[1 Sequencing and Assembly 7](#_Toc110089635)

[1.1 What is te formula for the Phred Score and what do the parameters mean. 7](#_Toc110089636)

[1.2 What does a FastQ entry look like? 7](#_Toc110089637)

[1.3 Define genome assembly according to Dear et al. 7](#_Toc110089638)

[1.4 Name issues when sequencing eukaryotic organisms. 7](#_Toc110089639)

[1.5 What is the formula for mean coverage? 7](#_Toc110089640)

[1.6 What is the lemma that relates the length of the target sequence with the number of reads and the average length of a read? 7](#_Toc110089641)

[1.7 What is the fragment assembly problem? 7](#_Toc110089642)

[1.8 What are the two laws of assembly? 8](#_Toc110089643)

[1.9 What is the shortest common superstring problem? 8](#_Toc110089644)

[1.10 Define overlap length and prefix length for two fragsments. 8](#_Toc110089645)

[1.11 Define prefix graph. 8](#_Toc110089646)

[1.12 Define the Hamiltonian Path and Travelling Salesman problem. 8](#_Toc110089647)

[1.13 How is the Shortest spanning tree failing for many biological genomes? 8](#_Toc110089648)

[1.14 Describe the three stages of OLC. 8](#_Toc110089649)

[1.15 What are the possible overlap relationships between two fragments with fixed orientation? 9](#_Toc110089650)

[1.16 What step of OLC is the computational bottleneck? Why? 9](#_Toc110089651)

[1.17 Reproduce the overlap alignment algorithm pseudocode. 9](#_Toc110089652)

[1.18 What is the overlap alignment runtime? 9](#_Toc110089653)

[1.19 Why is dynamic programming too slow for large sequencing projects? What is a better approach? 9](#_Toc110089654)

[1.20 We have found an overlap. What are the three options for its origin? 10](#_Toc110089655)

[1.21 How should one avoid repeat-induced overlaps? 10](#_Toc110089656)

[1.22 Describe the rules of the overlap graph 10](#_Toc110089657)

[1.23 What is the formal definition of the overlap graph? 11](#_Toc110089658)

[1.24 What are the four possibilities that an overlap can be drawn between the end points of two reads? 11](#_Toc110089659)

[1.25 What is the goal of the layout phase? 11](#_Toc110089660)

[1.26 Define spanning tree 11](#_Toc110089661)

[1.27 Define minimum spanning tree. 11](#_Toc110089662)

[1.28 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? 12](#_Toc110089663)

[1.29 What is the arrival statistic (A-statistic)? What do the values indicate? 12](#_Toc110089664)

[1.30 Define scaffolding and scaffold. 12](#_Toc110089665)

[1.31 What is done in the consensus phase? 12](#_Toc110089666)

[1.32 What is another approach for sequence assembly? 12](#_Toc110089667)

[2 Markov-Chains 13](#_Toc110089668)

[2.1 What are CpG islands and what are they used for in bioinformatics? 13](#_Toc110089669)

[2.2 Define the observed vs. expected ratio of a dinucleotide. 13](#_Toc110089670)

[2.3 What are the two main problems concerning CpG-islands? 13](#_Toc110089671)

[2.4 Define the Markov property. 13](#_Toc110089672)

[2.5 Define Markov model 13](#_Toc110089673)

[2.6 Show an example of a transition matrix for a coin, where thee coin is more likely (75%) to go to head than tail. 13](#_Toc110089674)

[2.7 Define Markov chain. 14](#_Toc110089675)

[2.8 Draw the state diagram off a Markov model with two states s1 and s2. 14](#_Toc110089676)

[2.9 What does one need to define for a Markov chain in addition to the state probabilities? 14](#_Toc110089677)

[2.10 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? 14](#_Toc110089678)

[2.11 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? 14](#_Toc110089679)

[2.12 Define a Hidden Markov model. 14](#_Toc110089680)

[2.13 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? 15](#_Toc110089681)

[2.14 Define a path in the context of HMMs. 15](#_Toc110089682)

[2.15 What are the three algorithms that we apply in the context of HMMs? What are they used for? 15](#_Toc110089683)

[2.16 What is the probability that a sequence x and a path pi are generated under Model M? 15](#_Toc110089684)

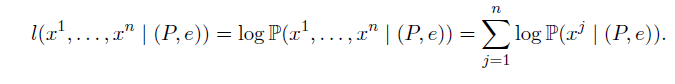
[2.17 Define the probability that the most probable path is in state k when it generates symbol x\_i at position i. 15](#_Toc110089685)

[2.18 Reproduce the Viterbi algorithm. 16](#_Toc110089686)

[2.19 Reproduce the Forward algorithm. 16](#_Toc110089687)

[2.20 What is the difference between the Viterbi and Forward algorithm? 16](#_Toc110089688)

[2.21 Describe the Log-Likelihood approach for estimating the model parameters P and 17](#_Toc110089689)

[ 17](#_Toc110089690)

[2.22 What are P and e defined as when using Maximum-Likelihood estimators? 17](#_Toc110089691)

[2.23 One can also estimate parameters by unsupervised training. What are the two approaches for this? 17](#_Toc110089692)

[2.24 Draw the scheme of a profile HMM. What are its states? What is the length of the HMM? 17](#_Toc110089693)

[2.25 Describe insert state. 18](#_Toc110089694)

[2.26 Describe delete-state. 18](#_Toc110089695)

[3 Gene finding 19](#_Toc110089696)

[3.1 What are different kinds of RNA? 19](#_Toc110089697)

[3.2 What is the C-value? What iss the C-value enigma? 19](#_Toc110089698)

[3.3 What are the three approaches to gene finding? 19](#_Toc110089699)

[3.4 Draw a sketch of the typical structure of prokaryotic genes. 19](#_Toc110089700)

[3.5 Define Open reading frame. 19](#_Toc110089701)

[3.6 What is the expected ORF length computed with a geometric distribution? 20](#_Toc110089702)

[3.7 Define codon usage. 20](#_Toc110089703)

[3.8 What is a fifth-order Markov chain? 20](#_Toc110089704)

[3.9 What is the pipeline for Markov Models for prokaryote gene finding? 20](#_Toc110089705)

[3.10 Name two Gene finding algorithms that use Markov Models. 20](#_Toc110089706)

[3.11 Why is gene prediction in eukaryotes so hard? 21](#_Toc110089707)

[3.12 Describe features of gene density in eukaryotes. 21](#_Toc110089708)

[3.13 Name three approaches to find genes in eukaryotes. 21](#_Toc110089709)

[3.14 How are eukaryotic genes distinguished in respect to exons? 21](#_Toc110089710)

[3.15 Define position specific weight matrix. 21](#_Toc110089711)

[3.16 What is the propensity value (odds score) and the log-propensity value? 21](#_Toc110089712)

[3.17 What is the problem with this definition? What is its solution? 21](#_Toc110089713)

[3.18 How can you compute the probability that sequence A is generated under PSWM? 22](#_Toc110089714)

[3.19 Describe the Genscan model. What does it look like? How many states and transition probabilities does it have? Which? 22](#_Toc110089715)

[3.20 GENSCAN uses a generalized HMM. What characterizes it? 22](#_Toc110089716)

[3.21 What does a generalized HMM produce? 23](#_Toc110089717)

[3.22 What are the four main components of Genscan? 23](#_Toc110089718)

[3.23 What is of central importance in comparative gene finding? 23](#_Toc110089719)

[3.24 Describe the model of an eukaryotic gene. 23](#_Toc110089720)

[3.25 Describe performance evaluation measures of genes on the nucleotide and the exon level. 24](#_Toc110089721)

[3.26 What are common start, stop, donor splice site and acceptor splice site codons? 24](#_Toc110089722)

[4 RNA Secondary Structure 25](#_Toc110089723)

[4.1 What is the process to make RNA? 25](#_Toc110089724)

[4.2 What does the RNA-world hypothesis suggest? 25](#_Toc110089725)

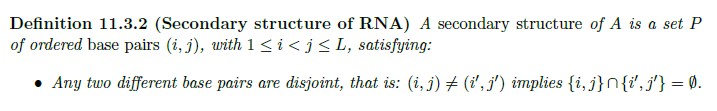
[4.3 Name some variants of RNA. 25](#_Toc110089726)

[4.4 Name some roles of RNA. 25](#_Toc110089727)

[4.5 What are the four nucleotides in RNA? 25](#_Toc110089728)

[4.6 Define primary structure of RNA. 25](#_Toc110089729)

[4.7 Define secondary structure of RNA. 25](#_Toc110089730)

[ 26](#_Toc110089731)

[4.8 Define nested secondary structure. 26](#_Toc110089732)

[4.9 Draw some structures that are not nested. 26](#_Toc110089733)

[4.10 Draw a pseudoknot structure. 26](#_Toc110089734)

[4.11 Why is the dot bracket notation unambiguous? 26](#_Toc110089735)

[4.12 What are two ways to visualize RNA structure? 27](#_Toc110089736)

[4.13 What are the different secondary structure elements? 27](#_Toc110089737)

[4.14 What are some approaches in predicting RNA secondary structure? 27](#_Toc110089738)

[4.15 What is the formula for the number of secondary structures for the sequence of length n? 27](#_Toc110089739)

[4.16 What are the four ways to get an optimal structure for A\_ij from smaller ones? 28](#_Toc110089740)

[4.17 What is the recursion formula for the Nussinov algorithm? 28](#_Toc110089741)

[4.18 Reproduce the pseudocode for the Nussinov fill stage. 28](#_Toc110089742)

[4.19 Reproduce the pseudocode for the Nussinov traceback. 30](#_Toc110089743)

[4.20 What things have to be taken into account when producing better predictions than Nussinov in regards to energy minimization? 30](#_Toc110089744)

[4.21 What is a compensatory base change? 30](#_Toc110089745)

[4.22 What is the description of the mutual information content measure? 30](#_Toc110089746)

[4.23 How is the mutual information content computed? What do its values mean? 31](#_Toc110089747)

[5 Protein Tertiary Structure 32](#_Toc110089748)

[5.1 What is the structure of an amino acid? 32](#_Toc110089749)

[5.2 What is a protein? How is it produced? 32](#_Toc110089750)

[5.3 What makes up the backbone of the protein. 32](#_Toc110089751)

[5.4 What are the two degrees of freedom in a polypeptide chain? 32](#_Toc110089752)

[5.5 Which third torsion angle is there? What value does it nearly always assume? 32](#_Toc110089753)

[5.6 What do the degrees of freedom look like in cartoon form? 33](#_Toc110089754)

[5.7 Define peptide, protein and start and end group of a protein. 33](#_Toc110089755)

[5.8 Draw an example where you can see the N- and C-terminus. 33](#_Toc110089756)

[5.9 Describe primary, secondary, tertiary and quarternary structure of proteins. 33](#_Toc110089757)

[5.10 Draw illustrations of primary, secondary, tertiary and quaternary structure. 34](#_Toc110089758)

[5.11 What are the two different approaches to determine 3D structure of proteins? 34](#_Toc110089759)

[5.12 What features can be used for the analysis of the primary structure? 34](#_Toc110089760)

[5.13 What are the two most common repetitive structure called? 34](#_Toc110089761)

[5.14 What is a Ramachandran plot? 34](#_Toc110089762)

[5.15 What are typical values for phi and psi for alpha helices and beta strands? 34](#_Toc110089763)

[5.16 When do helices arise? 35](#_Toc110089764)

[5.17 What two parameters are helices described by? 35](#_Toc110089765)

[5.18 What is the most common number of residues per turn for alpha helices? 35](#_Toc110089766)

[5.19 What is a specific functional role a of alpha-helices? 35](#_Toc110089767)

[5.20 What is the difference between beta strand and beta sheet` 35](#_Toc110089768)

[5.21 What are the two possible configurations of beta sheets? 35](#_Toc110089769)

[5.22 Name non-repetitive structures. 35](#_Toc110089770)

[5.23 What are the major levels in hierarchy of evolutionary relatedness? 35](#_Toc110089771)

[5.24 What are the four principal structural classes of protein structures based on the three-dimensional ensemble of secondary structures? 36](#_Toc110089772)

[5.25 What are the prediction classes for the secondary structure of amino sequences? 36](#_Toc110089773)

[5.26 In the DSSP-classification, how are the 8 states subsumed into three? Does DSSP predict protein structure? 36](#_Toc110089774)

[5.27 How is the Q3 defined? 36](#_Toc110089775)

[5.28 Name some different approaches that have been used to predict secondary structure from sequence. 37](#_Toc110089776)

[5.29 What are propensities of amino acids? How are they computed? 37](#_Toc110089777)

[5.30 What are the three helix and three beta sheets classes derived from the propensities? 38](#_Toc110089778)

[5.31 Describe the Chou-Fasman algorithm to find helices using the seed and extend algorithm. 39](#_Toc110089779)

[5.32 For the final assignment, what should one do when the helix and beta sheet prediction intervals overlap? 39](#_Toc110089780)

[5.33 What is a problem of the Chou-Fasman algorithm? 39](#_Toc110089781)

[5.34 Describe the function of the secondary structure prediction program. 39](#_Toc110089782)

[5.35 What are the 5 defined regions for a secondary structure? 39](#_Toc110089783)

[5.36 Which characteristics are linear discriminant functions combining? 39](#_Toc110089784)

[5.37 What is the singleton characteristic defined as? 40](#_Toc110089785)

[5.38 What is the doublet characteristic? 40](#_Toc110089786)

[5.39 Where do helices often lie on the protein? 40](#_Toc110089787)

[5.40 What is hydrophobicity defined as? What is the structural hydrophobic moment defined as? 40](#_Toc110089788)

[5.41 What are the steps of the SSP algorithm? 41](#_Toc110089789)

[5.42 What are the most successful methods for predicting secondary structure based on? 41](#_Toc110089790)

[5.43 Describe visible and hidden layers. 41](#_Toc110089791)

[5.44 Describe recurrent and feed forward architectures. 41](#_Toc110089792)

[5.45 Describe a simple perceptron. 41](#_Toc110089793)

[5.46 Describe backpropagation. 42](#_Toc110089794)

[5.47 How does the PHD neural network work? 42](#_Toc110089795)

[5.48 What are the levels that PHD processes information on? 43](#_Toc110089796)

[5.49 Describe the steps of PHD. 43](#_Toc110089797)

[6 Protein Tertiary Structure 44](#_Toc110089798)

[6.1 What does studying the structure of proteins help us to understand? 44](#_Toc110089799)

[6.2 What is the structure of a PDB entry? 44](#_Toc110089800)

[6.3 What are the three broad approaches to structure prediction? 44](#_Toc110089801)

[6.4 Formulate the fold recognition problem. 44](#_Toc110089802)

[6.5 What is inverse folding? 44](#_Toc110089803)

[6.6 Define threading and name three methods for it. 44](#_Toc110089804)

[6.7 Describe the 123D fold recognition method 44](#_Toc110089805)

[6.8 The free energy function is given by the sum of two terms. Which ones? 45](#_Toc110089806)

[6.9 What is defined as contact? 45](#_Toc110089807)

[6.10 What is the formula for CCP? 45](#_Toc110089808)

# 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

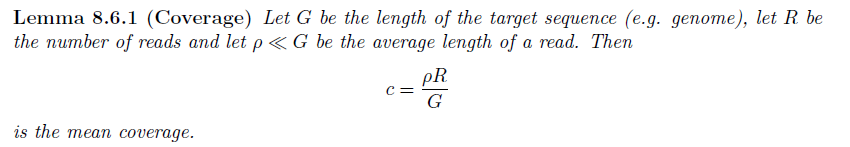
Computational

* 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?



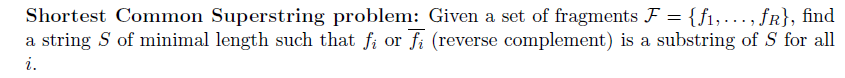
## What is the fragment assembly problem?

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

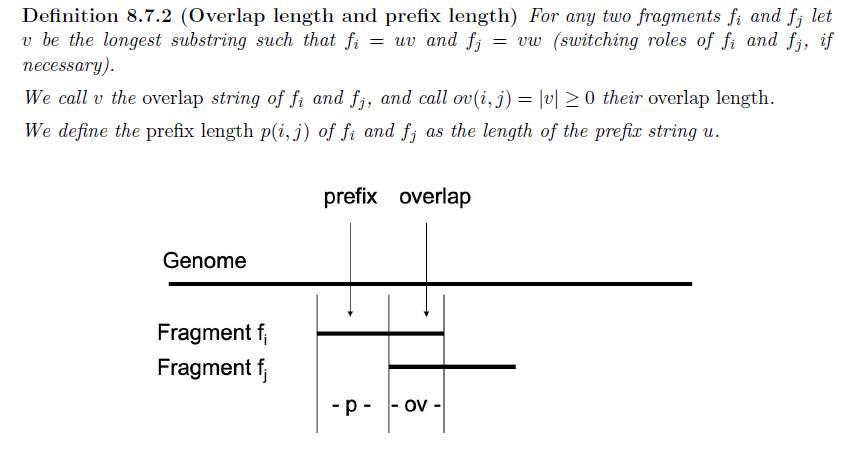
## What are the two laws of assembly?

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

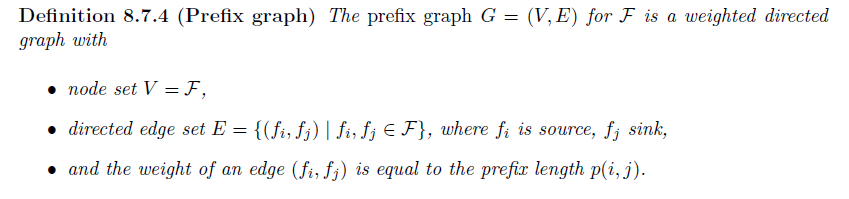
## What is the shortest common superstring problem?



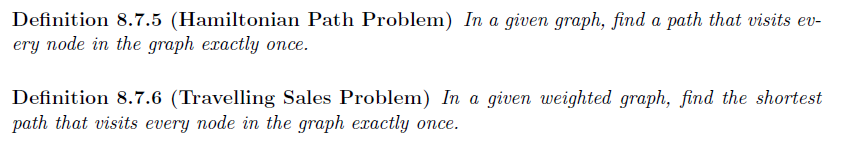
## Define overlap length and prefix length for two fragsments.



## Define prefix graph.



## Define the Hamiltonian Path and Travelling Salesman problem.



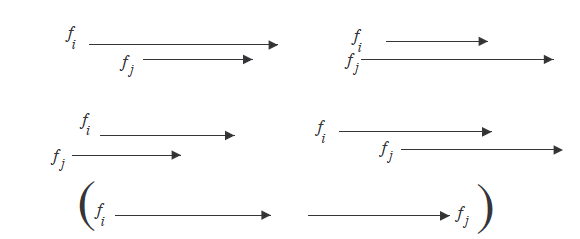
## How is the Shortest spanning tree failing for many biological genomes?

Interior parts of repeating regions are overcompressed

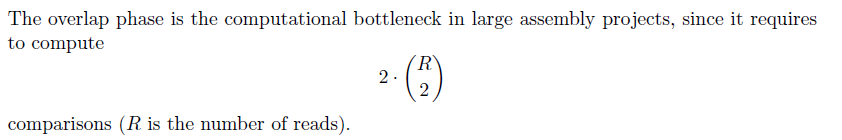
## Describe the three stages of OLC.

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.

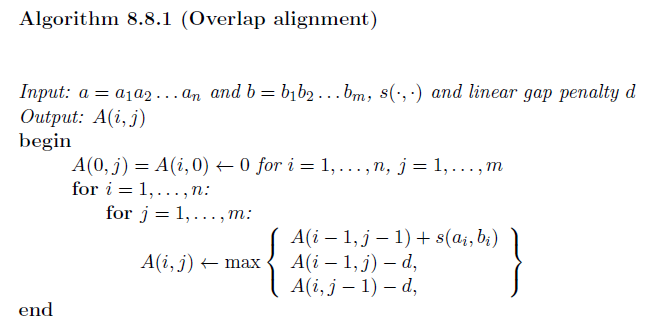
## What are the possible overlap relationships between two fragments with fixed orientation?



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



## Reproduce the overlap alignment algorithm pseudocode.

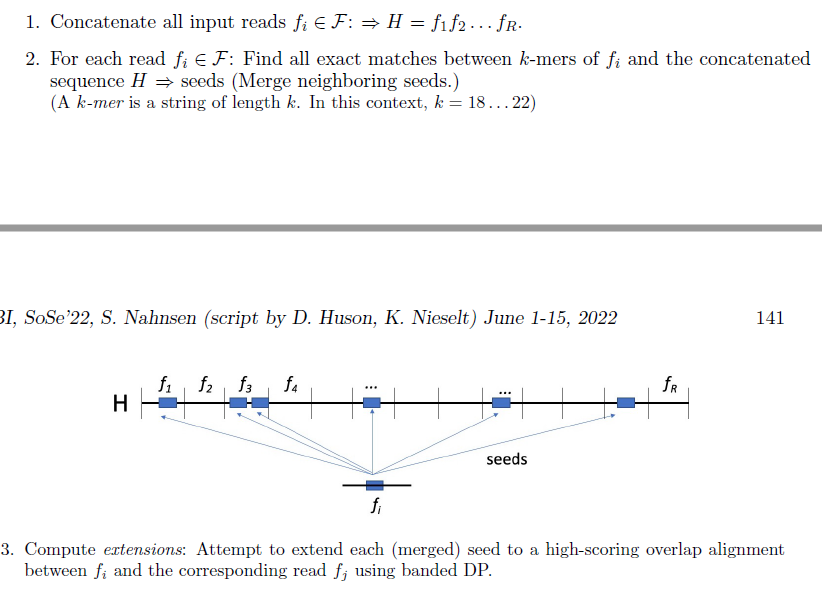


## What is the overlap alignment runtime?

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?

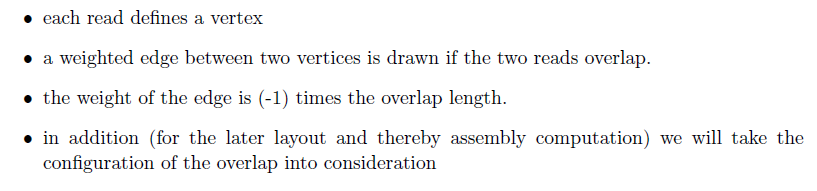
1. True origin
2. Repeat-induced
3. By chance

## How should one avoid repeat-induced overlaps?

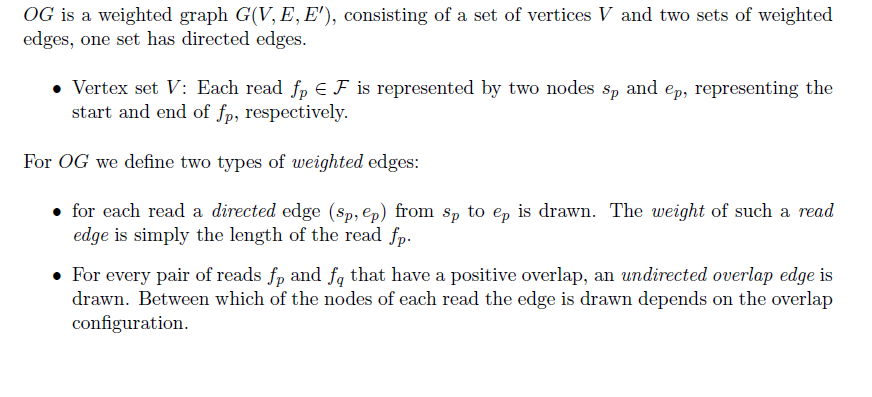
Only start seeds outside a repeat

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

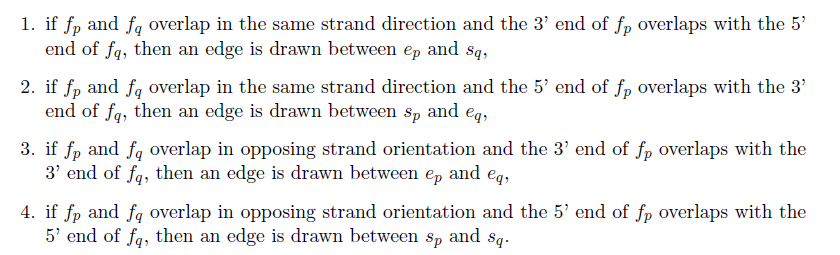
## Describe the rules of the overlap graph



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



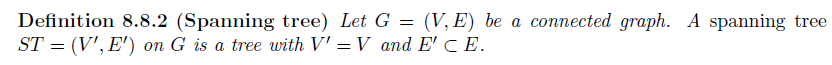
## What are the four possibilities that an overlap can be drawn between the end points of two reads?



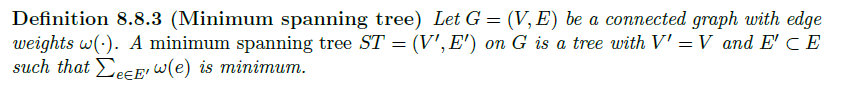
## What is the goal of the layout phase?

Arrange all reads into an approximate multi-alignment.

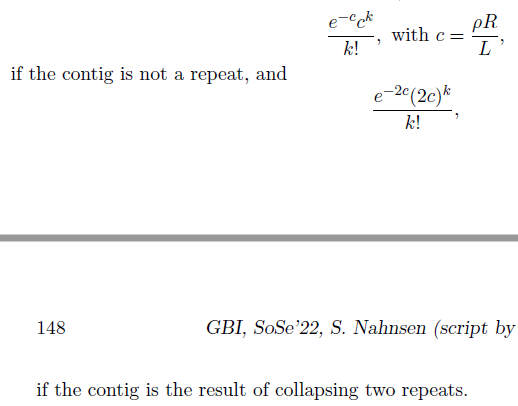
## Define spanning tree



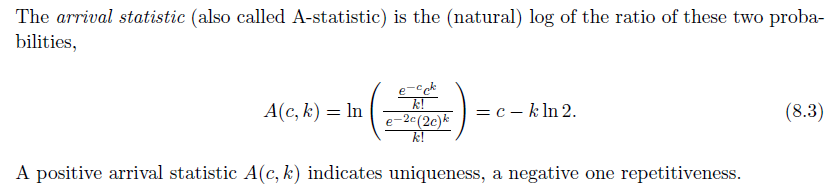
## Define minimum spanning tree.



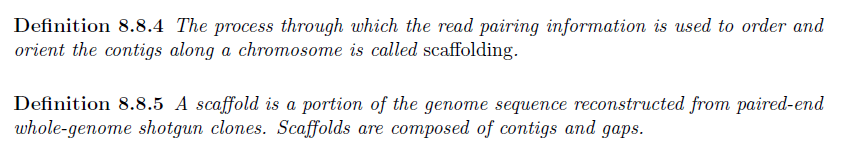
## What is the Poisson-based probability of the contig containing k reads, that is, seeing k-1 start positions in the interval of length p for a repeat and a non-repeat?



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



## Define scaffolding and scaffold.



## What is done in the consensus phase?

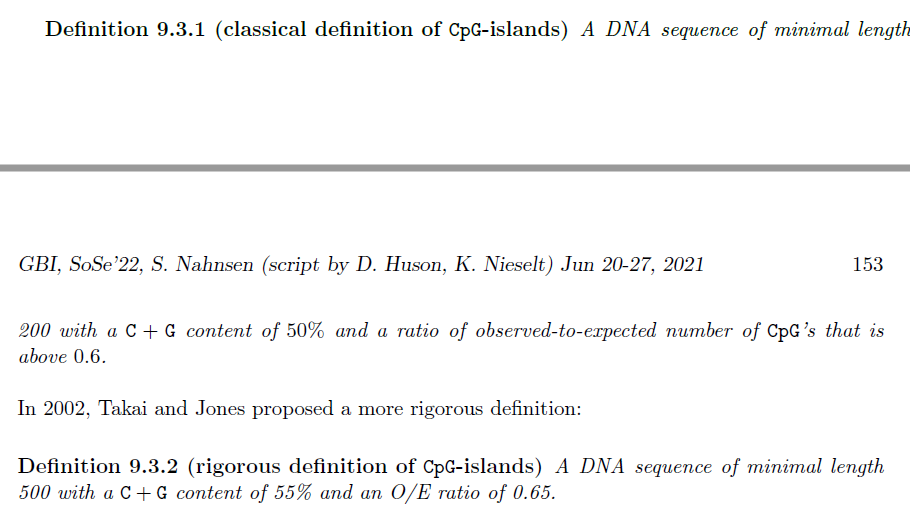
From the contig, we obtain a consensus sequence using a majority vote or a weighted decision taking the quality of reads into account.

## What is another approach for sequence assembly?

De Bruijn graph

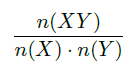
# Markov-Chains

## 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.



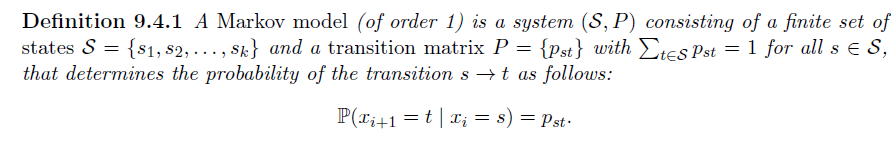
## What are the two main problems concerning CpG-islands?

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?

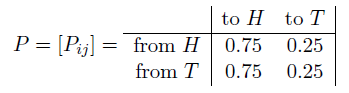
## Define the Markov property.

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

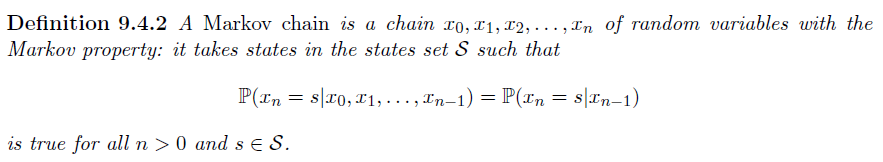
## Define Markov model



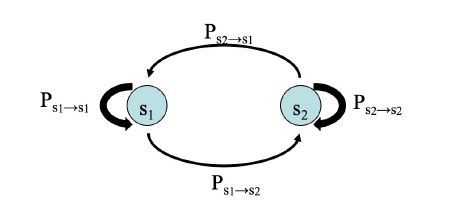
## Show an example of a transition matrix for a coin, where thee coin is more likely (75%) to go to head than tail.



## Define Markov chain.



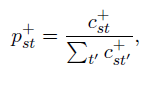
## Draw the state diagram off a Markov model with two states s1 and s2.



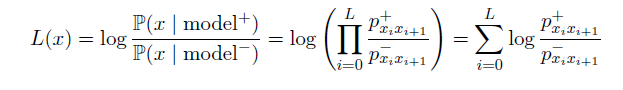
## What does one need to define for a Markov chain in addition to the state probabilities?

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

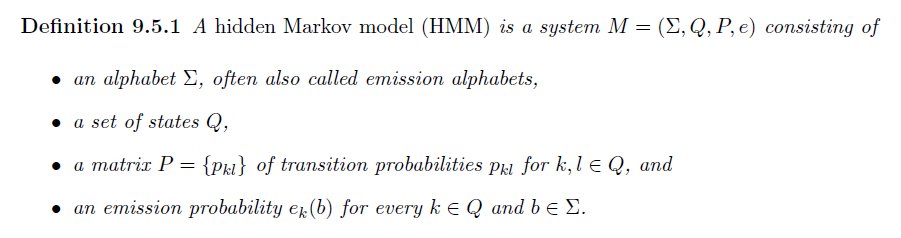
## We have + model and want to find out the transition probability that state s is followed by state t. What is the formula for this?



## 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?

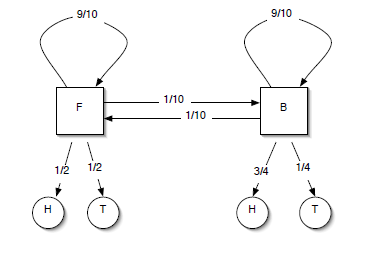


## Define a Hidden Markov model.

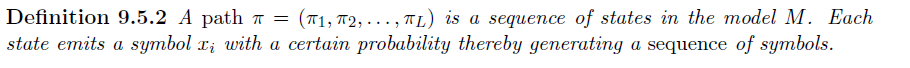


## Draw the Hidden Markov Model graph for a casino with a fair and biased coin (the numbers in the answer are placeholders). What do we need in addition to the transition matrix in the case of a HMM?

We also need the emission probabilities.



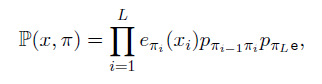
## Define a path in the context of HMMs.



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

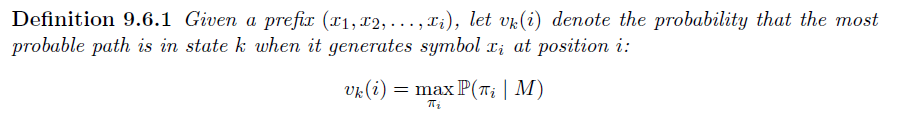
* Viterbi-algorithm
  + For x, determine the most probable sequence of states through M
* Forward algorithm
  + Determine the probability that M generated x
* Training algorithm
  + Given x and perhaps some additional sequences pi, estimate the parameters of M

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

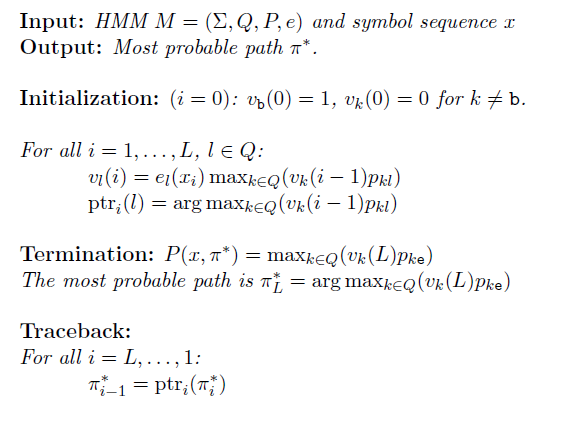


With pi0 = begin state b

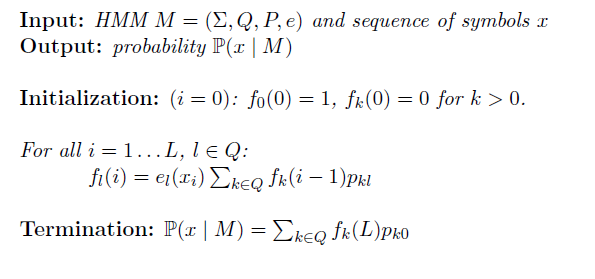
## Define the probability that the most probable path is in state k when it generates symbol x\_i at position i.



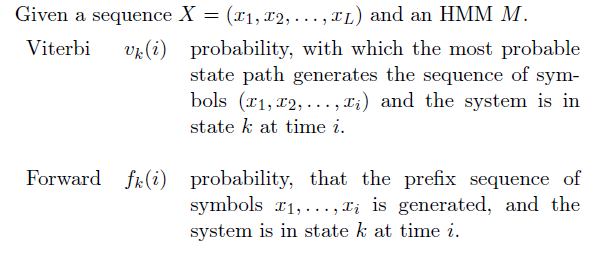
## Reproduce the Viterbi algorithm.



## Reproduce the Forward algorithm.

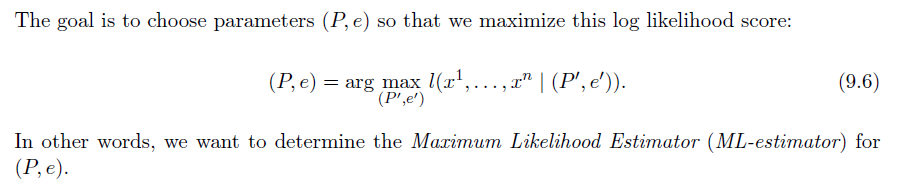


## What is the difference between the Viterbi and Forward algorithm?

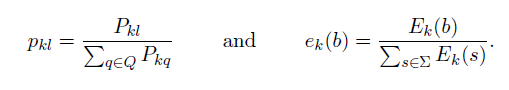


## Describe the Log-Likelihood approach for estimating the model parameters P and

## 



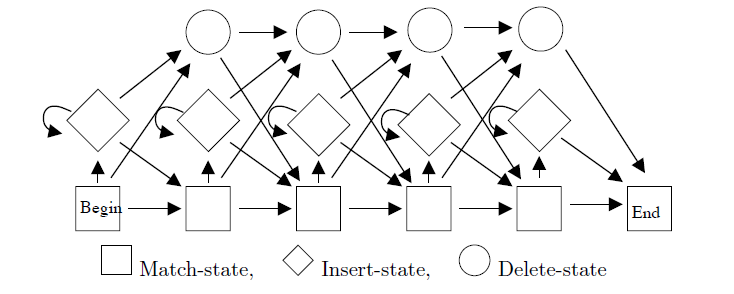
## What are P and e defined as when using Maximum-Likelihood estimators?



## 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.

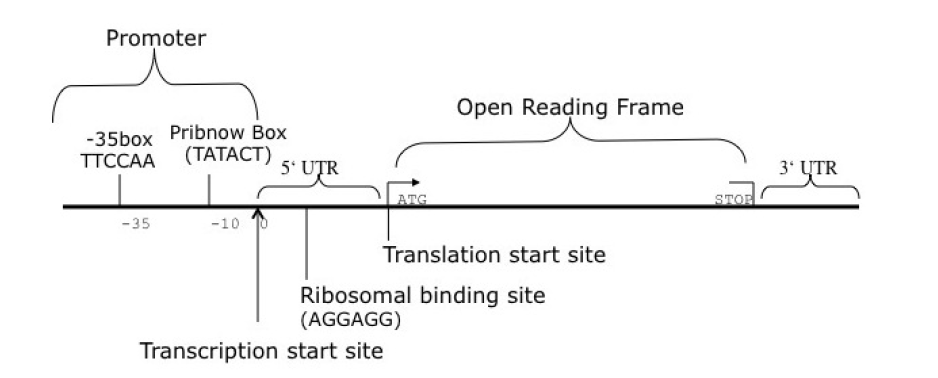
## What are the three approaches to gene finding?

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.

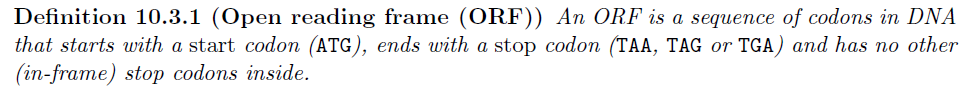
## Draw a sketch of the typical structure of prokaryotic genes.

UTR: Untranslated regions

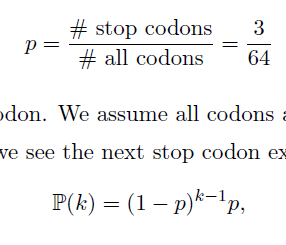
Transcriptase transcribes between them.



## Define Open reading frame.

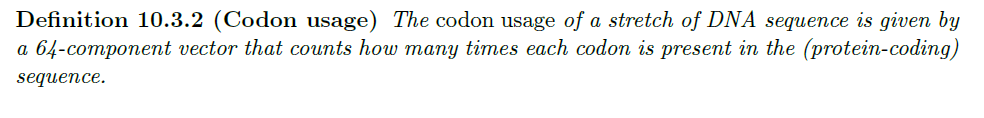


## What is the expected ORF length computed with a geometric distribution?





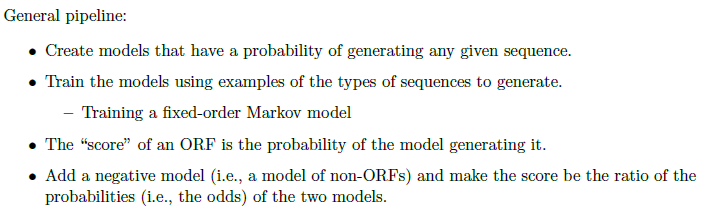
## Define codon usage.



## What is a fifth-order Markov chain?

A Markov chain where the current state does not only depend on the state just before it, but on the five previous states.

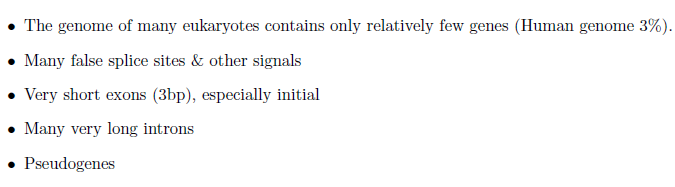
## What is the pipeline for Markov Models for prokaryote gene finding?



## Name two Gene finding algorithms that use Markov Models.

Genemark and Glimmer

## Why is gene prediction in eukaryotes so hard?



## Describe features of gene density in eukaryotes.

Largely dependent on GC content. In vertebrates, a gene has on average 6 exons spanning 30kb, with very long distances between any two genes.

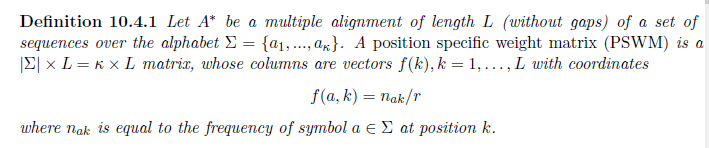
## Name three approaches to find genes in eukaryotes.

1. Statistical or ab initio methods
2. Homology methods (DNA sequence is compared with known proteins)
3. Cross-spec ies comparative methods

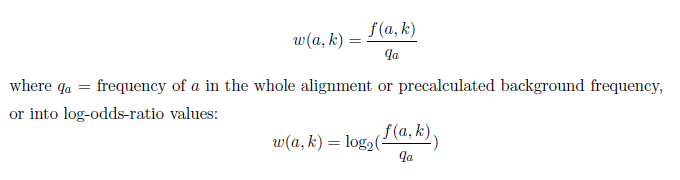
## How are eukaryotic genes distinguished in respect to exons?

* Single exon genes (no introns)
* Multi exon genes
  + Initial exons (region of start codon up to first donor site)
  + Internal exons (start after one acceptor site up to next donor site)
  + Terminal exons (start after last acceptor site up to stop codon)

## Define position specific weight matrix.



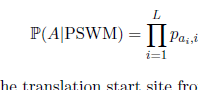
## What is the propensity value (odds score) and the log-propensity value?



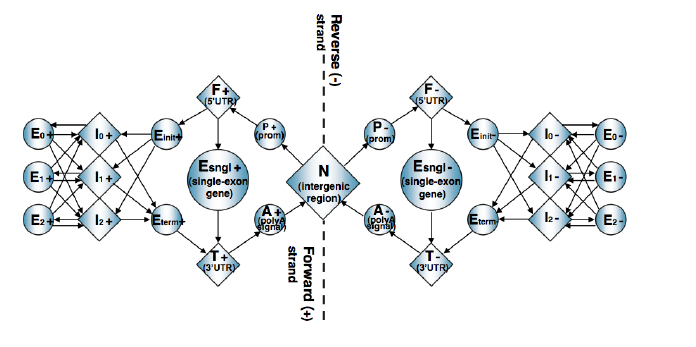
## What is the problem with this definition? What is its solution?

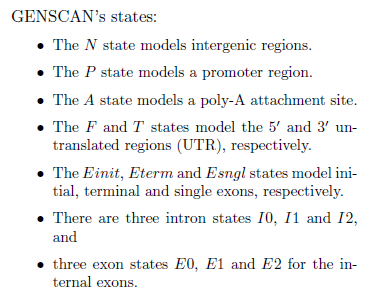
Log(0). Solution: Pseudocounts

## How can you compute the probability that sequence A is generated under PSWM?



## Describe the Genscan model. What does it look like? How many states and transition probabilities does it have? Which?





## GENSCAN uses a generalized HMM. What characterizes it?

A generalized HMM is one in which a duration period is explicitly modelled for each state, using a probability distribution.

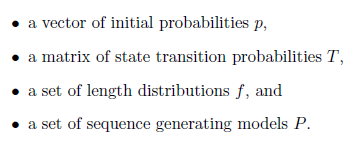
Upon entering a state:

1. Choose duration d, according to probability distribution
2. Generate d letters, each according to emission probabilities
3. Take a transition to next state according to transition probabilities

## What does a generalized HMM produce?

A parse, consisting of a sequence of states and an associated sequence of durations.

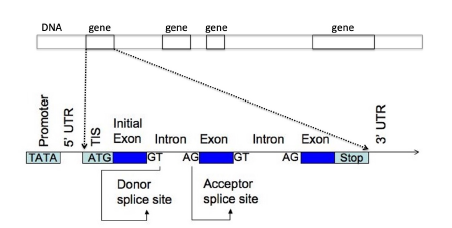
## What are the four main components of Genscan?

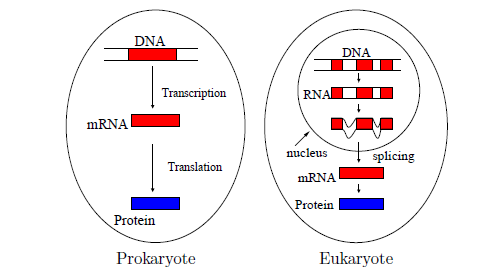


## What is of central importance in comparative gene finding?

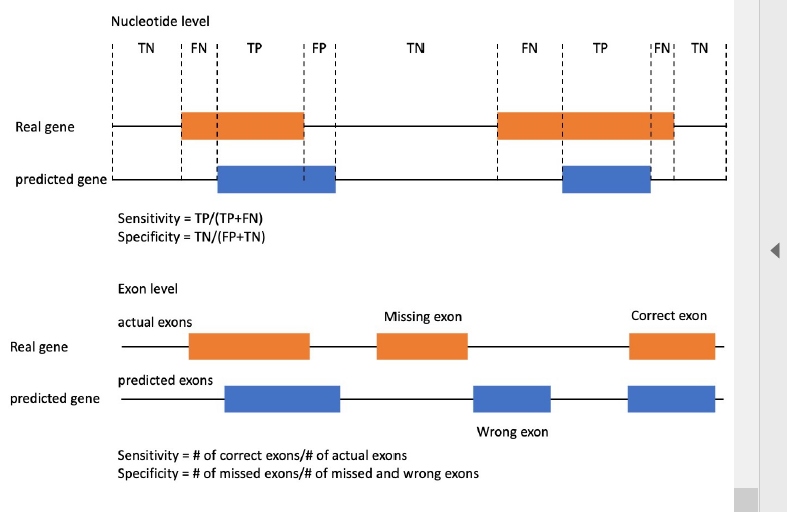
Exons are more conserved than introns. Between species, a lot of genes have correspondent genes. Unknown genes in one species may be compared to genes in some closely-related species which have been annotated

## Describe the model of an eukaryotic gene.





## Describe performance evaluation measures of genes on the nucleotide and the exon level.



## What are common start, stop, donor splice site and acceptor splice site codons?

ATG: Start

TAG: Stop

GT: Donor splice site

AG: Acceptor splice site

# RNA Secondary Structure

## What is the process to make RNA?

In eukaryotes, DNA is transcribed into pre-mRNA, from which introns are spliced to mature mRNA, which is then translated by ribosomes to produce proteins with the help of tRNAs.

## What does the RNA-world hypothesis suggest?

Life was originally based on RNA. Then, the data storage problem was delegated to DNA.

## Name some variants of RNA.

* Ribosomal RNA (rRNA)
* Micro RNA (miRNA)
* Small nucleolar RNA (snoRNA)
* Small interfering RNA (siRNA)
* Messenger RNA (mRNA)
* Transfer RNA (tRNA)

## Name some roles of RNA.

* Transmitter of genetic information (mRNA)
* Adaptor molecule (tRNA)
* Carrier of genetic information (RNA virus)
* Regulator of gene expression (siRNA)
* Catalyst (ribozyme)

## What are the four nucleotides in RNA?

Adenine, cytosine, guanine, uracil

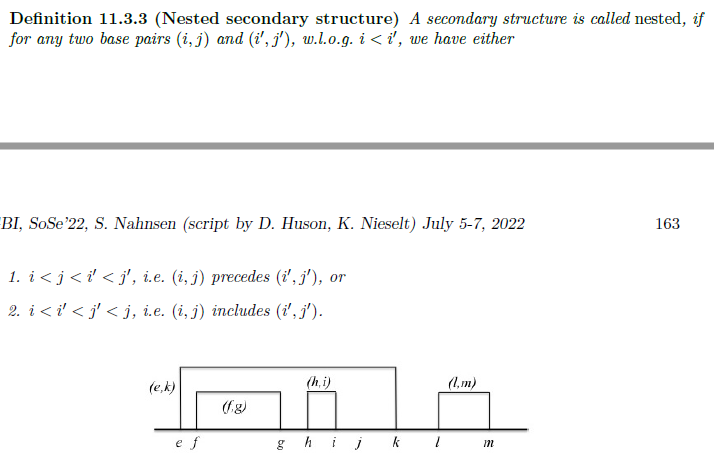
## Define primary structure of RNA.



## Define secondary structure of RNA.

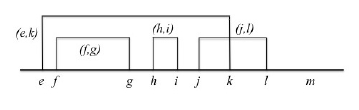
## 

## Define nested secondary structure.

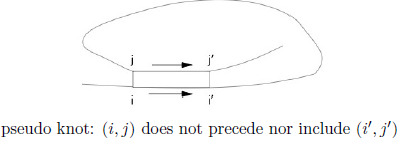


## Draw some structures that are not nested.

(e,k) and (j,l)



## Draw a pseudoknot structure.



## Why is the dot bracket notation unambiguous?

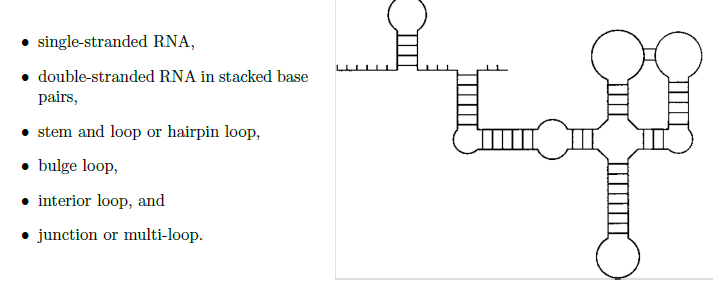
Base pairs cannot cross

## What are two ways to visualize RNA structure?

Secondary structure graph and drawn as in arc notation.

## What are the different secondary structure elements?

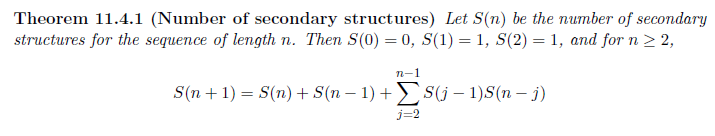
Insert a picture later where the things are actually labelled.



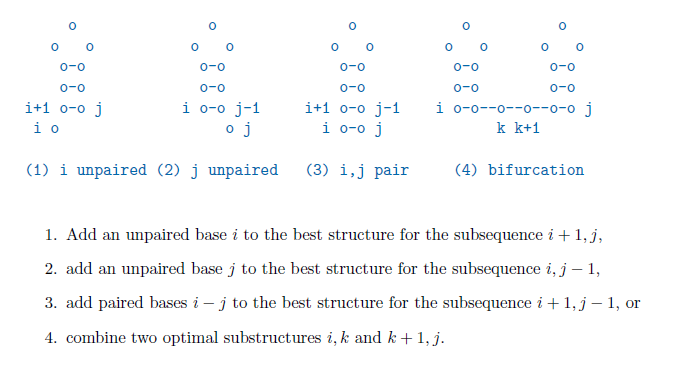
## What are some approaches in predicting RNA secondary structure?

1. Maximize the number of base pairs
2. Minimize the free energy
3. Optimize with respect to the “mutual information content”, when considering a comparison of RNA sequences

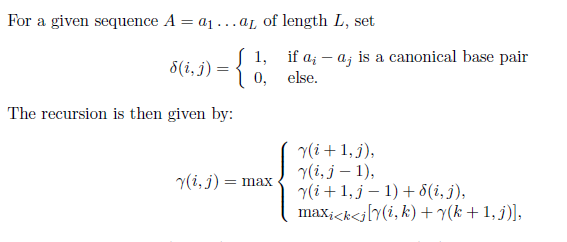
## What is the formula for the number of secondary structures for the sequence of length n?



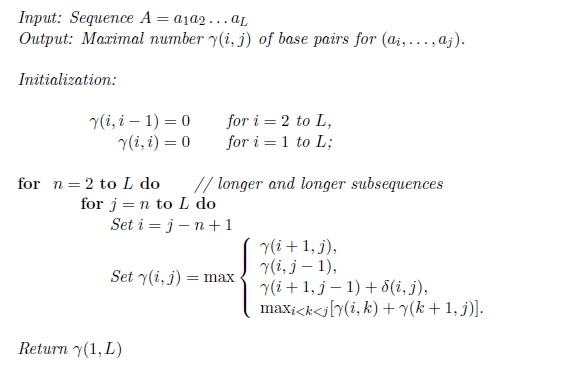
## What are the four ways to get an optimal structure for A\_ij from smaller ones?



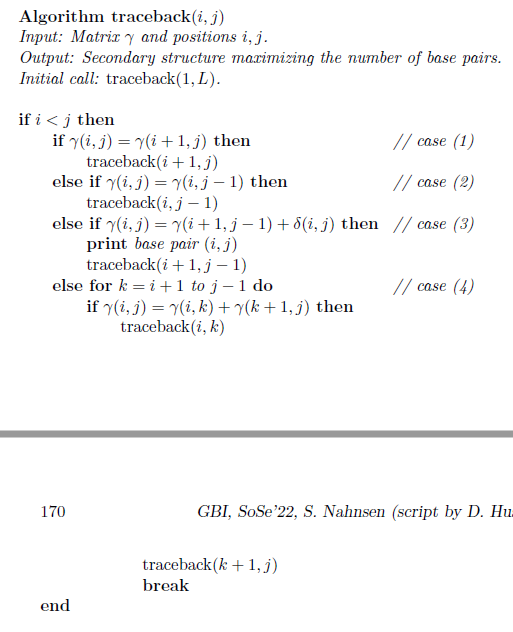
## What is the recursion formula for the Nussinov algorithm?



## Reproduce the pseudocode for the Nussinov fill stage.



## Reproduce the pseudocode for the Nussinov traceback.



## What things have to be taken into account when producing better predictions than Nussinov in regards to energy minimization?

* Helical stacks of base pairs have a stabilizing effect
* Loops have a destabilizing effect on the structure

## What is a compensatory base change?

* When one base of a pair changes, we usually find that its partner also changes so as to conserve that base pair.

## What is the description of the mutual information content measure?

* If you tell me the identity of position I, how much do I learn about the identity of position j?

## How is the mutual information content computed? What do its values mean?

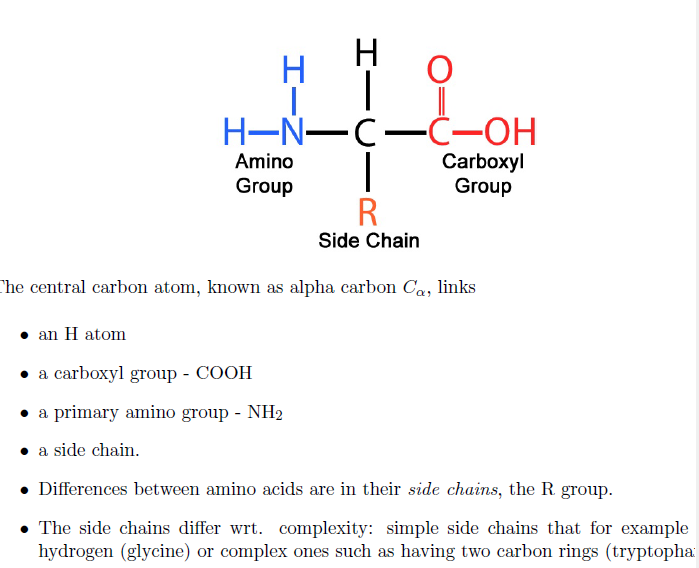
We calculate the odds of the relative frequency of nucleotides and joint frequency of nucleotides in columns i and j.



# Protein Tertiary Structure

## What is the structure of an amino acid?

## 



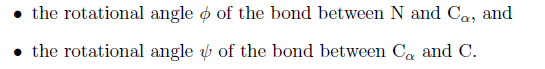
## What is a protein? How is it produced?

Chain of amino acids joined by peptide bonds. It is produced by a ribosome that moves along an mRNA and adds amino acis according to the codons that it observes.

## What makes up the backbone of the protein.

A chain of N-C\_alpha-C’s

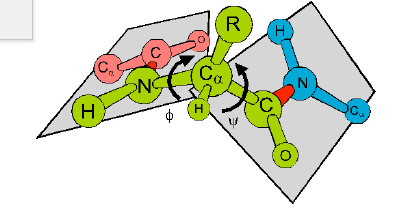
## What are the two degrees of freedom in a polypeptide chain?



## Which third torsion angle is there? What value does it nearly always assume?

Omega, the angle of the peptide bond between the C=0 and NH groups. It is nearly always 180 (planarity).

## What do the degrees of freedom look like in cartoon form?



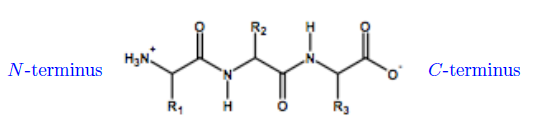
## Define peptide, protein and start and end group of a protein.

Peptide: 10-100 amino acids

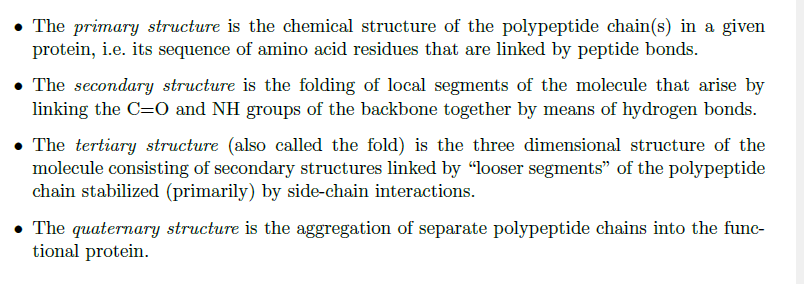
Protein: > 100 amino acids

Protein starts with a free Nh group (N-terminus) and ends with a free COOH group (the C-terminus)

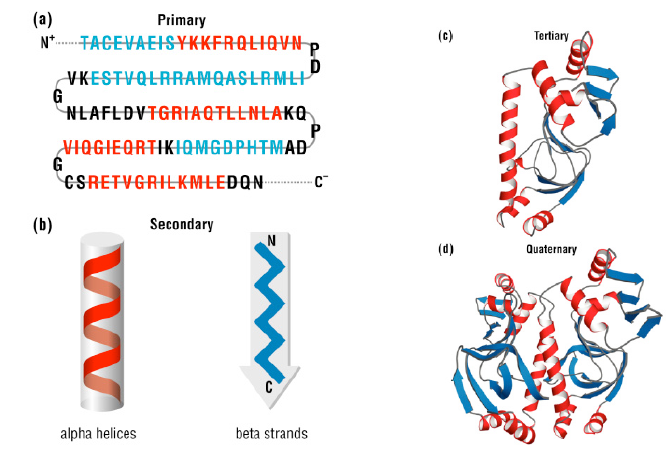
## Draw an example where you can see the N- and C-terminus.



## Describe primary, secondary, tertiary and quarternary structure of proteins.



## Draw illustrations of primary, secondary, tertiary and quaternary structure.



## What are the two different approaches to determine 3D structure of proteins?

Experimentally:

* X-ray crystallography
* Nuclear magnetic resonance

Homology Modelling

* Proteins with very similar primary structure have very similar tertiary structures

## What features can be used for the analysis of the primary structure?

* Amino acid composition
* Molecular weight
* Isoelectric point
* Average hydrophobicity
* Average charge

## What are the two most common repetitive structure called?

* Alpha-helix
* Beta-sheet

## What is a Ramachandran plot?

The pairs of angles phi and psi are plotted in a scatter plot

## What are typical values for phi and psi for alpha helices and beta strands?

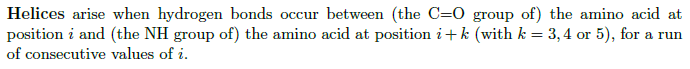
Alpha helix:

* Phi: -60^
* Psi: -40°

Beta strand:

* Phi: -90°
* Psi: 120°

## When do helices arise?



## What two parameters are helices described by?

1. Number of residues per turn
2. The rise per helical residue

## What is the most common number of residues per turn for alpha helices?

* 1. (phi, psi) = (-58,-47)

## What is a specific functional role a of alpha-helices?

DNA binding motifs, diameter of the alpha helix is 1.2 nanometres = width of the major groove in DNA

## What is the difference between beta strand and beta sheet`

Beta strand: single continuus stretch of amino acids adopting an extended conformation and involved in hydrogen bonds

Beta sheet: assembly of beta strands that are hydrogen-bonded to each other

## What are the two possible configurations of beta sheets?

Parallel: chains run in the same direction

Antiparallel: chains run in alternating directions

## Name non-repetitive structures.

Loops: Hairpin loops consitute a complete turn joining two anti-parallel beta-strands

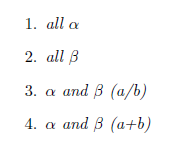
Turns: narrow 180° loops that contain at least 3 amino acids

Random coil: Region of secondary structure that is not a helix, a sheet or recognizable turn

## What are the major levels in hierarchy of evolutionary relatedness?

1. Families: clear evolutionary relationship
2. Superfamily: probable common evolutionary relationship
3. Fold: major structural similarity

## What are the four principal structural classes of protein structures based on the three-dimensional ensemble of secondary structures?



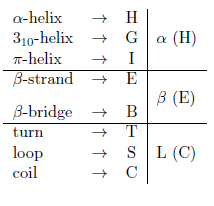
1. Bundle of alpha helices connected by loops on the surface of the protein
2. Two sheets in close contact forming a sandwich. (Enzymes, transport proteins and antibodies)
3. Mainly of parallel beta sheets with intervening alpha helices. Metabolic enzymes
4. Segregated alpha helices and mainly anti-parallel beta sheets
5. Small protein class: little or no secondary structure (usually dominated by metal ligand, heme, and/or disulfide bridges)

## What are the prediction classes for the secondary structure of amino sequences?

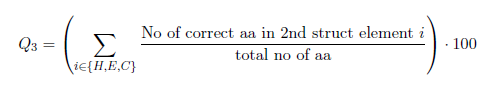
* E (extended, beta-strand)
* H (helical)
* C (coils or loops)
* - (no assigned secondary structure)

## In the DSSP-classification, how are the 8 states subsumed into three? Does DSSP predict protein structure?

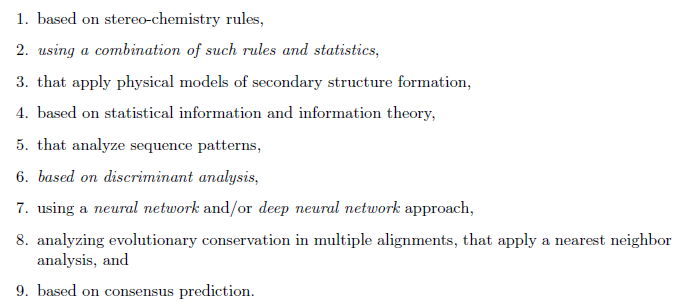
NO



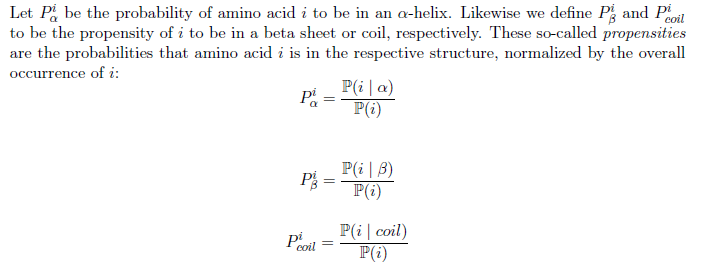
## How is the Q3 defined?

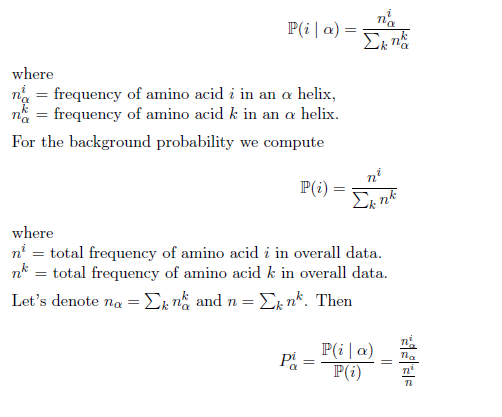


## Name some different approaches that have been used to predict secondary structure from sequence.

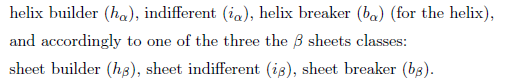


## What are propensities of amino acids? How are they computed?

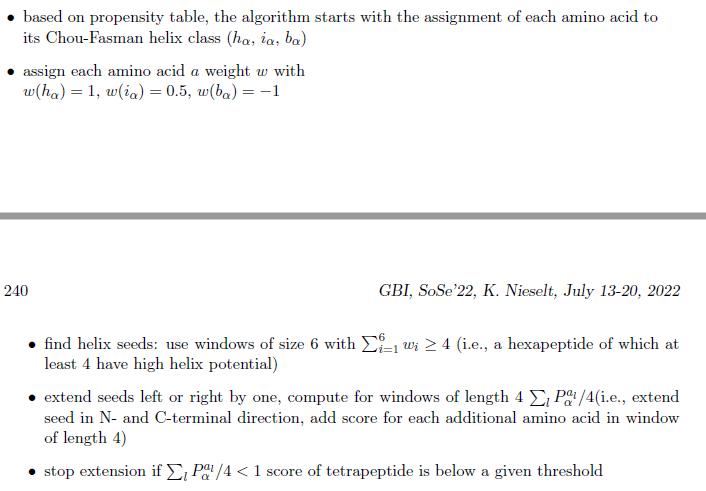




## What are the three helix and three beta sheets classes derived from the propensities?



## Describe the Chou-Fasman algorithm to find helices using the seed and extend algorithm.



## For the final assignment, what should one do when the helix and beta sheet prediction intervals overlap?

Compute average propensities, assign the one that is over 1 if the other one is smaller than one. Else, assign the larger one.

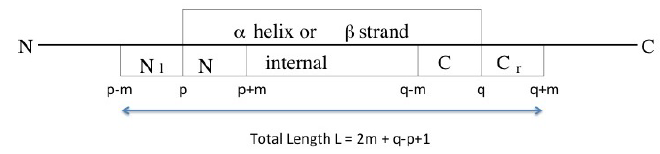
## What is a problem of the Chou-Fasman algorithm?

Does not take neighbourhood nor information from multiple sequence alignment into account.

## Describe the function of the secondary structure prediction program.

Locate secondary structure elements and use linear discriminant analysis to assign segments of a given amino acid sequence to a particular type of secondary structure.

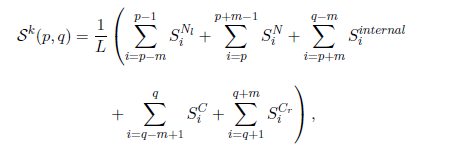
## What are the 5 defined regions for a secondary structure?



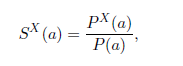
## Which characteristics are linear discriminant functions combining?

1. Singleton preference
2. Doublet preference
3. Hydrophobic moment

## What is the singleton characteristic defined as?



With



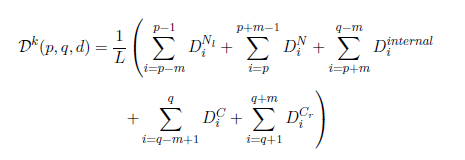
X element {N\_l, N-temrinal, internal, C-terminal and C\_r}

L = 2m + q -p + 1

For each type of secondary structure k a threshold c\_k is determined that is used to discriminate between true predictions and false ones.

## What is the doublet characteristic?

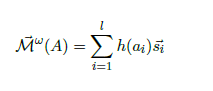




## Where do helices often lie on the protein?

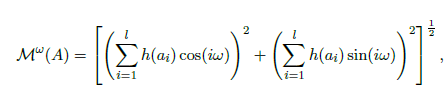
On the surface, there is a tendency for hydrophobi residues to face the core of the protein.

## What is hydrophobicity defined as? What is the structural hydrophobic moment defined as?

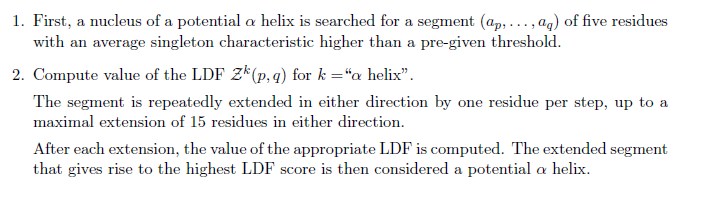


Hydrophobicity: Vector with magnitude and direction

Magnitude of the hydrophobic moment for angle omega:



## What are the steps of the SSP algorithm?



## What are the most successful methods for predicting secondary structure based on?

Neural networks

## Describe visible and hidden layers.

Hidden are layers that are not connected with the external world.

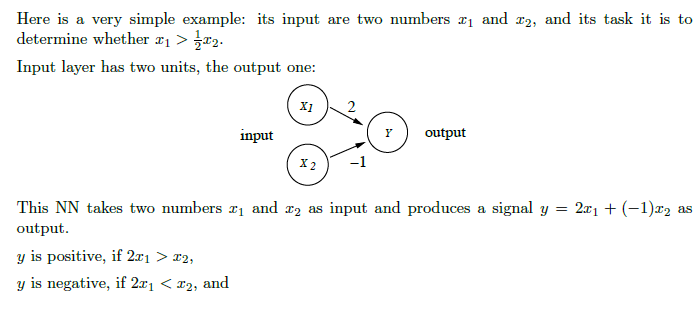
## Describe recurrent and feed forward architectures.

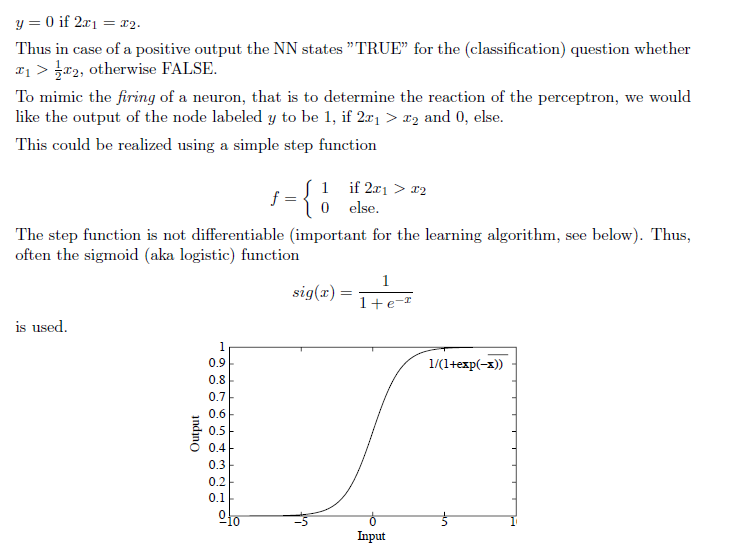
Recurrent: contain directed loops

Feed forward: contain no directed loops

## Describe a simple perceptron.

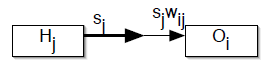
Very simple neural network.





## Describe backpropagation.

How to learn the weights? Iteratively the weights are optimised with training data. We have an output node O\_i. The signal arriving at O\_i is s\_j\*w\_ij.



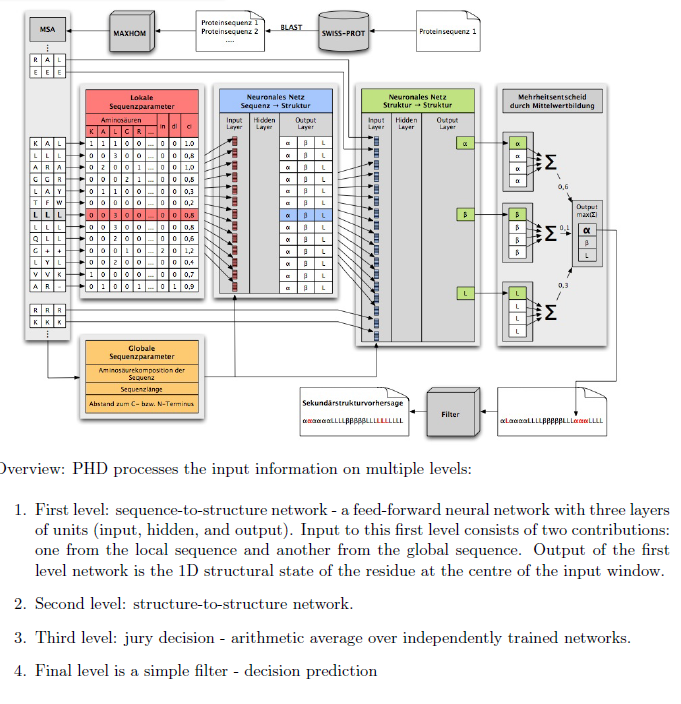
How do we alter w\_ij so as to bring the value s\_i of O\_i cloaer to the desired (known) value d\_i? The error will decrease the better we choose w\_ij.

We choose the best option with gradient descent.

## How does the PHD neural network work?

Two connected artificial neural networks. Each ANN is a feed-forward network with three layers: input layer, output layer and a hidden layer.

## What are the levels that PHD processes information on?



## Describe the steps of PHD.

1. Get MSA
2. Get global and local parameters from MSA present it to the input layer of the first NN
3. In the first NN, predict the secondary structure of a central residue within a sliding window of 13 residues 🡪 outputs three values foor the central residue that predict the values for alpha-helix, beta-sheet and loop
4. For the same residue again a sliding window, now of length 17. Local parameters are the three values (alpha, beta, L) + value for gap and conservation. 🡪 output three values (alpha, beta, L)
5. Presentation is done four times, four neural networks. Mean value is computed. Residue is assigned to that secondary structure class with the largest mean.

# Protein Tertiary Structure

## What does studying the structure of proteins help us to understand?

* Processes of protein synthesis
* Enzyme function
* Molecular causes for disease
* How cells use chemical energy, light energy, electrical energy and mechanical energy to power the processes of life
* Complex infrastructure of molecules for support and communication
* How scientists can create new molecular machines with novel functions

## What is the structure of a PDB entry?

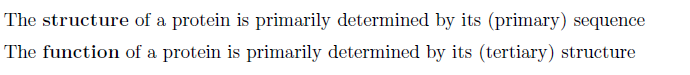
* Coordinates of each atom as given in its respective PDB entry.

## What are the three broad approaches to structure prediction?

* Fold recognition
* Comparative modelling
* Ab initio (de novo) modelling

## Formulate the fold recognition problem.

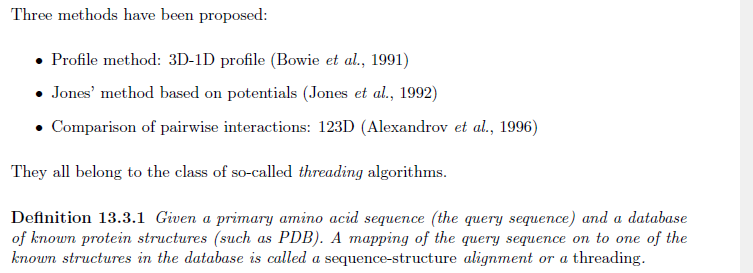
Given a protein sequence of unknown structure and a database of representative folds, identify the most plausible fold for the sequence, if there is one, and assess the quality or reliability of the proposed structure.



## What is inverse folding?

Instead of trying to predict the structure from the sequence, a structure is sought to be fit to a sequence.

## Define threading and name three methods for it.



## Describe the 123D fold recognition method

Given a query amino acid sequence A and a proposed fold f, 123D uses a dynamic programming algorithm to find an optimal threading of A through f, i.e. an alignment of A to f.

The scoring function is the free energy function (which needs to be minimized) . The scoring function for the matching uses a weighted sum of sequence, secondary structure preference and contact capacity potential contributions.

## The free energy function is given by the sum of two terms. Which ones?

Secondary structure preferences: Like propensity

Contact capacity potentials: When the structure of the input sequence and the template share the same fold, then the residues in both structures also share a similar chemical environment. Characterizes the ability of a residue to make a certain number of contacts with other residues.

## What is defined as contact?

If the distance between their C\_beta atoms is less than 7 Ängström.

## What is the formula for CCP?

