[1 Chapter 2 - Algorithms 2](#_Toc105427073)

[1.1 Name 8 algorithm strategies 2](#_Toc105427074)

[1.2 Describe the branch and bound algorithm 2](#_Toc105427075)

[1.3 Describe the steps of the divide-and-conquer algorithm 3](#_Toc105427076)

[1.4 Describe the principle of dynamic programming 3](#_Toc105427077)

[1.5 Describe the pre-order tree-traversal. 3](#_Toc105427078)

[1.6 Describe the post-order tree-traversal 3](#_Toc105427079)

[1.7 Describe in-order tree-traversal 3](#_Toc105427080)

[1.8 Describe breadth-first traversal of a tree 3](#_Toc105427081)

[2 Chapter 3 – Pairwise Alignment 4](#_Toc105427082)

[2.1 Define symbols and alphabet 4](#_Toc105427083)

[2.2 Define concatenation 4](#_Toc105427084)

[2.3 Define substring 4](#_Toc105427085)

[2.4 Reverse complement 4](#_Toc105427086)

[2.5 What are the building blocks of a fasta-file? 4](#_Toc105427087)

[2.6 How is a dot matrix composed? 4](#_Toc105427088)

[2.7 How are common substrings (reversed, displaced, repeated) visible in the dot matrix? 4](#_Toc105427089)

[2.8 What parameters are used to reduce noise in a dot matrix? 4](#_Toc105427090)

[2.9 What runtime does the dot matrix have? 4](#_Toc105427091)

[2.10 Define pairwise global alignment. 4](#_Toc105427092)

[2.11 What two characters CAN’T be aligned in a pairwise alignment? 5](#_Toc105427093)

[2.12 What assumption is made with scoring such that the scoring scheme can be additive? 5](#_Toc105427094)

[2.13 What is the formula of the Minkowski metric? What are its first two instances called? 5](#_Toc105427095)

[2.14 What is the formula of the Hamming distance? What is an intuitive explanation for it? 5](#_Toc105427096)

[2.15 What is the definition of the Levenshtein (or edit) distance? 5](#_Toc105427097)

[2.16 What is the definition of an edict transcript? 5](#_Toc105427098)

[2.17 What are the three essential components of dynamic programming? 5](#_Toc105427099)

[2.18 What is a prefix? 5](#_Toc105427100)

[2.19 What is the distance matrix defined as? In which direction goes an insertion, in which a deletion? 5](#_Toc105427101)

[2.20 What is the formula for the score S of a sequence A? 6](#_Toc105427102)

[2.21 What is the definition of a substitution matrix? 6](#_Toc105427103)

[2.22 What is the formula for the conditional probability of A given B? 6](#_Toc105427104)

[2.23 What is the formula for the joint probability? 6](#_Toc105427105)

[2.24 What are the entries of a substitution matrix S representing? How are they computed? 6](#_Toc105427106)

[2.25 Kochrezept: How can you compute your own substitution matrix? 6](#_Toc105427107)

[2.26 Name two commonly used substitution matrices. 6](#_Toc105427108)

[2.27 What does BLOSUM stand for? What is the requirement if there is an integer behind the word? 6](#_Toc105427109)

[2.28 How are the entries in the BLOSUM matrix computed? 6](#_Toc105427110)

[2.29 What are some possible ways to classify amino acids? 7](#_Toc105427111)

[2.30 What is the formula for a function that assigns a constant gap penalty of gap length g? 7](#_Toc105427112)

[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? 7](#_Toc105427113)

[2.32 Define subsequence of X and Y. 7](#_Toc105427114)

[2.33 What is the formula for the number of possible alignments between X and Y? 7](#_Toc105427115)

[2.34 What is another way to write the formula for the number of possible alignments between X and Y? 7](#_Toc105427116)

[2.35 What does Stirling’s formula look like? For which n and m does it hold? 7](#_Toc105427117)

[2.36 Define optimal global alignment. 7](#_Toc105427118)

[2.37 What are the three cases for the Fitch algorithm (in words and in formula)? What other conditions are defined for its initialization? 7](#_Toc105427119)

[2.38 Reproduce the pseudocode for the Needleman-Wunsch algorithm. 8](#_Toc105427120)

[2.39 What is the complexity of the Needleman-Wunsch algorithm, and why? 8](#_Toc105427121)

[2.40 Where do we apply pairwise local alignment? 8](#_Toc105427122)

[2.41 Define local alignment. 8](#_Toc105427123)

[2.42 What is the difference between the Smith-Waterman algorithm to the Needleman-Wunsch-algorithm? What is the corresponding formula? 8](#_Toc105427124)

[2.43 Reproduce the pseudocode for the Smith-Waterman algorithm. 9](#_Toc105427125)

[2.44 What is the complexity of the Smith-Waterman algorithm? 9](#_Toc105427126)

[2.45 Under what condition does the Smith-Waterman algorithm work? Include the formula. 9](#_Toc105427127)

# 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

### 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. Divide a larger problem into smaller subproblems of the same type

### 2. Solve each subproblem recursively

### 3. Combine smaller solutions into a solution for the whole problem

### Divide-and-conquer is a top-down approach.

## Describe the principle of dynamic programming

### 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.

### It consists of four steps:

### 1. Characterisation of the structure of the optimal solution

### 2. Recursive definition of the value of an optimal solution

### 3. Computation of the optimum using recursion

### 4. Construction of an optimal solution through the computed optimal value

## Describe the pre-order tree-traversal.

### 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

### 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

### 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 symbols and alphabet

### 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

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

## Define substring

### 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 edict 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 a sequence 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

### 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.

## 