## BIOINF101 - Aufgabe 6

## Aufgabe 2) Proteinsequenzen:

Human Hemoglobin Subunit Alpha HBA1	Human Hemoglobin Subunit Beta HBB
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQV	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAA	PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLA
HLPAEFTPAVHASLDKFLASVSTVLTSKYR	HHFGKEFTPPVQAAYQKVVAGVANALAHKYH

## Aufgabe 3) Unterschiede zwischen Global/Local Alignment

Globales Alignment	Lokales Alignment
<ul> <li>Needleman-Wunsch-Algorithmus für optimalen, globalen Alignment-Score</li> <li>Alignment der gesamten Sequenz</li> <li>möglichst ähnliche Sequenzen (Länge, Proteinfamilie,), da Score auf gesamte Sequenz maximiert werden muss</li> <li>Komplexität: O(max(n,m)³)</li> <li>Varianten: einheitliche "gap-Kosten", Free-Shift Alignment</li> </ul>	<ul> <li>Smith-Waterman-Algorithmus für optimalen, lokalen Alignment-Score</li> <li>Alignment von Teilsequenzen beider Sequenzen</li> <li>Einfüge- und Löschoperationen am Anfang und Ende können zugunsten der Optimierung ignoriert werden</li> <li>für z. T. sehr unterschiedliche Sequenzen (Motive in verschiedenen Proteinen)</li> <li>Score der Teilsequenzen muss maximiert werden</li> <li>Komplexität: O(nm)</li> <li>Modifikationen des Needleman-Wunsch-Algorithmus: Initialisierung erste Zeile, erste Spalte mit 0</li> </ul>
	Maximierung über vierten Fall: 0
	<ul> <li>backtracking vom Matrixeintrag mit größtem Wert bis 0- Eintrag</li> </ul>

## Aufgabe 4) Alignments

\_\_\_\_\_\_ # Program: needle # Rundate: Tue 3 Jul 2018 10:42:35 Vergleich Folie 11: gaps treten # Commandline: needle -auto auch hier eher gesammelt auf, auf -stdout -asequence emboss needle-I20180703-104231-0157-25232106-plm.asequence Folie mehr gaps, weil viel mehr -bsequence emboss needle-I20180703-104231-0157-25232106-p1m.bsequence Sequenzen zum Alignment -datafile EBLOSUM62 -gapopen 10.0 herangezogen werden -gapextend 0.5 -endopen 10.0 BLOSUM62-Matrix: BLOcks -endextend 0.5 -aformat3 pair SUbstitution Matrix 62 -sprotein1 Matrix enthält einen Score/Wert -sprotein2 # Align format: pair für jedes Aminosäurepaar # Report file: stdout beim Vergleich wird eine Gewichtung vorgenommen, welcher Aminosäureaustausch # Aligned\_sequences: 2 (1) Globales Alignment (default) # 1: HBA vertretbar ist, wenn kein match # 2: HBB # Matrix: EBLOSUM62 zustande kommt **BLOSUM 62** # Gap penalty: 10.0 match bringt höheren Score als # Extend\_penalty: 0.5 nicht-match # Length: 149 # Identity: 65/149 (43.6%) manche Abzüge sind geringer, # Similarity: 90/149 (60.4%) # Gaps: 9/149 ( 6.0%) weil eine Mutation im Triplettcode # Score: 292.5 ausreicht, um bestimmte Aminosäuren ineinander zu überführen (Mutation häufiger als HBA 1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D || |:|.:|:.|.|.||| :..|.|.|||.|::::|.|:::|.| durch Zufall zu erwarten) HBB 1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTORFFESFGD alternative Matrix: PAM (Point HBA 49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR Accepted Mutation) HBB 49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH HBA 94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH HBB

(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-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D
                                                                48
                  HBB
                1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTORFFESFGD
                                                                48
               49 LSH-----GSAOVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR
                       _ [..:||.||||..|.:..||.|::....:.||:||..||.
               49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH
HBB
                                                                98
HBA
               94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
                  |||-||:||--|--||-|--||||-|:|:--|--|-|
HBB
               99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

- 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

# Program: needle # Rundate: Tue 3 Jul 2018 11:02:45 # Commandline: needle -auto -stdout -asequence emboss\_needle-I20180703-110242-0211-3562419-p2m.asequence -bsequence emboss\_needle-I20180703-110242-0211-3562419-p2m.bsequence -datafile EBLOSUM62 -gapopen 1.0 -gapextend 0.5 -endopen 10.0 -endextend 0.5 -aformat3 pair -sprotein1 -sprotein2 # Align\_format: pair # Report file: stdout # Aligned\_sequences: 2 # 1: HBA (3) Globales Alignment # 2: HBB # Matrix: EBLOSUM62 (GAP OPEN penalty change) # Gap penalty: 1.0 # Extend\_penalty: 0.5 **GAP OPEN 1.0** # Length: 165 # Identity: 72/165 (43.6%) 90/165 (54.5%) # Similarity: # Gaps: 41/165 (24.8%) # Score: 347.0 #-----HBA 1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLS--FP-TTKTYFPH HBB 1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRL-L-VVYPWTOR-FFES HBA 47 F-DLSH-----GSAQVKGHGKKV--A--DALTNAVAHVDDMPNAL--S-1:::11.11111 | 1.1 | 11:1 | 1.1 : 46 FGDLS-TPDAVMGNPKVKAHGKKVLGAFSDGL----AHLD---N-LKGTF HBB 83 A-LSDLHAHKLRVDPVNFKLLSH---CLLVTLAAHLPA-EFTPAVHASLD HBA 1 ||:||..||.||.||:||.: | | | ||.|. . ||||.|.|:..

HBB

HBB

- 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 gapsgap open 1: 41 gaps

127

87 ATLSELHCDKLHVDPENFRLLGNVLVC--V-LAHHF-GKEFTPPVQAAYQ

147

128 KFLASVSTVLTSKYR

133 KVVAGVANALAHKYH

\* # Program: matcher # Rundate: Tue 3 Jul 2018 10:57:57 # Commandline: matcher -auto -stdout -asequence emboss\_matcher-I20180703-105753-0556-63275701-p1m.asequence -bsequence emboss matcher-I20180703-105753-0556-63275701-plm.bsequence -datafile EBLOSUM62 -gapopen 14 -gapextend 4 -alternatives 1 -aformat3 pair -sprotein1 -sprotein2 # Align\_format: pair # Report file: stdout Vergleich Folie 11: beginnt nicht mit (M)V wie vorherige Alignments und auf Folie 11, endet auf KY # Aligned\_sequences: 2 (4) Lokales Alignment (default) statt KYR/H (gekürzt) # 1: HBA HUMAN # 2: HBB HUMAN LALIGN liefert separat viele # Matrix: EBLOSUM62 # Gap penalty: 14 Matcher (LALIGN) verschiedene alignierte # Extend penalty: 4 Teilsequenzen unterschiedlicher # Length: 145 Länge (zu viele für Screenshot) # Identity: 63/145 (43.4%) # Similarity: 88/145 (60.7%) siehe Aufgabe 3) 8/145 ( 5.5%) # Gaps: # Score: 264 #-----3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH HBA HUMAN HBB HUMAN 4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST HBA\_HUMAN 52 ----GSAOVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDP HBB\_HUMAN 52 PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP 101 HBA HUMAN 97 VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY HBB HUMAN 102 ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY