

Aufgabe 2) Proteinsequenzen:

Human Hemoglobin Subunit Alpha HBA1	Human Hemoglobin Subunit Beta HBB
MVLSPADKTNVKAAWGKVGGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQV KGGHGGKVVADALTNVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAA HLPAAFTPAVHASLDKFLASVSTVLTISKYR	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGN PKVKAHGKKVLGAFSDGLAHLDNLKGTFTATSELHCDKLHVDPENFRLLGNVLVCVLA HHFGKEFTPPVQAAYQKVVAGVANALAHKYH

Aufgabe 3) Unterschiede zwischen Global/Local Alignment

Globales Alignment	Lokales Alignment
<ul style="list-style-type: none"> – Needleman-Wunsch-Algorithmus für optimalen, globalen Alignment-Score – Alignment der gesamten Sequenz – möglichst ähnliche Sequenzen (Länge, Proteinfamilie, ...), da Score auf gesamte Sequenz maximiert werden muss – Komplexität: $O(\max(n,m)^3)$ – Varianten: einheitliche „gap-Kosten“, Free-Shift Alignment 	<ul style="list-style-type: none"> – Smith-Waterman-Algorithmus für optimalen, lokalen Alignment-Score – Alignment von Teilsequenzen beider Sequenzen – Einfüge- und Löschooperationen am Anfang und Ende können zugunsten der Optimierung ignoriert werden – für z. T. sehr unterschiedliche Sequenzen (Motive in verschiedenen Proteinen) – Score der Teilsequenzen muss maximiert werden – Komplexität: $O(nm)$ – Modifikationen des Needleman-Wunsch-Algorithmus: Initialisierung erste Zeile, erste Spalte mit 0 Maximierung über vierten Fall: 0 – backtracking vom Matrixeintrag mit größtem Wert bis 0-Eintrag

Aufgabe 4) Alignments

(1) Globales Alignment (default)

BLOSUM 62

```
#####
# Program: needle
# Rundate: Tue 3 Jul 2018 10:42:35
# Commandline: needle
# -auto
# -stdout
# -asequence emboss_needle-I20180703-104231-0157-25232106-p1m.asequence
# -bsequence emboss_needle-I20180703-104231-0157-25232106-p1m.bsequence
# -datafile EBLOSUM62
# -gapopen 10.0
# -gapextend 0.5
# -endopen 10.0
# -endextend 0.5
# -aformat3 pair
# -sprotein1
# -sprotein2
# Align_format: pair
# Report_file: stdout
#####

#
# Aligned_sequences: 2
# 1: HBA
# 2: HBB
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 292.5
#
#
#####

HBA      1  MV-LSPADKTNVKAANGKVGAHAGEYGAELERMFLSFPTTKTYFPHF-D      48
          |||.:.|.:.|.|.|||.:.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
HBB      1  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

HBA      49  LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR      93
          |||.:.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
HBB      49  LSTPDAMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTLSELHCDKLH      98

HBA      94  VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      142
          |||.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
HBB      99  VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH      147
```

- Vergleich Folie 11: gaps treten auch hier eher gesammelt auf, auf Folie mehr gaps, weil viel mehr Sequenzen zum Alignment herangezogen werden
- BLOSUM62-Matrix: BLOcks SUBstitution Matrix 62
- Matrix enthält einen Score/Wert für jedes Aminosäurepaar
- beim Vergleich wird eine Gewichtung vorgenommen, welcher Aminosäureaustausch vertretbar ist, wenn kein match zustande kommt
- match bringt höheren Score als nicht-match
- manche Abzüge sind geringer, weil eine Mutation im Tripletcode ausreicht, um bestimmte Aminosäuren ineinander zu überführen (Mutation häufiger als durch Zufall zu erwarten)
- alternative Matrix : PAM (Point Accepted Mutation)

(2) Globales Alignment
(Substitution MATRIX change)

BLOSUM 90

```
#####
# Program: needle
# Rundate: Tue  3 Jul 2018 10:52:03
# Commandline: needle
#   -auto
#   -stdout
#   -asequence emboss_needle-I20180703-105202-0621-57105068-p1m.asequence
#   -bsequence emboss_needle-I20180703-105202-0621-57105068-p1m.bsequence
#   -datafile EBLOSUM90
#   -gapopen 10.0
#   -gapextend 0.5
#   -endopen 10.0
#   -endextend 0.5
#   -aformat3 pair
#   -sprotein1
#   -sprotein2
# Align_format: pair
# Report_file: stdout
#####

#=====
#
# Aligned_sequences: 2
# 1: HBA
# 2: HBB
# Matrix: EBLOSUM90
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    83/149 (55.7%)
# Gaps:          9/149 ( 6.0%)
# Score: 311.5
#
#
#=====

HBA      1  MV-LSPADKTNVKAANGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-D    48
|| :|.:|.:|.|.||||| ...|.|.|||||:|.:|.:|.:|.|.|.
HBB      1  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD    48

HBA      49  LSH-----GSAQVKGHGKKVADALTNAAHVDDMPNALSA LSDLHAHKLR    93
||.      |.:|.|.|||||.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
HBB      49  LSTPDAMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH    98

HBA      94  VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
|||.||:|.|.:|.:|.:|.:|.:|.:|.:|.:|.:|.:|.:|.:|.:|
HBB      99  VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

- Vergleich Folie 11: viele uneindeutige Zuordnungen (ein Punkt), weniger gaps als auf Folie, siehe (1)
- similarity steigt, wenn BLOSUM-Zahl sinkt
- BLOSUM 30: 102/149 (68,5%)
- BLOSUM 90: 83/149 (55,7%)
- hohe Zahl für evolutionär nah verwandte Proteine
- Score der Aminosäurepaare verändert sich
- hohe Zahl: Aufbau der Matrix so konstruiert, dass nicht-matches mit immer stärkeren Abzügen beim Score geahndet werden

- Vergleich Folie 11: viel mehr gaps als auf der Folie, teils auch vereinzelt auftretend, hier weniger uneindeutige „matches“ (durch gaps vermieden)
- gap open penalty beschreibt die Toleranz des Algorithmus gegenüber dem Einfügen von gaps an Stellen, wo kein match zustande kommt, wenn darauf aber ein match folgen würde
- setzt man den Wert herunter, so neigt das System dazu, zugunsten der similarity viele gaps zu setzen
- gap open 10: 9 gaps
gap open 1: 41 gaps

- Vergleich Folie 11: beginnt nicht mit (M)V wie vorherige Alignments und auf Folie 11, endet auf KY statt KYR/H (gekürzt)
- LALIGN liefert separat viele verschiedene alignierte Teilsequenzen unterschiedlicher Länge (zu viele für Screenshot)
- siehe Aufgabe 3)