[1 Chapter 2 - Algorithms 6](#_Toc105511034)

[1.1 Name 8 algorithm strategies 6](#_Toc105511035)

[1.2 Describe the branch and bound algorithm 6](#_Toc105511036)

[1.3 Describe the steps of the divide-and-conquer algorithm 6](#_Toc105511037)

[1.4 Describe the principle of dynamic programming 6](#_Toc105511038)

[1.5 Describe the pre-order tree-traversal. 6](#_Toc105511039)

[1.6 Describe the post-order tree-traversal 6](#_Toc105511040)

[1.7 Describe in-order tree-traversal 6](#_Toc105511041)

[1.8 Describe breadth-first traversal of a tree 7](#_Toc105511042)

[2 Chapter 3 – Pairwise Alignment 8](#_Toc105511043)

[2.1 Define string and alphabet 8](#_Toc105511044)

[2.2 Define concatenation 8](#_Toc105511045)

[2.3 Define substring 8](#_Toc105511046)

[2.4 Reverse complement 8](#_Toc105511047)

[2.5 What are the building blocks of a fasta-file? 8](#_Toc105511048)

[2.6 How is a dot matrix composed? 8](#_Toc105511049)

[2.7 How are common substrings (reversed, displaced, repeated) visible in the dot matrix? 8](#_Toc105511050)

[2.8 What parameters are used to reduce noise in a dot matrix? 8](#_Toc105511051)

[2.9 What runtime does the dot matrix have? 8](#_Toc105511052)

[2.10 Define pairwise global alignment. 8](#_Toc105511053)

[2.11 What two characters CAN’T be aligned in a pairwise alignment? 9](#_Toc105511054)

[2.12 What assumption is made with scoring such that the scoring scheme can be additive? 9](#_Toc105511055)

[2.13 What is the formula of the Minkowski metric? What are its first two instances called? 9](#_Toc105511056)

[2.14 What is the formula of the Hamming distance? What is an intuitive explanation for it? 9](#_Toc105511057)

[2.15 What is the definition of the Levenshtein (or edit) distance? 9](#_Toc105511058)

[2.16 What is the definition of an edict transcript? 9](#_Toc105511059)

[2.17 What are the three essential components of dynamic programming? 9](#_Toc105511060)

[2.18 What is a prefix? 9](#_Toc105511061)

[2.19 What is the distance matrix defined as? In which direction goes an insertion, in which a deletion? 9](#_Toc105511062)

[2.20 What is the formula for the score S of an alignment A? 10](#_Toc105511063)

[2.21 What is the definition of a substitution matrix? 10](#_Toc105511064)

[2.22 What is the formula for the conditional probability of A given B? 10](#_Toc105511065)

[2.23 What is the formula for the joint probability? 10](#_Toc105511066)

[2.24 What are the entries of a substitution matrix S representing? How are they computed? 10](#_Toc105511067)

[2.25 Kochrezept: How can you compute your own substitution matrix? 10](#_Toc105511068)

[2.26 Name two commonly used substitution matrices. 10](#_Toc105511069)

[2.27 What does BLOSUM stand for? What is the requirement if there is an integer behind the word? 10](#_Toc105511070)

[2.28 How are the entries in the BLOSUM matrix computed? 10](#_Toc105511071)

[2.29 What are some possible ways to classify amino acids? 11](#_Toc105511072)

[2.30 What is the formula for a function that assigns a constant gap penalty of gap length g? 11](#_Toc105511073)

[2.31 What is the formula for a function that assigns a gap penalty that increases with gap extension? Is e usually smaller or bigger than d? 11](#_Toc105511074)

[2.32 Define subsequence of X and Y. 11](#_Toc105511075)

[2.33 What is the formula for the number of possible alignments between X and Y? 11](#_Toc105511076)

[2.34 What is another way to write the formula for the number of possible alignments between X and Y? 11](#_Toc105511077)

[2.35 What does Stirling’s formula look like? For which n and m does it hold? 11](#_Toc105511078)

[2.36 Define optimal global alignment. 11](#_Toc105511079)

[2.37 What are the three cases for the Fitch algorithm (in words and in formula)? What other conditions are defined for its initialization? 11](#_Toc105511080)

[2.38 Reproduce the pseudocode for the Needleman-Wunsch algorithm. 12](#_Toc105511081)

[2.39 What is the complexity of the Needleman-Wunsch algorithm, and why? 12](#_Toc105511082)

[2.40 Where do we apply pairwise local alignment? 12](#_Toc105511083)

[2.41 Define local alignment. 12](#_Toc105511084)

[2.42 What is the difference between the Smith-Waterman algorithm to the Needleman-Wunsch-algorithm? What is the formula for F(i,j) with the Smith-Waterman algorithm? 12](#_Toc105511085)

[2.43 Reproduce the pseudocode for the Smith-Waterman algorithm. 13](#_Toc105511086)

[2.44 What is the complexity of the Smith-Waterman algorithm? 13](#_Toc105511087)

[2.45 Under what condition does the Smith-Waterman algorithm work? Include the formula. 13](#_Toc105511088)

[3 Chapter 4 – BLAST 14](#_Toc105511089)

[3.1 What is the formula for sensitivity? 14](#_Toc105511090)

[3.2 What is the formula for specificity? 14](#_Toc105511091)

[3.3 What does BLAST stand for? 14](#_Toc105511092)

[3.4 What are the two components of BLAST? 14](#_Toc105511093)

[3.5 What is the algorithm strategy of BLAST? 14](#_Toc105511094)

[3.6 Describe the seed lemma by Owolabi and McGregor. Does this describe the worst or the best case? 14](#_Toc105511095)

[3.7 Describe the goal of BLAST. 14](#_Toc105511096)

[3.8 Describe the alignment score using the BLAST terms “query”, “database”, “segment” and “scoring matrix”. 14](#_Toc105511097)

[3.9 What three parameters are used in BLAST? 14](#_Toc105511098)

[3.10 Describe the preprocessing for BLAST. 15](#_Toc105511099)

[3.11 What are the steps of the BLAST search algorithm? 15](#_Toc105511100)

[3.12 What are typical default parameters for word size W for proteins and DNA sequences? 15](#_Toc105511101)

[3.13 Describe the keyword tree used in BLAST. 15](#_Toc105511102)

[3.14 Name and describe the different variants of the BLAST program. 15](#_Toc105511103)

[3.15 What is the underlying distribution under BLASTs statistical considerations? What is the formula? 15](#_Toc105511104)

[3.16 What is the formula of the E-value? What is its description? 15](#_Toc105511105)

[3.17 What is the formula of the p-value in the BLAST context? 15](#_Toc105511106)

[3.18 Describe the BLAST bit score. What is its formula? 16](#_Toc105511107)

[3.19 How are E-value and bit score S’ related (formula)? 16](#_Toc105511108)

[4 Chapter 5 – Multiple Alignments 17](#_Toc105511109)

[4.1 What constitutes a multiple sequence alignment? What are its goals? 17](#_Toc105511110)

[4.2 What is the advantage of MSA in comparison to pairwise alignments? 17](#_Toc105511111)

[4.3 How is an MSA obtained? What are the requirement concerning the resulting and original sequences in it? 17](#_Toc105511112)

[4.4 How is an MSA scored? 17](#_Toc105511113)

[4.5 Given a column of length r, how many different columns can be obtained by inserting a gap symbol at certain places instead of the original character? Give the formula. 17](#_Toc105511114)

[4.6 Why do we need the sum of pairs score? What are the ‘reasonable asumptions’ so that we can compute the score in this way? 17](#_Toc105511115)

[4.7 For the sum of pairs score, hat are the output values of the scoring function for a given combination of residues and gaps? 18](#_Toc105511116)

[4.8 Reproduce the formal definition of the sum of pairs score. 18](#_Toc105511117)

[4.9 How would you compute the Dynamic Programming/Fitch algorithm for a three-sequence MSA? How many cases of the scoring function for a certain column composition would there be? 18](#_Toc105511118)

[4.10 What is the space complexity of the generalized Dynamic Programming algorithm? What is the time complexity? Is it NP-complete? 18](#_Toc105511119)

[4.11 What are the steps of progressive alignment? 18](#_Toc105511120)

[4.12 Where do different progressive alignment implementations differ? 18](#_Toc105511121)

[4.13 Reproduce the pseudocode for progressive alignments 19](#_Toc105511122)

[4.14 Describe the process of pair-guided alignment. 19](#_Toc105511123)

[4.15 What is the main difference between profile alignment and pair-guided alignment? 19](#_Toc105511124)

[4.16 Provide the formula for the sum of pairs score for an MSA A\* computed from two profiles A1\* and A2\*. Explain each resulting term. 19](#_Toc105511125)

[4.17 Define guide trees. 19](#_Toc105511126)

[4.18 Describe the steps of the Feng-Doolittle algorithm. What is its purpose? 19](#_Toc105511127)

[4.19 What is the formula for the distance score as defined by Feng-Doolittle for two sequences X and Y? 20](#_Toc105511128)

[4.20 In the Feng-Doolittle formula, how is Sobs defined? 20](#_Toc105511129)

[4.21 In the Feng-Doolittle formula, how is Sid defined? What is the formula? 20](#_Toc105511130)

[4.22 In the Feng-Doolittle formula, how is Srand defined? What is the formula? 20](#_Toc105511131)

[4.23 What are the steps of the ClustalW algorithm? 20](#_Toc105511132)

[4.24 What distance functions does ClustalW use? Provide formulas if applicable. 20](#_Toc105511133)

[4.25 What is one distinctive feature of ClustalW? 21](#_Toc105511134)

[4.26 Which of the steps in ClustalW is the most time-consuming? 21](#_Toc105511135)

[4.27 Name other programs for MSAs. 21](#_Toc105511136)

[5 Chapter 6 – Phylogeny 22](#_Toc105511137)

[5.1 What is the goal of phylogenetic analysis? 22](#_Toc105511138)

[5.2 Describe the basic structure of phylogenetic trees. 22](#_Toc105511139)

[5.3 Provide the formal definition off phylogenetic trees. 22](#_Toc105511140)

[5.4 When are two species related by definition? 22](#_Toc105511141)

[5.5 Define a Cluster 22](#_Toc105511142)

[5.6 Define a Clade. 22](#_Toc105511143)

[5.7 Name different ways to draw a dendrogram. 22](#_Toc105511144)

[5.8 How many nodes and edges does a phylogenetic tree with n taxa have? 22](#_Toc105511145)

[5.9 How many unrooted and rooted trees can be constructed from n taxa? 22](#_Toc105511146)

[5.10 Name some strategies to construct phylogenetic trees 22](#_Toc105511147)

[5.11 What is the normalized Hamming distance defined as? 22](#_Toc105511148)

[5.12 Define the tree distance of a phylogenetic tree (with formula). 23](#_Toc105511149)

[5.13 Describe the steps that are taken in UPGMA. 23](#_Toc105511150)

[5.14 What is the formula for the average distance between pairs of taxa from each cluster? 23](#_Toc105511151)

[5.15 What is the update formula for a Cluster k (union of clusters i and j) and any other cluster l? 23](#_Toc105511152)

[5.16 What are the three cases that are distinguished for determining edge lengths in a tree? What are the corresponding formulas? 23](#_Toc105511153)

[5.17 Provide the pseudocode for the UPGMA algorithm. 24](#_Toc105511154)

[5.18 What is the molecular clock hypothesis? 24](#_Toc105511155)

[5.19 Is UPGMA suitable or unsuitable for data that doesn’t fulfil the molecular clock hypothesis? What about NJ? 24](#_Toc105511156)

[5.20 What is the main difference in the algorithm process between UPGMA and NJ? 24](#_Toc105511157)

[5.21 How is the neighboring matrix made up (with formulas)? 24](#_Toc105511158)

[5.22 Provide the pseudocode for the Neighbor-Joining algorithm 25](#_Toc105511159)

[5.23 What is the complexity of UGMA and Neighbor-joining? 25](#_Toc105511160)

[5.24 What are the two assumptions made for maximum parsimony? 25](#_Toc105511161)

[5.25 How is the parsimony score of a tree computed? 25](#_Toc105511162)

[5.26 Define the small parsimony problem. 25](#_Toc105511163)

[5.27 What algorithm is used to solve the small parsimony problem? 25](#_Toc105511164)

[5.28 What are the two stages of the Fitch algorithm and what is their purpose? Are they bottom-up or top-down? 25](#_Toc105511165)

[5.29 What are the steps of the forward pass? 26](#_Toc105511166)

[5.30 Provide the pseudocode for the forward-pass 26](#_Toc105511167)

[5.31 Provide the pseudocode of the backward-pass. 26](#_Toc105511168)

[5.32 What is the run time complexity of the Fitch algorithm? 26](#_Toc105511169)

[5.33 What is the Large Parsimony Problem? 26](#_Toc105511170)

[5.34 What is the problem with the Large Parsimony Problem? 27](#_Toc105511171)

[5.35 What is the process of the heuristic for choosing a parsimony tree? 27](#_Toc105511172)

[5.36 What is the difference between orthologous genes and paralogous genes? 27](#_Toc105511173)

# Chapter 2 - Algorithms

## Name 8 algorithm strategies

### Recursive

### Dynamic programming

### Backtracking

### Heuristic

### Branch and bound

### Divide and conquer

### Greedy

### Brute force

## Describe the branch and bound algorithm

* + 1. Find a solution of a combinatorial problem with each solution associated with a cost, such that the whole solution cannot be smaller than any partial solution. Remember the lowest-cost solution and keep that as a lower bound for other solutions. Problems: Search can still become exponential

## Describe the steps of the divide-and-conquer algorithm

* + 1. 1. Divide a larger problem into smaller subproblems of the same type
    2. 2. Solve each subproblem recursively
    3. 3. Combine smaller solutions into a solution for the whole problem
    4. Divide-and-conquer is a top-down approach.

## Describe the principle of dynamic programming

* + 1. Divide each problem into a succession of stages, in which the first stage only consists of trivial solutions and each partial solution in a later stage can be calculated by recurring on only a fixed number of partial solutions in an earlier stage. DP is bottom-up.
    2. It consists of four steps:
    3. 1. Characterisation of the structure of the optimal solution
    4. 2. Recursive definition of the value of an optimal solution
    5. 3. Computation of the optimum using recursion
    6. 4. Construction of an optimal solution through the computed optimal value

## Describe the pre-order tree-traversal.

* + 1. Pre-order tree-traversal is top-down. Examine the root of a tree T. Then traverse its left sub-tree in pre-order, then traverse its right sub-tree in pre-order.

## Describe the post-order tree-traversal

* + 1. Post-order tree-traversal is bottom-up. First traverse the left subtree of the root in post-order then traverse the right subtree of the root in post-order. Finally examine the root.

## Describe in-order tree-traversal

* + 1. First examine thee left subtree of a tree in in-order, then examine the root, then examine the right subtree of a tree in in-order.

## Describe breadth-first traversal of a tree

### Put the root in a queue. Repeat the following until the queue is empty: Pop a node off the beginning of the queue and examine it. Then, if the node has any children, push them onto the end of the queue. First, always examine the root and then the left and right child-node of the root. Then examine the left-and right child node of the left node, then of the right. Stop if a node has no child-nodes.

# Chapter 3 – Pairwise Alignment

## Define string and alphabet

* + 1. An alphabet is a finite, non-empty set of symbols and letters. A string is obtained by writing a sequence of elements s1s2s3… of symbols of the alphabet. The length of a string is the number of symbols in it denoted as |S|.

## Define concatenation

* + 1. Let S and T be two strings on an alphabet. We will use ST to denote the concatenation of S and T.

## Define substring

* + 1. Let S and T be two strings on an alphabet. We call S a substring of T if there exists two strings U, V from the alphabet such that T = USV. We denote a substring of S as S[i,j].

## Reverse complement

Let S be a sequence of length n on the DNA alphabet {A,G,C,T}. The sequence S bar is computed with s\_bar\_i, which is:

A if s\_n-i+1 = T

G if s\_n-i+1 = C

T if s\_n-i+1 = A

C if s\_n-i+1 = G

## What are the building blocks of a fasta-file?

Header line starts with “>” and contains identifier for sequence. Successive lines contain the sequence.

## How is a dot matrix composed?

### Write Sequence S in row zero. Write sequence T in column zero. Place a dot at each cell for which the corresponding symbols match. M[i,j] is 1 where s\_i = t\_i, and 0 everywhere else

## How are common substrings (reversed, displaced, repeated) visible in the dot matrix?

### Common: Diagonal line from top left to bottom right

### Reversed common: Diagonal line from bottom let to top right

### Displaced common: Diagonal line NOT on the main diagonal

### Repeated common: diagonal lines at the same horizontal/vertical position as other diagonal lines

## What parameters are used to reduce noise in a dot matrix?

Window size w (how many characters are looked at?) and stringency s (how many characters need to be equal in the next w position?).

## What runtime does the dot matrix have?

Quadratic.

## Define pairwise global alignment.

A pairwise alignment is obtained by inserting dashes into two strings X and Y such that: (1) The resulting strings X’ and Y’ are of the same length (2) They can be written on top of each other so each character in one string is opposite a unique character in the other string and (3) removing all gaps results in the original strings (X’\{-}=X).

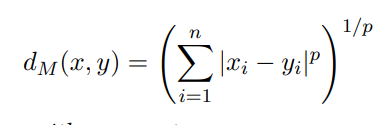
## What two characters CAN’T be aligned in a pairwise alignment?

Two dashes.

## What assumption is made with scoring such that the scoring scheme can be additive?

Evolution of all sites within each sequence is independent of each other.

## What is the formula of the Minkowski metric? What are its first two instances called?



Manhattan distance, Euclidean distance

## What is the formula of the Hamming distance? What is an intuitive explanation for it?

### 

The Hamming distance is the number of characters between two strings that are not equal at the same positions.

## What is the definition of the Levenshtein (or edit) distance?

The minimum number of replacements, deletions or insertions that are needed to transform sequence X into sequence Y.



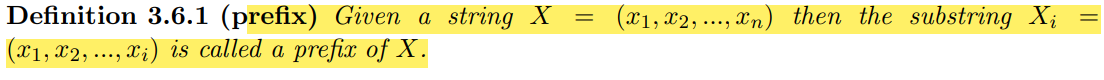
## What is the definition of an edit transcript?

Using M for match, R for replacement, D for deletion and I for insertion, the edit transcript is a string that describes the transformation of a string X into a string Y.

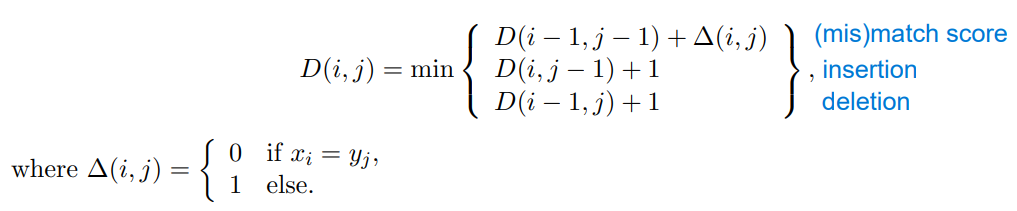
## What are the three essential components of dynamic programming?

Recurrence relation, tabular computation and traceback

## What is a prefix?

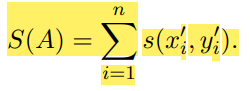


## What is the distance matrix defined as? In which direction goes an insertion, in which a deletion?



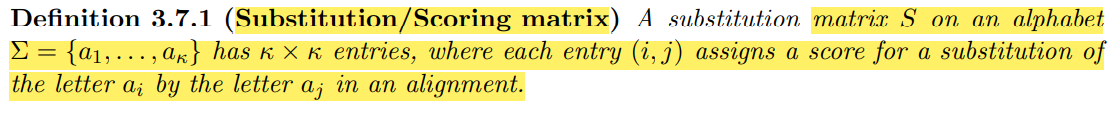
In the first column of D, we assume all deletions. In the first row of D, we assume all insertions.

## What is the formula for the score S of an alignment A?



Where X’ and Y’ are strings in which there have possibly been inserted gaps.

## What is the definition of a substitution matrix?



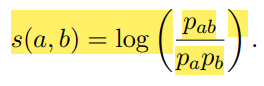
## What is the formula for the conditional probability of A given B?



## What is the formula for the joint probability?



## What are the entries of a substitution matrix S representing? How are they computed?



They represent the log-odds-ratio of the probability that the two elements are derived from a common ancestor divided by the probability that they occur randomly. Because it is the log, we can add up the elements of the scoring matrix for a score.

## Kochrezept: How can you compute your own substitution matrix?

### Compute match model: Derive the frequency of each substitution in the given alignments.

### Compute random model: Derive the frequency of each symbol

* + 1. Compute each entry s(x,y) with the formula given above (joint probability divided by single probabilities).

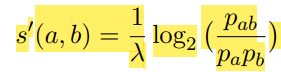
## Name two commonly used substitution matrices.

BLOSUM(x), PAM

## What does BLOSUM stand for? What is the requirement if there is an integer behind the word?

BLOSUM stands for BLOcks Substitution Matrix. The integer means that if two sequences have at least this amount of sequence identity they are placed in the same cluster. All counts are based on comparisons between pairs of sequences in different clusters. The more similar the sequences should be, the higher the integer.

## How are the entries in the BLOSUM matrix computed?



Rounded to the nearest integer value.

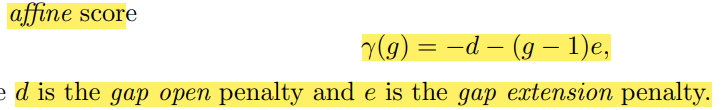
## What are some possible ways to classify amino acids?

Side chain polarity (polar vs. nonpolar), side chain charge (positive, neutral or negative), hydropathy index, size..

## What is the formula for a function that assigns a constant gap penalty of gap length g?

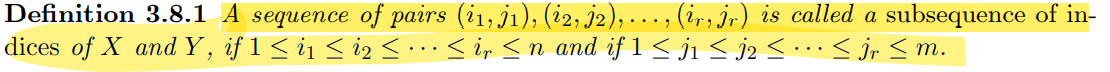


## What is the formula for a function that assigns a gap penalty that increases with gap extension? Is e usually smaller or bigger than d?



Usually, e < d, which means that less isolated gaps are produced.

## Define subsequence of X and Y.

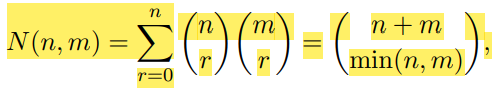


## What is the formula for the number of possible alignments between X and Y?

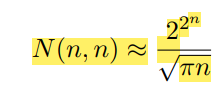


This equals the number of possible subsequences, ergo the combinatorial possibilities to choose r mismatch/match positions in each sequence.

## What is another way to write the formula for the number of possible alignments between X and Y?

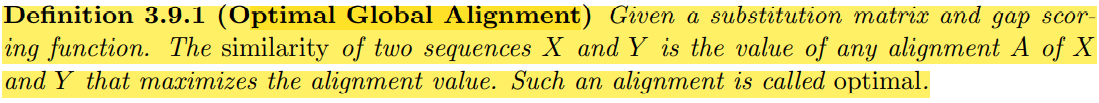


## What does Stirling’s formula look like? For which n and m does it hold?

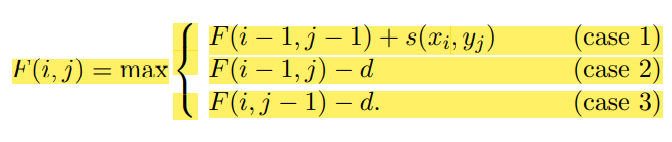


If n = m

## Define optimal global alignment.



## What are the three cases for the Fitch algorithm (in words and in formula)? What other conditions are defined for its initialization?



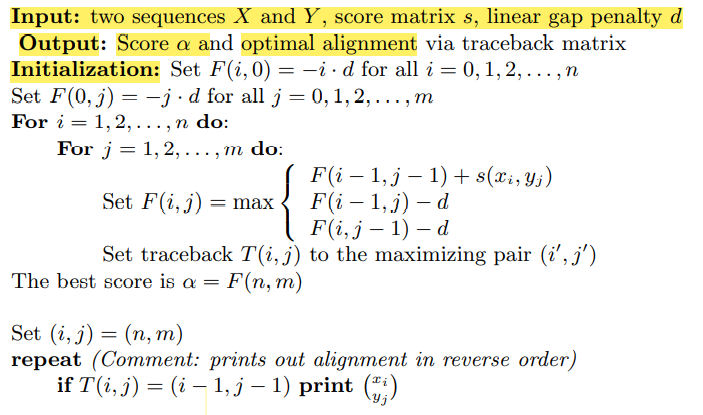
Case 1: Add score for the two aligned characters.

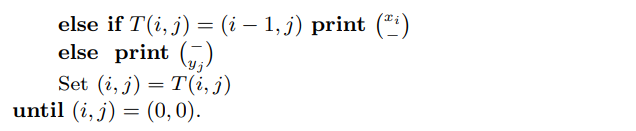
Case 2: Add score for gap score aligned with x

Case 3: Add score for gap score aligned with y

Additional initialization conditions: 1. F(0,0) = 0 2. F(i,0) > 0 for all i = 1,…,n and F(j, 0) > 0 for all j=1,…,m.

## Reproduce the pseudocode for the Needleman-Wunsch algorithm.





## What is the complexity of the Needleman-Wunsch algorithm, and why?

The algorithm takes O(nm) time and memory. This is because we need to store (n+1)x(m+1) numbers and each number takes three additions and a max to compute.

## Where do we apply pairwise local alignment?

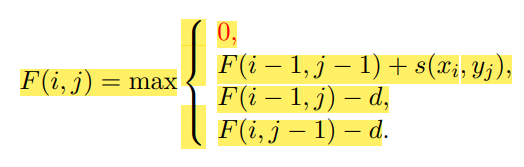
When we find out the alignment of two substrings that have a higher score than the score between the full strings.

## Define local alignment.

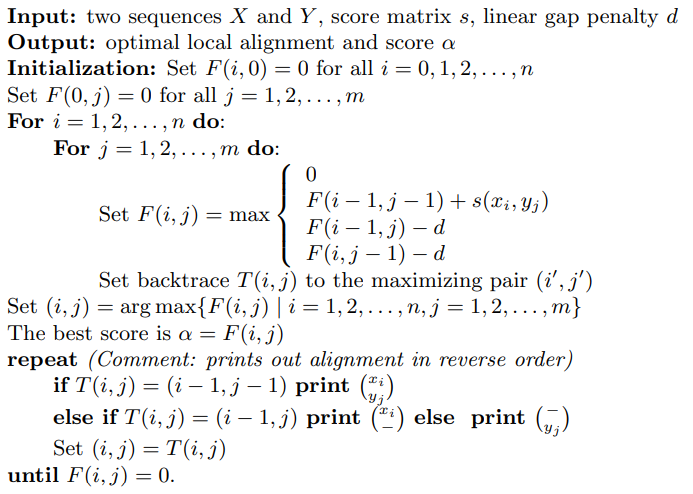
### 

## What is the difference between the Smith-Waterman algorithm to the Needleman-Wunsch-algorithm? What is the formula for F(i,j) with the Smith-Waterman algorithm?

The optimal score is never negative, thus we set the score to zero if all attainable values of NW at position (i, j) are negative. For the traceback, we start at the cell with the highest score and end at any cell with score 0.



## Reproduce the pseudocode for the Smith-Waterman algorithm.

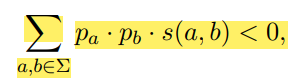


## What is the complexity of the Smith-Waterman algorithm?

O(nm) (like Needleman-Wunsch)

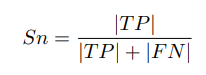
## Under what condition does the Smith-Waterman algorithm work? Include the formula.

The expected score of a random match to be negative.

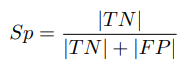


# Chapter 4 – BLAST

## What is the formula for sensitivity?



## What is the formula for specificity?



## What does BLAST stand for?

Basic Local Alignment Search Tool

## What are the two components of BLAST?

1. Search algorithm 2. Computation of statistical significance of solutions

## What is the algorithm strategy of BLAST?

Seed and extend. First find small exact matches (hits), then possibly extend to find long inexact ones (High-scoring Segment Pairs). Extending to the left or right of an HSP would lead to a lower score.

## Describe the seed lemma by Owolabi and McGregor. Does this describe the worst or the best case?

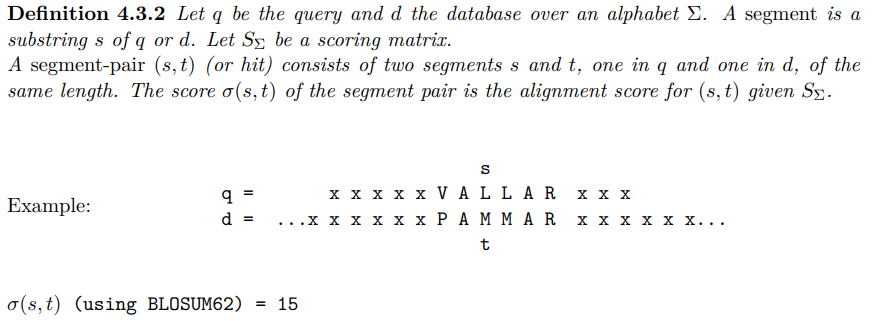


It describes the worst case.

## Describe the goal of BLAST.

Goal: Compute all High-scoring segment Pairs between two sequences with score >= S

## Describe the alignment score using the BLAST terms “query”, “database”, “segment” and “scoring matrix”.



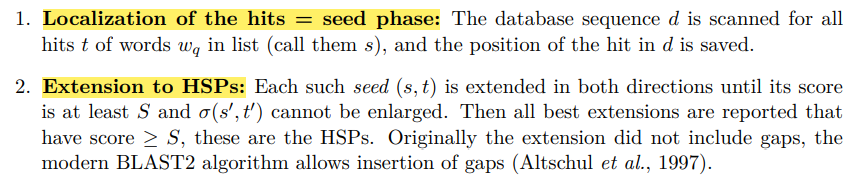
## What three parameters are used in BLAST?

Word size W, word similarity threshold T and HSP score S.

## Describe the pre-processing for BLAST.

Generate all words of length W from a query sequence q. Generate a list of all W-mers out of the alphabet that have similarity >= T to some word in the query sequence q.

## What are the steps of the BLAST search algorithm?



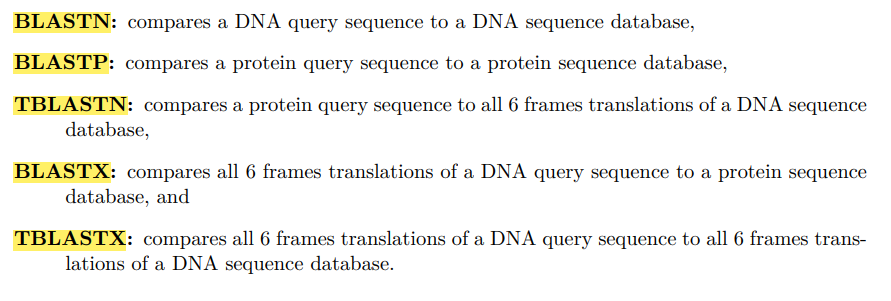
## What are typical default parameters for word size W for proteins and DNA sequences?

DNA: W= 11, Proteins: W = 3

## Describe the keyword tree used in BLAST.

A keyword tree is a finite state automaton used in pattern matching. It’s a rooted tree built from all words, edges are labelled with exactly one letter, only one edge per letter is leaving each node. Each word therefore maps to one node in the tree. We move forward through the tree if we can with a specific word, and jump back to the root if we can’t.

## Name and describe the different variants of the BLAST program.

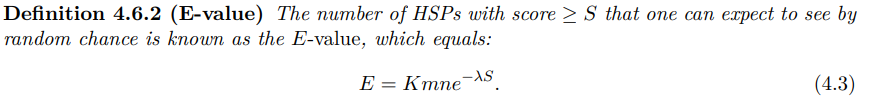


## What is the underlying distribution under BLASTs statistical considerations? What is the formula?

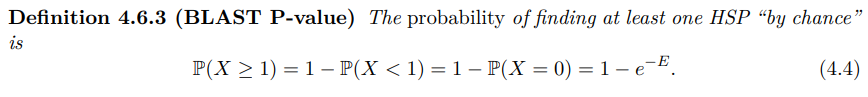
The Poisson distribution.



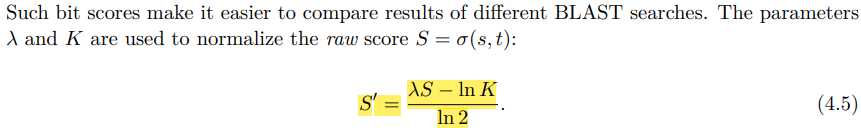
## What is the formula of the E-value? What is its description?



## What is the formula of the p-value in the BLAST context?



## Describe the BLAST bit score. What is its formula?



## How are E-value and bit score S’ related (formula)?



# Chapter 5 – Multiple Alignments

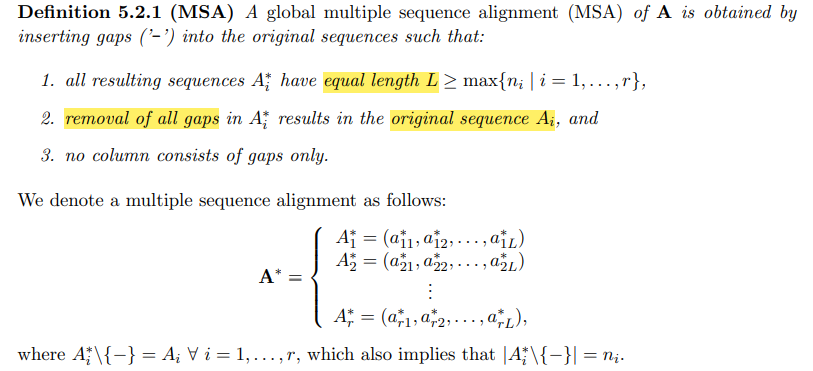
## What constitutes a multiple sequence alignment? What are its goals?

A multiple sequence alignment is an alignment of more than two sequences. The MSA arranges protein sequences into rectangular array with the goal that the residues in a given column are homologous (derived from a single position in an ancestral sequence), superposable (in a rigid local structural alignment) or play a common functional role. Detect homologous residues and place them in the same column of the MSA.

## What is the advantage of MSA in comparison to pairwise alignments?

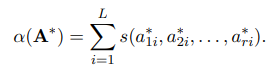
The chance off random similarities decreases with the increasing number of aligned sequences.

## How is an MSA obtained? What are the requirements concerning the resulting and original sequences in it?

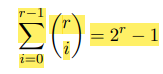


## How is an MSA scored?

Because we can assume independence of the different columns of an MSA, the score alpha can be defined as a sum off column scores. The function returns a score for every combination of r symbols (including the gap symbol).



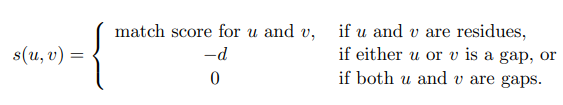
## Given a column of length r, how many different columns can be obtained by inserting a gap symbol at certain places instead of the original character? Give the formula.



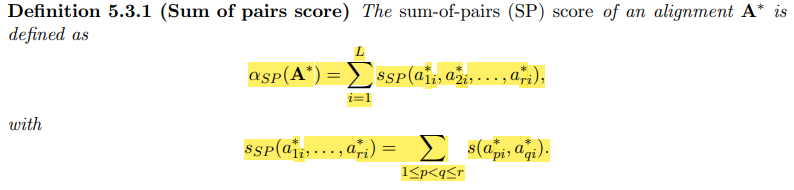
## Why do we need the sum of pairs score? What are the ‘reasonable assumptions’ so that we can compute the score in this way?

We need the sum of pairs score because constructing a substitution matrix for a whole specific column would be to laboursome. Instead, we can add up the single scores for each combination. Assumption 1: The score must be independent of order of gaps in the column. Assumption 2: Similar or identical aligned residues should get a high score, while unrelated or gaps a negative one.

## For the sum of pairs score, hat are the output values of the scoring function for a given combination of residues and gaps?



## Reproduce the formal definition of the sum of pairs score.



## How would you compute the Dynamic Programming/Fitch algorithm for a three-sequence MSA? How many cases of the scoring function for a certain column composition would there be?

You would construct a three-dimension cube, progressing in the three-dimensional direction of the optimal score. There would be seven cases for three dimensions.

## What is the space complexity of the generalized Dynamic Programming algorithm? What is the time complexity? Is it NP-complete?

Space complexity: O(nr). Run time complexity: O(r2n22r). Computing an MSA with optimal SP-score is NP-complete.

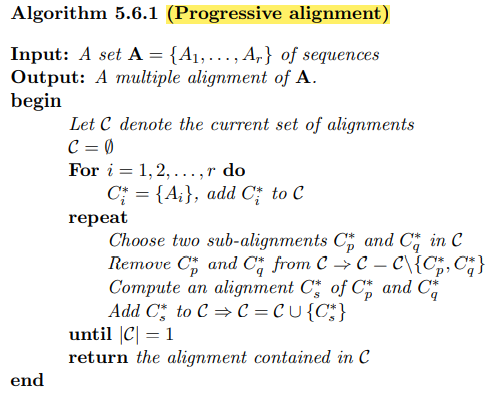
## What are the steps of progressive alignment?

1. Calculate a distance matrix, representing the distance between each pair of sequences
2. From this matrix, build a phylogenetic tree
3. Use this tree as guide to progressively align the sequences

## Where do different progressive alignment implementations differ?

1. Order of the aligned sequences
2. Whether a single multiple alignment is generated or several ones, following a tree structure
3. Which scoring function is used

## Reproduce the pseudocode for progressive alignments



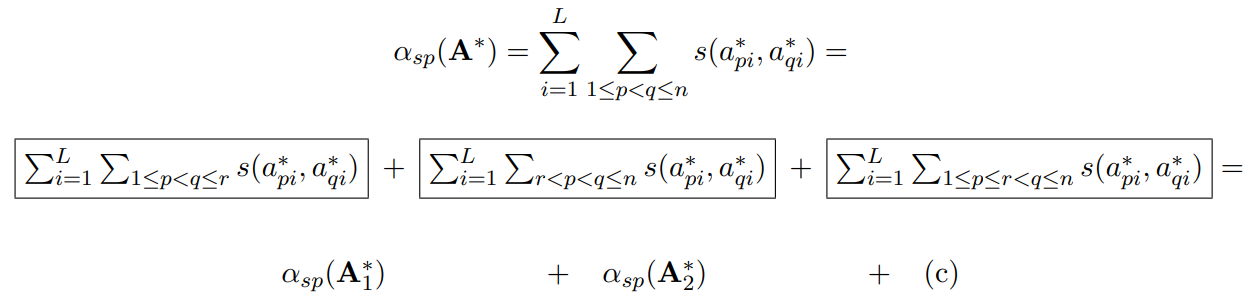
## Describe the process of pair-guided alignment.

Choose two sequences, one from each profile. Align these. By adjusting the gaps in the other sequences, the full alignment is produced by following this alignment.

## What is the main difference between profile alignment and pair-guided alignment?

In profile alignment, whole columns of gaps are inserted into either profile, without further changing the alignment of either of the two profiles.

## Provide the formula for the sum of pairs score for an MSA A\* computed from two profiles A1\* and A2\*. Explain each resulting term.



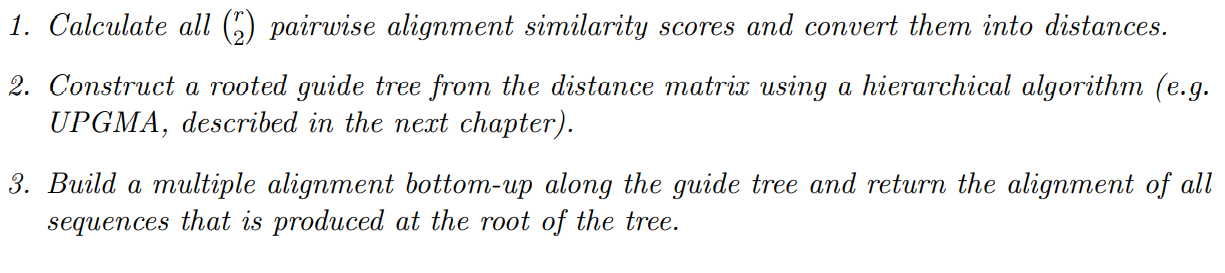
The first two scores are the alignment scores of each of the profiles. The term c) is the optimized alignment score between one sequence in the first and second profile respectively.

## Define guide trees.

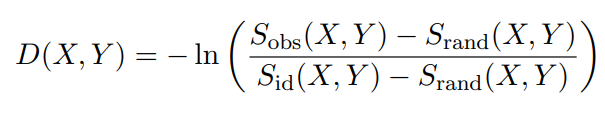
A guide tree is a phylogenetic tree whose leaves are labelled by the sequences that are to be aligned.

## Describe the steps of the Feng-Doolittle algorithm. What is its purpose?

Its purpose is building a multiple sequence alignment.



## What is the formula for the distance score as defined by Feng-Doolittle for two sequences X and Y?

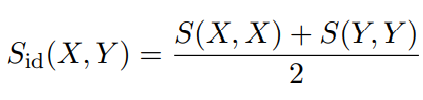


## In the Feng-Doolittle formula, how is Sobs defined?

Here, Sobs is the observed similarity score computed by pairwise alignment of X and Y.

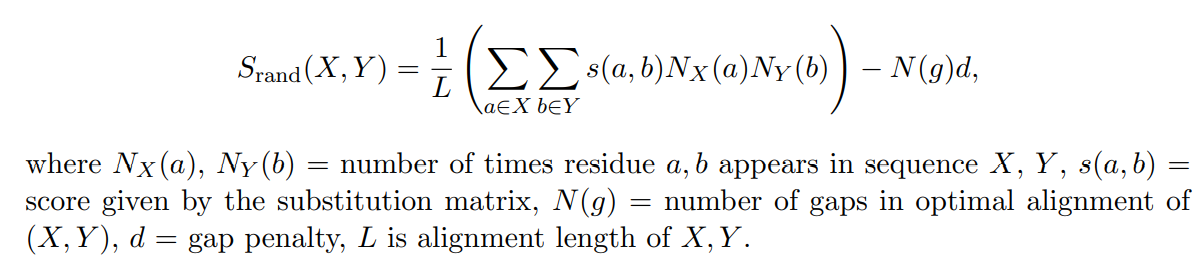
## In the Feng-Doolittle formula, how is Sid defined? What is the formula?

Sid is the average score of both aligned with each other, given by

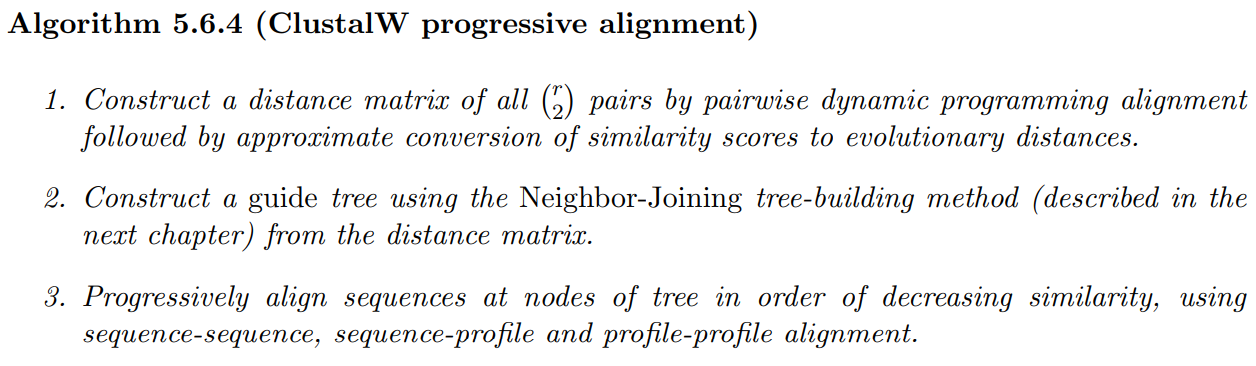


## In the Feng-Doolittle formula, how is Srand defined? What is the formula?

And Srand is the expected score of an alignment of two random sequences of the same length and composition as X and Y.

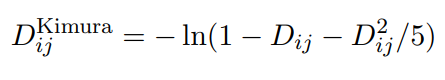


## What are the steps of the ClustalW algorithm?



## What distance functions does ClustalW use? Provide formulas if applicable.

1. Kimura distance.



Here, Dij is the number of mismatches in an optimal pairwise alignment divided by the length of the sequence.

1. K-tuple distance measure of Wilbur and Lipman for unaligned sequences. Computes the number of exactly matching k-tuples (for proteins k=2) using a hash-table approach.

## What is one distinctive feature of ClustalW?

Once a gap, always a gap.

## Which of the steps in ClustalW is the most time-consuming?

Computation of the initial pairwise alignments (> 90% of time)

## Name other programs for MSAs.

Muscle, T-Coffee, Clustal Omega, MAFFT

# Chapter 6 – Phylogeny

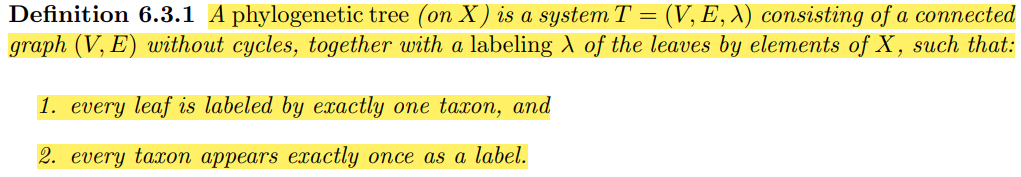
## What is the goal of phylogenetic analysis?

Describe the evolutionary relationship between a collection of extant species.

## Describe the basic structure of phylogenetic trees.

A tree consists of nodes connected by edges. Terminal nodes (leaves) represent sequences or species for which we have data. Internal nodes represent hypothetical ancestors.

## Provide the formal definition of phylogenetic trees.



Where V is the set of nodes and E is the set off edges.

## When are two species related by definition?

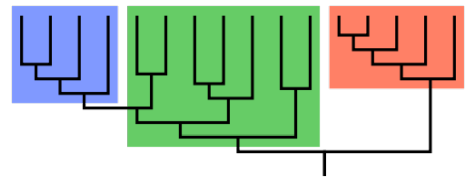
Two species are related if they share a recent common ancestor.

## Define a Cluster

A set of species.

## Define a Clade.

A cluster C is called a clade or monophyletic group if all species in C are more closely related to each other than to any species outside of C; in other words: all species in C are descended from a common ancestor, which is not an ancestor of any other species under consideration. In the example below, the green cluster is not a clade.



## Name different ways to draw a dendrogram.

Rooted, unrooted, slanted or traditional, circular.

## How many nodes and edges does a phylogenetic tree with n taxa have?

### 

## How many unrooted and rooted trees can be constructed from n taxa?

Unrooted: (2n – 5)!! Rooted: (2n-3)!!

## Name sequence-based strategies to construct phylogenetic trees

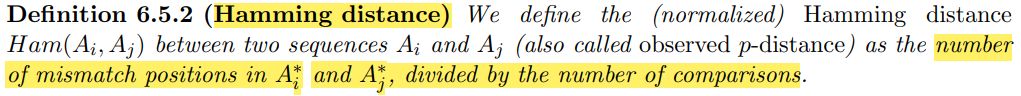
Distance method, maximum parsimony, Maximum Likelihood, Bayesian approaches.

## Name distance-based strategies to construct phylogenetic trees

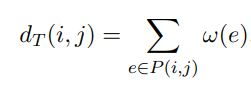
UPGMA, Neighbor-Joining

## What is the normalized Hamming distance defined as?

Columns where both sequences contain a gap are ignored. If only one sequence has a gap we can ignore it, or treat it as a match or a mismatch, depending on the data.



## Define the tree distance of a phylogenetic tree (with formula).

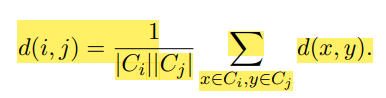


This is with weighted edges, the edge weight is omega.

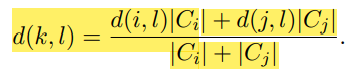
## Describe the steps that are taken in UPGMA.

* Input it the distance matrix D on all taxa, they form n clusters
* Starting from the leaves, identify the off-diagonal entry that is minimal
* These two leaves are merged into a new cluster with one internal new node
* Update the distance matrix by recomputing the distances of the not yet merged elements with the new cluster.
* At the next step the two elements with minimal distance are clustered and merged to a single cluster, distances are updated, etc., until only one cluster is left

## What is the formula for the average distance between pairs of taxa from each cluster?



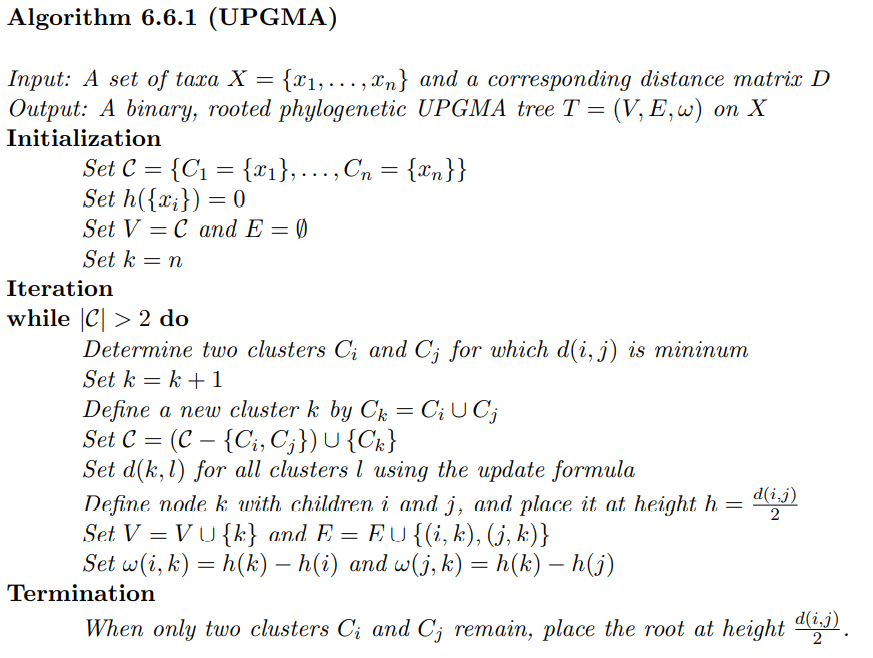
## What is the update formula for a Cluster k (union of clusters i and j) and any other cluster l?



## What are the three cases that are distinguished for determining edge lengths in a tree? What are the corresponding formulas?

1. Both child-nodes are leaves: ω(i, k) = ω(j, k) = h(k) = d(i,j)/2
2. One or two of the two children is an internal node: For the leave: see above, for the other one: ω(j, k) = h(k) – h(j) = d(i, j)/2 − d(a, b)/2
3. Both children are internal nodes: See number 2

## Provide the pseudocode for the UPGMA algorithm.



## What is the molecular clock hypothesis?

The rate of evolutionary events of sequences was constant over time and equal for all lineages in the tree.

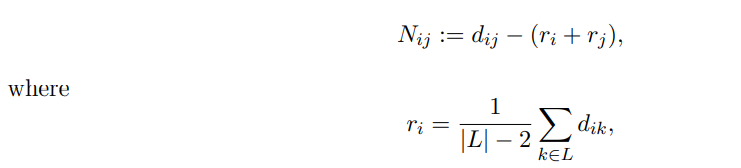
## Is UPGMA suitable or unsuitable for data that doesn’t fulfil the molecular clock hypothesis? What about NJ?

UPGMA: Unsuited. NJ: Suitable for all.

## What is the main difference in the algorithm process between UPGMA and NJ?

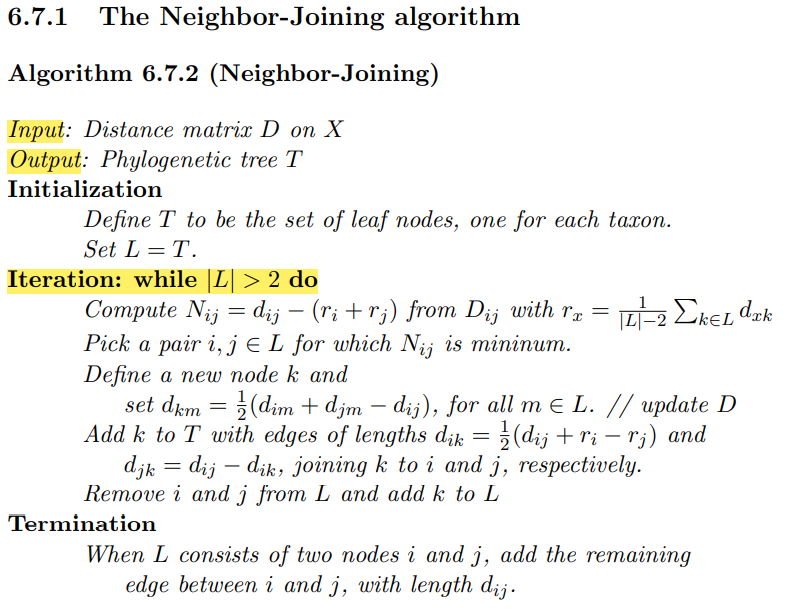
NJ joins branches which minimize the total branch length while UPGMA just chooses the minimal branch length between any two clusters.

## How is the neighboring matrix made up (with formulas)?



L is the set of clusters still present at the respective iteration step.

## Provide the pseudocode for the Neighbor-Joining algorithm



## What is the complexity of UGMA and Neighbor-joining?

UPGMA: runtime complexity O(n3). There are O(n) iteration steps (n-1) internal nodes in a rooted tree, each iteration involves joining a cluster-pair and reducing the input distance matrix (needs O(n) time). Choosing a cluster pair (global minimum) takes O(n2) time.

NJ: Also O(n3). Compute the N matrix to find the pair of clusters to join takes O(n2) time. There are O(n) iteration steps (there are n-2 internal nodes in an unrooted tree).

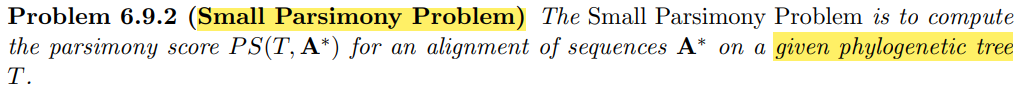
## What are the two assumptions made for maximum parsimony?

Changes in different sites are independent (each column is treated independently). Changes in different lineages across the phylogeny are independent (changes on each edge can be summed up for the overall score).

## How is the parsimony score of a tree computed?

This is done by taking the minimum of all possible labelings of the sum of all non-normalized Hamming distances

## Define the small parsimony problem.



## What algorithm is used to solve the small parsimony problem?

The Fitch algorithm.

## What are the two stages of the Fitch algorithm and what is their purpose? Are they bottom-up or top-down?

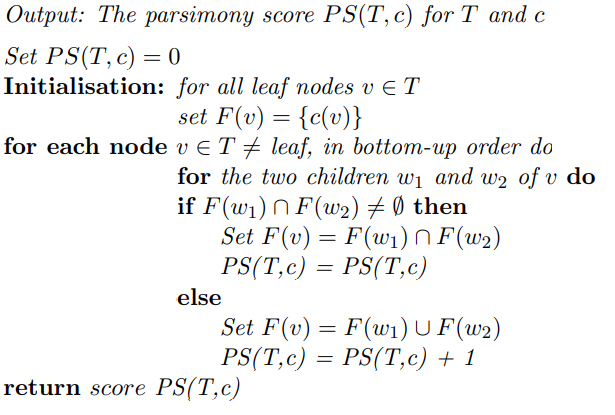
Forward pass is bottom up. It computes the parsimony score and a set F of possible states for each node. Backward pass is top down. It chooses a specific state for each node.

## What are the steps of the forward pass?

* Set parsimony score to zero.
* Start at the leaves and go along the tree up the root
* At each internal node, check if the intersection of states is empty. If not, take intersection and make it the label of the new parent node. If yes, add +1 to parsimony score, label parent node with union set of states.

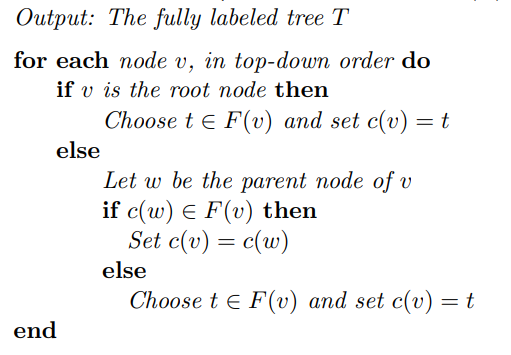
## Provide the pseudocode for the forward-pass

Input: A binary phylogenetic tree T, a state c(w) for each leaf (v)) w of T



## Provide the pseudocode of the backward-pass.

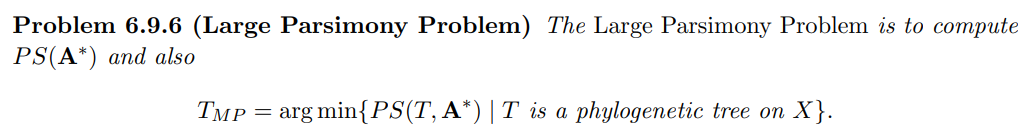
Input: A phylogenetic tree T and the set F(w) for every node w in T



## What is the run time complexity of the Fitch algorithm?

O(nL), because it requires O(n) steps for every column, for the forward-pass as well as for the backward-pass.

## What is the Large Parsimony Problem?



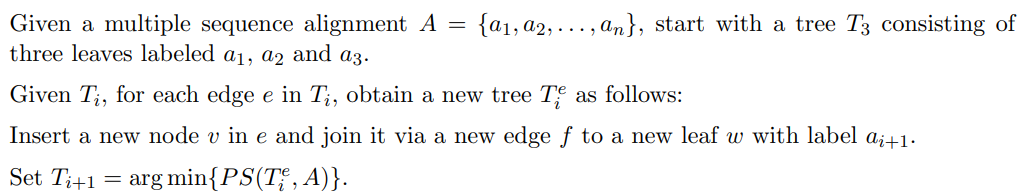
## What is the problem with the Large Parsimony Problem?

It is NP-hard. You must compute the parsimony score for every unrooted tree. Therefore, we need a heuristic.

## What is the process of the heuristic for choosing a parsimony tree?

We start with three nodes, choose the one with the best parsimony score and then add one node. We join it to the tree and compute the best parsimony score for all options again. Repeat until we have the desired number of nodes.

Formally:



## What is the difference between orthologous genes and paralogous genes?

Orthologous genes are genes with a common ancestor, diverged after a speciation event. Paralogous genes are genes that arose from a duplication event (multiple copies in the genome)

# Chapter 7 – Population Genetics

## Define the study objectives of population genetics.

In population genetics, the genetic composition of populations, including distributions and changes in genotype and phenotype frequency are studied in response to the processes of natural selection, gene drift, mutation, and gene flow.

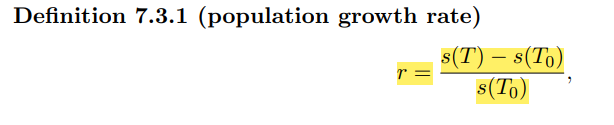
## Define single-nucleotide variant and single-nucleotide polymorphism.

SNV is a variation in a single nucleotide that occurs at a specific position in the genome. A SNV is an SNP if it occurs in a significant proportion of a population (typically > 1%).

## What is the difference between polymorphism and divergence?

Polymorphism describes sites that are variable within a species, divergence describe sites variable between species.

## Provide the formula for population growth rate



With s(T0) the size of the initial population at time T0 and s(T) the population size at time T.

## What are the three models for population growth?

Constant, polynomial, exponential.

## Define allele.

An allele is the variant form of a given gene found at the same chromosomal location.

## Define Ploidy.

Ploidy is the number of sets of chromosomes in a cell and hence the number of possible alleles for genes.

## What is the difference between haploid and diploid cells? How do they originate respectively?

Haploid cells have half the number of chromosomes (n) as diploid cells. They originate of meiosis, while diploid cells originate from mitosis.

## How does inheritance work in a haploid vs. diploid model?

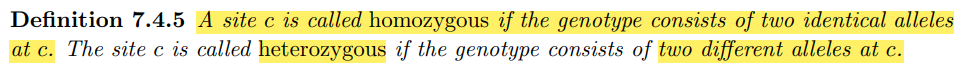
Haploid model: One parent gene from one parent

Diploid model: One parent gene per gene, an individual has two parents. A gene is chosen for each parent, with equal odds.

## Define haplotype.



## Define homozygous and heterozygous.



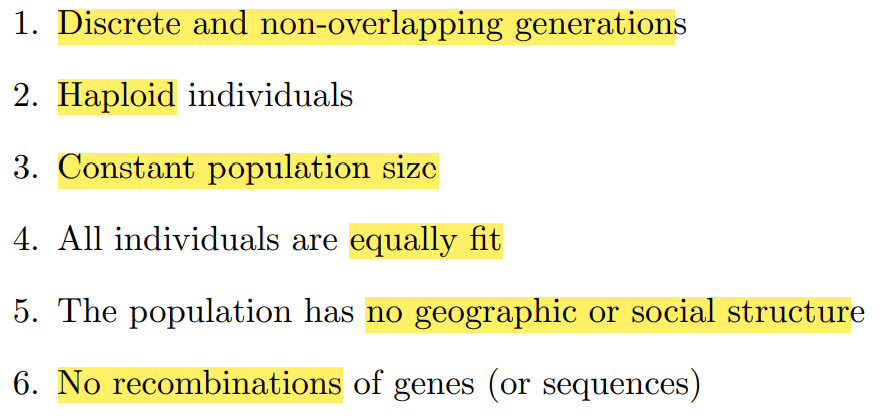
## Define a coalescent event.

If, traversing the sequence-transmission paths backward in time, two sequence-transmission paths intersect at some sequence, the paths coalesce at that intersection point. This is called a coalescent event.

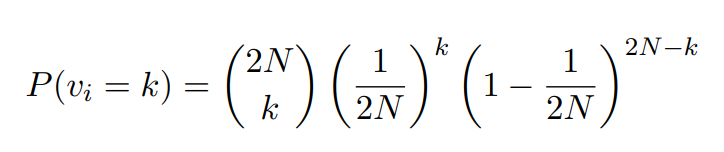
## Describe the coalescent process.

The goal of the process is to make inferences by moving backward in time. By randomly choosing parents in the previous generation, starting with the present-day generation, one constructs previous generations.

## What simplifying assumptions does the Wright-Fisher model make?



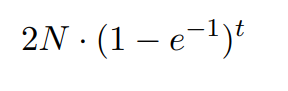
## What is the formula that a given gene has exactly k descendants, with a population size of 2N?



## Which underlying distribution can we use in the Wright-Fisher model? Why?

We can use the Poisson-distribution, because p=1/2N becomes very small and thus the binomial distribution approaches the Poisson distribution.

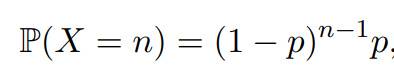
## What is the formula for the size of the gene pool at any generation t?



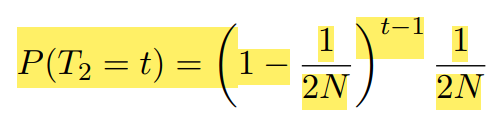
## How many coalescent events are there for 2N genes?

n-1 coalescent events

## What is the formula of the geometric distribution?



## What is the formula for the coalescence time for two genes, i. e. the probability that two lineages find a common ancestor t generations back?

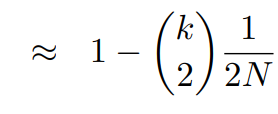


## For the coalescence time for k genes, what is one simplification that needs to be assumed?

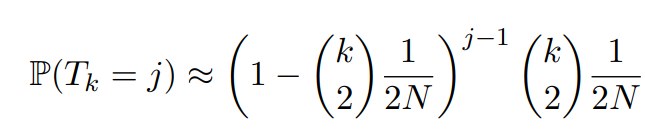
The probability of more than one simultaneous coalescent event is zero.

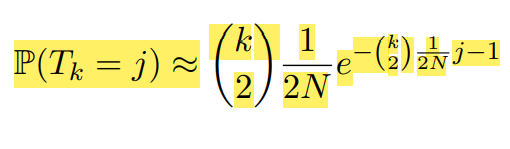
## What is the formula of the probability that k genes have k different parents? Under what conditions does the formula hold?

The formula holds if 2N is large and k is small. This is because the complicated, higher-order terms of 1/N can be discarded

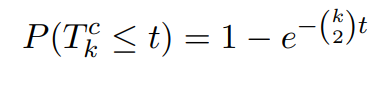


## Following from the previous formula, what is the probability for two genes out of the k genes to find a common ancestor j generations ago? What is the formula approximated by the exponential function?

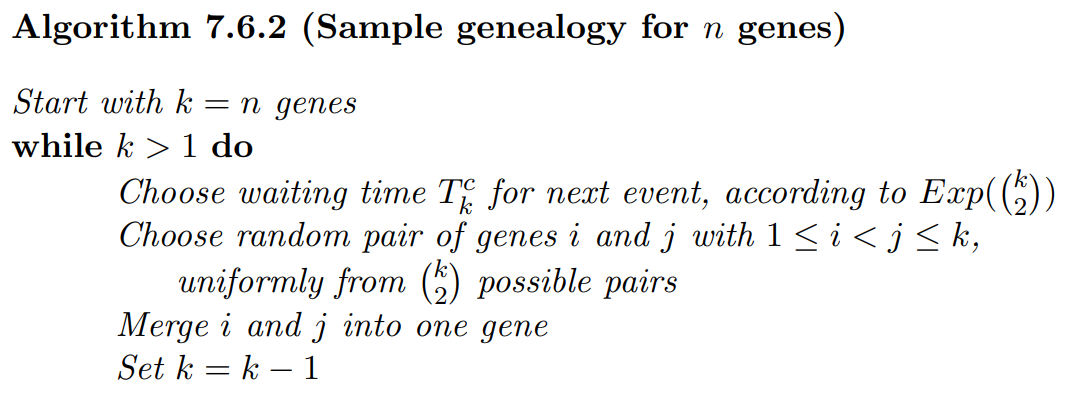




## When converting the time from distinct generations to continuous time, what is the continuous waiting time for k genes to have k-1 ancestors?



## What is the algorithm for a sample genealogy for n genes?



## What is the formula for the effective population size according to the Wright-Fisher model?

