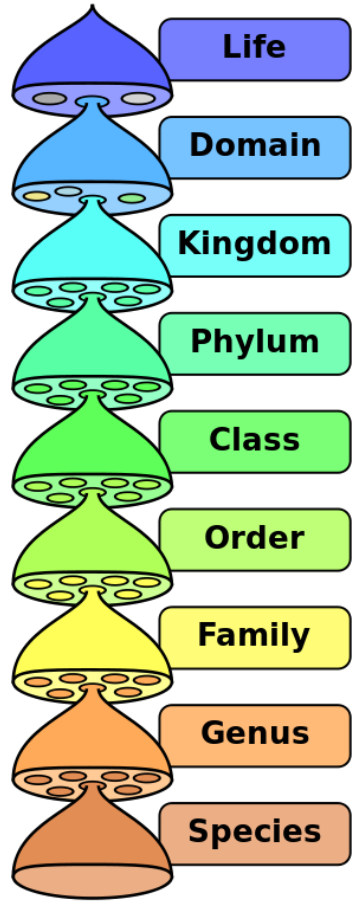
Hidden Markov Models: lecture 9

Application to genetic data

Xavier Didelot

Introduction to biology

- Biology is the science that studies living organisms
- Life on Earth is classified hierarchically
- The highest level is the three domains (archaea, bacteria, eukarya)
- The lowest level is the species
- Viruses are not included, but may still be of interest

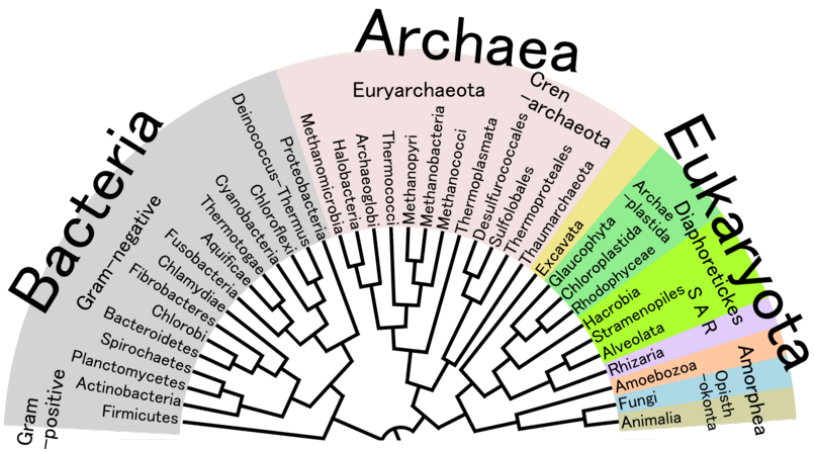


A famous example...

Classification	Example
Domain	Eukarya
Kingdom	Animalia
Phylum	Chordata
Class	Mammalia
Order	Primates
Family	Hominidae
Genus	Homo
Species	sapiens

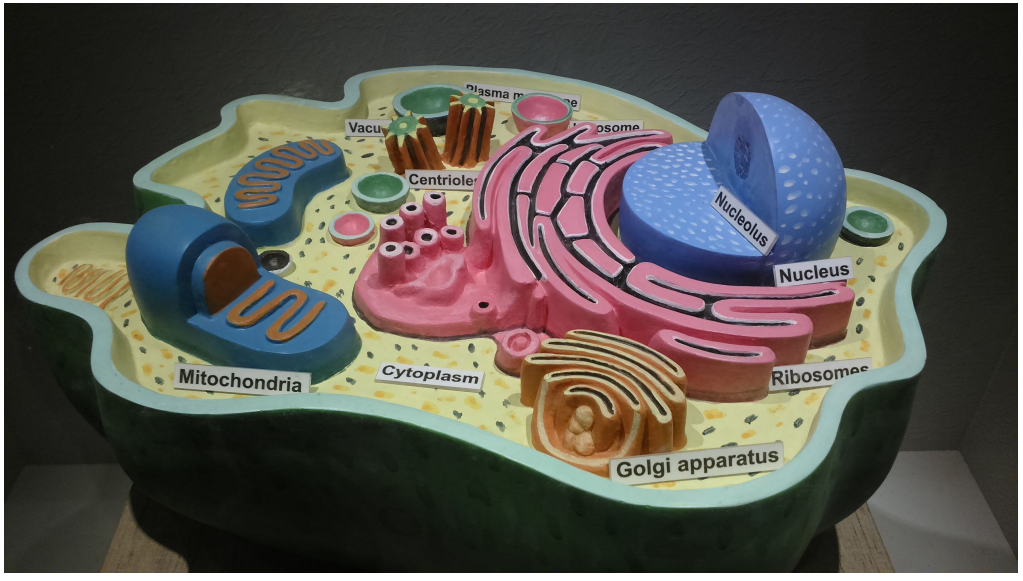
The tree of life

- The hierarchical classification can be represented as a tree
- How can we reconstruct this tree of life?
- First classifications were based on observable traits
- Nowadays we have a much more powerful tool: genetic data
- Can we use genetic data to "count" the differences between organisms and deduce an evolutionary tree?



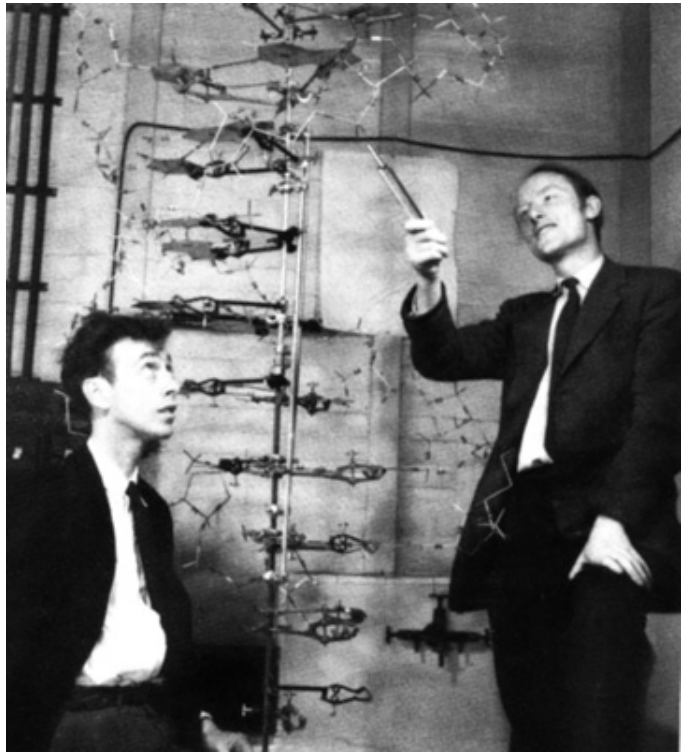
Cells are the basic unit of life

- All living organisms on Earth are made of cells
- Some organisms are made of a single cell (bacteria, archaea, and unicellular eukaryotes) and others are made of multiple cells which may be differentiated in their functions
- A model of an animal cell:



DNA

- Each cell contains a copy of the whole genetic inherited information allowing it to perform biochemical functions
- In 1953, Watson and Crick discovered that the genetic information is stored in molecules of DNA (except for some viruses using RNA instead)



The structure of DNA

- DNA is a double helix
- Each part is made of a backbone (dark blue) supporting a sequence of nucleotides
- There are four nucleotides: Adenine (blue), Thymine (yellow), Guanine (green) and Cytosine (red)
- A on one side is always aligned with T, and C is always aligned with G
- The double helix can split into two parts, each of which contains the full information
- This is useful for both reproduction and growth (in multicellular organisms)

Transcription

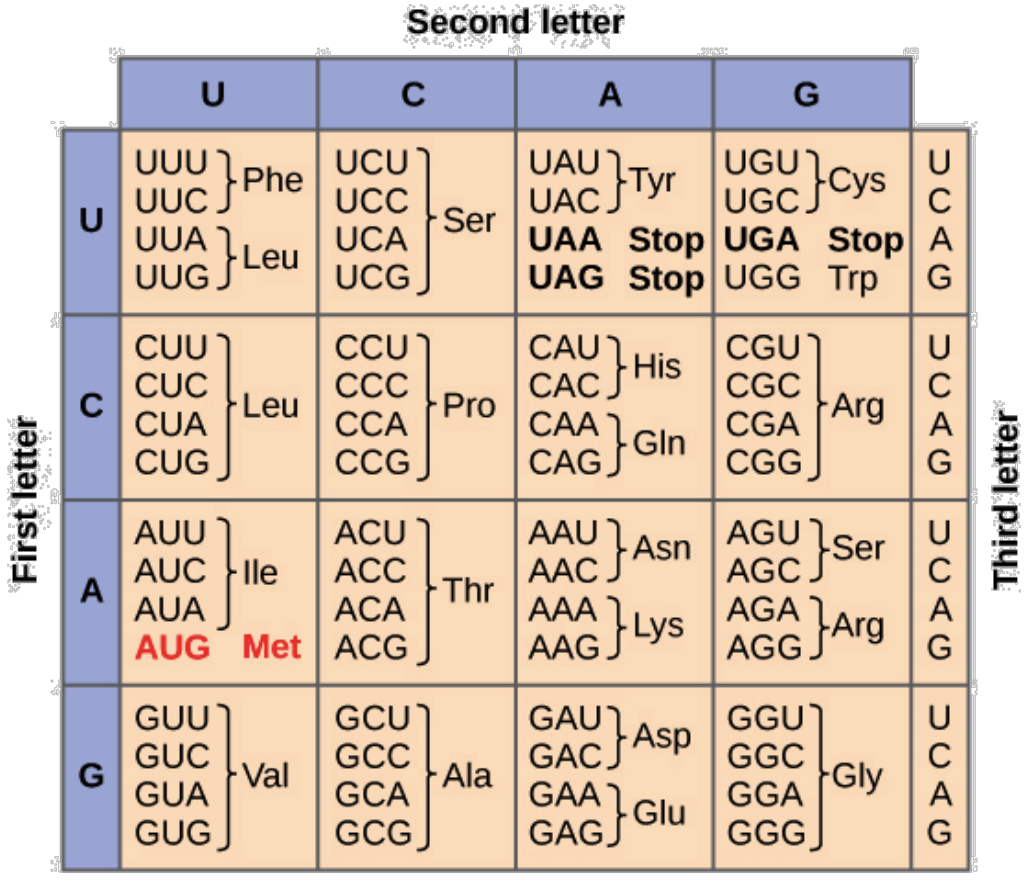
- Genes are specific subsets of the DNA sequence of nucleotides
- A gene is transcribed into a RNA molecule called messenger RNA or mRNA
- The mRNA molecule is a chain of nucleotides similar to DNA but encoding just a given gene
- The mRNA molecule is of the exact same length as the gene
- RNA uses a Uracil (U) nucleotide instead of Thymine (T)

Translation

- The mRNA molecule is translated into a protein, which has a certain function
- Proteins are chains of amino acids
- Each triplet of nucleotides encodes a given amino acid, determined by the genetic code which is (almost) universal

- Together, transcription and translation represent the central dogma of molecular biology

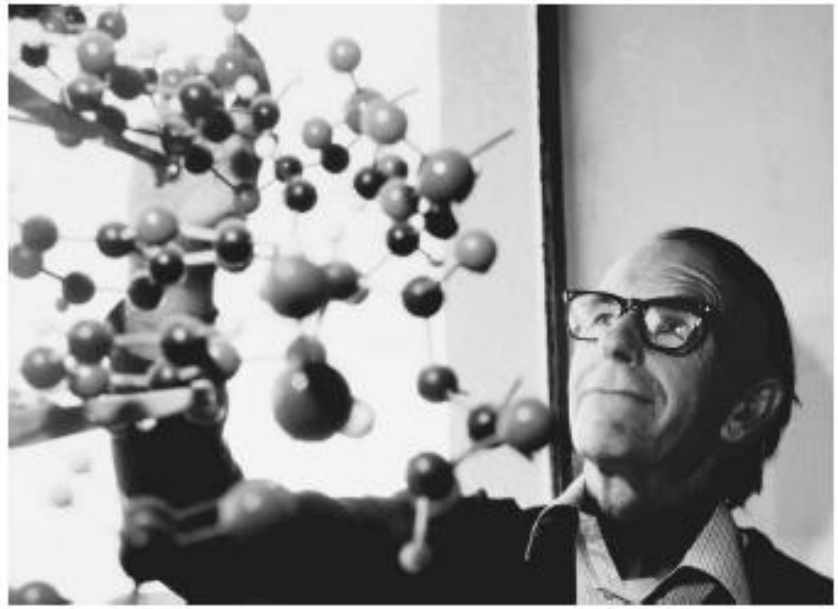
Genetic code



- A gene starts (usually) with a start codon (AUG)
- A gene ends (usually) with a stop codon (UAA, UAG or UGA)

Sequencing of DNA

- DNA sequencing was very tedious until the discovery by Frederick Sanger in 1975 of a new method called Sanger sequencing
- Sequencing techniques have steadily improved since, and current methods (eg Illumina) are fast and cheap
- Sequencing a whole human genome costs about $1000
- Sequencing a whole bacterial genome costs about $10



Let's take a first look at some genetic data!

- This is an extract of the gene coding for insulin in humans:

```
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGC
AGGTCTGTTCCAAGGGCCTTTGCGTCAGGTGGGCTCAGGA
TTCCAGGGTGGCTGGACCCCAGGCCCCAGCTCTGCAGCAG
GGAGGACGTGGCTGGGCTCGTGAAGCATGTGGGGGTGAGC
CCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCC
TCAGCCCTGCCTGTCTCCCAGATCACTGTCCTTCTGCCAT
GGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTG
GCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACC
AACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCT
AGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACC
```

- What sort of statistics shall we use to describe this sequence?
- Can we determine that this is a human gene based on the sequence alone?
- Can we determine that it is a gene?

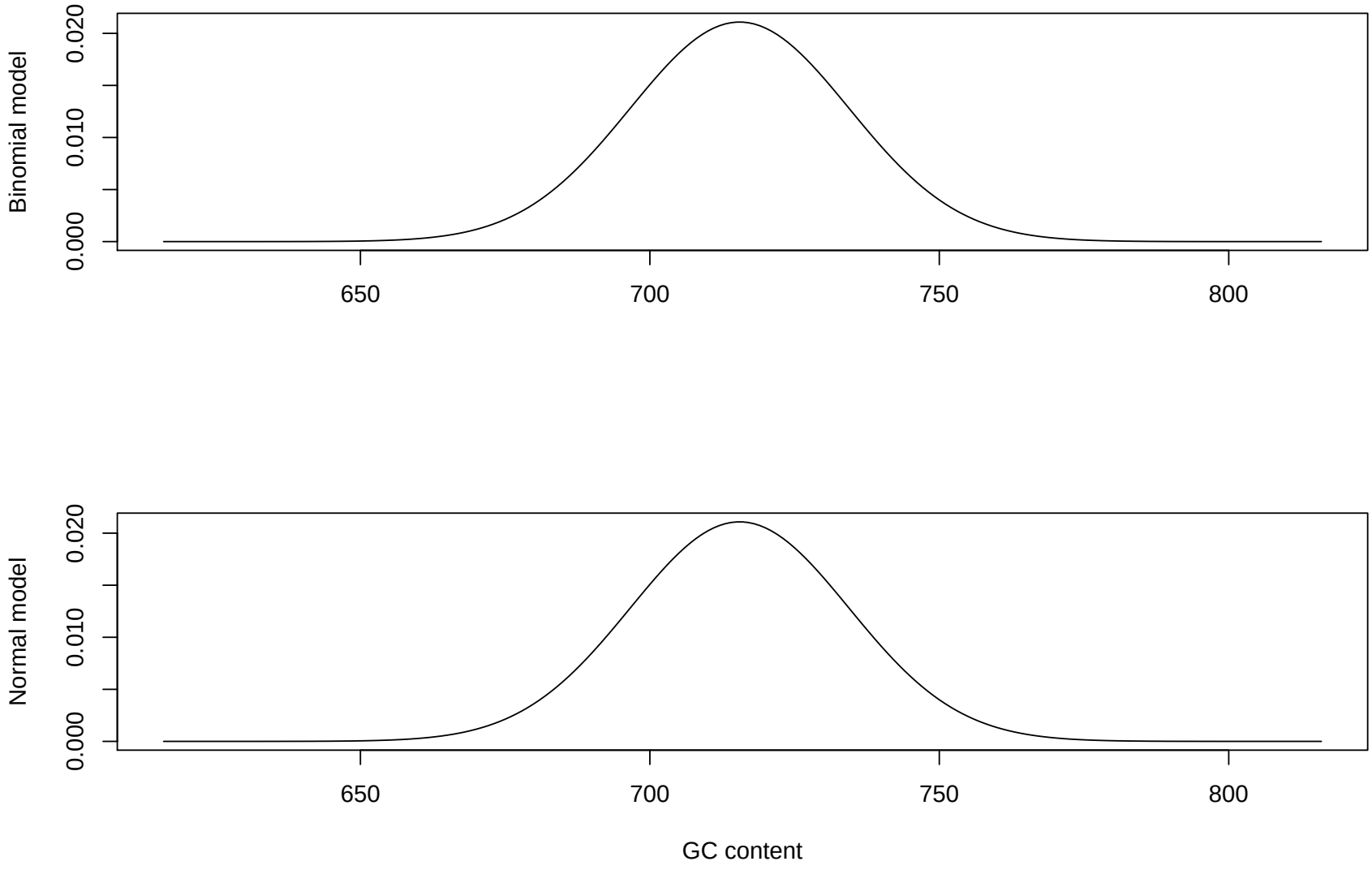
GC content

- An important summary statistic for any DNA sequence is its GC content
- GC content is defined as the proportion of Gs and Cs observed in the sequence as opposed to As and Ts
- Since Gs and Cs are paired in DNA, this statistic is the same for both strands
- The full insulin gene is of length 1431 bp. bp stands for base pair and is the unit used to measure DNA length
- There are 925 Gs and Cs in the sequence
- So the GC content is 0.6464
- Is this surprising? What would you expect?

GC content model

- A null model would have all four nucleotides occurring with equal probability 0.25 at each site.
- The distribution for the number of observed Gs and Cs in a sequence of length n would then be Binomial$(n, 0.5)$
- The probability of observing at least x Gs and Cs in a sequence of length n is equal to $p = 1 - F_{\text{Binomial}(n,0.5)}(x)$
- For this insulin gene, $n = 1431$ and $x = 925$ and the probability is $p =$ 2.43e-29
- We could also approximate using the central limit theorem to a Normal$(0.5n, 0.25n)$ which gives $p =$ 8.18e-29
- Clearly there is a lot more Gs and Cs in this gene than would be expected just by chance...

Binomial and Normal distributions



GC content examples

Species	GC content	Genome size
Homo sapiens	41%	3.2 Gbp
Sheep	42.4%	2.6 Gbp
Chicken	42.0%	1.2 Gbp
Turtle	43.3%	2.2 Gbp
Salmon	41.2%	3.0 Gbp
Sea urchin	35.0%	0.8 Gbp
A. thaliana (plant)	35.0%	125 Mbp
Malaria (protozoan)	20.0%	22.9 Mbp
E. coli (bacteria)	50.7%	4.6 Mbp
M. genitalium (bacteria)	31.6%	0.6 Mbp
P. aeroginosa (bacteria)	66.4%	6.3 Mbp
T. volcanium (archaea)	39.9%	1.6 Mbp

Base composition

- GC content varies a lot between species
- GC content also varies within species, from one genomic region to another
- Why are the frequencies of As, Cs, Gs and Ts not 0.25? Partly for the same reason that English does not use all 26 letters equally frequently! And other languages have other distributions...
- GC rich regions tend to contain more genes and are therefore of special interest
- Many more interesting properties can be uncovered by considering the marginal distribution of nucleotides
- This is also known as the study of base composition

A simple HMM for GC content

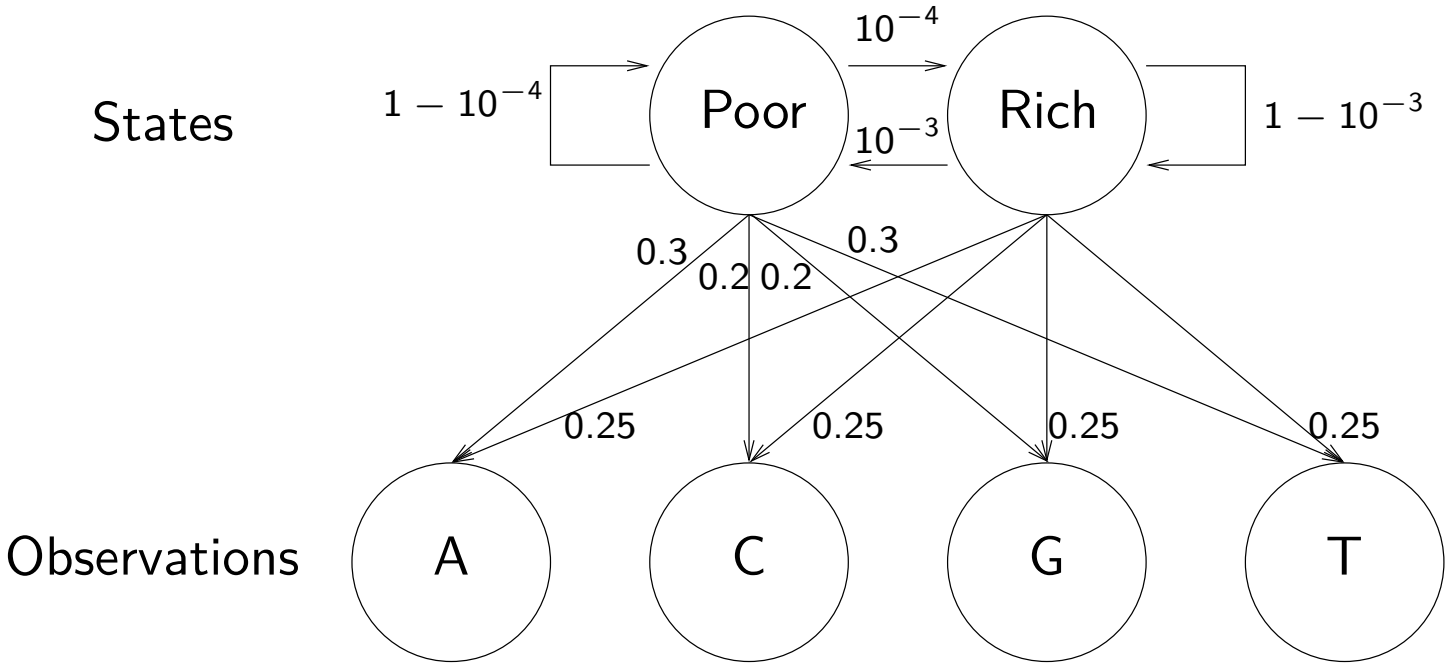
- Some regions of the human genome are more GC rich than others
- We consider a HMM with $m = 2$ states rich (R) and poor (P)
- Let's say that the emission probabilities of A, C, G and T are (0.3,0.2,0.2,0.3) in GC poor regions and (0.25,0.25,0.25,0.25) in GC rich regions. So the emission matrix is:

$$p_k(x) = \begin{array}{c} \\ \mathrm{P} \\ \mathrm{R} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} 0.3 & 0.2 & 0.2 & 0.3 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \end{array}$$

- We expect GC rich region to be of length 1000 bp on average, and GC poor regions to be of length 10,000 bp on average. This gives the following transition matrix:

$$\gamma_{kl} = \begin{array}{c} \\ \mathrm{P} \\ \mathrm{R} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{P} & \mathrm{R} \end{array} \\ \begin{pmatrix} 1 - 10^{-4} & 10^{-4} \\ 10^{-3} & 1 - 10^{-3} \end{pmatrix} \end{array}$$

A simple HMM for GC content



A simple HMM for GC content

- The parameters of the model can be learnt more exactly using the Baum-Welch algorithm
- The Viterbi algorithm can be used to detect regions of high GC content, which are likely to contain genes
- For example the insulin gene is clearly in a GC rich region
- Many other similar applications of HMM to detect patterns in single sequences
- Detection of genes in sequences: can train an HMM using many coding vs non-coding sequences
- How can we analyse multiple sequences to compare them and learn about evolution?

Types of mutations

- Point mutation introduces punctual changes

```
ACTTGCTTTCGACCCTAGGATTTTACTTCTTT
|||||||||||||||| ||||||||||||||
ACTTGCTTTCGACCCTTGGATTTTACTTCTTT
```

- Insertion adds some bases

```
ACTTGCTTTCGACCCTA--GGATTTTACTTCTTT
|||||||||||||||||  ||||||||||||||
ACTTGCTTTCGACCCTAAAGGATTTTACTTCTTT
```

- Deletion removes some bases

```
ACTTGCTTTCGACCCTAGGATTTTACTTCTTT
|||||||||||||||  ||||||||||||||
ACTTGCTTTCGACCCT--GATTTTACTTCTTT
```

Sequence alignment to reveal mutations

- Raw sequences

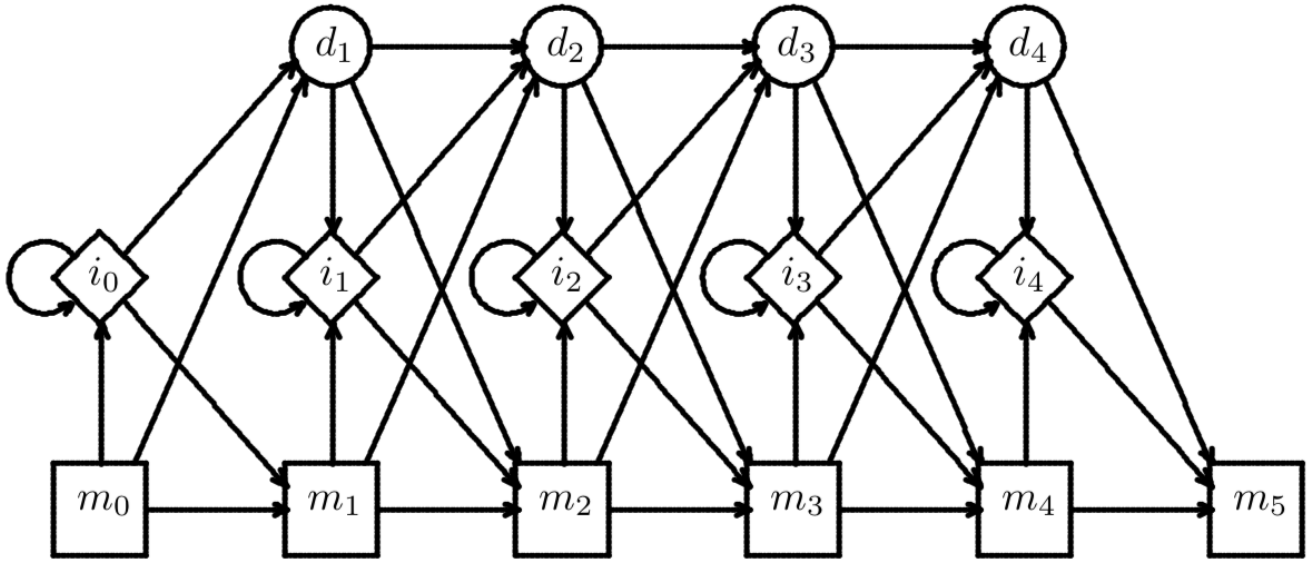
```
ACTTGCTCGACCCTGATTTACTTACTTCTTT
ACTTGCTTTCGACCCAGATTTTACTTCTTT
```

- Impossible to say, based on the raw sequences, what evolutionary events happened
- Aligned sequences

```
ACTTGC--TCGACCCTGATTTACTTACTTCTTT
||||||  ||||||| |||||   |||||||||
ACTTGCTTTCGACCCAGATTT---TACTTCTTT
```

- Based on the aligned sequences, it is clear that there was an insertion, a point mutation and a deletion
- We can now measure how much evolution took place
- Problem: how to align two or more sequences?

HMM to align sequences



- $m_{0..5}$ are match states
- $i_{0..4}$ are insertion states
- $d_{1..4}$ are deletion states

Example

- Let's say we start with a single sequence:
  `ACTTGCTCGACCCTGATTTACTTACTTCTTT`

- We build a HMM with the same number of m states as the length of this sequence
- Each m state emits the corresponding letter with high probability (eg 0.9) or something else with lower probability (eg 0.1, to allow point mutations)
- Each i state emits random sequences
- Each d state emits the special gap character: -
- Transition into m states happens with high probability (to penalise insertions and deletions)
- Transition into i and d states happens with low probability

Example

- We want to align the sequences:
  `ACTTGCTCGACCCTGATTTACTTACTTCTTT`
  `ACTTGCTTTCGACCCAGATTTTACTTCTTT`

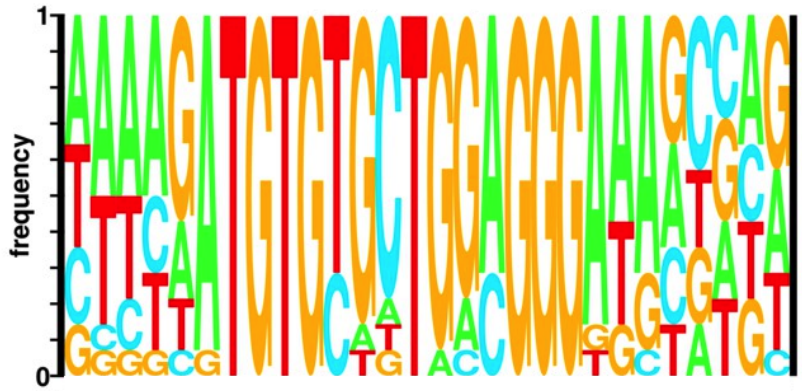
- First build a HMM corresponding to the first sequence
- Then use the Viterbi algorithm on the second sequence and find the optimal path:

$$m_0\, m_1\, m_2\, m_3\, m_4\, m_5\, m_6\, i_6\, i_6\, m_7\, m_8\, m_9\, m_{10}\, m_{11}\, m_{12}\, m_{13}\, m_{14}\, m_{15}\, m_{16}$$
$$m_{17}\, m_{18}\, m_{19}\, d_{20}\, d_{21}\, d_{22}\, m_{23}\, m_{24}\, m_{25}\, m_{26}\, m_{27}\, m_{28}\, m_{29}\, m_{30}\, m_{31}\, m_{32}$$

- This global decoding corresponds to the alignment:
  `ACTTGC--TCGACCCTGATTTACTTACTTCTTT`
  `ACTTGCTTTCGACCCAGATTT---TACTTCTTT`

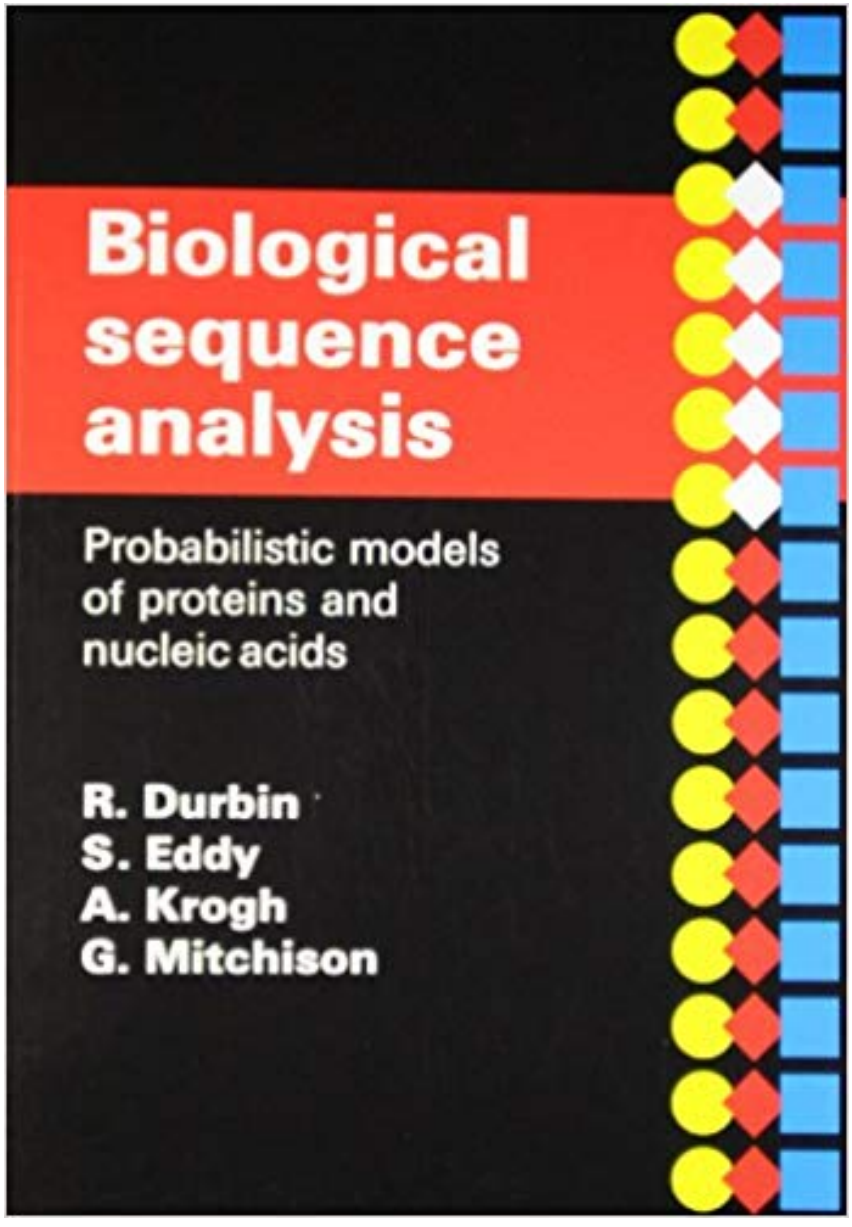
In practice

- We can repeat this procedure to align more and more sequences
- Instead of making a model based on just the first sequence, it is better to build the model based on the **profile** of multiple sequences



- Progressive alignment of increasingly large number of sequences requires to carefully consider the order in which the sequences are added to the alignment
- Another application of the model above is to identify when a gene belongs to a family of genes
- Automatic annotation of genes

To find out more...



Conclusions

- A DNA sequence is a string of four nucleotides A, C, G and T
- A gene is encoded in DNA, transcribed into mRNA and translated into a functional protein
- The sequence composition in terms of these four nucleotides is not uniform, but varies between species and within genomes
- Hidden Markov Models are very useful when analysing genetic data
- Analysis of sequence composition
- Detection of biologically relevant patterns
- Detection of genes
- Annotation of genes
- Alignment of sequences
- Application of HMM in bioinformatics became popular in the 1990s, and is still an area of active research