

## 2. Proteinsequenzen

HBA\_HUMAN:

```

      10      20      30      40      50
MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS
      60      70      80      90     100
HGSAQVKGHG KKVADALTNA VAHVDDMPNA LSALSDLHAH KLRVDPVNFK
      110     120     130     140
LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVSTVLTSK YR
    
```

HBB\_HUMAN:

```

      10      20      30      40      50
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS
      60      70      80      90     100
TPDAVMGNPK VKAHGKKVLG AFSDGLAHLN NLKGTFTATLS ELHCDKLHVD
      110     120     130     140
PENFRLLGNV LVCVLAHHFG KEFTPPVQAA YQKVVAGVAN ALAHKYH
    
```

## 3. Arten von Alignment

Globales Alignment:

Erzeugt ein end-to-end alignment der zu vergleichenden Sequenzen. Dass heißt es wird die Übereinstimmung der beiden Sequenzen über die vollständige Länge der Sequenzen geprüft. Ein globales Alignment macht Sinn wenn die Sequenzen ähnlich und ungefähr gleich lang sind.

```

Zielsequenz: 5'ACTACTAGATTACTTACGGATCAGGTACTCCAGAGGCTTGCAACCG 3'
              ||||| ||||| ||||| ||||| |||||
Suchsequenz:  5'ACTACTAGATT---ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
    
```

Lokales Alignment:

Findet eine oder mehrere Übereinstimmungen in den Sequenzen, welche die ähnlichsten Sequenzteile beschreiben. Ein globales Alignment bietet sich bei Sequenzen an die größere Unterschiede zueinander aufweisen und wo erwartet wird, dass sie übereinstimmende Regionen oder ähnlich Sequenzmotife innerhalb der Gesamtsequenz aufweisen.

```

Zielsequenz: 5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
              ||||| ||||| ||||| ||||| |||||
Suchsequenz:          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
    
```

(Sequenzen von <https://www.majordifferences.com/2016/05/difference-between-global-and-local.html> mit leichter Modifizierung)

## 4. Alignments

### (1) default Global Alignment

```
#=====
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# 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
#
#
#=====

EMBOSS_001      1  MV-LSPADKTNVKAAWGKVGAGAHAGEYGAEALERMFLSFPTTKTYFPHF-D      48
                  || |:|.:|.:|.|.||||  :..|.|.||||.:|.:|.:|.|.|.|.
EMBOSS_001      1  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

EMBOSS_001     49  LS-----HGSAQVKGHGKKVADALTNAAVAHVDDMPNALSALSDLHAHKLR      93
                  || .|.:|.|.|||||.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001     49  LSTPDVAVMGNPVKVKAHGKKVLGAFSDGLAHLNLLKGTFTATLSELHCDKLH      98

EMBOSS_001     94  VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      142
                  |||.||:|.|.:|.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001     99  VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH      147

#-----
#-----
```

### (2) andere Substitutionsmatrix: PAM30

```
#####
# Program: needle
# Rundate: Tue 10 Jul 2018 06:34:16
# Commandline: needle
#   -auto
#   -stdout
#   -asequence emboss_needle-I20180710-063414-0691-13292990-plm.asequence
#   -bsequence emboss_needle-I20180710-063414-0691-13292990-plm.bsequence
#   -datafile EPAM30
#   -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: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EPAM30
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    74/149 (49.7%)
# Gaps:          9/149 ( 6.0%)
# Score: 236.5
#
#
#=====

EMBOSS_001      1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D      48
      || |.|.:|..|.|.|||| |...|.|.||||.|:....|.|.:.|..| |
EMBOSS_001      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

EMBOSS_001     49 LSH-----GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR      93
      ||.      |...|.|||||.|.:.|..|.|.:.|..|.|.:.|..|.|.
EMBOSS_001     49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLNLDNLKGTFTATLSELHCDKLH      98

EMBOSS_001     94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      142
      |||.|||.||...|...|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001     99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH      147

#-----
#-----

```

### (3) Gap open penalty = 1

```

#####
# Program: needle
# Rundate: Tue 10 Jul 2018 20:23:44
# Commandline: needle
# -auto
# -stdout
# -asequence emboss_needle-I20180710-202342-0933-10021742-plm.asequence
# -bsequence emboss_needle-I20180710-202342-0933-10021742-plm.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: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 1.0

```

```

# Extend_penalty: 0.5
#
# Length: 165
# Identity:      72/165 (43.6%)
# Similarity:    90/165 (54.5%)
# Gaps:          41/165 (24.8%)
# Score: 347.0
#
#
#=====
EMBOSS_001      1 MV-LSPADKTNVKAAGKVGAGAHAGEYGAEALERMFSL--FP-TTKTYFPH      46
      || |:|.:|.:|.|.||| | :..|.|.|||. |: | :| |.: :|..
EMBOSS_001      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRL-L-VVYPWTQR-FFES      45

EMBOSS_001     47 F-DLSH-----GSAQVKGHGKKV--A--DALTNAVAHVDDMPNAL--S-      82
      | ||| |.:|.|.|||| | |. | |:| | | :
EMBOSS_001     46 FGDLS-TPDAVMGNPKVKAHGKKVLGAFSDGL----AHLD---N-LKGTFF      86

EMBOSS_001     83 A-LSDLHAHKLRVDPVNFKLLSH---CLLVTLAAHLPA-EFTPAVHASLD      127
      | ||:|.|.||||.|||.||:|. | | |||. . ||||.|.|.:..
EMBOSS_001     87 ATLSSELHCDKLHVDPENFRLLGNVLVC--V-LAHHF-GKEFTPPVQAAYQ      132

EMBOSS_001    128 KFLASVSTVLTSKYR      142
      |.:|.|.:..|..|||.
EMBOSS_001    133 KVVAGVANALAHKYH      147

```

```

#-----
#-----

```

#### (4) default Local Alignment

```

#####
# Program: water
# Rundate: Tue 10 Jul 2018 06:38:40
# Commandline: water
#   -auto
#   -stdout
#   -asequence emboss_water-I20180710-063838-0410-76045746-plm.asequence
#   -bsequence emboss_water-I20180710-063838-0410-76045746-plm.bsequence
#   -datafile EBLOSUM62
#   -gapopen 10.0
#   -gapextend 0.5
#   -aformat3 pair
#   -sprtein1
#   -sprtein2
# Align_format: pair
# Report_file: stdout
#####

#=====
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 145
# Identity:      63/145 (43.4%)
# Similarity:    88/145 (60.7%)
# Gaps:          8/145 ( 5.5%)

```

```

# Score: 293.5
#
#
#=====
EMBOSS_001      3  LSPADKTNVKAAGKVGGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-      50
    |:|.:|:|.|.|.||||  :..|.|.|||.|:~.:|.|:~.:|.|.  |||
EMBOSS_001      4  LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST      51
EMBOSS_001     51  ----HGSAQVKGHGKKVADALTNVAHVDDMPNALSALSDLHAHKLRVDP      96
    .|:~.:|.|.|||||.|.~.:|:|:~.:~.:|.|:|~.:|.|.|||
EMBOSS_001     52  PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTLSELHCDKLHVDP     101
EMBOSS_001     97  VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY      141
    .||:|.|:~.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001    102  ENFRLLGNVLVLCVLAHFFGKEFTPPVQAAYQKVVAGVANALAHKY      146

#-----
#-----

```

#### Vergleich mit Alignment aus Vorlesung:

Bei dem Alignment in der Vorlesung ist weder die verwendete Matrix noch irgendwelche Parameter (z.B. Gap Penalties) angegeben. Dies macht es nicht sinnvoll die Scores zu vergleichen. Der Vergleich in der Vorlesung ist zwischen 6 verschiedenen Sequenzen wohingegen wir hier nur 2 verschiedene vergleichen. Bei dem Vergleich in der Vorlesung scheint es sich um Globales Alignment von zwei verschiedenen Bereichen der Sequenz eines Major prion proteins zu handeln, es könnten aber auch nur die übereinstimmenden Sequenzen eines Local Alignments sein

(b)

(1) BLOSUM62 Matrix: Eine evidenzbasierende Substitutionsmatrix, die für Sequenzanalysen von Proteinen benutzt wird. Sie wurde 1992 von Jorja G. Henikoff und Steven Henikoff entwickelt. Die Matrix wurde aus einer Database auf bestätigten Alignments erzeugt. Die 62 gibt den Schwellenwert an der für die Erzeugung der Matrix mithilfe der Database verwendet wurde. So wurden für die BLOSUM62 Matrix nur Alignments aus der Database genommen die weniger als 62% identisch waren. Dieser Schwellenwert erlaubt die Anpassung der Matrix an die jeweiligen Umstände. So gibt der 80% Schwellenwert der BLOSUM80 höher konservierte Zielfrequenzen und ein Schwellenwert von 45% gibt die mehr divergierende BLOSUM45 matrix

Gap OPEN penalty = 10 (default) (Score -10 für eröffnen einer Gap)

(2) PAM30 Matrix: Die PAM (Point Accepted Mutation) Matrix wurde in den 1970ern von Margaret Dayhoff. Als Basis für die PAMXX Matrixes dient die PAM1 Matrix welche angibt mit welcher Rate eine Substitution zu erwarten wäre, wenn sich 1% der Aminosäuren verändert hätten (entspricht einer Ähnlichkeit von 99%). Für das Erzeugen anderer Matrix wird angenommen das mehrfache Mutationen dem Muster der PAM1 Matrix folgen und das mehrere Substitutionen an der selben Stelle erfolgen können. Die bedeutet das PAM Matrixen über 100 hinausgehen können. Die PAM250 Matrix entspricht einer Übereinstimmung von ~20%.

(3) Gap OPEN Penalty: Um ein sinnvolles Alignment zu erhalten muss die Einbringung von Gaps limitiert werden. Dies geschieht mithilfe von Gap Penalties, welche die Einbringung von Gaps in Bezug auf ihre Nummer und Länge bestrafen. Die Gap OPEN Penalty ist ein Typ von Gap Penalty, die eine Strafe für das Initiieren einer Gap. Für die Erweiterung einer Gap gilt dann die Gap Extend Penalty

Gap OPEN penalty = 1 (Score -1 für eröffnen einer Gap, erzeugen von Gaps wird weniger schwer bestraft als mit default Einstellungen -> % Gaps von 6 zu 24.8)

(4) Siehe (1)