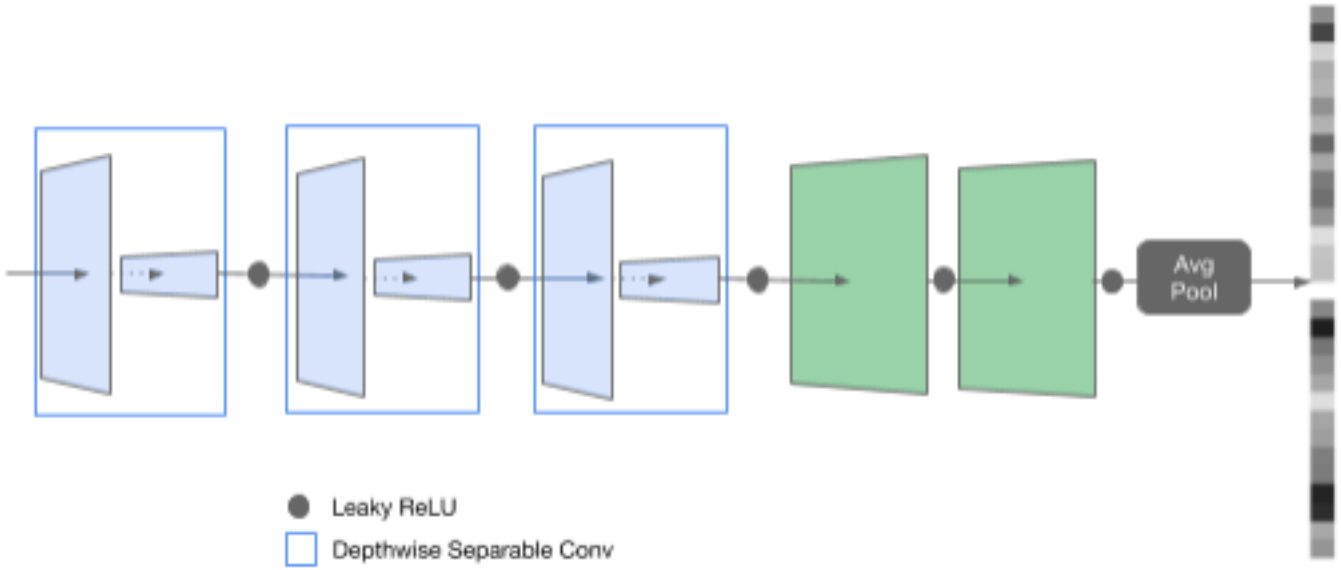
- Pairs of matching ChIP-seq datasets of the same histone modifications in the same cell type
  - one high-quality a
  - the other noisy
- The noisy dataset used in training can be derived:
  - computationally (e.g. by subsampling the high-quality data)
  - experimentally (e.g. by conducting the same ChIP-seq experiment with fewer input cells).

Architecture

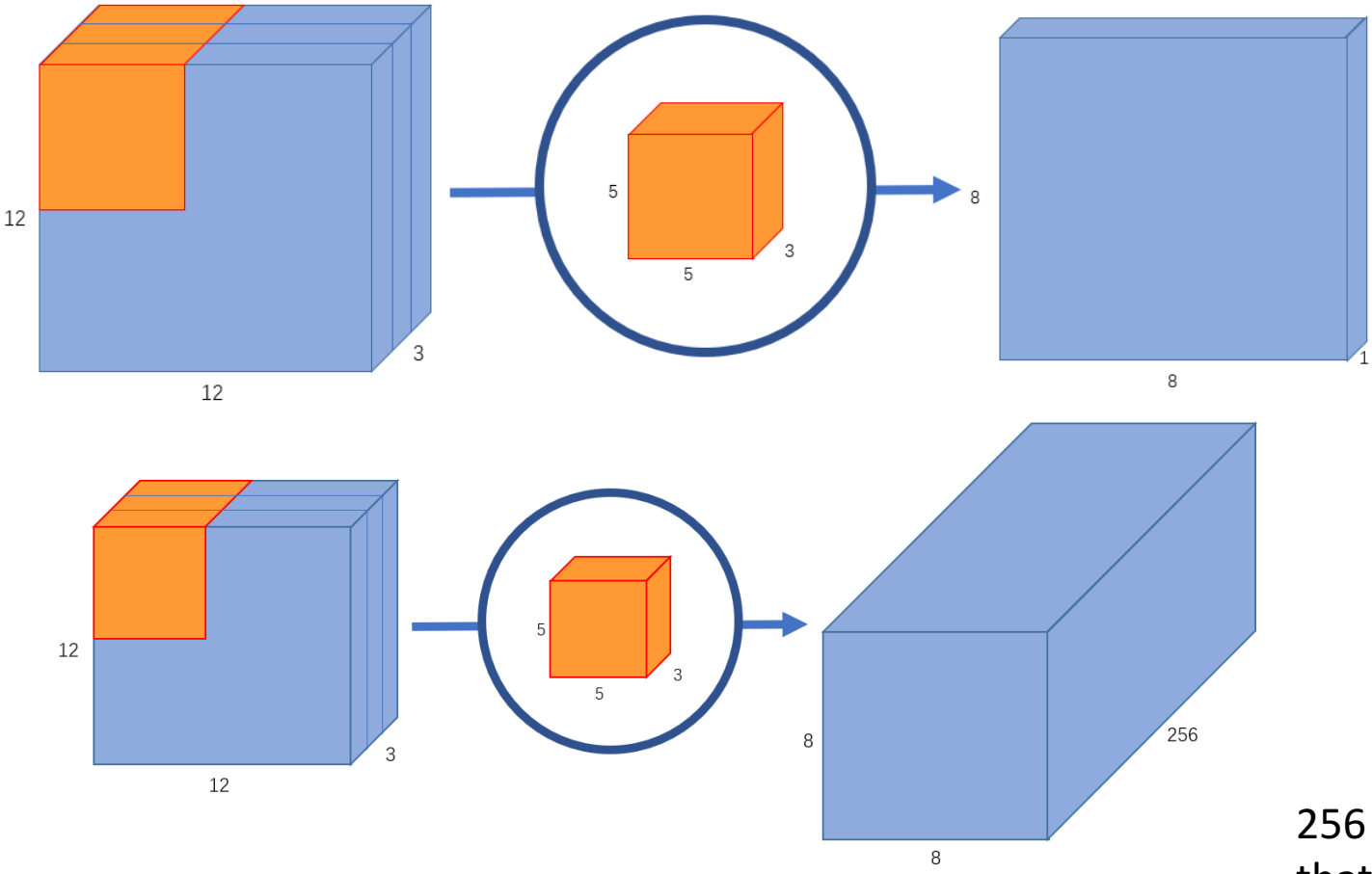


Alternative architectures for this task?

- Many important problems in bioinformatics can be framed as determining a mapping from short biological query sequences to salient categorical or numerical labels
- Examples include:
  - Taxonomic classification or binning;
  - prediction of protein function,
  - gene properties, or pathogenicity;
  - read filtering for contaminants;
  - RNA-seq quantification

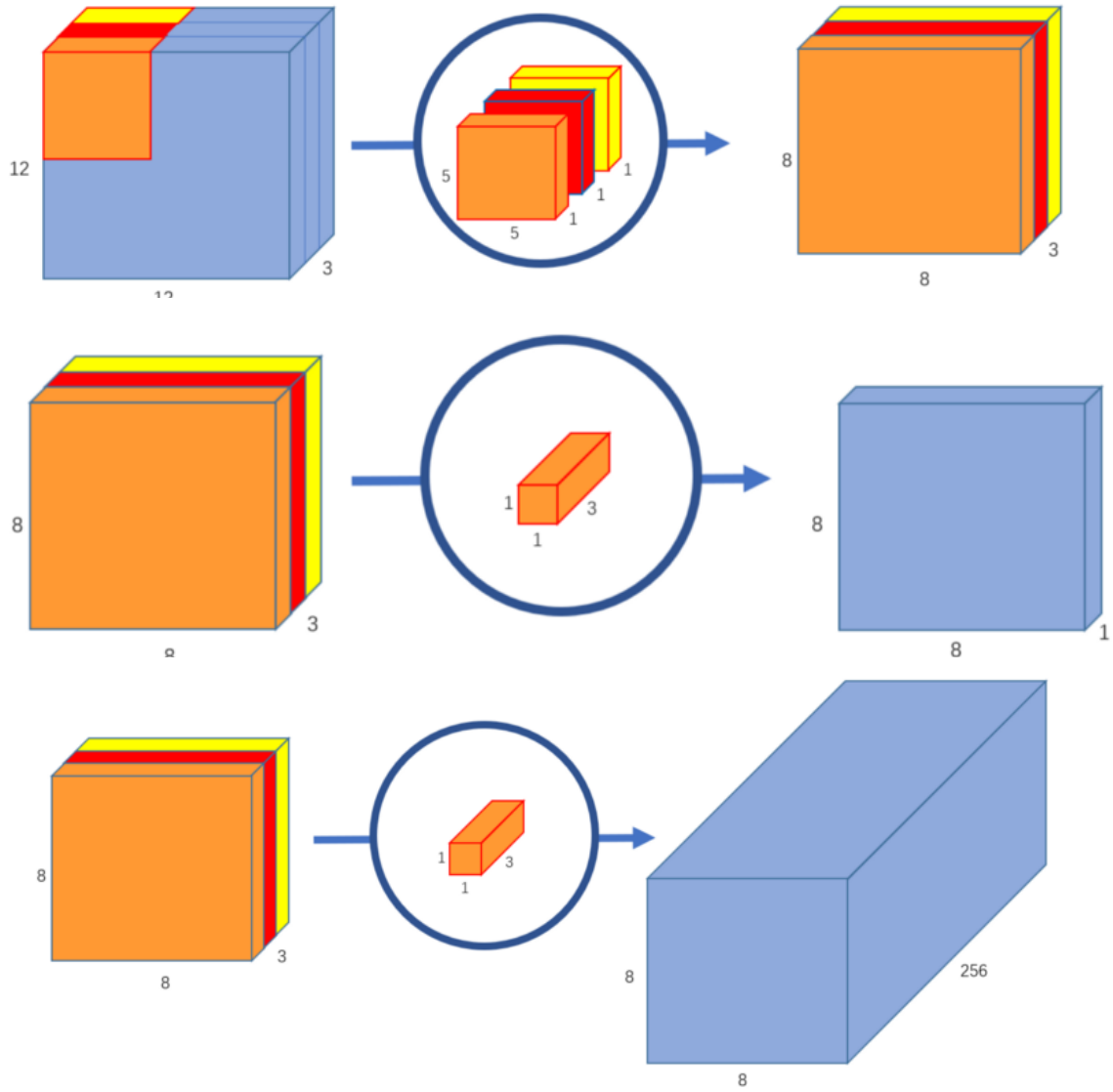

Avg
Pool
Leaky ReLU
Depthwise Separable Conv

Normal Convolution



256 5x5x3 kernels
that move 8x8 times
1228800 multiplications

Depthwise Convolution



[Howard 2017]

Depthwise

3 5x5X1 kernels that move
8x8 times = 4800

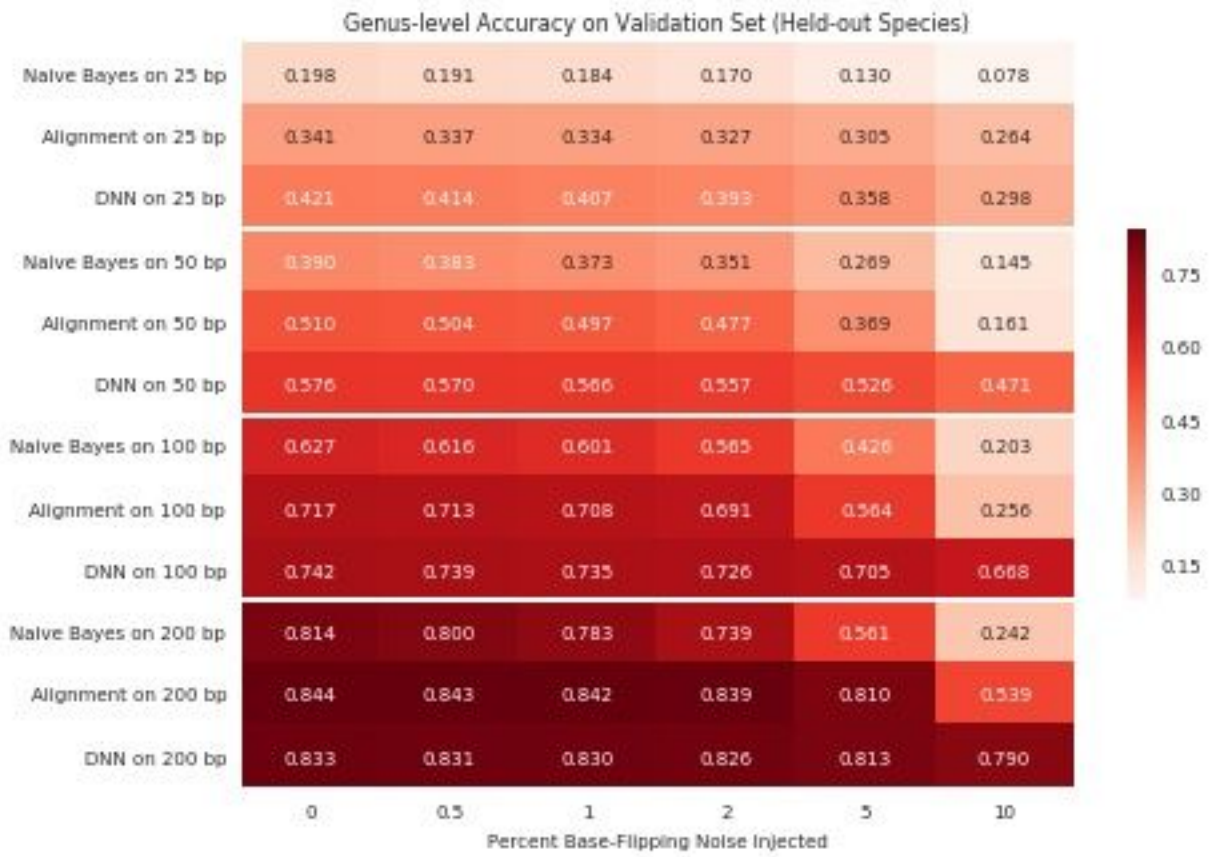
Pointwise

256 1x1x3 kernels moving
8x8 times = 49152

Total of 53952 multiplications

Ribosomal RNA

- 16S sequencing data
- Predicts species level taxonomy of microbiota
- Human Microbiome Project contains more than 14 terabytes of data and thousands of taxonomically characterized communities
- Database Project (RDP) SeqMatch and QIIME typically rely on explicit sequence matching against identified genomic sequences via a k-mer
  - Basic Local Alignment Search Tool


Genus-level Accuracy on Validation Set (Held-out Species)
Naïve Bayes on 25 bp
0.198
0.191
0.184
0.170
0.130
0.078
Alignment on 25 bp
0.341
0.337
0.334
0.327
0.305
0.264
DNN on 25 bp
0.421
0.414
0.407
0.393
0.358
0.298
Naïve Bayes on 50 bp
0.390
0.383
0.373
0.351
0.269
0.145
Alignment on 50 bp
0.510
0.504
0.497
0.477
0.369
0.161
DNN on 50 bp
0.576
0.570
0.566
0.557
0.526
0.471
Naïve Bayes on 100 bp
0.627
0.616
0.601
0.565
0.426
0.203
Alignment on 100 bp
0.717
0.713
0.708
0.691
0.564
0.256
DNN on 100 bp
0.742
0.739
0.735
0.726
0.705
0.668
Naïve Bayes on 200 bp
0.814
0.800
0.783
0.739
0.561
0.242
Alignment on 200 bp
0.844
0.843
0.842
0.839
0.810
0.539
DNN on 200 bp
0.833
0.831
0.830
0.826
0.813
0.790
0
0.5
1
2
5
10
Percent Base-Flipping Noise Injected
0.75
0.60
0.45
0.30
0.15

Further reading

- CODA https://academic.oup.com/bioinformatics/article/33/14/i225/3953958
- DeepBind https://www.nature.com/articles/nbt.3300
- MobileNet https://arxiv.org/abs/1704.04861
- Separable convolutions: https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728