BCH 519
Introduction to Bioinformatics

Genome Annotation

Dr. Tao Liu

March 24, 2015

Outline

- Learn what is '**genome**' and types of '**genome annotations**'.
- Focus on bioinformatics approaches to **predict genes** (or gene finding) and to **predict transcription factor binding sites**.
- Thursday: In-class exercises for homework. Practice some of the programs we discuss in class today.
- Next week: **Motif discovery** and **regulatory module prediction**
- Supplemental reading if you want more information:

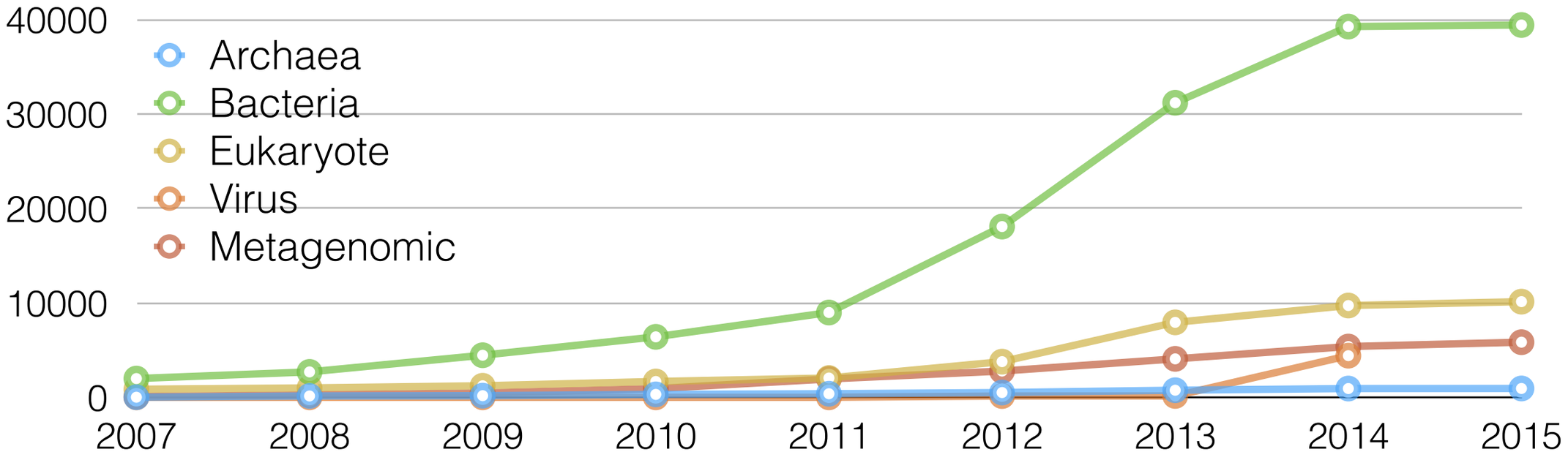
  ANNs: Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*(2), 195
  HMMs: Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315
  Gene Prediction: Brent, M. R. (2007). How does eukaryotic gene prediction work? *Nature Biotechnology*, *25*(8), 883
  Motif: D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, *24*(4), 423

Genome and Genes

- Genome is the genetic material in an organism, encoded in DNA or RNA (many virus), including ‘genes’ and ‘non-coding sequences’.

- Gene is a locatable region of genomic sequence, a unit of hereditary information, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions. It can code for a peptide or for an RNA product.

Project Totals in GOLD(Genomes Online Database)



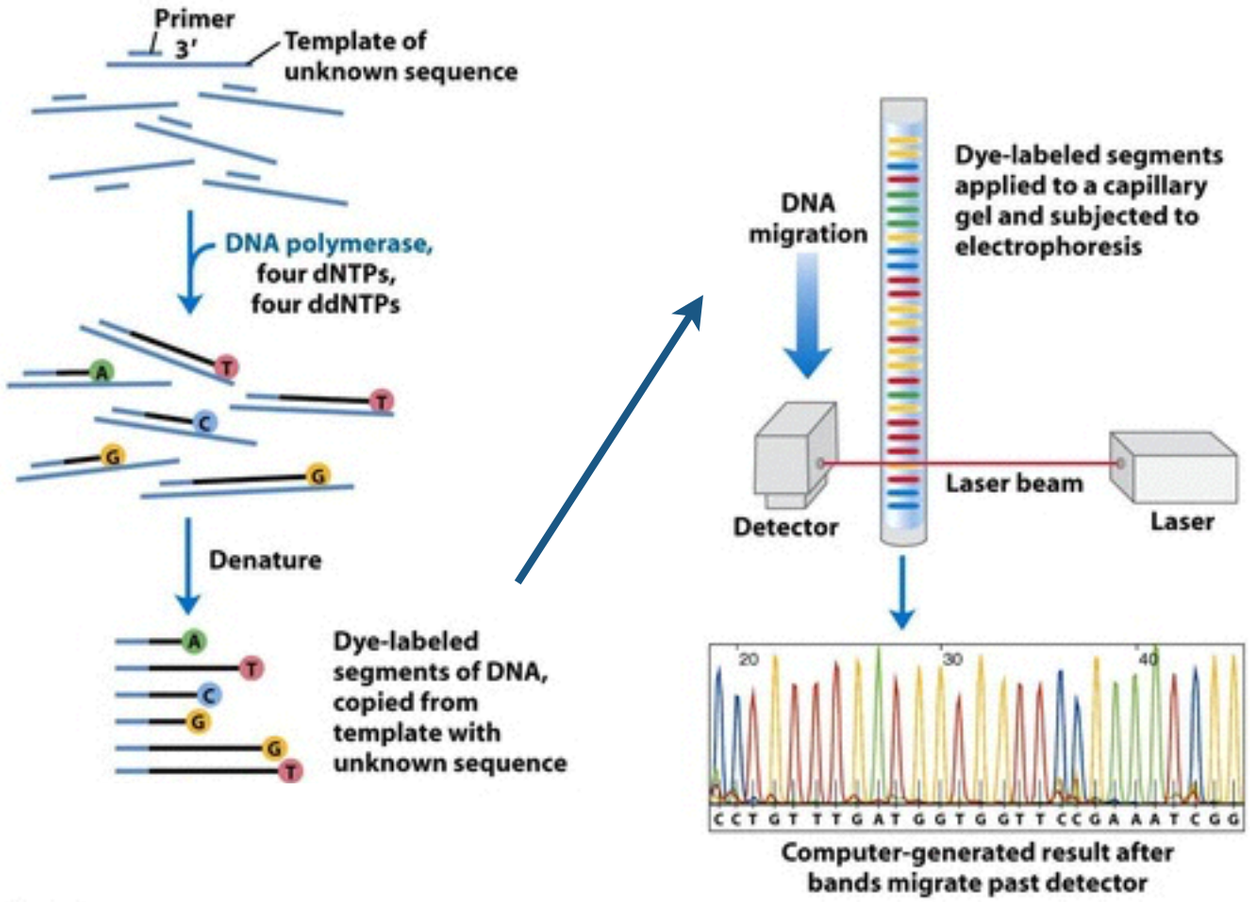
Buchnera aphidicola BCc (bacterium)	**422 kb**	*Oryza sativa* (rice)	**420,000 kb**
Bacillus anthracis (anthrax)	**5228 kb**	*Mus musculus* (mouse)	**2,493,000 kb**
S. cerivisiae (yeast)	**12,069 kb**	*Homo sapiens* (human)	**2,900,000 kb**
Drosophila melanogaster (fruit fly)	**137,000 kb**	*Amoeba dubia* (amoeba)	**670,000,000 kb**

DNA sequencing

- Determine the order of nucleotides in a DNA molecule
- Common applications:
  - Studying genome
  - Studying evolutionary biology: compare different organisms
  - Studying population genetics: compare different individuals
- Old sequencing technology, **Sanger's method**:
  - 1) make DNA fragments; 2) Construct libraries to store DNA (optional); 3) Chain Termination sequencing
- Current technology, **illumina** high-throughput sequencer:
  - 1) make DNA fragments; 2) ligate linker sequences; 3) amplify by PCR; 4) sequencing by synthesis
- Developing technologies: PacBio SMRT, Nanostring …

Sanger's method

- a.k.a "Chain Termination Method"
- **ddNTP** is mixed with dNTPs, lacking a 3′-OH for formation of phosphodiester bond
- ddNTP can be radioactively or fluorescently labeled
- Start **DNA synthesis** using **primers from known sequence** to 3′ end, dNTP and ddNTP
- Gel electrophoresis separate synthesized DNA fragments **by their sizes**
- Either:
  - label the dNTP or ddNTP or even primer with the same radioactive or fluorescent tag, then separate the reaction in four lanes with different ddNTP
  - label the ddNTP with four different fluorescent dyes then run in one lane (**dye-terminator sequencing**)



dye-terminator sequencing shown here

In Informatics Point of View

Each chromosome is just a sequence of A/T/C/G (or A/U/C/G of RNA).

Beware that DNA has two reversely complementary strands — technically ‘plus’ or ‘Watson’ strand and ‘minus’ or ‘Crick’ strand.

Thumb of rule: if you see different genome-wide statistics on plus vs minus strand, then the analysis must be wrong.

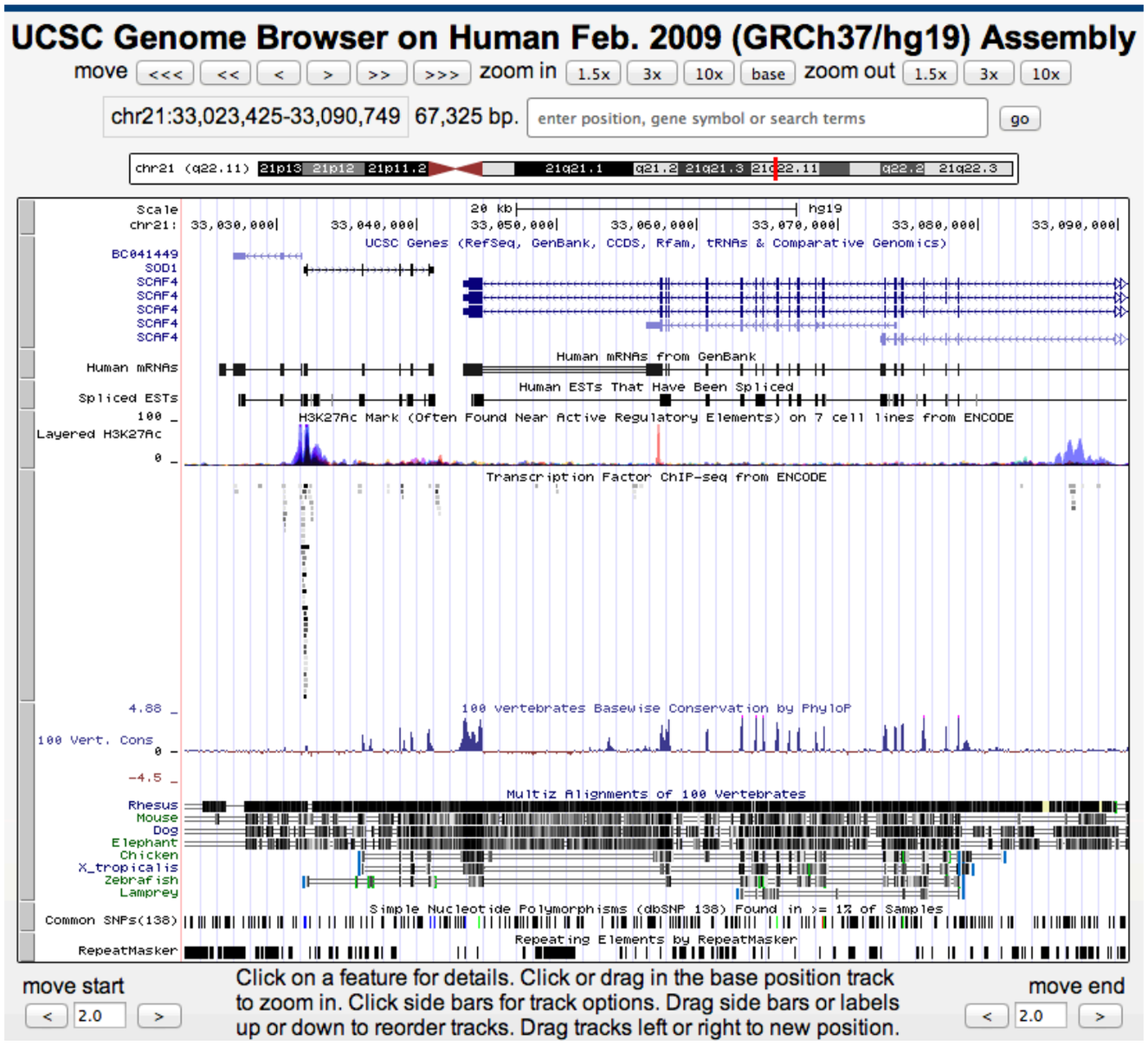
From human chr. 16

(<6.4 kb of 3x10⁶ kb, 0.0002%)

Genome Annotation

- Having the raw genome sequence is therefore just a beginning. To make use of the sequence, we must annotate it—describe the function of each basepair of sequence.

- Similar to natural language, in order to understand a book, we have to know **1)** the language (dictionary); **2)** the grammar (verb/noun/adjective…); **3)** the meaning, the story. "There are a thousand hamlets in a thousand people's eyes."



Elements in a genome

- Note: these are based on what we have learned so far.
- Functional **genetic** information/Genes:
  - **Protein coding genes: those can be transcribed to mRNA then be translated into proteins.**

    only 2% of human genome
  - **Non-protein coding genes: those RNA can't be translated into proteins.**

    aka ncRNA: rRNA, tRNA, miRNAs, lncRNA and etc.
- Functional **epigenetic** information**:**
  - Structural sequences (scaffold attachment regions)
  - Regulatory elements (promoter, enhancer, repressor and so on)
  - Other (including repeats, transposons, retroviral insertions, etc.)

Human Genome Project

- We got the human genome sequence.
    - Note: It's so-called 'reference genome' and can't be associated to a simple person
- We got the human genes annotated.
- We can predict novel protein-coding genes, and non-protein coding genes.
- We got the statistics of human genome, e.g. GC content, conservation with other species and so on.
- **BUT**, we don't know how genes work together and how they are regulated. i.e. we have a book and a dictionary, but can't get the ideas of it.

The ENCODE Project

- The ENCyclopedia Of DNA Elements (ENCODE) Project was established to **identify all functional elements in the human genome sequence.**

- ENCODE 1/Pilot phase (2004~07) focused on 30 Mb (~ 1%) of the genome

    - International consortium of computational and laboratory-based scientists worked to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function

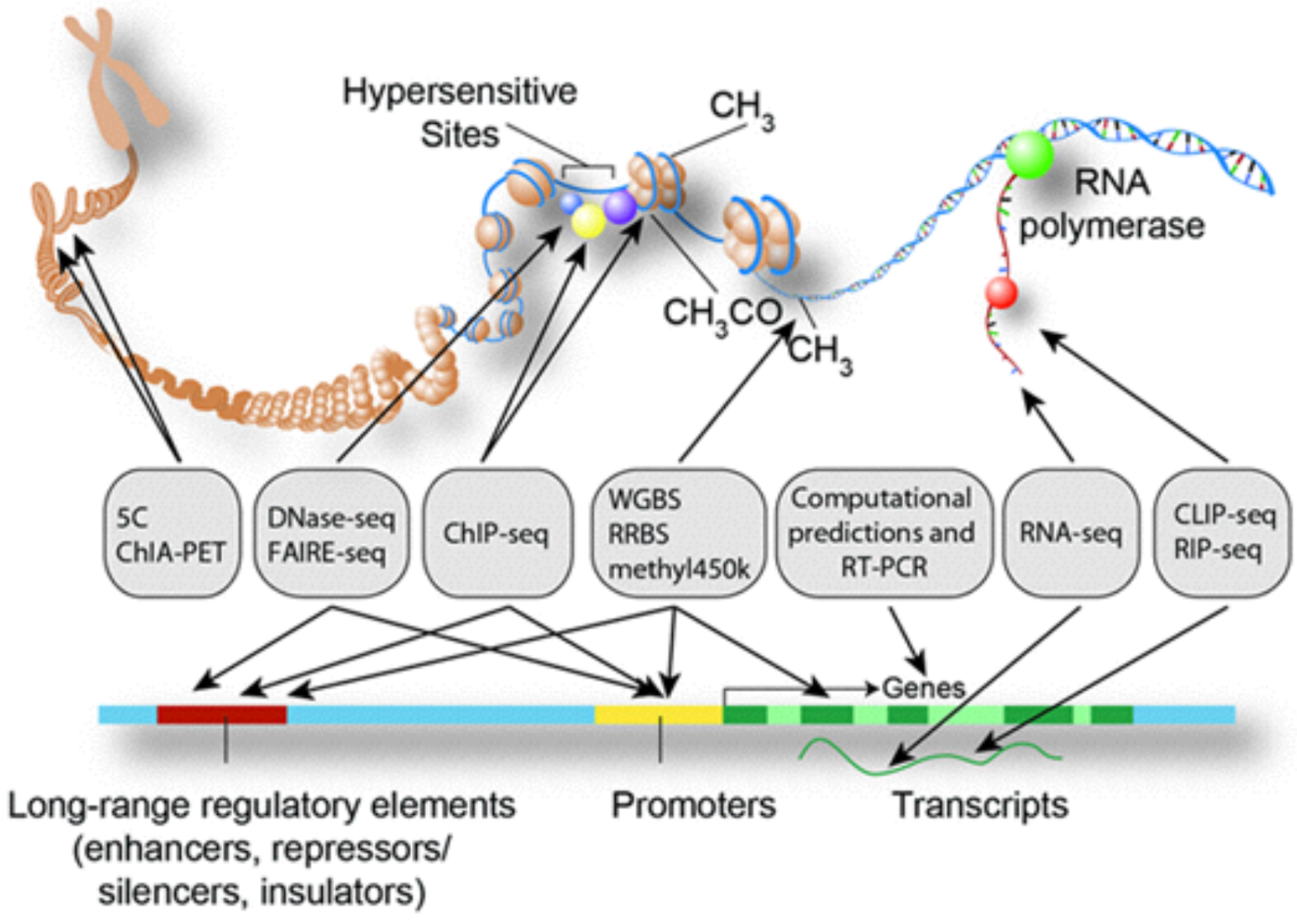
- ENCODE 2/production phase (2007~2012) extended study to entire human genome

    - modENCODE is a similar effort for the fly and worm genomes

- ENCODE 3 (2012~) extends study to hundreds of cell lines, human disease, human tissues, and weights more on computational methodologies

    - Mouse ENCODE joins third phase (2012~).

The scope of ENCODE



Protein-coding genes

- Protein coding genes are easier to find than other elements
- Why? Because DNA -> mRNA -> protein
- We can identify them by looking at
  - **RNA** sequence(e.g. cDNA or EST sequencing, microarray, next-gen sequencing, etc.)
  - or **protein** sequence (e.g. mass-spec, etc)

Protein-coding genes

- We can also predict genes *ab initio* using computational methods
- Option 1: **Conservation**. High homology in protein function domains (especially at level of BLASTp, BLASTx, tBLASTx)
- Option 2: Scan for **recognizable features**
  - open reading frames (ORFs)
  - codon usage bias – 64 codons -> ~ 20 amino acids
  - known transcription and translational start and stop motifs (promoters, 3′ poly-A sites)
  - splice consensus sequences at intron-exon boundaries

ab initio gene discovery

- Protein-coding genes have **recognizable features**
- We can design software to scan the genome and identify these features
    - Simple method: ORF finder (we will practice it on Thursday). Scan for the potential ORF
    - Machine learning approaches to integrate multiple features: common ones are **Artificial Neural Networks** and **Hidden Markov Models**
        - Input: features measured in numbers of certain DNA or RNA sequence
        - Output: possibility that the DNA or RNA encode a protein-coding gene

ORF Finder

- Implementation:
  - Given a DNA/RNA sequence, scan for potential ORFs by translating DNA/RNA into protein sequence using the **codon table**.
  - DNA has **6** possible reading frames
  - RNA has **3** possible reading frames
  - The length of translated protein sequence should be **larger** than certain cutoff.
- Limitations:
  - Work quite well, especially in bacteria and simpler eukaryotes with smaller and more compact genomes
  - It's a lot harder for the higher eukaryotes where there are a lot of long introns, genes can be found within introns of other genes, etc.
  - We tend to do OK finding CDSs, but miss a lot of non-coding 5' exons and the like -- UTRs

Reading frame

- The message is read in a **nonoverlapping** triplet code. Furthermore, the mRNA is normally read in only **one of three** possible reading frames. The one with **start codon** and **stop codon** is called **open reading frame**.

5’-AGGTGACACCGCAAGCCTTATATTAGC-3’

AGGITGAICACICGCIAAGICCTITATIATTIAGC
A**IGGTIGACIACCIGCAIAGCICTTIATAITTA**IGC
AG**IGTGIACAICCGICAAIGCCITTAITATITAGI**C

- If we plan to decode information on DNA without knowing the actual mRNA (e.g. we have the whole genome sequenced for a new species), there will be **six** possible reading frames.

Codon Table

- Start codon: AUG (in most of cases) which also encodes methionine.
- Stop/nonsense codons: UAG (amber), UAA (ochre), UGA (opal)

1st position (5' end) ↓	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end) ↓
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

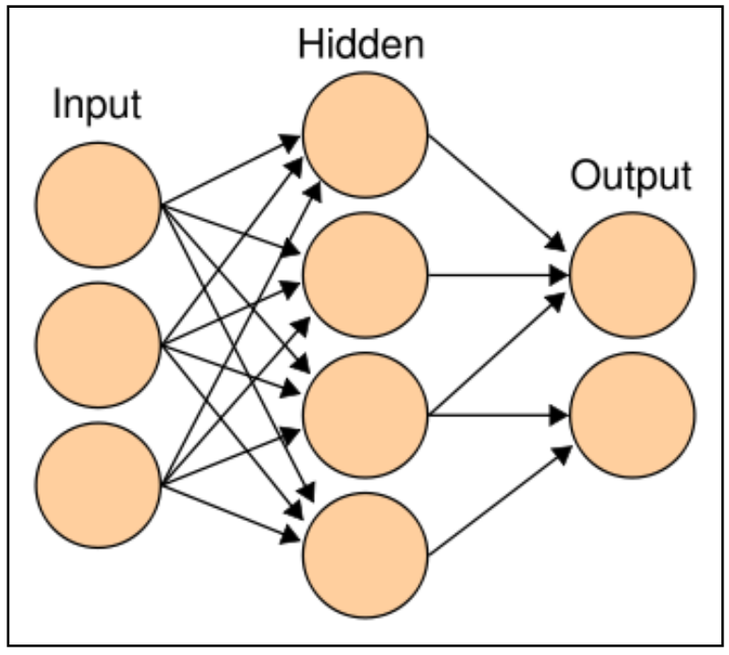
Machine Learning Approaches

- Most gene-discovery programs makes use of some form of machine learning algorithm. A machine learning algorithm requires a **training set** of input data that the computer uses to "learn" how to find a **pattern**.
- Two common machine learning approaches used in gene discovery (and many other bioinformatics applications) are **Artificial Neural Networks** (ANNs) and **Hidden Markov Models** (HMMs).

Artificial Neural Networks

- Neural networks “mimic” the brain in the sense that **connections** between **nodes** (“neurons”) vary in **strength (weight)**, with strength increasing or decreasing during **training**. One or more “hidden layers” learn relationships among the inputs.

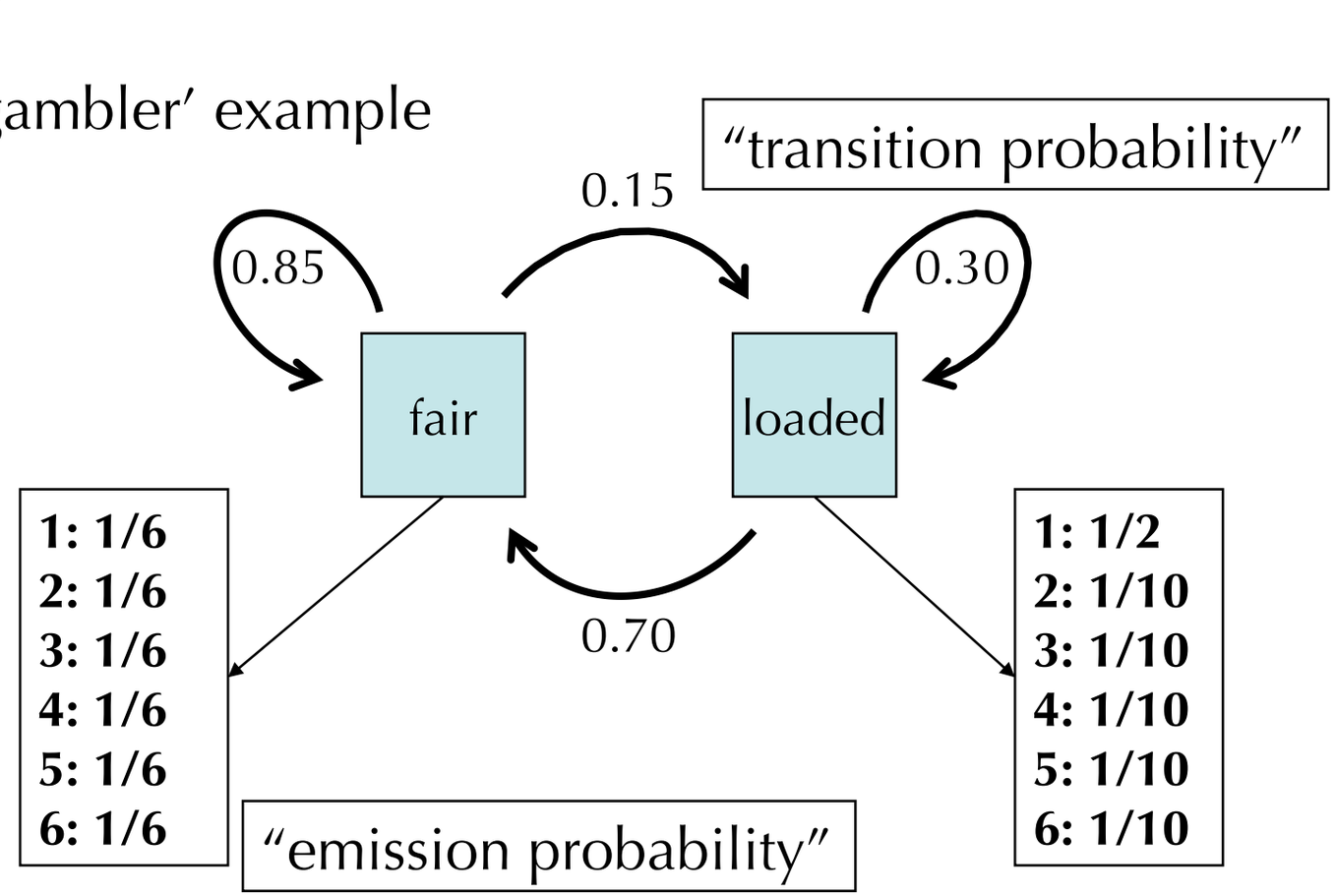
- e.g. Grail tool. Input features include **GC-compositions**, **scores for splice donor and acceptor sites**, **length of region**, and **6-mer frequencies**. Output is in the form of “**exon**” or “**not exon.**”



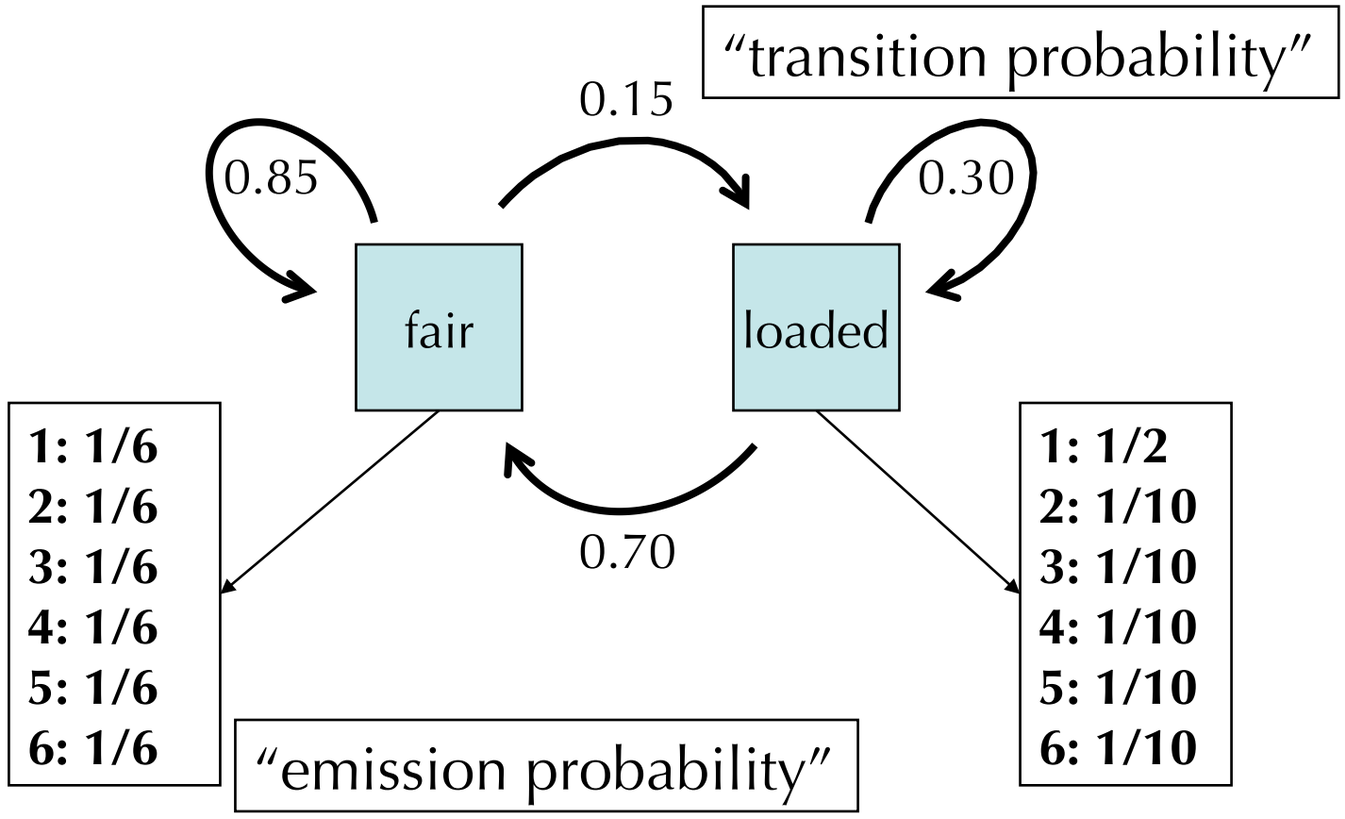
Hidden Markov Models

- HMM is a probabilistic framework in which we move from **state** to **state** (e.g., exon to intron) with certain probabilities. Moving to a new state depends only on the **current** or **a defined number of immediately proceeding states** (this is known as a Markov process). Sometimes the **states themselves are not known**, in which case we have a hidden Markov process.

- The 'dishonest gambler' example

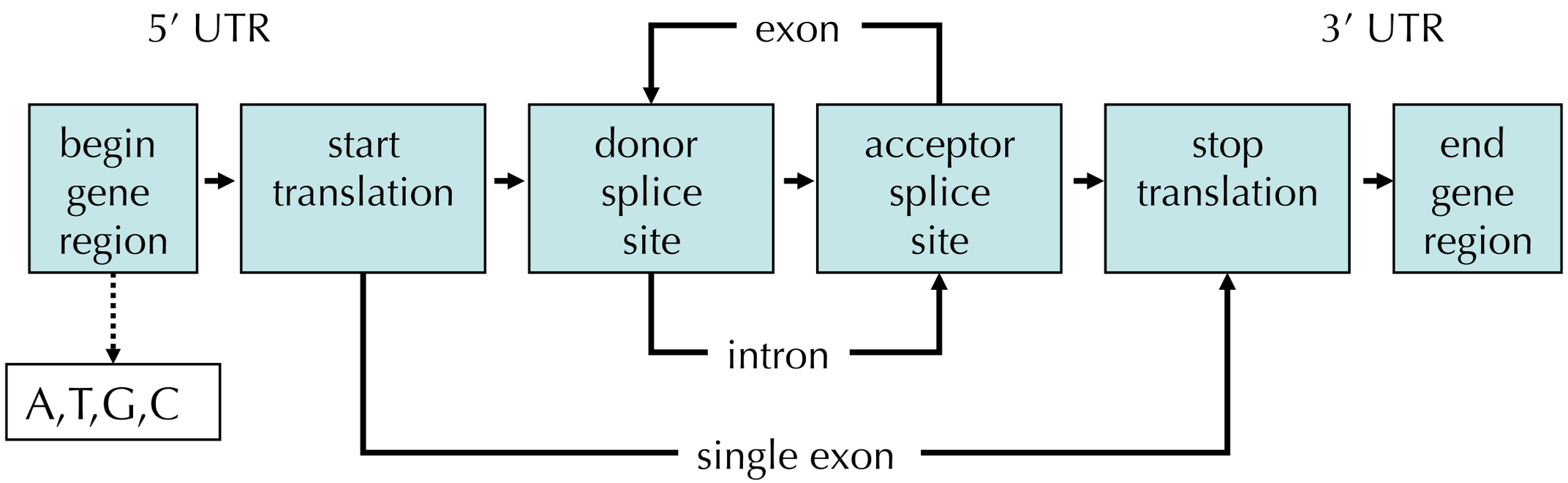


Hidden Markov Models



- Q: if the observed numbers are: **1-1-1-4**, then when the loaded dice is used?
- **Probability/Likelihood of a certain state path** needs to be calculated from emission and transition probabilities.
- e.g. all fair dice: 1/6*0.85*1/6*0.85*1/6*0.85*1/6 = 0.00047
  all loaded dice: 1/2*0.3*1/2*0.3*1/2*0.3*1/10 = 0.00034
- Think: how to answer our Q? We need to test every possible state paths! (**Vertibi**)

HMMs in gene discovery

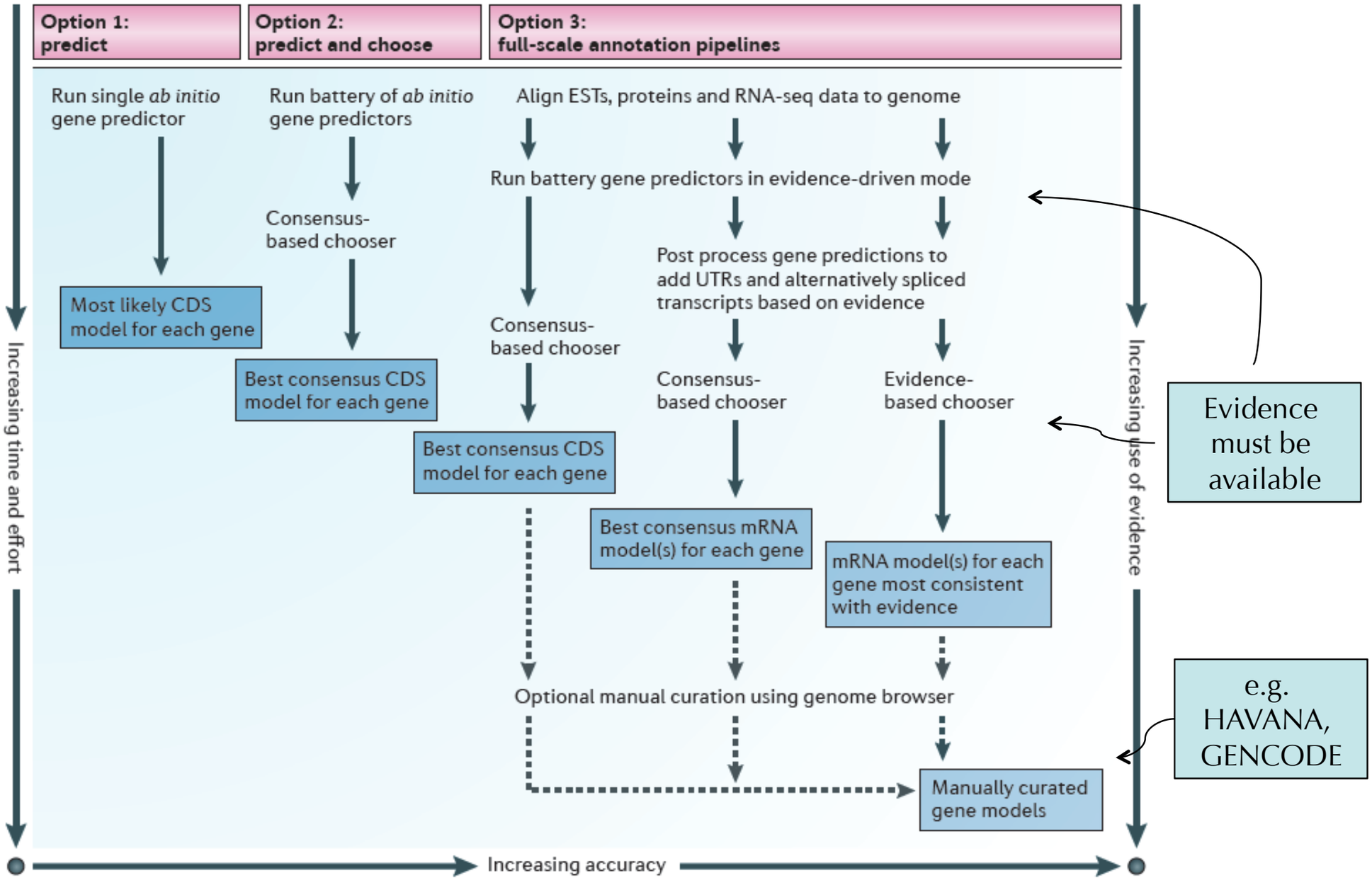


adapted from Gibson and Muse, *A Primer of Genome Science*

- A simple HMM for predicting gene, used by Genscan tool. Each box and arrow has associated transition probabilities, and emission probabilities for observation of nucleotides (dotted arrow). These are **learned** from set of known gene models.

Validate Gene Annotation

- More complex than **gene prediction**
- Need to combine experimental evidence and manual curation!



Elements in a genome

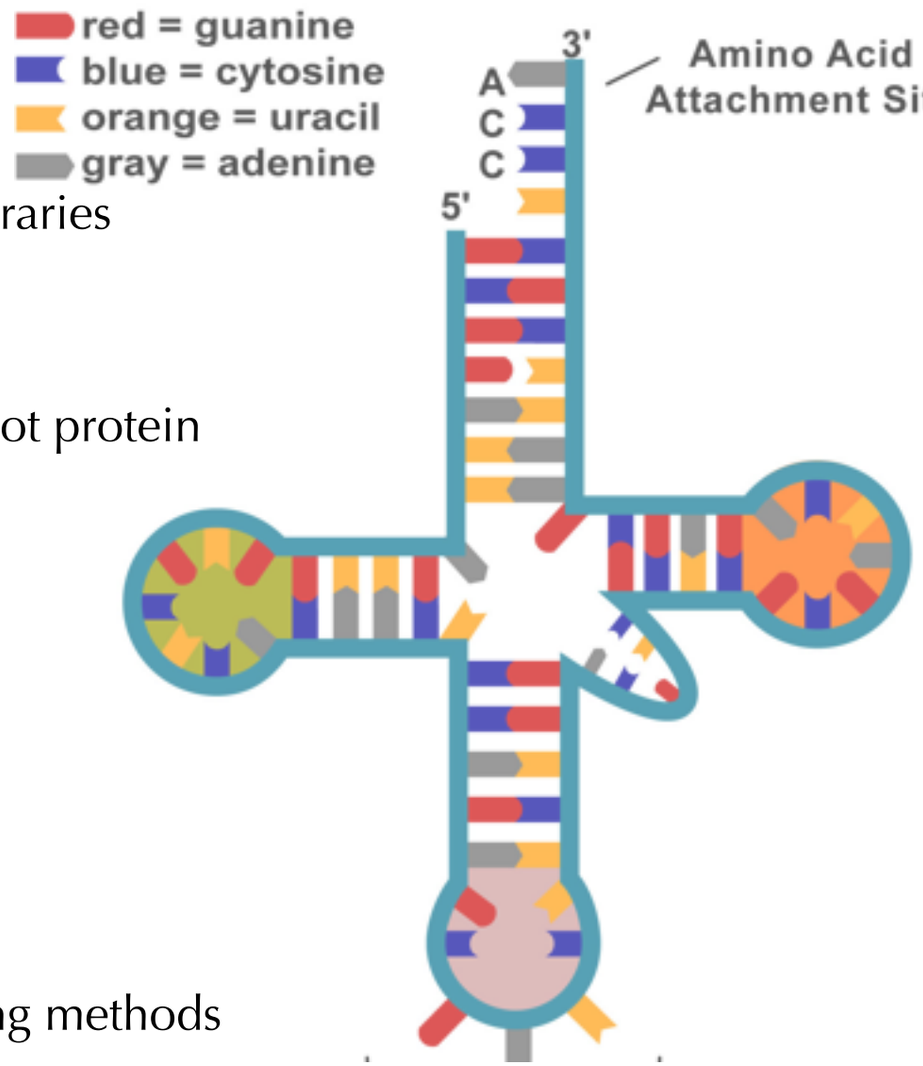
- Note: these are based on what we have learned so far.
- Functional **genetic** information/Genes:
  - **Protein coding genes: those can be transcribed to mRNA then be translated into proteins.**

    only 2% of human genome
  - **Non-protein coding genes: those RNA can't be translated into proteins.**

    aka ncRNA: rRNA, tRNA, miRNAs, lncRNA and etc.
- Functional **epigenetic** information**:**
  - Structural sequences (scaffold attachment regions)
  - Regulatory elements (promoter, enhancer, repressor and so on)
  - Other (including repeats, transposons, retroviral insertions, etc.)

Predict non-protein coding genes

- Including tRNA, rRNA, snoRNA, miRNA, lncRNA, and various other ncRNAs
- Harder to predict than protein-coding genes
- Why?
  - often not poly-A tailed—don't end up in cDNA libraries
  - no ORF
  - constraint on sequence divergence at nucleotide not protein level, so homology is harder to detect
- So, how do we find these?
  - **secondary structure**
  - **homology**, especially alignment of related species
  - **experimentally.** Steps**:**
    - isolation through non-polyA dependent cloning methods
    - microarrays/next-gen sequencing
    - filter out those has coding potentials



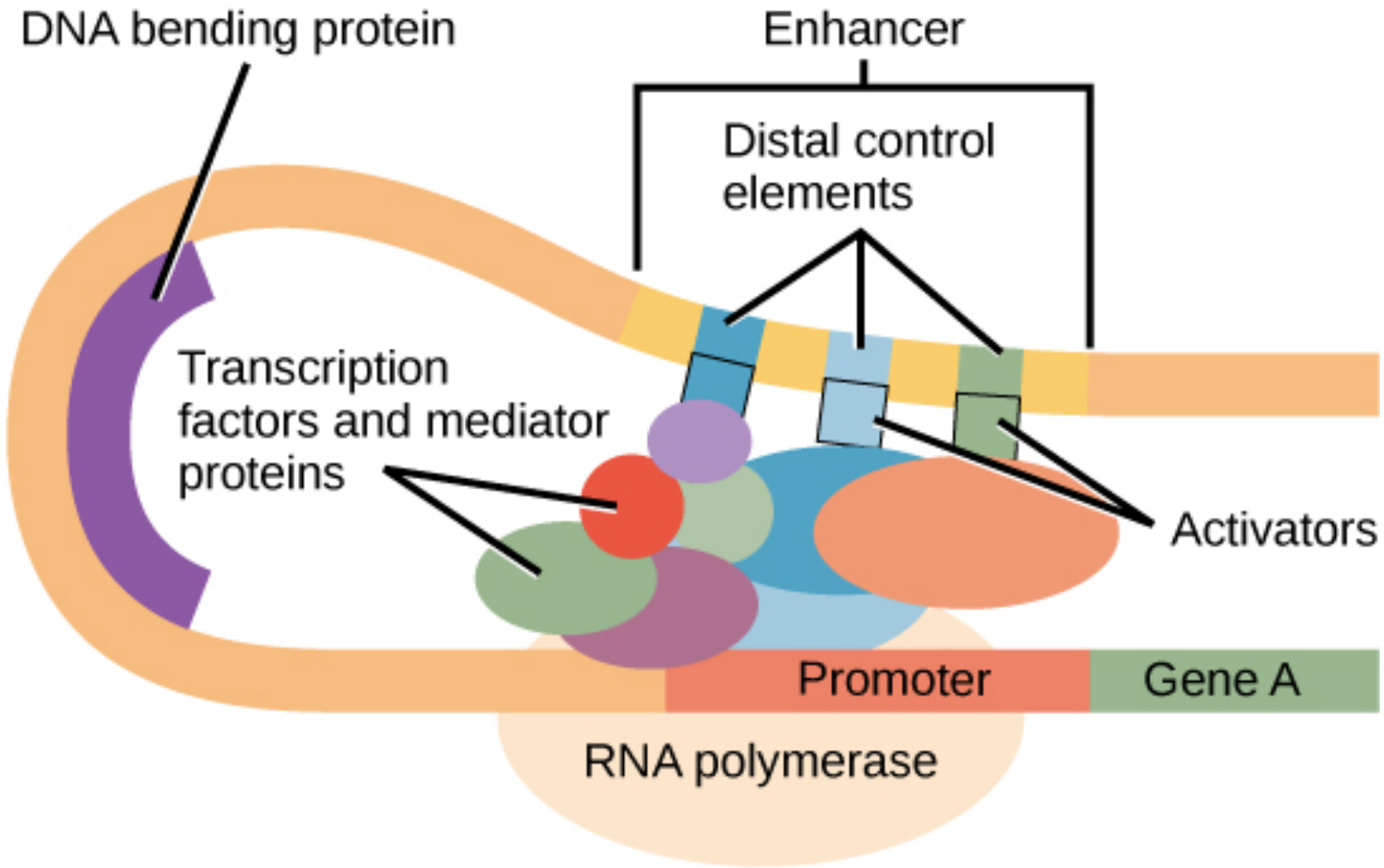
Elements in a genome

- Note: these are based on what we have learned so far.
- Functional **genetic** information/Genes:
  - **Protein coding genes: those can be transcribed to mRNA then be translated into proteins.**

    only 2% of human genome
  - **Non-protein coding genes: those RNA can't be translated into proteins.**

    aka ncRNA: rRNA, tRNA, miRNAs, lncRNA and etc.
- Functional **epigenetic** information**:**
  - Structural sequences (scaffold attachment regions)
  - **Regulatory elements** (promoter, enhancer, repressor and so on)
  - Other (including repeats, transposons, retroviral insertions, etc.)

Regulatory elements are bound by transcription factors



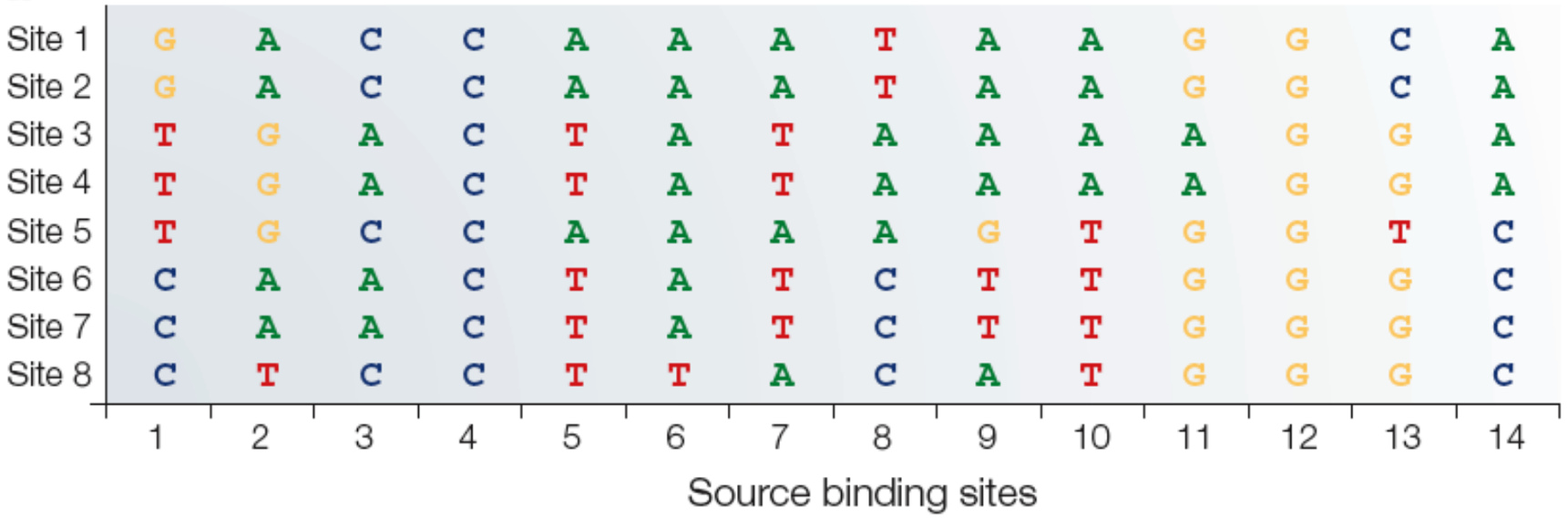
- Special DNA sequence is bound at regulatory element.
- Note, some regulatory elements can repress gene expression.

Identify Transcription factor binding sites

- Usually, binding sites are first determined empirically:
  - - EMSA
  - - SELEX (Systematic Evolution of Ligands by EXponential enrichment)
    - - protein-binding microarrays — aka PBM
    - - DNaseI footprinting/ATAC-Seq
    - - ChIP-chip/ChIP-seq
- Then Q is: what is the special DNA pattern at binding sites, and how to present it?

Sequence motif

a



b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

Calculation

Box 2 | **Formulae linked to methods for the analysis of regulatory sequences**

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation:

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

*pseudocount

$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion:

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

$p(b)$ = background probability of base b; $p(b,i)$ = corrected probability of base b in position i; $W_{b,i}$ = PWM vaue of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3).

Evaluation of sequences:

$$S = \sum_{i=1}^{w} W_{l_i,i} \quad (3)$$

l_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

Information content calculation:

$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \quad (4)$$

D_i = information content in position i; $p(b,i)$ = corrected probability of base b in position i

Wasserman, W. W. and A. Sandelin (2004). *Nat Rev Genet* **5**(4): 276-287.

Sequence Logo

b

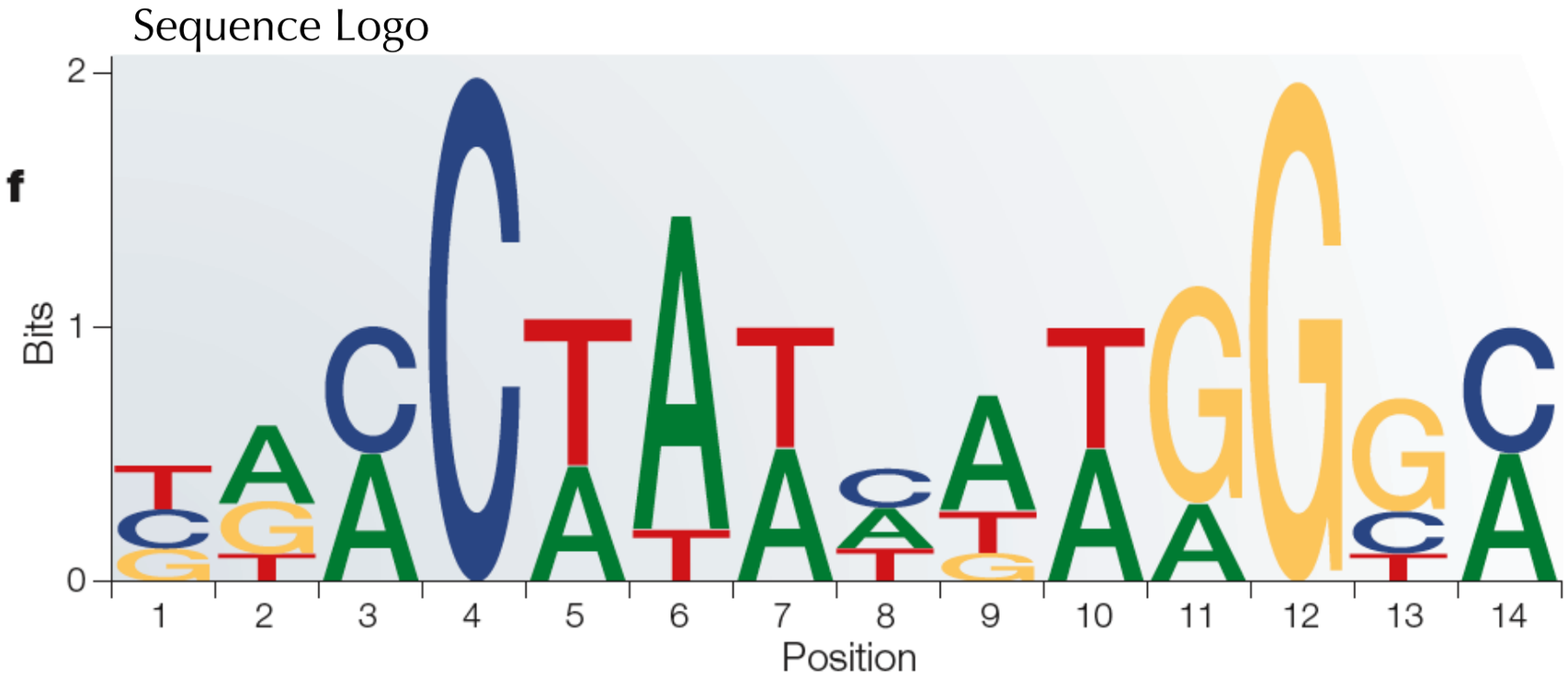
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

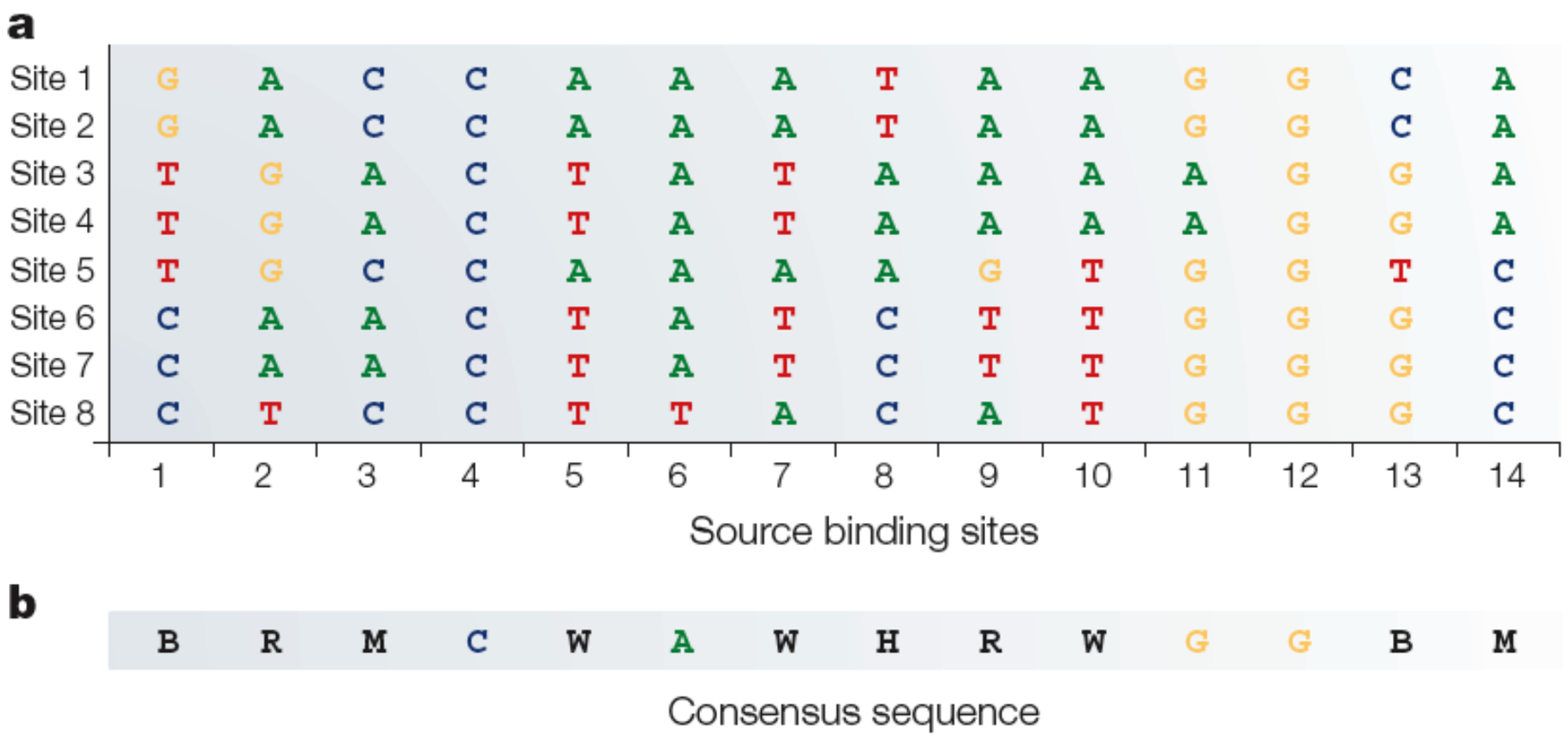
c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

f



Consensus sequence



IUPAC nucleotide code	
	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

- Consensus sequence is easy to be applied. We will practice it on Thursday.
  - A simple Python code:

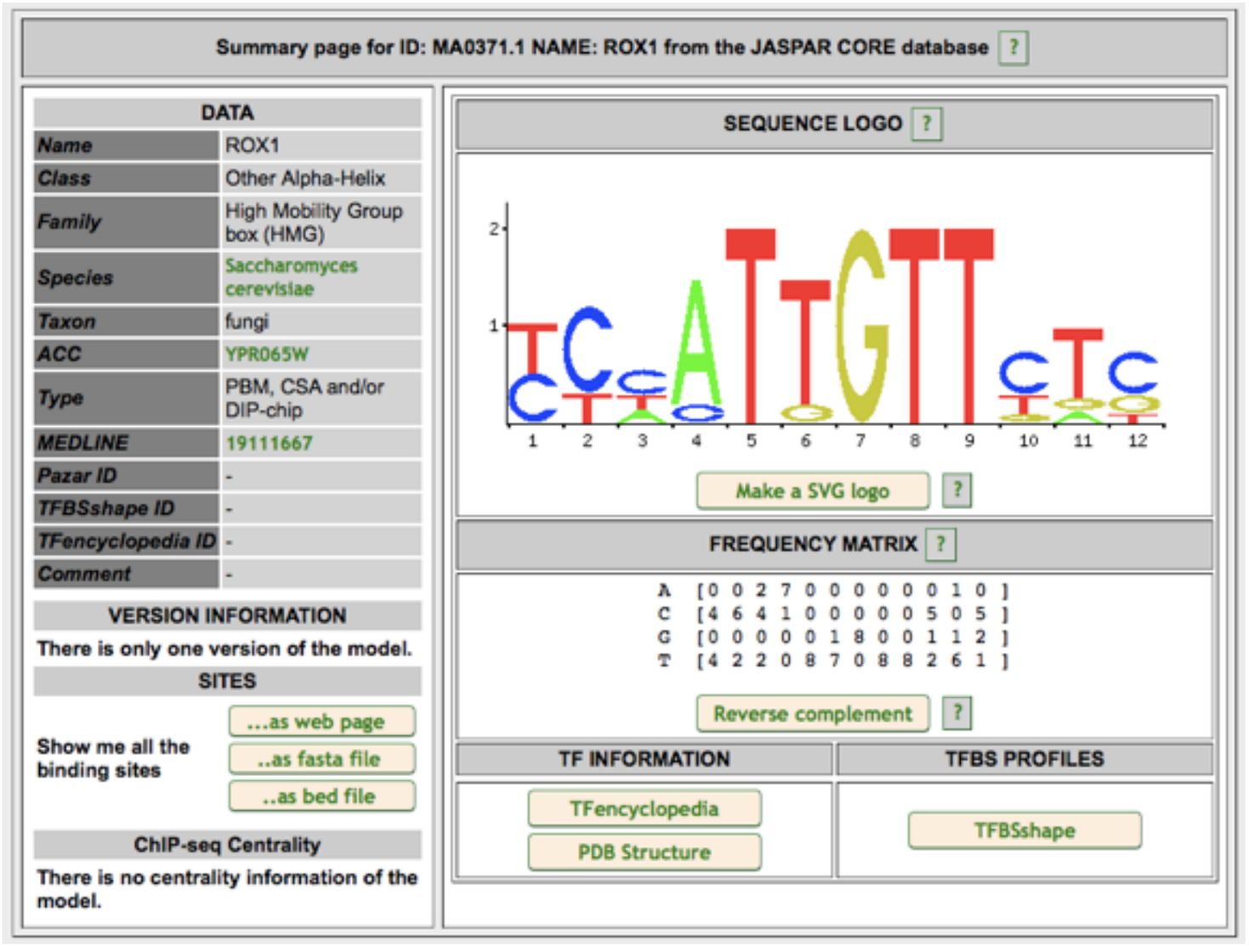
```
import re
for s in sequences:
    if re.match("[TC]C[TC]GGA[TA][GC][CT]", s):
        do_something(s)
```

PFM/PWM

- However PFM/PWMs are generally more useful than consensus:
  - they allow us to assign more importance to more invariant positions
  - they are related to the binding energy of the DNA-protein interaction
  - we can compare PWMs and we can score PWMs
- Scores are based on the **probability** of a given nucleotide being in a given position. e.g. the final score = sum of all log2 scaled weights of observed base at all positions.
- However, a cutoff needs to be selected, such as the minimum score. But “high scoring” does not necessarily mean “biologically relevant” and that “stronger binding” does not automatically imply “better function.”

Motif databases

- Matrices for known TFs have been collected into the **TRANSFAC** (http://www.gene-regulation.de/) and **JASPAR** (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl) databases.
- **TRANSFAC** is bigger and commercial; **JASPAR** is curated and free



Data integration to define functional elements

- Implemented by researchers of modENCODE and ENCODE projects.
- Capture the pattern of epigenetic features (mainly histone modifications) at different types of functional elements such as: gene, promoter, enhancer, repressor, heterochromatin and so on.
- Epigenetic features can be regarded as variables at each basepair along the whole genome.
- Use an unsupervised machine learning, mainly **HMMs** given a sometimes arbitrary number of states, to learn the patterns and then ‘segment’/‘label’ the whole genome.