BIO214 Lecture 10

Bioinformatics-II

Sequence Modeling

Zhen Wei; 2023-Feb-14

Outline

- Motif discovery
- Genomic predictive modeling
- Evaluating model performance

Motif discovery

What computational techniques can be used to interpret biological sequences?

Human genome sequence: ~ **3 billion bp**

5’ - ATGCTAACGTACAGCGCTAGGATCGTG - 3’

Transcription

5’ - AUGCUAACGUACAGCGCUAGGAUCGUG - 3’

Translation

MLTYRARIV

- With the advancement of NGS techniques, DNA & RNA & Protein sequences are massively measured by researchers.

- How to gain insights from the primary biological sequences?

  Motif discovery: finding repetitive patterns

  Genomic predictive modeling: predict genomic markers & conservation scores directly from sequences.

Sequence motif

- The motifs can be discovered from:
  - Sequences of common function (e.g. Zinc-Finger DNA binding domain, phosphorylation sites).
  - From antibody pull down experiments (e.g. CHIP-Seq).
  - Comparative genomics by multiple-sequence alignment.

- What we can do with the motifs:
  - Predict DNA / RNA binding protein binding preferences.
  - Predict covalent-modification sites on protein / DNA / RNA.
  - Recover the network of gene expression regulation. (Know which protein / RNA / DNA is regulated by which regulator at what residue)

Zinc-finger protein motif

Nucleotide epigenetic modification motif

Human 5’ splice site motif

Computational representation of motif

Visualization

5 bp flanking sequences of 9536 epigenetic modification sites (m6A)

```
DNAStringSet object of length 9536:
       width seq
   [1]     5 AGACT
   [2]     5 GGACT
   [3]     5 GAACG
   [4]     5 TGACA
   [5]     5 GGACT
   ...   ... ...
[9532]     5 TAACT
[9533]     5 GGACT
[9534]     5 CGACG
[9535]     5 CGACA
[9536]     5 ACACT
```

Per base frequency summary

$$PPM = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{pmatrix} 0.27 & 0.27 & 1 & 0 & 0.30 \\ 0.06 & 0.06 & 0 & 1 & 0.12 \\ 0.51 & 0.58 & 0 & 0 & 0.06 \\ 0.17 & 0.09 & 0 & 0 & 0.52 \end{pmatrix}$$

Position Probability Matrix

Per base probabilities Calculation

$$\begin{pmatrix} 2548 & 2551 & 9536 & 0 & 2819 \\ 593 & 553 & 0 & 9536 & 1131 \\ 4821 & 5567 & 0 & 0 & 607 \\ 1574 & 865 & 0 & 0 & 4979 \end{pmatrix} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

Consensus Matrix

- Motif is often described by **PPM** (position probability matrix), which summarizes the probabilities of observing different nucleotides (rows) at each positions (columns) of the motif sequences.

How to discover motifs over a set of long genomic sequences?

Known set of functional relevant sequences (e.x. context of single based resolution epigenetic modification sites)

TA**A**CT

AG**A**CT GG**A**CT

CG**A**CG GG**A**CT CG**A**CA

TG**A**CA

GA**A**CG GG**A**CT

AC**A**CT

Directly calculate motif PPM

Set of longer sequences that contain potential motifs (e.x. Peaks from CHIP-Seq experiment)

5’ - AGGATCGAG**GACAA**CGT - 3’

5’ - AGACTAAC**GTACA**ACGCTA**GGACT**GTG - 3’

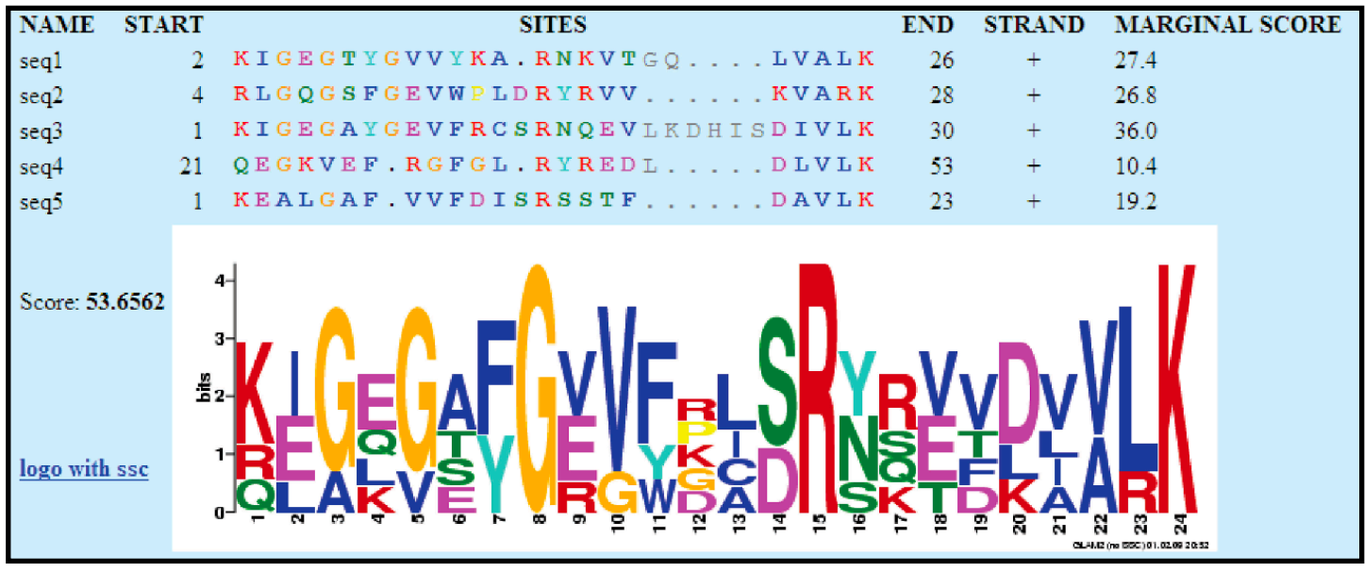
5’ - ATGCTATAGGGTAAA**CGACG**T - 3’

5’ - GTACAGC**GGACA**CGT - 3’

Discover potential motifs using EM algorithm

Top 1 enriched motif:

MEME: motif discovery software



NAME	START	SITES	END	STRAND	MARGINAL SCORE
seq1	2	KIGEGTYGVVYKA.RNKVTGQ....LVALK	26	+	27.4
seq2	4	RLGQGSFGEVWPLDRYRVV......KVARK	28	+	26.8
seq3	1	KIGEGAYGEVFRCSRNQEVLKDHISDIVLK	30	+	36.0
seq4	21	QEGKVEF.RGFGL.RYREDL.....DLVLK	53	+	10.4
seq5	1	KEALGAF.VVFDISRSSTF......DAVLK	23	+	19.2

- **MEME** is a bioinformatic tool to identify unknown short motifs over long input sequences (e.g. > 10000 bp).
- Its core method is based on the following EM algorithm:
  - Randomly initialize motif PPM.
  - Iterate:
    - **E-step**: Infer expected counts of the motif over long sequences, given the current motif PPM.
    - **M-step**: Calculate updated motif PPM from the expected counts.
  - Repeat until convergence.

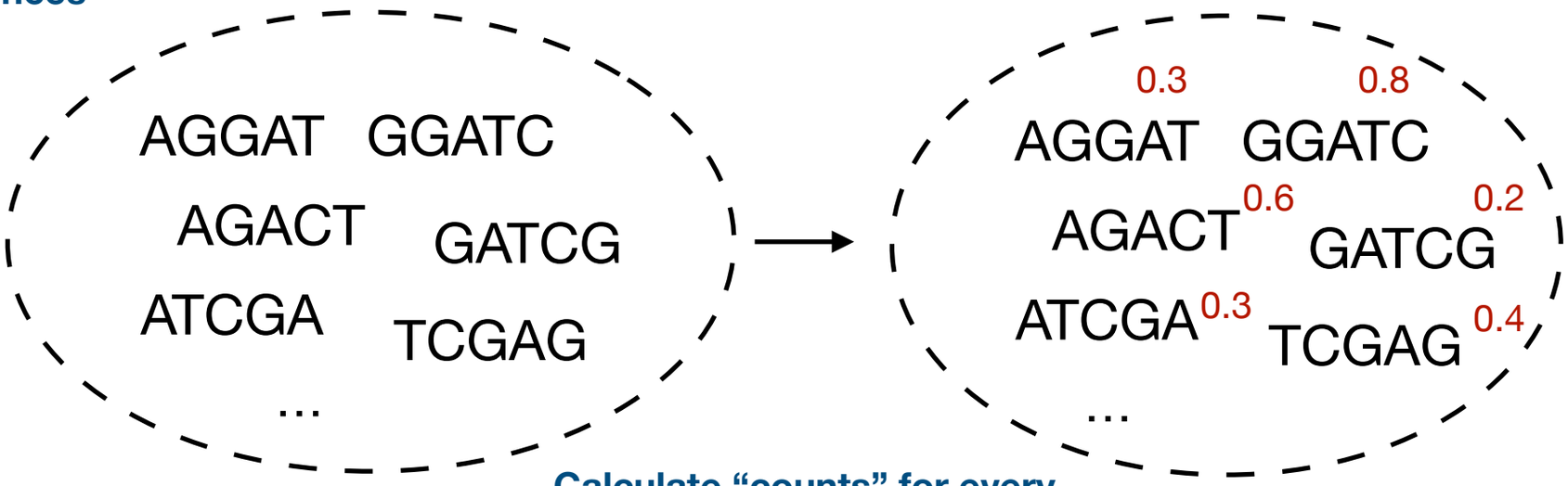
How to discover motif with EM algorithm?

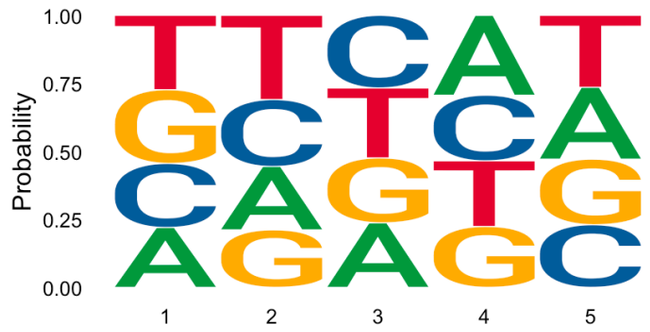
E.x. For a motif with length = 5:

5’ - AGGATCGAGGACAACGT - 3’

5’ - AGACTAACGTACAACGCTAGGACTGTG - 3’

Scan all potential 5-mers in the sequences





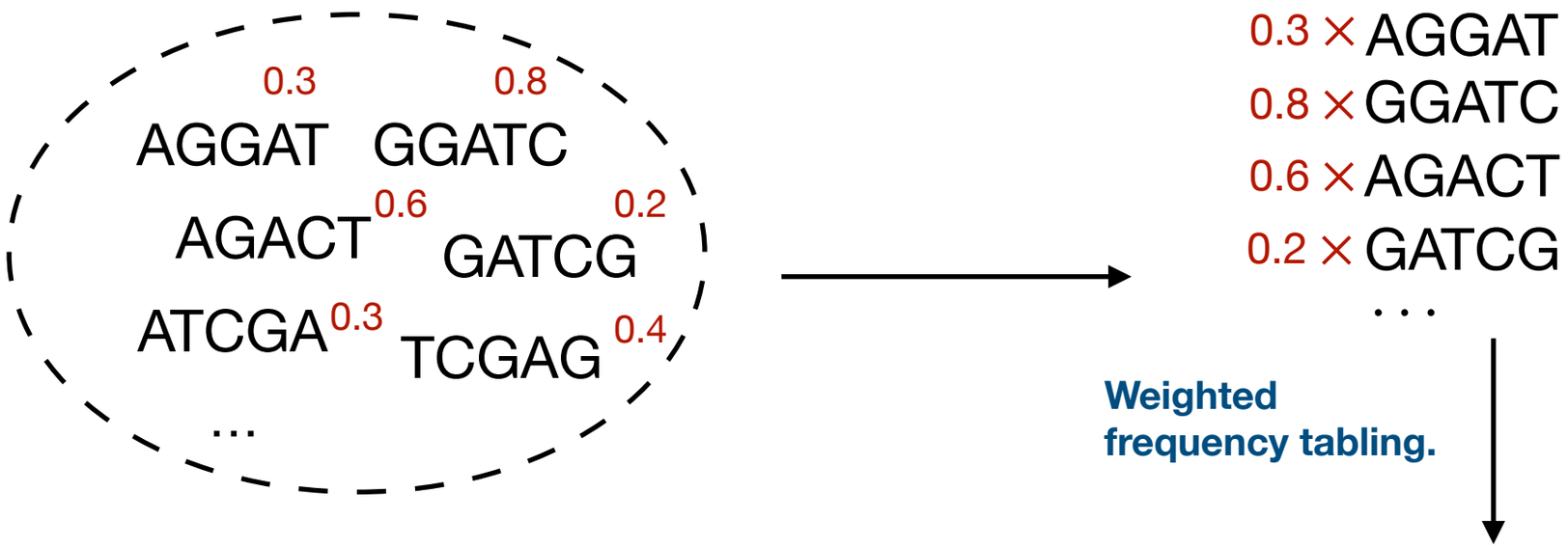
Calculate “counts” for every 5-mers using PPM

E-Step:

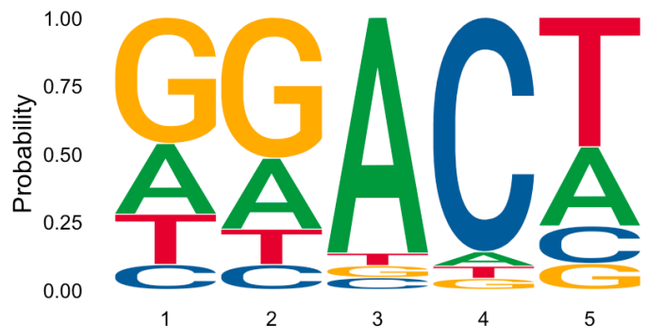
Given a five-mer, e.g. AGGAT, its count on a given PPM is calculated as:

$p_{A,1} * p_{G,2} * p_{G,3} * p_{A,4} * p_{T,5}$; where $p_{i,j}$ is the probability of j th position in the PPM equal to nucleotide $i \in \{A, T, C, G\}$.

How to discover motif with EM algorithm?



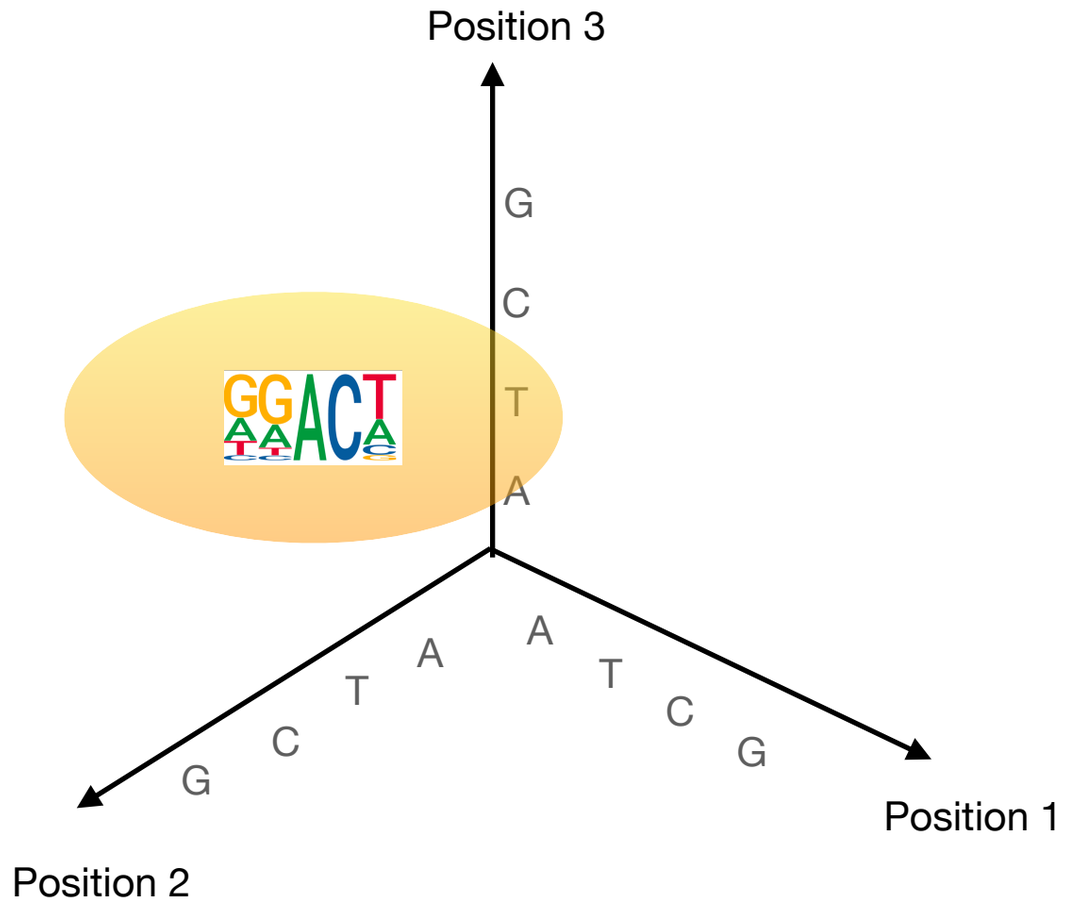
Updated PPM



M-step:

Using the associated counts/weights of K-mers, recalculate PPM by the weighted nucleotide frequencies at each position.

Motif finding is a soft clustering



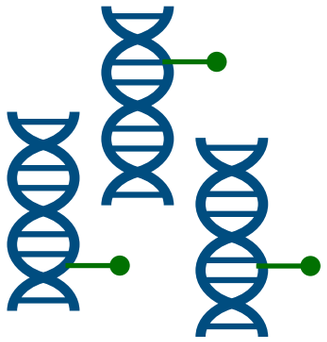
- The motif finding process is essentially a soft clustering on discrete variable space.
- Like gaussian distributions are fitted in GMM, the fitted probabilistic models here are the multinomial distributions (rolling dices with 4 faces).

Genomic predictive modeling

How to predict epigenetic markers from DNA sequence automatically?

Motif based prediction

Functional relavent DNA sequences (e.g. CHIP-Seq peaks)



Discover motifs

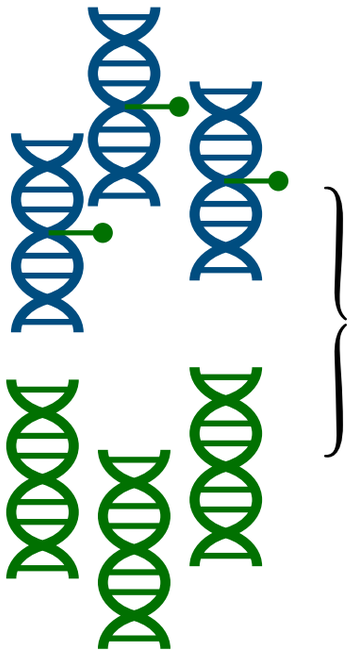
m6A motif

Given a new DNA sequence, scan for motif as candidate prediction.

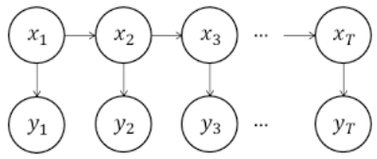
Supervised machine learning modeling

Positive sequences (e.g. flanking region of epigenetic markers)

Negative sequences (e.g. genome background)

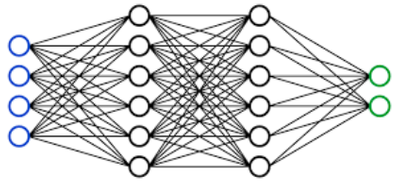


HMM



Or

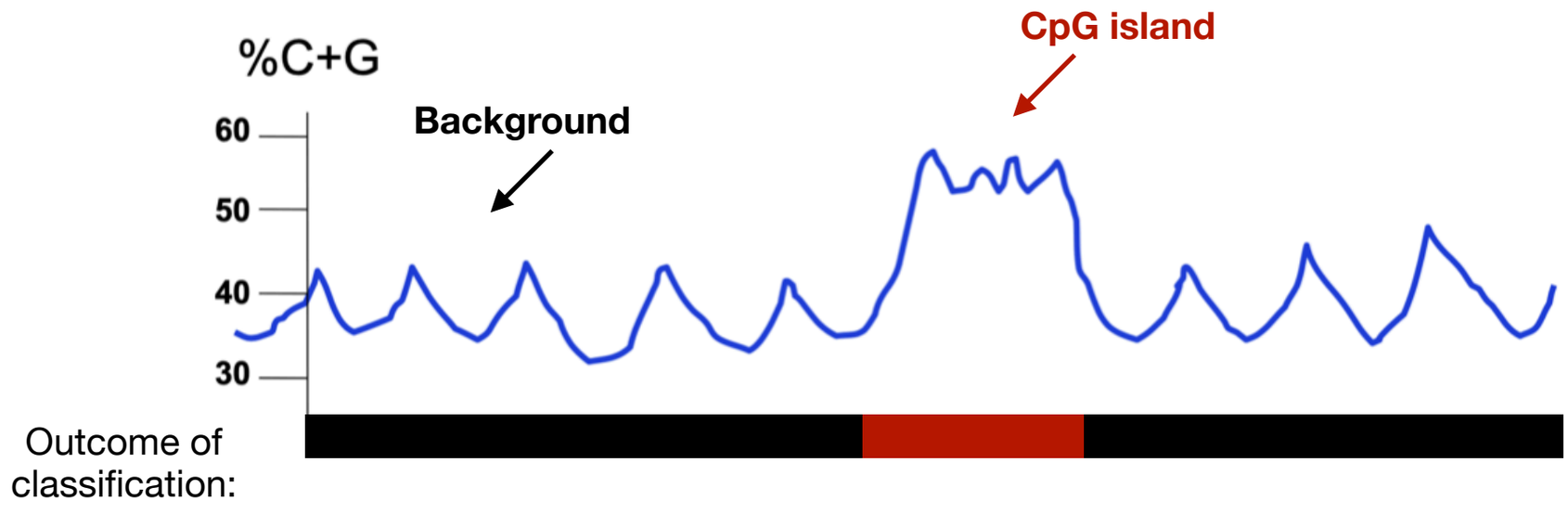
Deep learning



Inference over new sequence using the trained prediction model.

- **Often more accurate and specific than the motif based method.**

Case: finding CpG island from DNA sequence



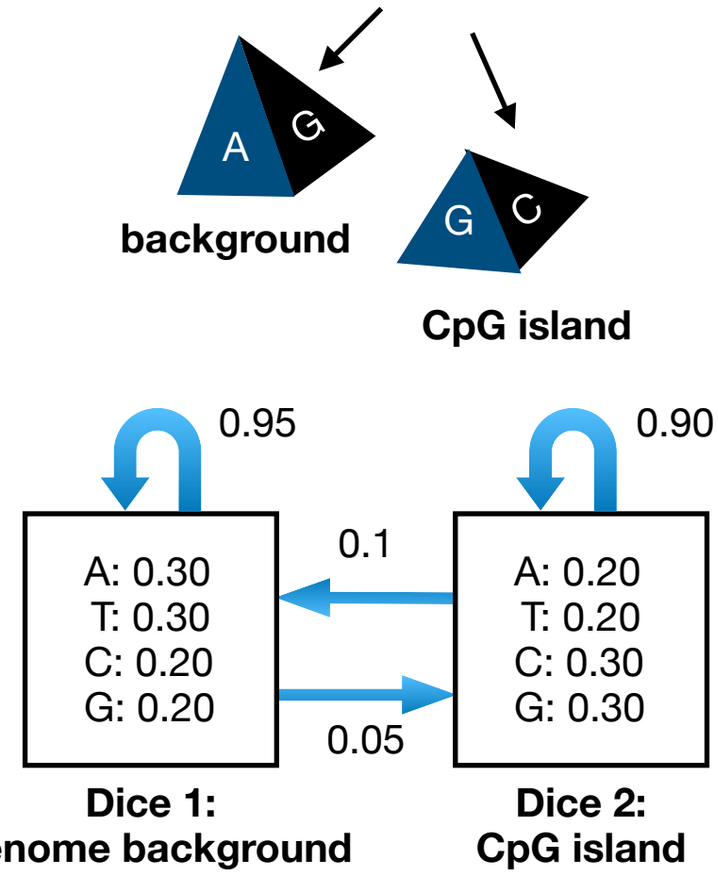
- GC content (the fraction of letters that are a C or a G) can be used to classify the genome into high-GC regions (on average 60% G or C) and low-GC regions (on average 60% A or T).

- The high and low GC regions have different melting temperatures, different replication times across the cell cycle, and different gene density. They have also been hypothesized to have different evolutionary origins.

- How to encode the properties of GpG island in a probabilistic model?

Hidden Markov model for CpG island

Two dices (states), each with outcomes corresponding to the four nucleotides.



"Transition" parameters

$$\text{From}\ \begin{matrix} & \text{Bg} & \text{CpG} \\ \text{Bg} & 0.95 & 0.05 \\ \text{CpG} & 0.10 & 0.90 \end{matrix}\ \text{To}$$

Specify probabilities of dice switching

"Emission" parameters

$$\begin{matrix} & \text{A} & \text{T} & \text{C} & \text{G} \\ \text{Bg} & 0.30 & 0.30 & 0.20 & 0.20 \\ \text{CpG} & 0.20 & 0.20 & 0.30 & 0.30 \end{matrix}$$

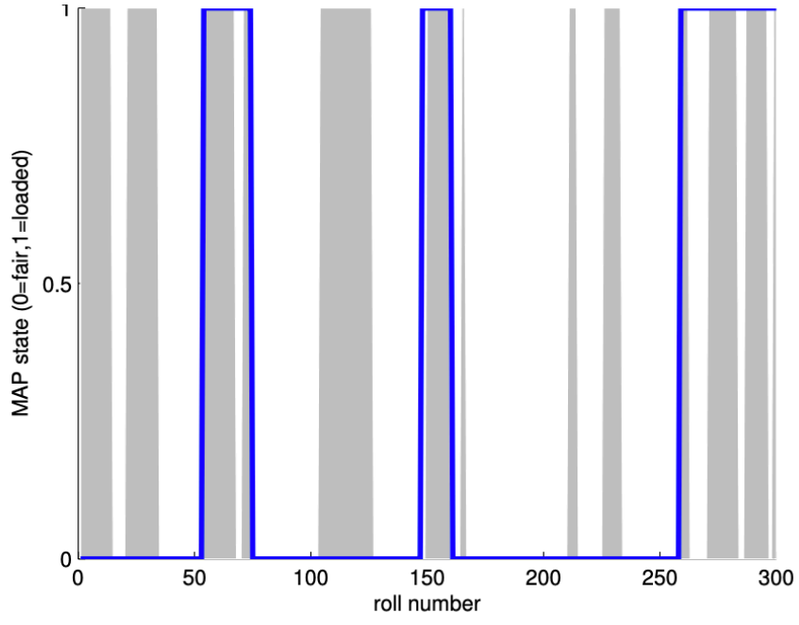
Specify parameters of each dice

- **HMM** is a commonly used machine learning model for biological sequences.
- Considering 2 unfair dices, each with 4 faces of {A, T, C, G); one is for genome background and another is for CpG-island.
- At each roll, we will either keep the current dice, or switch to the other one. The initial roll is selected evenly between the 2 dices.
- After rolling a series of outcomes, we have generated a DNA string, in which the CpG island properties are encoded by the transition and emission parameters.

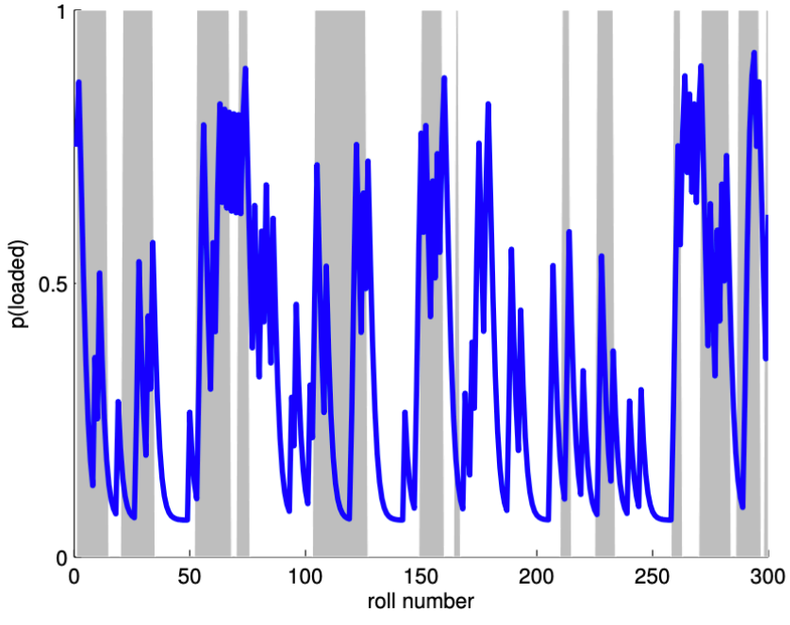
State inference (prediction)

- After estimating the transition & emission parameters from the data, one can compute the **state posterior** along the genome using Bayesian inference.
- State posterior := P(state at position i | the entire observed sequence)
- Two inference algorithms are often used: Viterbi algorithm and forward backward algorithm.

Viterbi algorithm (Return binary classification):



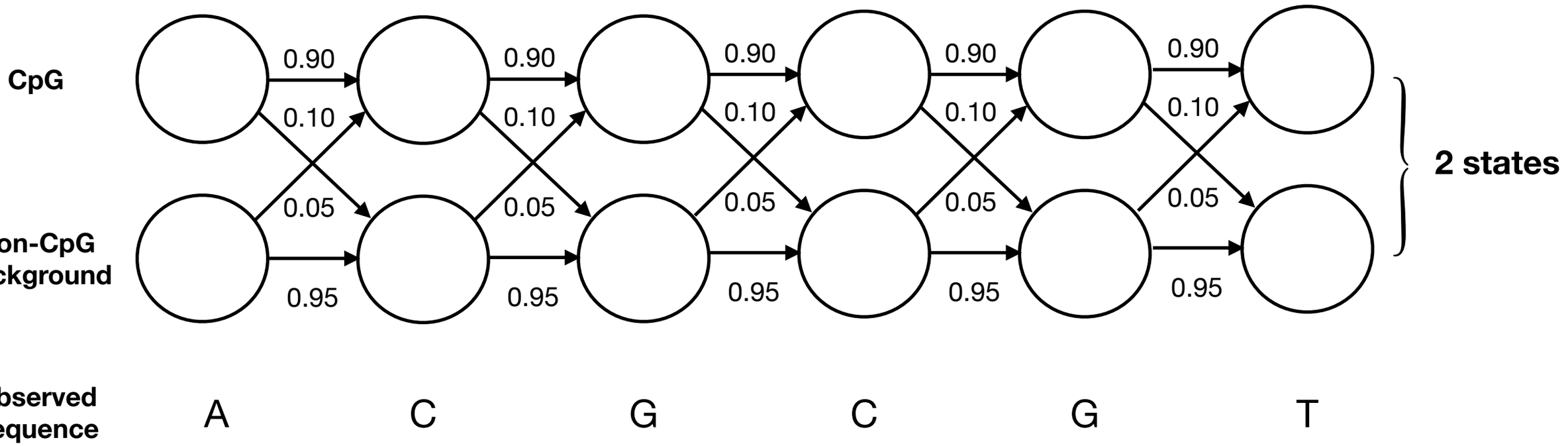
Forward backward algorithm (Return probabilities):



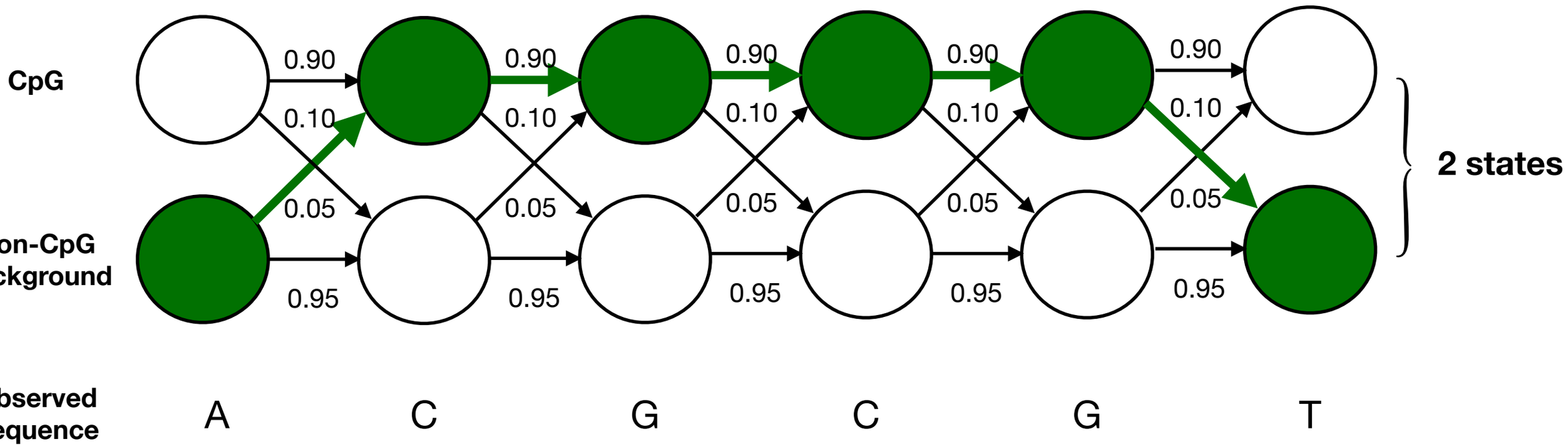
Real case applications:

- Classify the regions of CpG island from background on genome.
- Predict protein coding genes.
- Estimating a score for evolutionary conservation along the genome. (e.g. **phastCons score** in phylo-HMM)

State inference from a graphical perspective

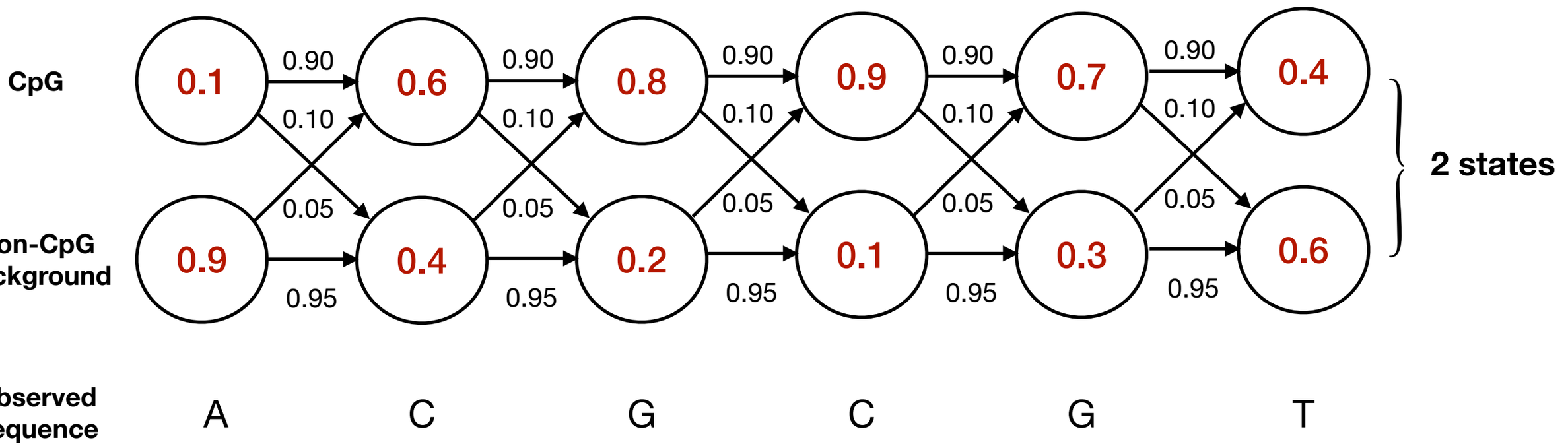


State inference from a graphical perspective



- Viterbi algorithm is estimating **the most likely pathway of states** given the observed sequence.

State inference from a graphical perspective



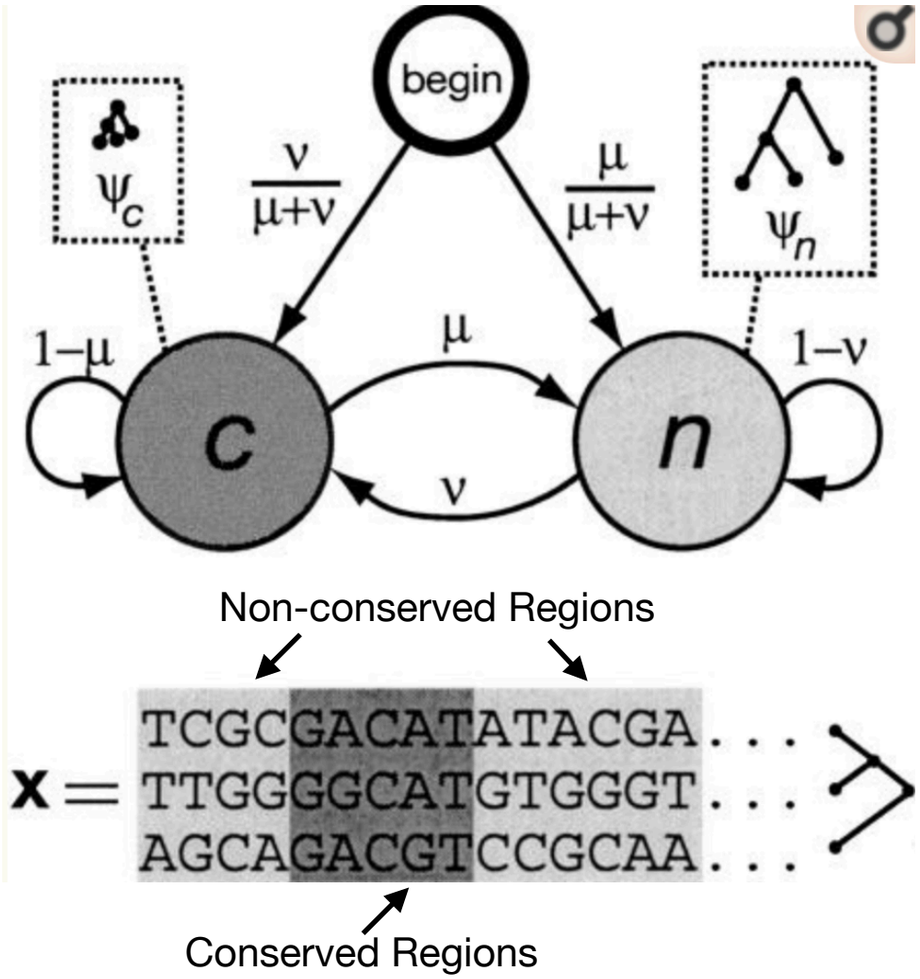
- Forward backward algorithm is estimating **the state probabilities** given the observed sequence.

More applications of HMM in genomics

Application	Detection of GC-rich region	Detection of Conserved region	Detection of Protein coding exons	Detection of Protein coding conservation	Detection of Protein coding gene structures (Gene Prediction)	Detection of chromatin states
Topology / Transitions	2 states, different nucleotide composition	2 states, different conservation levels	2 states, different tri-nucleotide composition	2 states, different evolutionary signatures	~20 states, different composition / conservation, specific structure	40 states, different chromatin mark combinations
Hidden States / Annotation	GC-rich / AT-rich	Conserved / non-Conserved	Coding (exon) / non-Coding (intron or intergenic)	Coding (exon) / non-Coding (intron or intergenic)	First / last / middle coding exon, UTRs, intron 1/2/3, intergenic, *(+,-) strand	Enhancer / Promoter / Transcribed / Repressed / Repetitive
Emissions / Observations	Nucleotides	Level of conservation (PhastCons Score)	Triplets of nucleotides	64 x 64 matrix of codon substitution frequencies	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies

- HMM is often used to decode or parse a genome into its biological components: genes, exons, introns, regulatory regions.
- In addition, conservation states of nucleotides and regions can be learned (often in the form of conservation scores).

Phylo-HMM (PhastCons score)



Transition parameters

$$A = \begin{pmatrix} \mu & 1-\mu \\ v & 1-v \end{pmatrix}$$

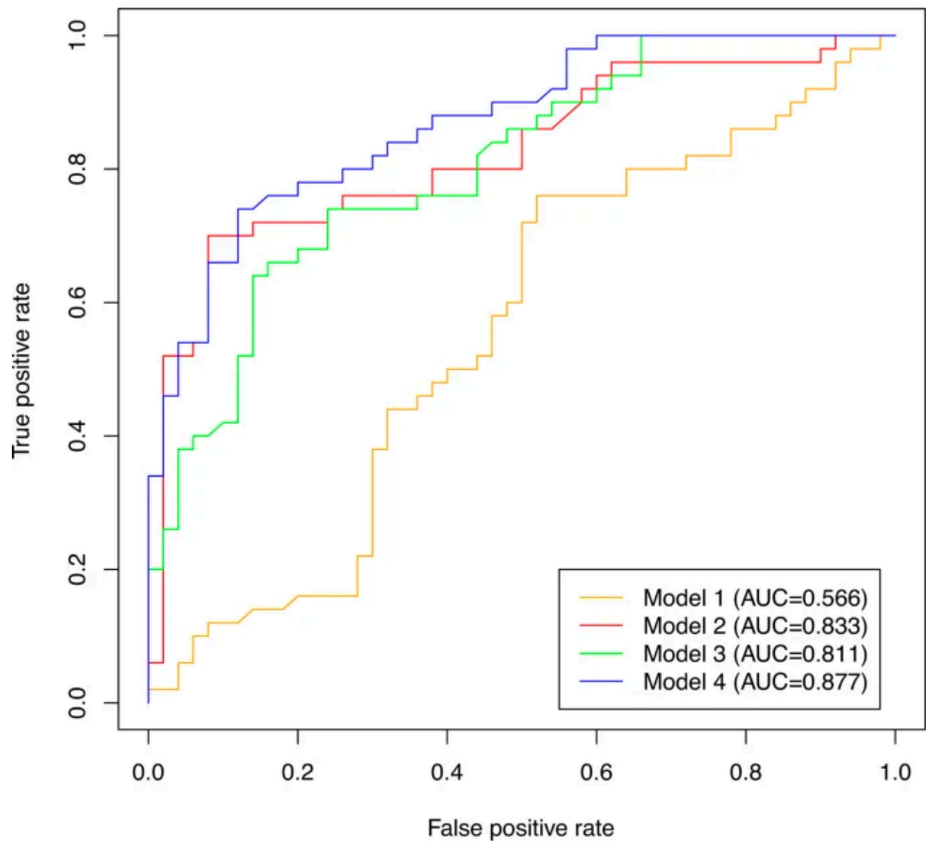
Emission parameters

$$\pi = \begin{pmatrix} \frac{v}{\mu+v} & \frac{\mu}{\mu+v} \end{pmatrix}$$

- The Phylo-HMM aims to predict level of evolutionary conservation (quantified by PhastCons score) over genomes.
- 2 latent states are defined: conserved regions (*c*) and non-conserved regions (*n*).
- Observation is the **multiple sequence alignment** result**.**

Evaluating model performance

How to know which genomic predictor perform better?



ROC curves of 4 HMM based model between positive and negative sequences from Masato Yano et al 2014.

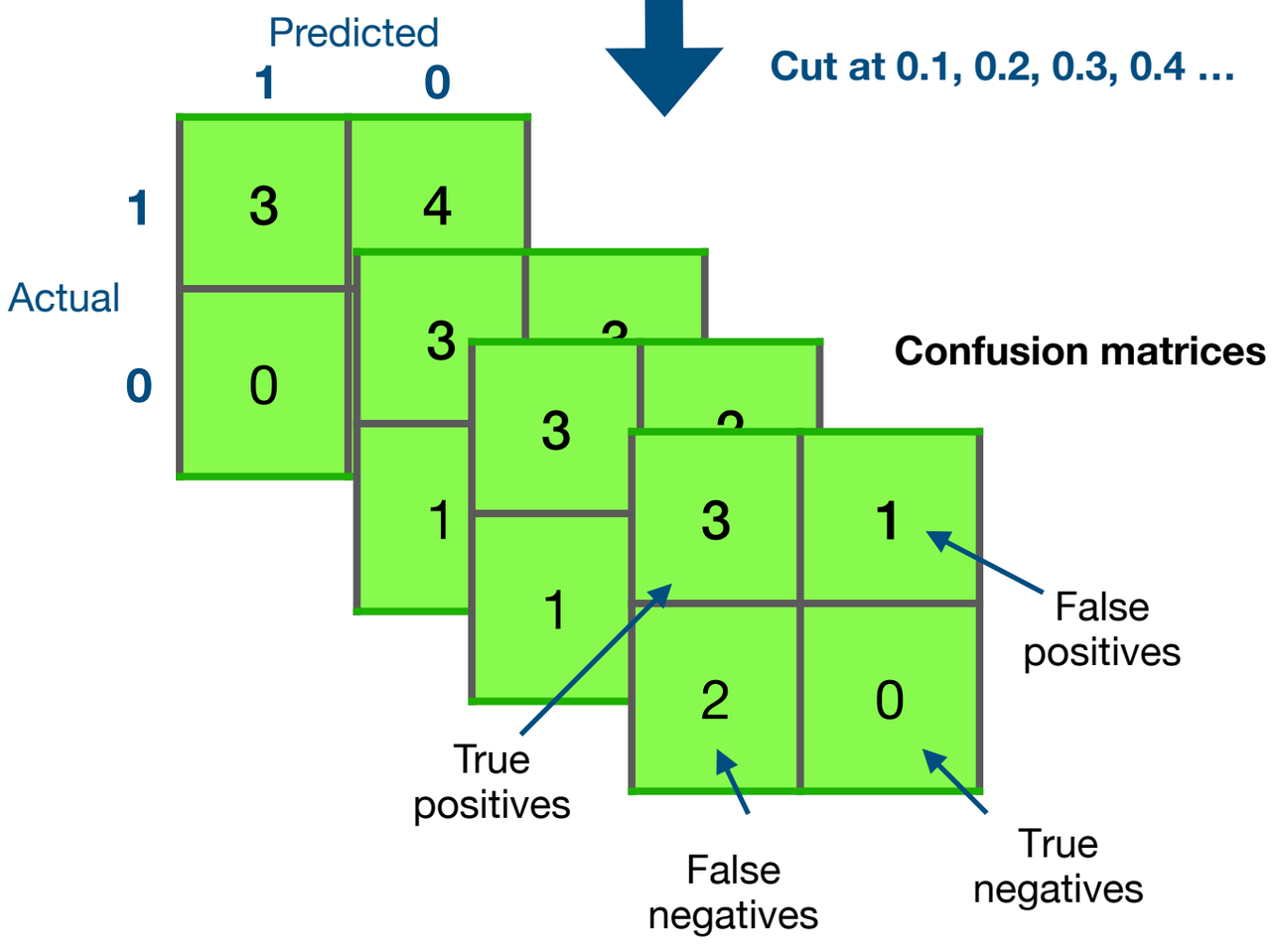
- Different genomic predictors often compete in their performances on the same end application.
- To avoid overfitting, the performances are required to be evaluated “**out of sample”.** In other words, the final prediction accuracy should be reported over the **test set** never revealed to the model before.

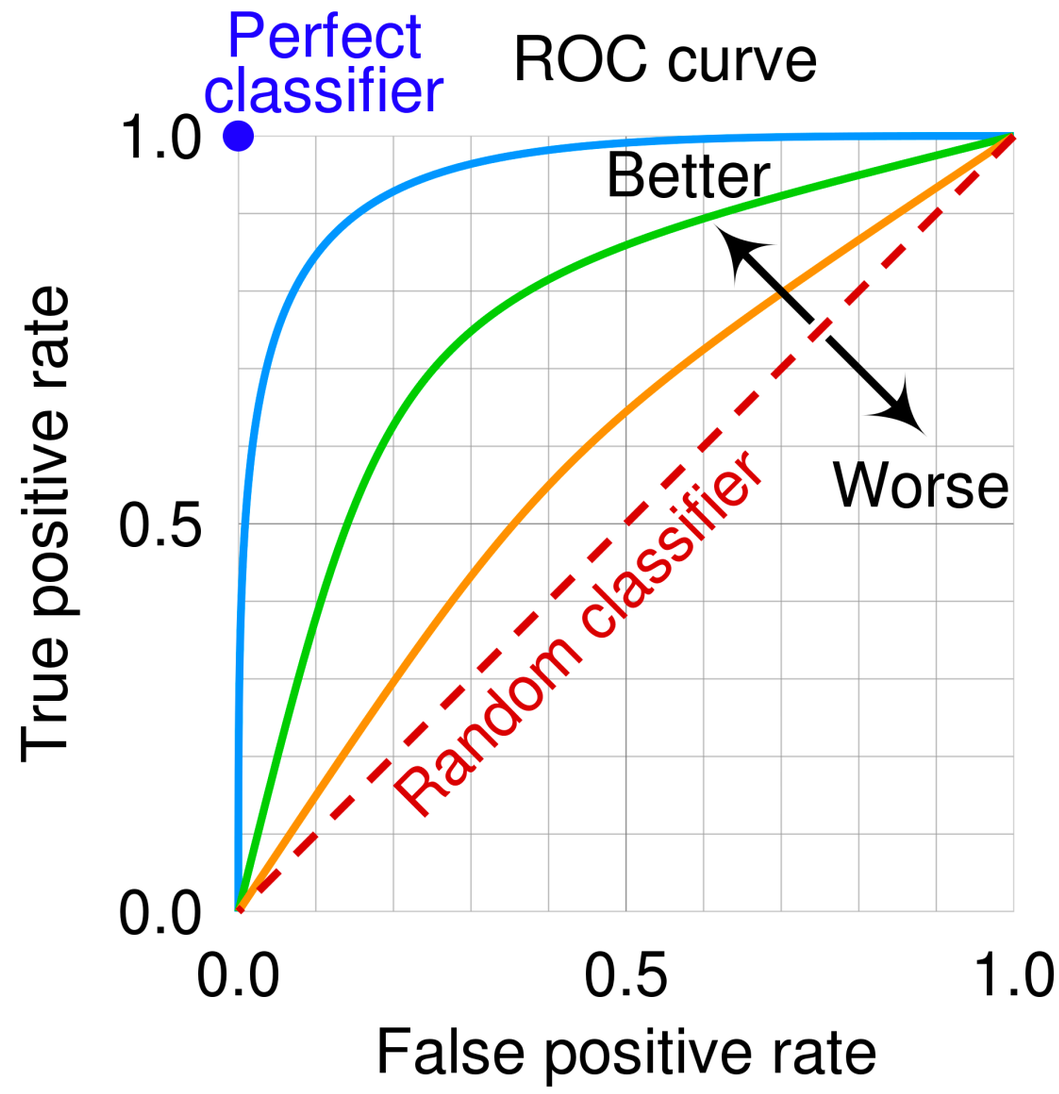
Classification evaluation metric: AUROC

On test set

Ground truth (real label)	1	1	0	1	0	0
Predicted probabilities	0.8	0.6	0.3	0.6	0.2	0.7

Cut at 0.1, 0.2, 0.3, 0.4 …

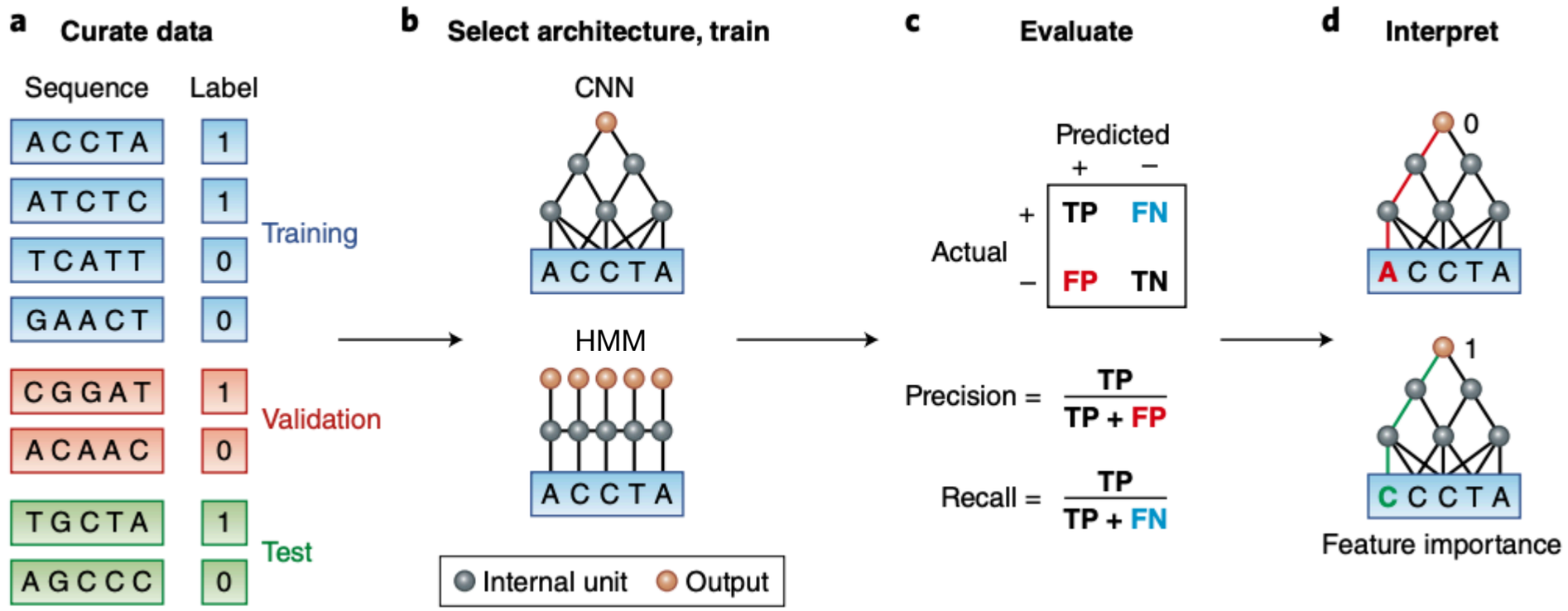




X axis: FPR = False positives / (True negatives + False positives)

Y axis: TPR = True positives / (True positives + False negatives)

Workflow of sequence based supervised learning



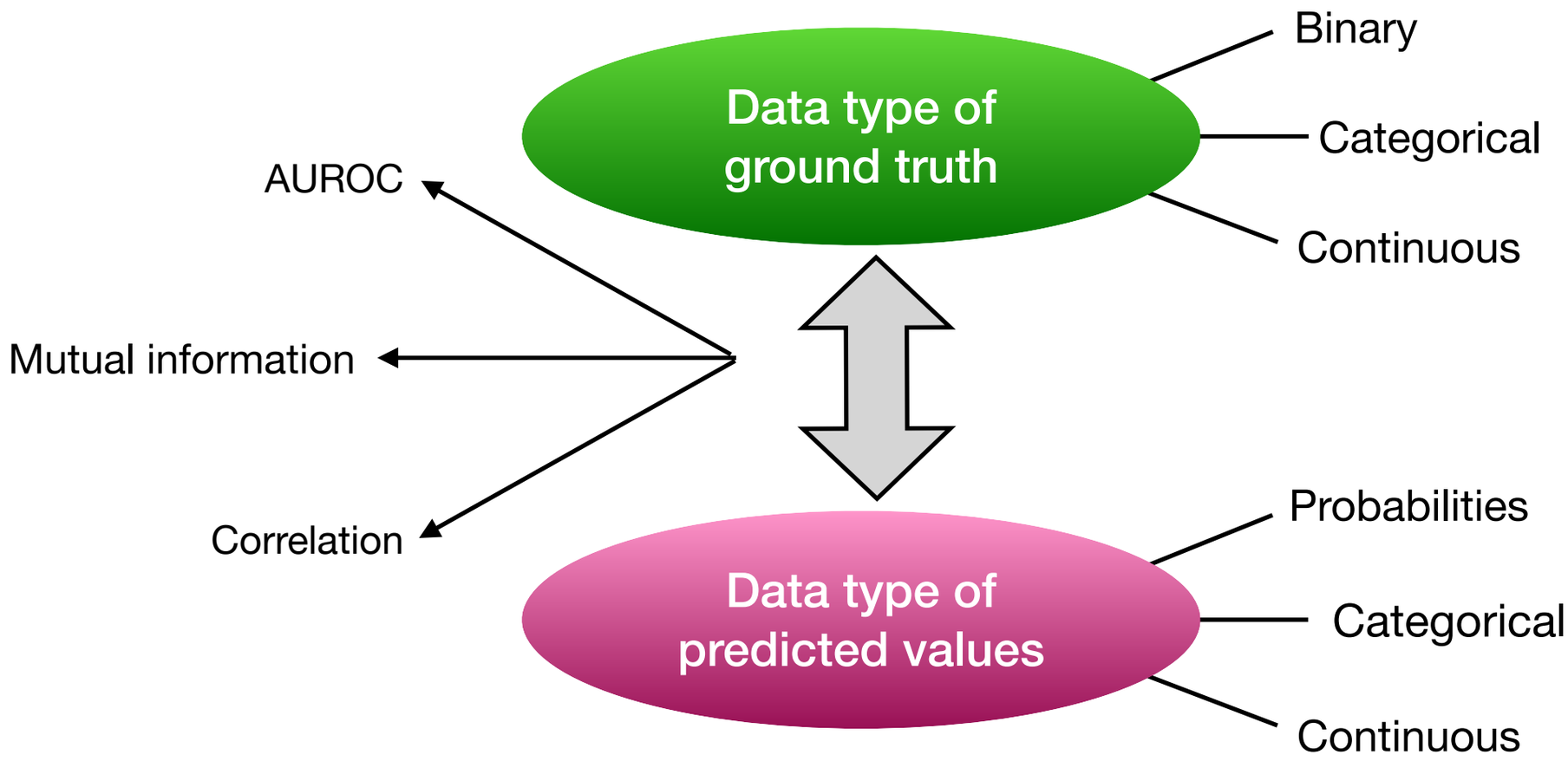
a. A dataset should be randomly split into training, validation and test sets. The positive and negative examples should be balanced for potential confounders (for example, sequence content and location) so that the predictor learns salient features rather than confounders.

b. The appropriate machine learning algorithm is selected and trained on the basis of domain knowledge. For example, CNNs (Convolutional Neural Networks) capture translation invariance, and HMMs capture more flexible spatial interactions.

c. True positive (TP), false positive (FP), false negative (FN) and true negative (TN) rates are evaluated. When there are more negative than positive examples, precision and recall are often considered.

d. The learned model is interpreted by computing how changing each nucleotide in the input affects the prediction.

Performance evaluation: general scheme



- The evaluation statistics we choose depend on the forms of the ground truth labels and the predicted values.

Summary of performance evaluation methods used by different data types

Metric	Description	ground truth data type	predicted value data type	Example in bioinformatics application
FDR	Proportion of false positively predicted instances.	Binary	Binary	**Differential gene expression analysis**
AUROC	The area under the fall-out (x-axis) and recall (y-axis) curve.	Binary	Class probabilities	**Supervised classification model**
Mutual information	Information lost when encoding the 2 categorical labels independently.	Categorical (> 2 classes)	Categorical. (> 2 classes)	**Clustering**
Pearson Correlation	Linear correlation between the 2 sets of values.	Countinous	Countinous	**Gene expression level quantification**