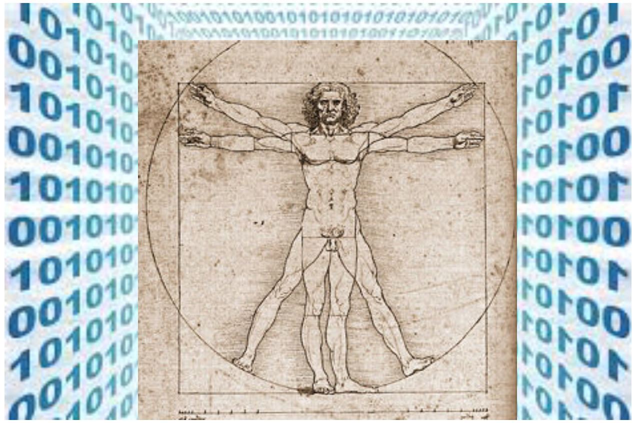


Introduction to deep learning applications in biomedical data

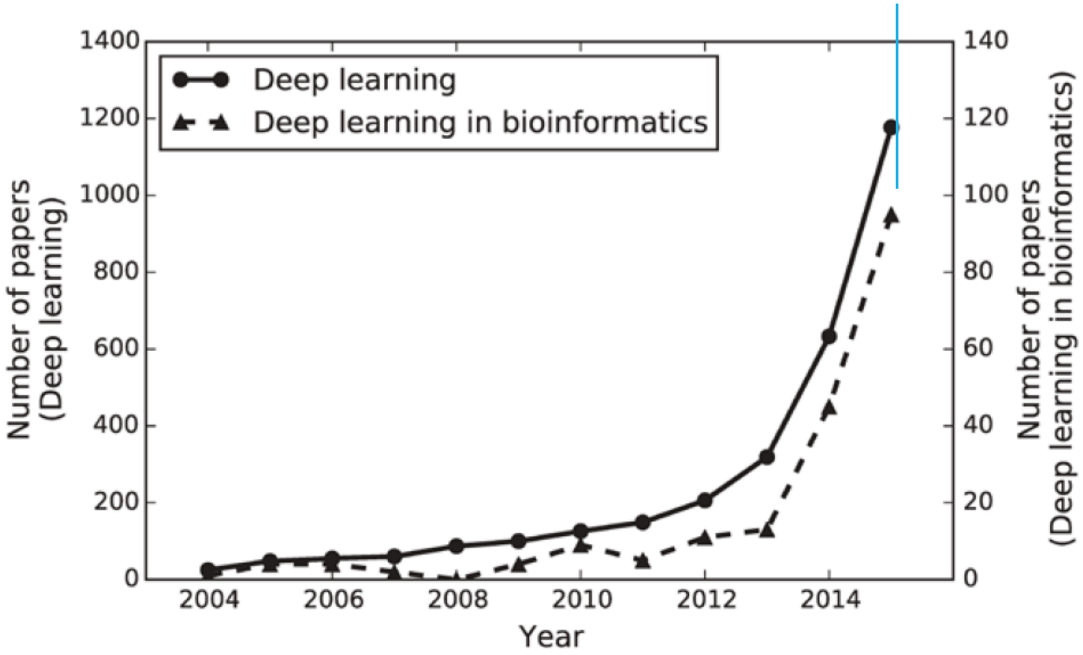
MERJA HEINÄNIEMI

ASSOCIATE PROFESSOR IN BIOINFORMATICS

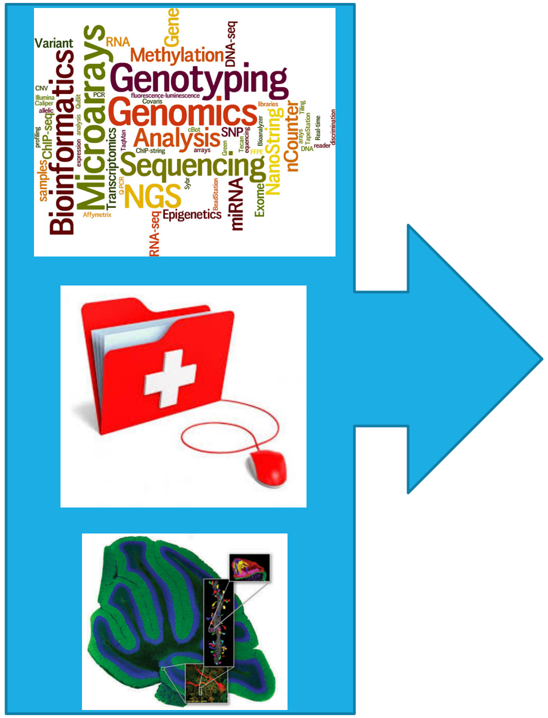
INSTITUTE OF BIOMEDICINE, SCHOOL OF MEDICINE

UNIVERSITY OF
EASTERN FINLAND

Short history – interesting future

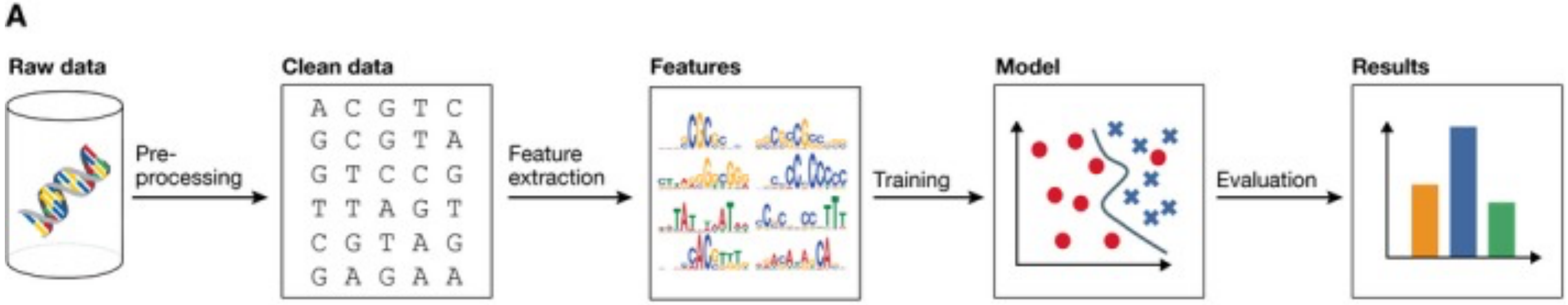


convergence?



Interesting new problems!

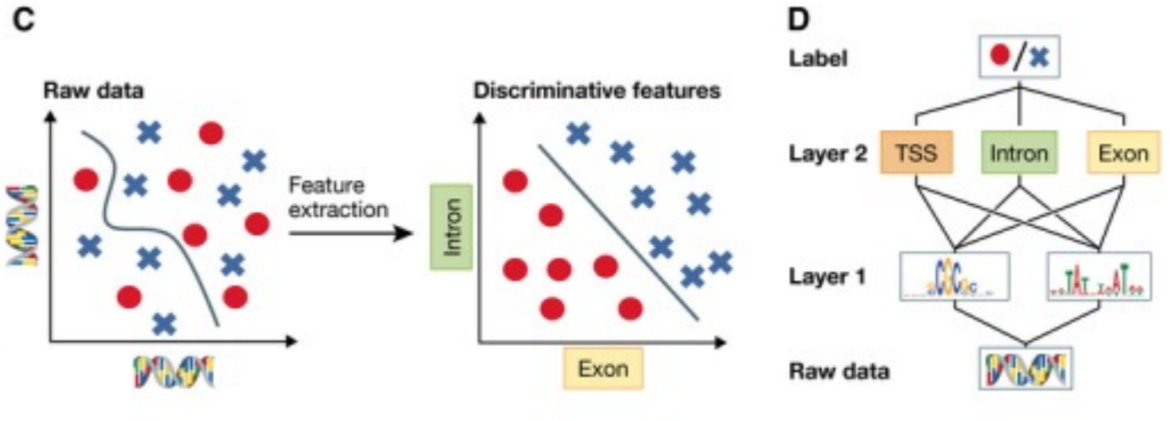
Motivation: higher level features may better discriminate between classes



Classical workflow: data pre-processing, feature extraction, model learning and model evaluation

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965871/

Motivation: higher level features may better discriminate between classes



Deep neural network workflow: use a hierarchical structure to learn increasingly abstract feature representations from the raw data -> higher-level features to better discriminate between classes

Data pre-processing: general

Some minimal pre-processing to consider:

- Numerical features zero-centred by subtracting their mean value

- Image pixels jointly by subtracting the mean pixel intensity per colour channel

- Another normalization is to standardize features to unit variance

- Skewed distribution -> log transformation or similar may be appropriate

! Validation and test data need to be normalized consistently with the training data. For example, features of the validation data need to be zero-centred by subtracting the mean computed on the training data, not on the validation data.

Model architecture

Let's focus on three often encountered:

- a feedforward neural network with fully connected hidden layers – a good starting point for many problems

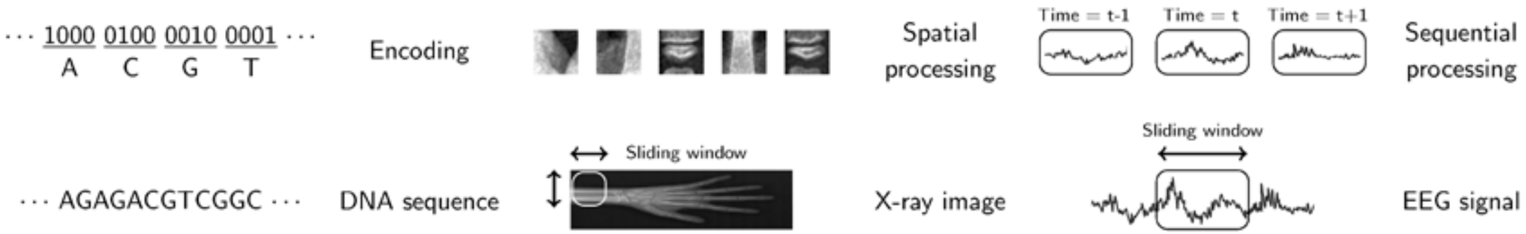
- convolutional architecture - well suited for multi- and high-dimensional data, such as two-dimensional images or abundant genomic data

- recurrent neural networks - capturing long-range dependencies in sequential data (text, DNA, RNA or protein sequence)

Model architecture: simple features

Like language, DNA/RNA/protein sequence can be presented to the model using one hot encoding

Sequential models

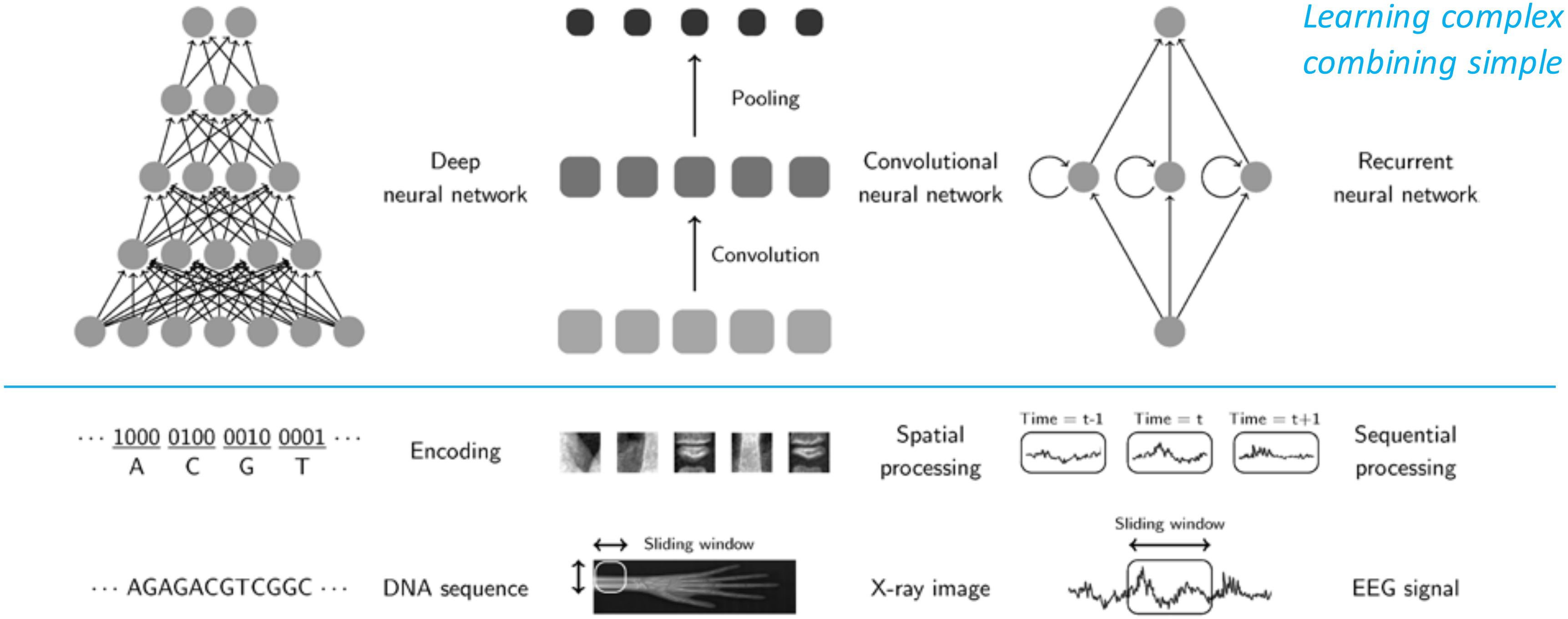


Extracting sub-parts of image

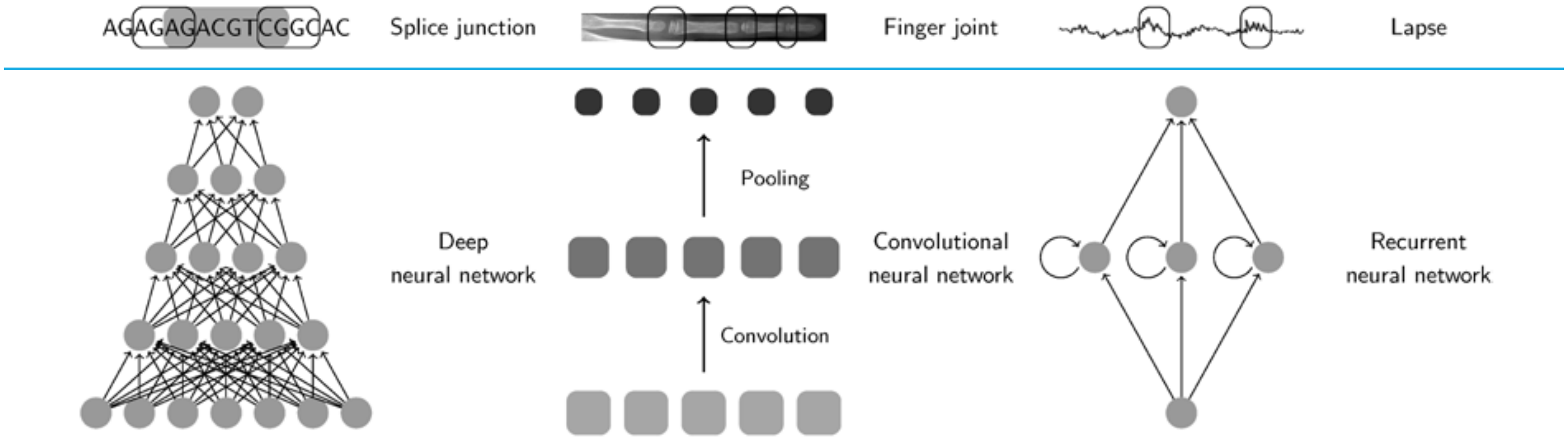
https://arxiv.org/pdf/1603.06430.pdf

Model architecture: network alternatives

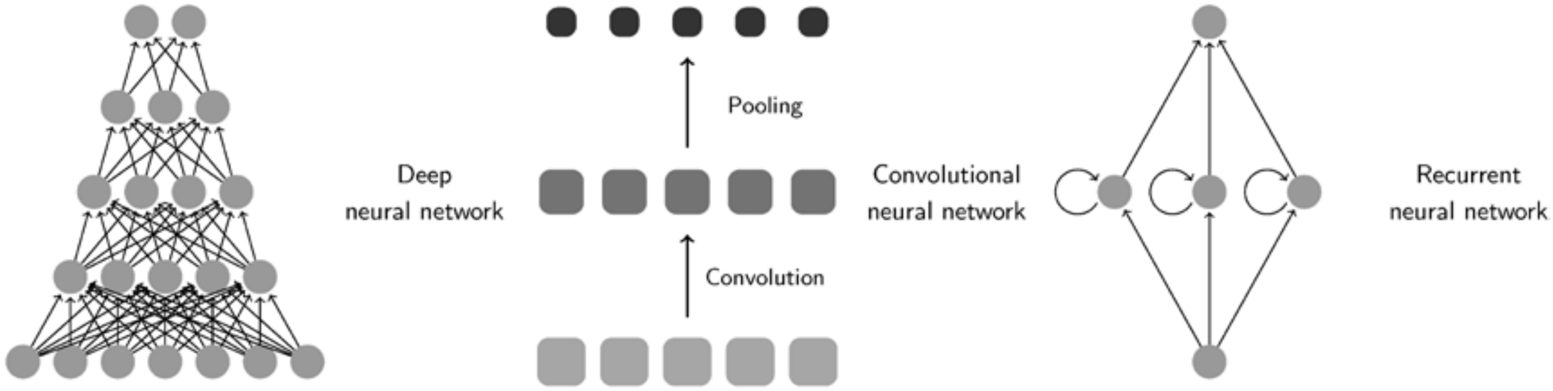
Learning complex features by combining simple features



Model architecture: output



Model architecture: building blocks

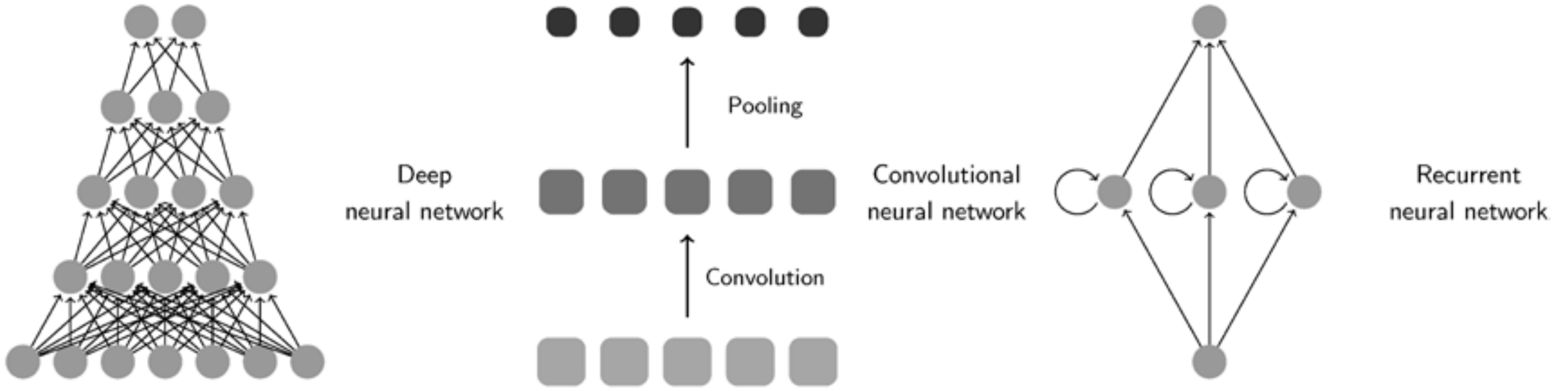


Perceptrons (+++ labelled data!)
Autoencoders
Restricted Bolzmann machines

Convolution layers
Nonlinear layers
Pooling layers

Perceptrons
Long short-term memory units (LSTM)
Restricted Bolzmann machines
Gated recurrent units

Model architecture: examples



Various tasks
Splice signal detection
Protein secondary structure
Cancer classification

Imaging data +++
also Transcription factor motif
DNA accessibility (multitask)

*Multitask joint learning

Sequence data +++
Protein secondary structure
microRNA motif

Model training: general

-Objective function

-Parameter initialization

-Learning rate and batch size

-Learning rate decay

-Momentum

-Adaptive learning rate

-Batch normalization

-Analyzing the learning curve

-Monitoring performance

Imbalanced data – issue in bioinformatics

most deep learning algorithms assume sufficient and balanced data BUT

- cost can be high to perform measurement

- unequal class distribution due to sample availability (cases vs controls)

Solutions that could be relevant in independent task:

- unsupervised pre-training

- transfer learning (pre-training with sufficient data from similar but different domains and fine-tuning with real data)

Interpretation

- In biomedical domains, **it is not enough to simply produce good outcomes**

- visualizing a trained deep learning model:

e.g. Transcription factor motif:

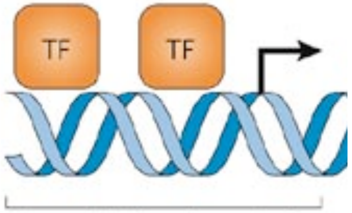
DBN: choose the most class discriminative weight vector among those in the first layer

CNN (DeepBind): count nucleotide frequencies of positive input subsequences with high activation values

Example DNA application 1

Data explained here, network solution in next lecture

A protein that **binds to DNA** can be measured from cells – signal across the genome with peaks at bound sites (ChIP-seq signal)



Q: Which DNA sequence does it recognize?

Chromatin immunoprecipitation (ChIP)-seq

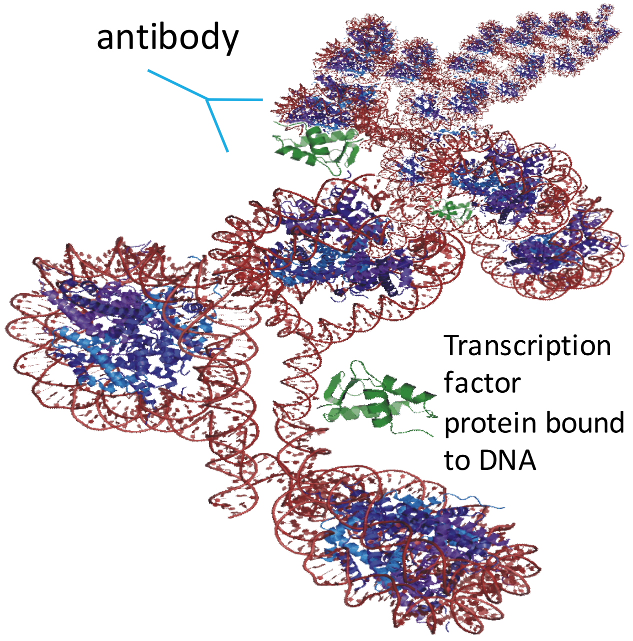
Signal (read counts) represents levels of immunoprecipitated DNA

- DNA bound by transcription factors

or

- DNA at histones that are marked with a specific modification

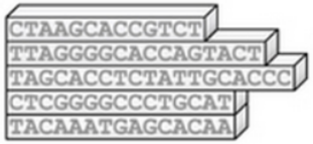
Signal informative about *gene regulation*



Example DNA application 1

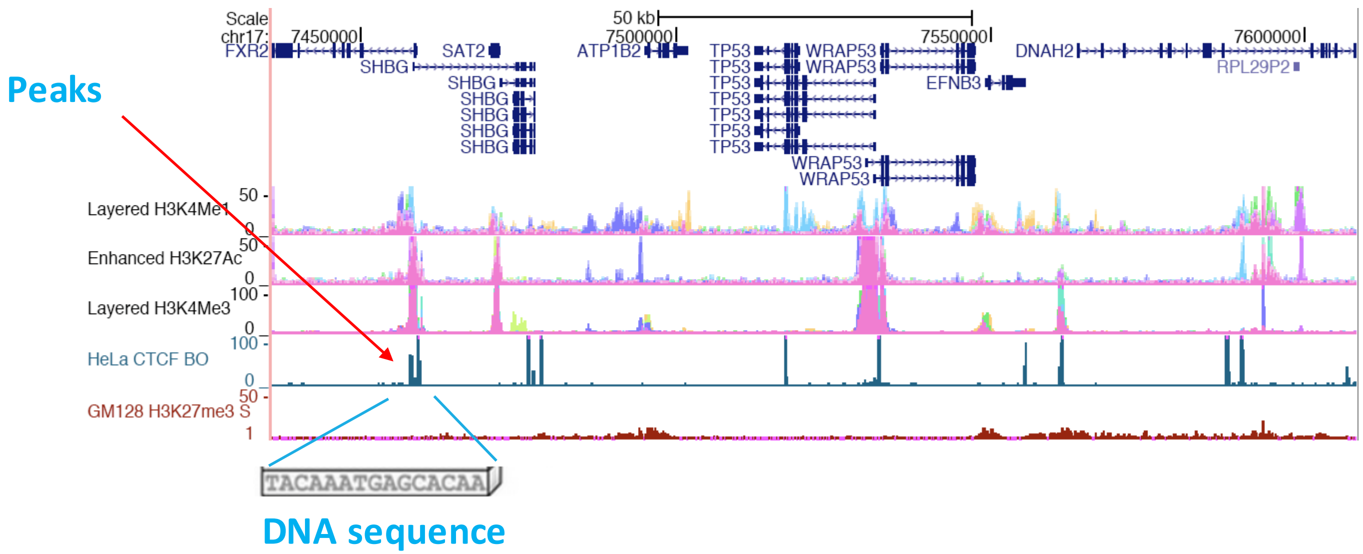
A protein that **binds to DNA** can be measured from cells – signal across the genome with peaks at bound sites (ChIP-seq signal)

Measure where it bound and extract DNA sequence as input data



Q: Which DNA sequence does it recognize?

ChIP-seq signal



Peaks

DNA sequence

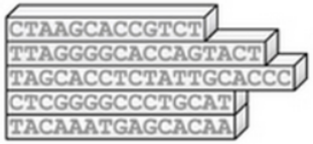
Data from the ENCODE project – one of the first large-scale ChIP-seq efforts

Example DNA application 1

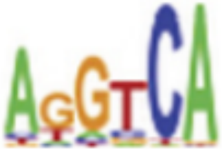
Data explained here, network solution in next lecture

A protein that **binds to DNA** can be measured from cells – signal across the genome with peaks at bound sites (ChIP-seq signal)

Measure where it bound and extract DNA sequence as input data



Q: Which DNA sequence does it recognize?

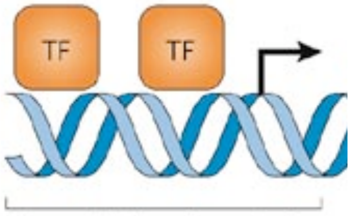


Example DNA application 2

Data explained here, network solution in next lecture

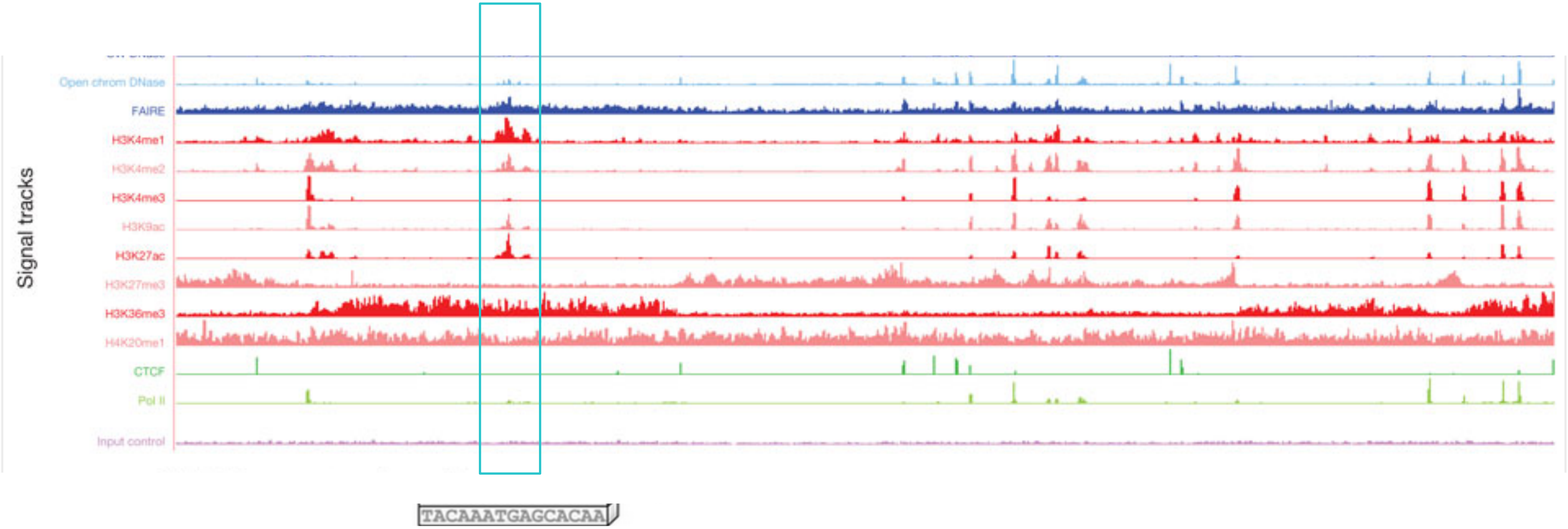
Many proteins **binds to DNA** can be measured from cells – signal across the genome with peaks at bound sites (ChIP-seq signal)

Also so-called histone-markers (labels of active / inactive regions) and DNA accessibility markers can be measured generating similar signal across genome (ChIP-seq, DNAse-seq aka DHS, ATAC-seq)



Q: Can the binding profile (i.e. signal profile) at a given DNA location be predicted?

Example DNA application 2



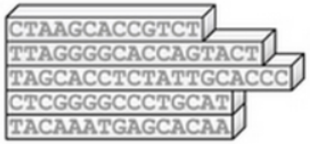
Example DNA application 2

Data explained here, network solution in next lecture

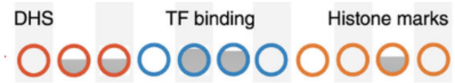
Many proteins **binds to DNA** can be measured from cells – signal across the genome with peaks at bound sites (ChIP-seq signal)

Also so-called histone-markers (labels of active / inactive regions) and DNA accessibility markers can be measured generating similar signal across genome (ChIP-seq, DNAse-seq aka DHS, ATAC-seq)

Measure each protein/marker level and extract DNA sequence as input data



Q: Can the binding profile (i.e. signal profile) at a given DNA location be predicted?

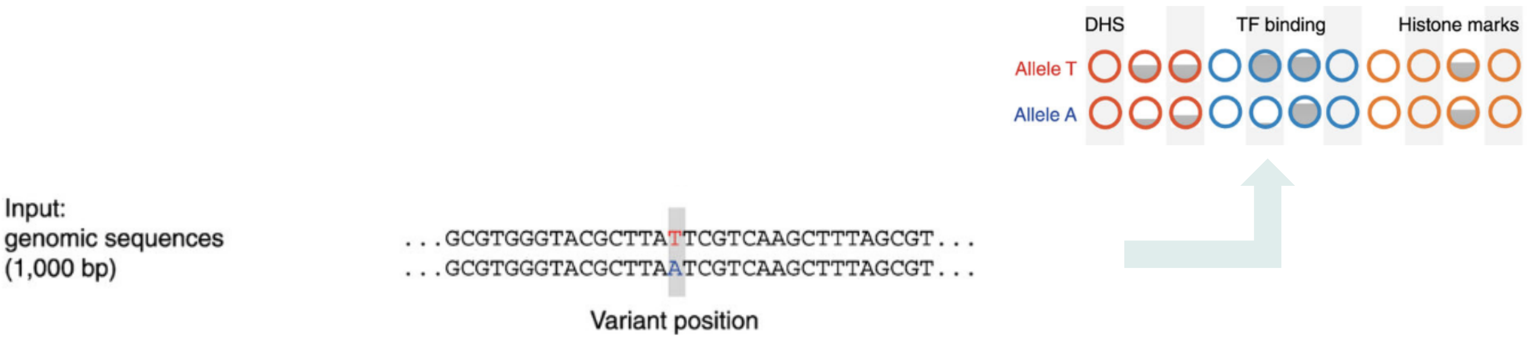


Example DNA application 2

Data explained here, network solution in next lecture

- Using trained model: **compare** DNA sequences that differ at certain position (mutation / natural variant i.e. SNP)

Q: Can the binding profile (i.e. signal profile) at a given DNA location be predicted?


DHS
TF binding
Histone marks
Allele T
Allele A
Input:
genomic sequences
(1,000 bp)
...GCGTGGGTACGCTTATTCGTCAAGCTTTAGCGT...
...GCGTGGGTACGCTTAATCGTCAAGCTTTAGCGT...
Variant position