[1 Chapter 2 - Algorithms 2](#_Toc109996350)

[1.1 Name 8 algorithm strategies 2](#_Toc109996351)

# Sequencing and Assembly

## What is te formula for the Phred Score and what do the parameters mean.





Q is the score

P is the probability that a base was incorrectly called

## What does a FastQ entry look like?

@ Sequence identifier

Sequence

+ Optional, identifier again

Base quality

## Define genome assembly according to Dear et al.

A sequence assembly is essentially a set of contigs, each contig being a multiple alignment of reads.

## Name issues when sequencing eukaryotic organisms.

Technical

* Incomplete coverage
* Sequencing errors
* DNA is double stranded (two orientations)

Biological

* Eukaryotic genes contain many repetitive duplications that are longer than typical read lengths

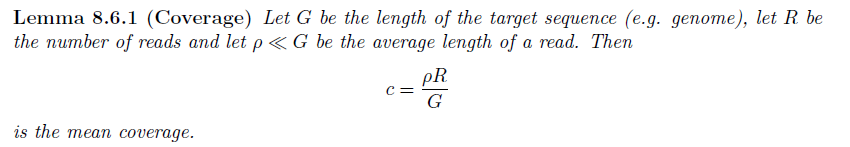
Computational

* Huge amounts of data

## What is the formula for mean coverage?

Sum of all read lengths divided by the length of the target sequence

## What is the lemma that relates the length of the target sequence with the number of reads and the average length of a read?



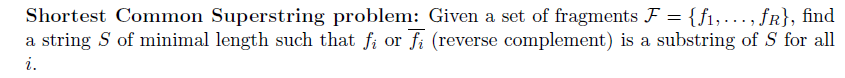
## What is the fragment assembly problem?

Reconstructing the sequence A of a DNA molecule from a set of fragments F.

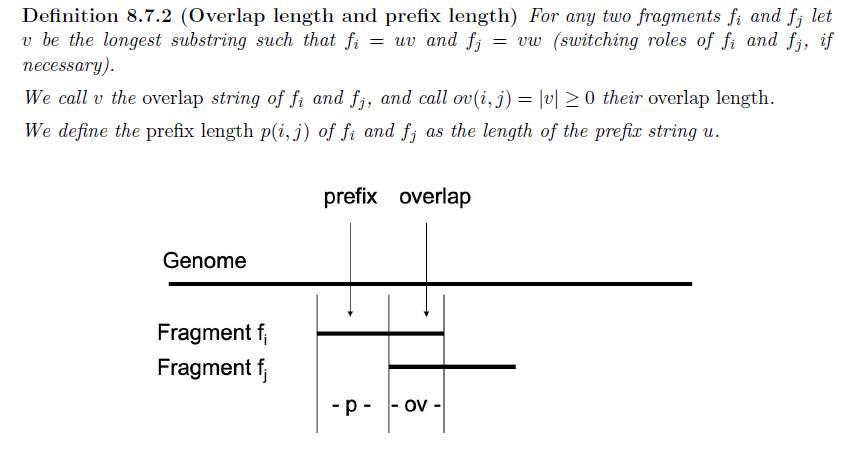
## What are the two laws of assembly?

1. If a suffix of read A is similar to a prefix of read B, then A and B might overlap in the genome
2. More coverage leas to more and longer overlaps and contigs

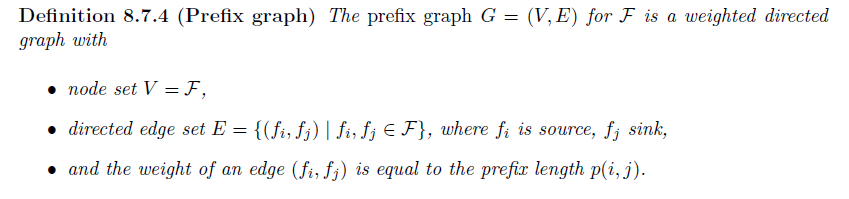
## What is the shortest common superstring problem?



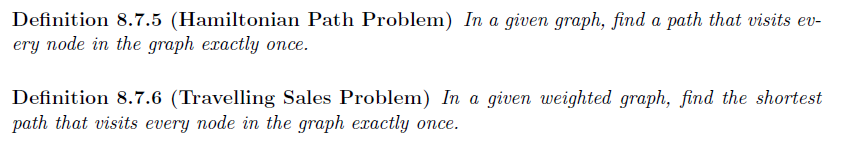
## Define overlap length and prefix length for two fragsments.



## Define prefix graph.



## Define the Hamiltonian Path and Travelling Salesman problem.



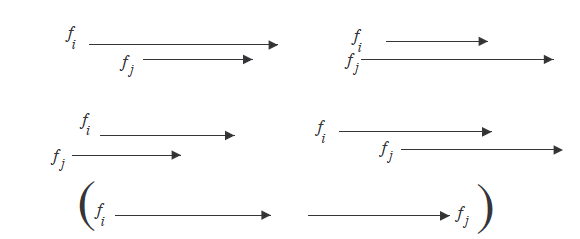
## How is the Shortest spanning tree failing for many biological genomes?

Interior parts of repeating regions are overcompressed

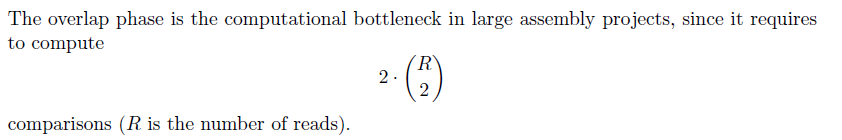
## Describe the three stages of OLC.

1. Overlap: Each read is compared with all other reads, the overlap graph is computed
2. Layout: Pairs of numbers are determined that place all reads in the assembly. From this phase, contigs and their lengths are derived. If paired-end sequencing was performed, the scaffolding phase may follow
3. Consensus: A multi-alignment of all the placed reads is produced to obtain the final contig sequences.

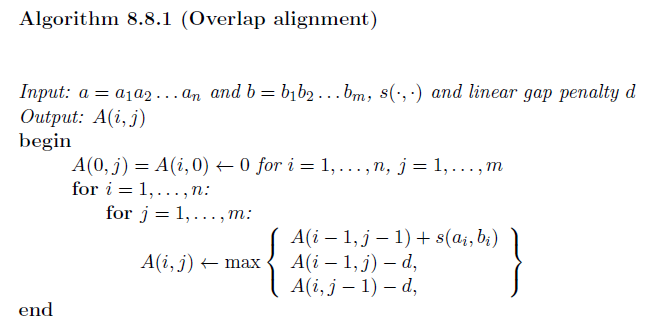
## What are the possible overlap relationships between two fragments with fixed orientation?



## What step of OLC is the computational bottleneck? Why?



## Reproduce the overlap alignment algorithm pseudocode.

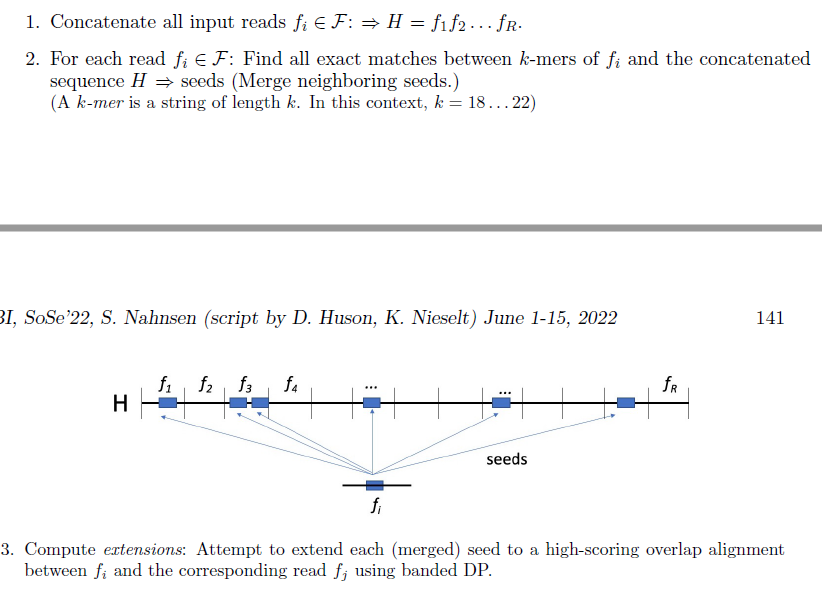


## What is the overlap alignment runtime?

O(nm)

## Why is dynamic programming too slow for large sequencing projects? What is a better approach?

Too slow, only overlaps with 97% identity at least play a role. Instead, seed and extend approach.



## We have found an overlap. What are the three options for its origin?

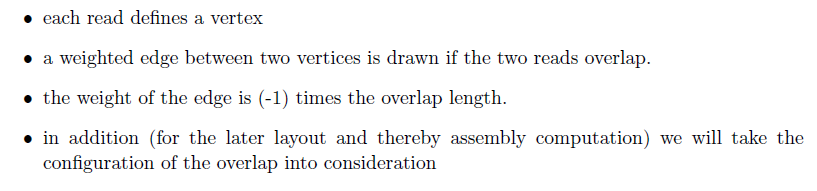
1. True origin
2. Repeat-induced
3. By chance

## How should one avoid repeat-induced overlaps?

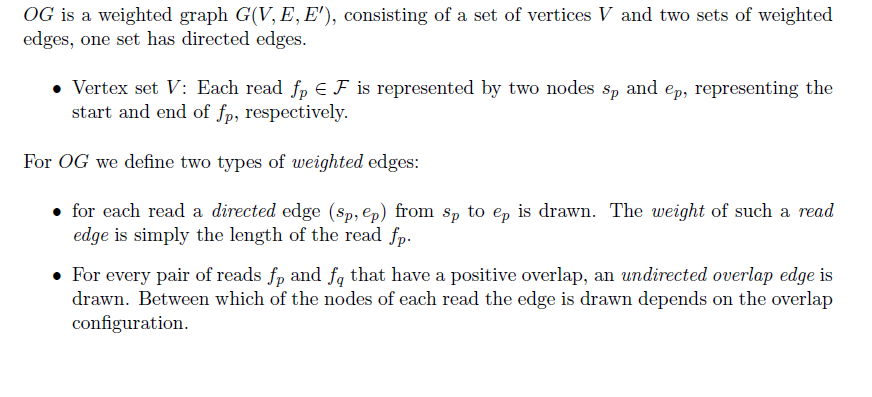
Only start seeds outside a repeat

1. Screening known repeat
2. We screen for each kmer, how often it occurs in H

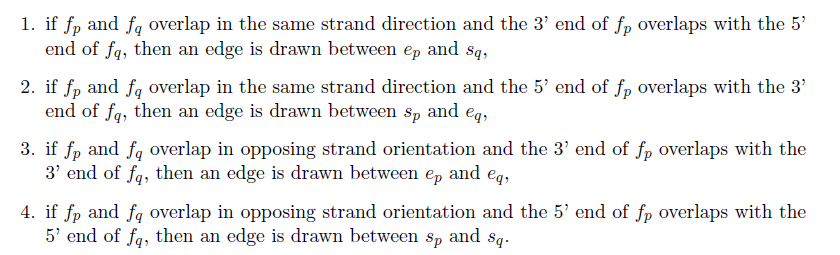
## Describe the rules of the overlap graph



## What is the formal definition of the overlap graph?



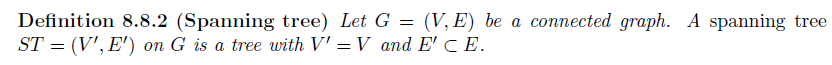
## What are the four possibilities that an overlap can be drawn between the end points of two reads?



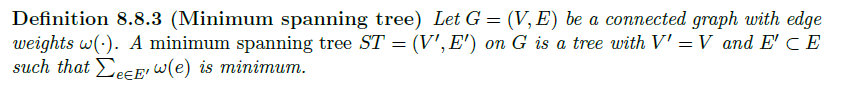
## What is the goal of the layout phase?

Arrange all reads into an approximate multi-alignment.

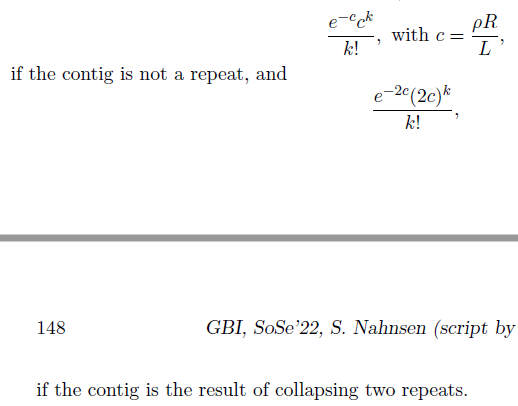
## Define spanning tree



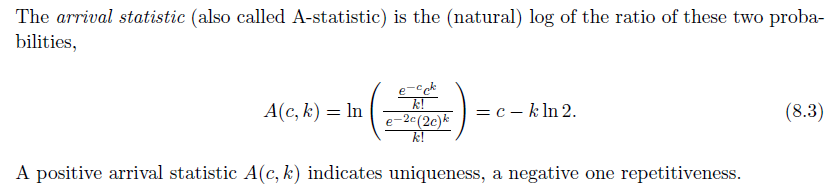
## Define minimum spanning tree.



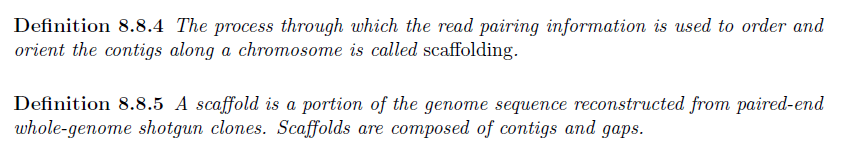
## What is the Poisson-based probability of the contig containing k reads, that is, seeing k-1 start positions in the interval of length p for a repeat and a non-repeat?



## What is the arrival statistic (A-statistic)? What do the values indicate?



## Define scaffolding and scaffold.



## What is done in the consensus phase?

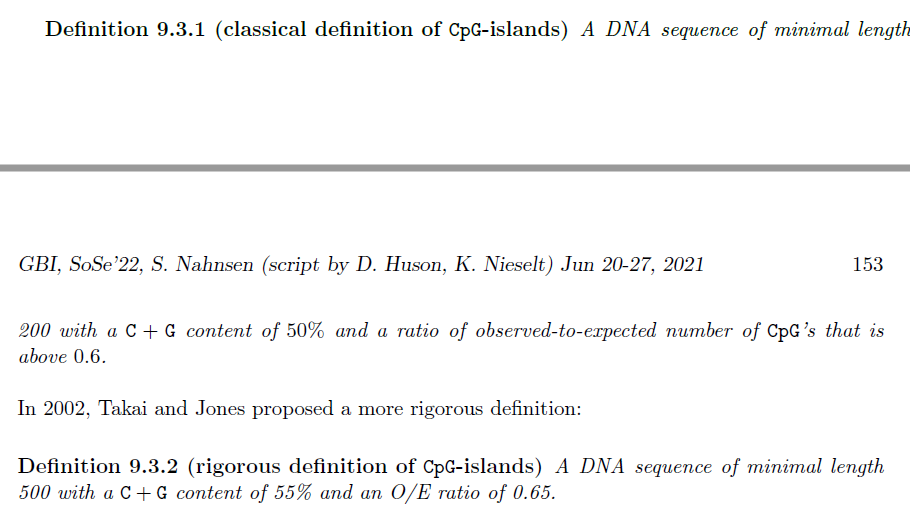
From the contig, we obtain a consensus sequence using a majority vote or a weighted decision taking the quality of reads into account.

## What is another approach for sequence assembly?

De Bruijn graph

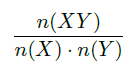
# Markov-Chains

## What are CpG islands and what are they used for in bioinformatics?



Areas with more CG-dinucleotides. They are used to separate individual genes.

## Define the observed vs. expected ratio of a dinucleotide.



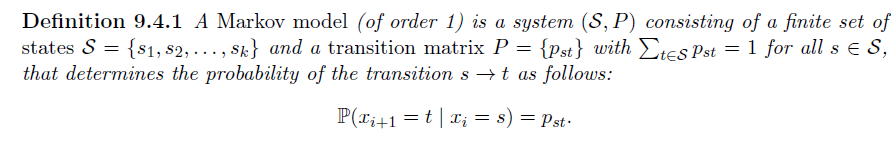
## What are the two main problems concerning CpG-islands?

1. Discrimination problem: Given a short sequence, how do we know whether it comes from a CpG-island?
2. Localisation problem: Given a long sequence, how can we find all contained CpG-islands?

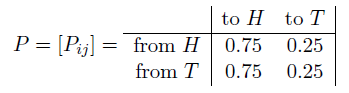
## Define the Markov property.

The future state only depends on the present state and not on preceding states.

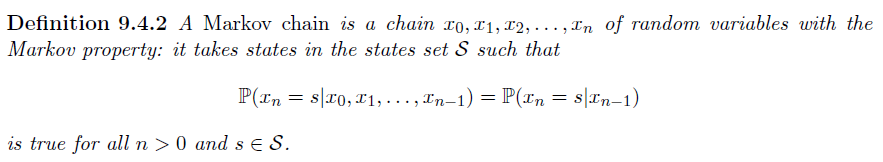
## Define Markov model



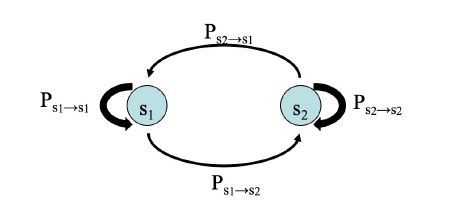
## Show an example of a transition matrix for a coin, where thee coin is more likely (75%) to go to head than tail.



## Define Markov chain.



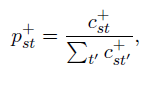
## Draw the state diagram off a Markov model with two states s1 and s2.



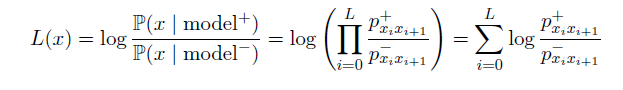
## What does one need to define for a Markov chain in addition to the state probabilities?

Probabilities for states b (begin) and e (end).

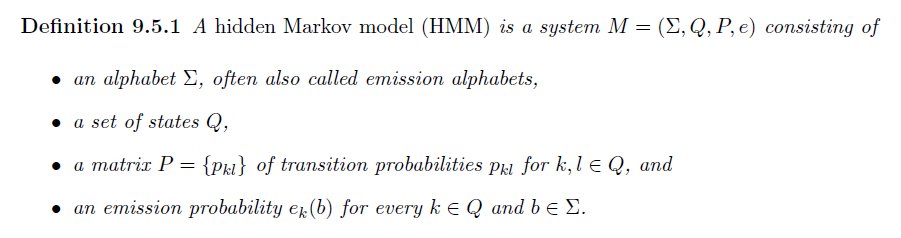
## We have + model and want to find out the transition probability that state s is followed by state t. What is the formula for this?



## How does one compute the probability that a sequence x comes from a plus or minus model? What is the ratio defined as? What is the log ratio defined as?

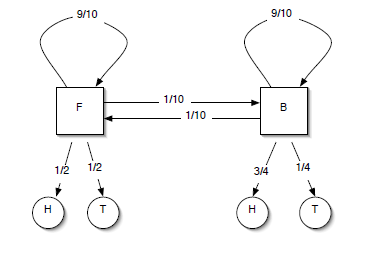


## Define a Hidden Markov model.

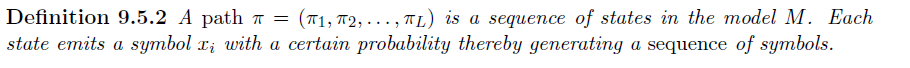


## Draw the Hidden Markov Model graph for a casino with a fair and biased coin (the numbers in the answer are placeholders). What do we need in addition to the transition matrix in the case of a HMM?

We also need the emission probabilities.



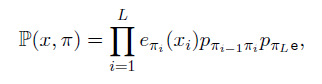
## Define a path in the context of HMMs.



## What are the three algorithms that we apply in the context of HMMs? What are they used for?

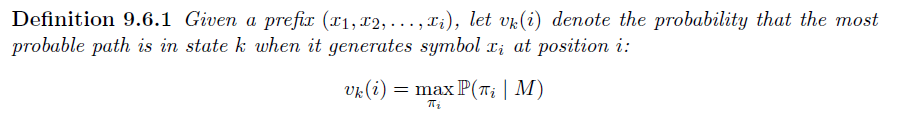
* Viterbi-algorithm
  + For x, determine the most probable sequence of states through M
* Forward algorithm
  + Determine the probability that M generated x
* Training algorithm
  + Given x and perhaps some additional sequences pi, estimate the parameters of M

## What is the probability that a sequence x and a path pi are generated under Model M?

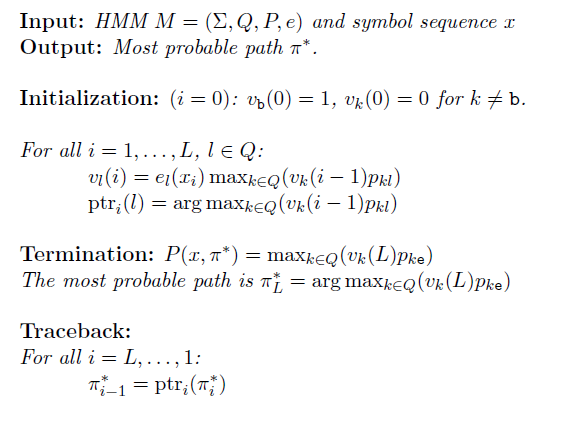


With pi0 = begin state b

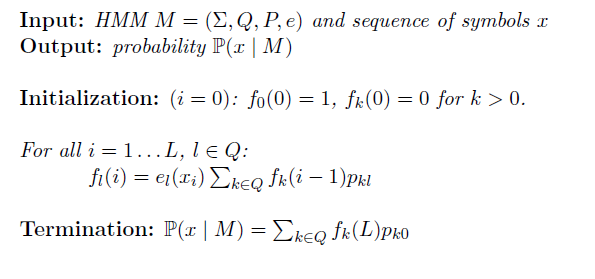
## Define the probability that the most probable path is in state k when it generates symbol x\_i at position i.



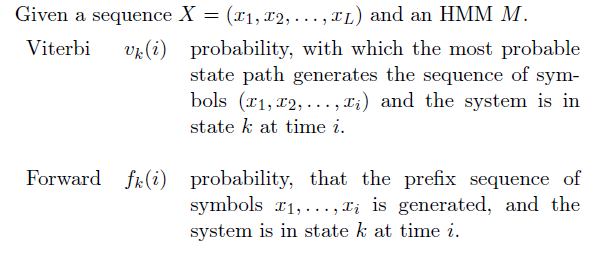
## Reproduce the Viterbi algorithm.



## Reproduce the Forward algorithm.

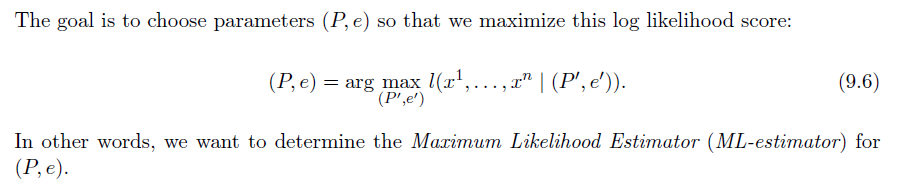


## What is the difference between the Viterbi and Forward algorithm?

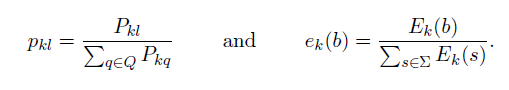


## Describe the Log-Likelihood approach for estimating the model parameters P and

## 



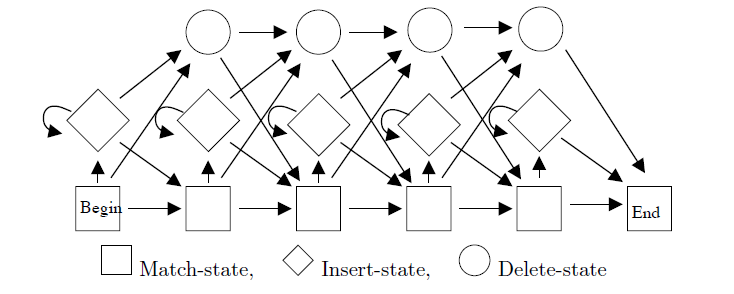
## What are P and e defined as when using Maximum-Likelihood estimators?



## One can also estimate parameters by unsupervised training. What are the two approaches for this?

Iterative improvement by Viterbi training or from the forward-backward algorithm (Baum-Welch-training)

## Draw the scheme of a profile HMM. What are its states? What is the length of the HMM?

The length is equal to the length of the MSA.

## Describe insert state.

Insert states emit symbol from the amino acide alphabet. They are used for sequences that have a letter at a position and all the other members a gap.

## Describe delete-state.

Silent states, emit a gap. They model the fact that positions in an MSA can be ‘skipped’ by individual sequences.

# Gene finding

## What are different kinds of RNA?

rRNA, tRNA, snRNA, snoRNA, miRNA

## What is the C-value? What iss the C-value enigma?

C-value: amount of DNA in picograms in a haploid nucleus.

C-value enigma: number of genes and genome sizes do not correlate: Single cell amoebae have genomes up to 100 times larger than humans.

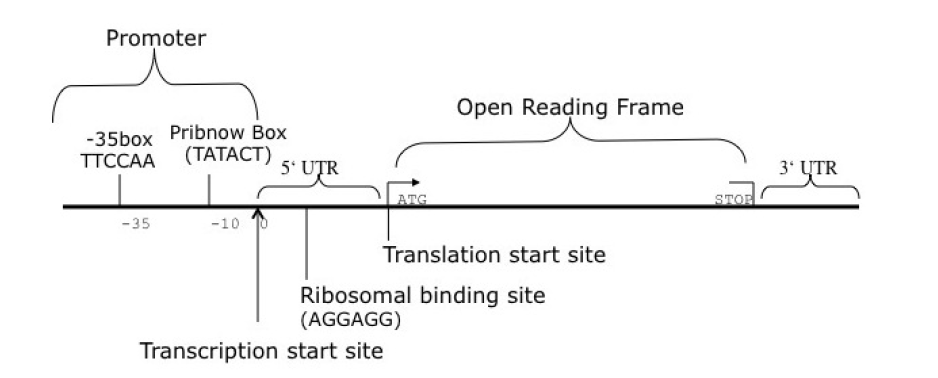
## What are the three approaches to gene finding?

1. Intrinsic: Statistical or ab initio models (e.g. Hidden Markov Models). Attempt to predict genes solely based on statistical properties of the given DNA sequence
2. Extrinsic: Similarity searches. DNA sequence is compared with known proteins
3. Cross-species comparative methods. The given DNA sequence is compared with a similar DNA sequence from a different species.

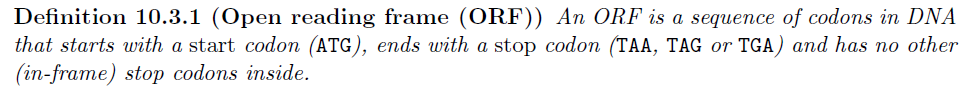
## Draw a sketch of the typical structure of prokaryotic genes.

UTR: Untranslated regions

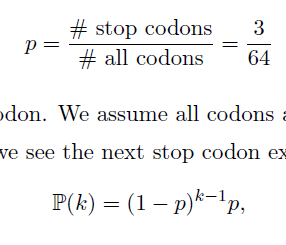
Transcriptase transcribes between them.



## Define Open reading frame.



## What is the expected ORF length computed with a geometric distribution?





## Define codon usage.

