

# Bericht zur Alignierung zweier Protein-Sequenzen

Johanna Böttger

## 2) Human Hemoglobin subunit alpha (HBA\_HUMAN)

10	20	30	40	50
MVLSPADKTN	VKAAWGKVGA	HAGEYGAEAL	ERMFLSFPTT	KTYFP HFDLS
60	70	80	90	100
HGSAQVKGHG	KKVADALTNA	VAHVDDMPNA	LSALS DLHAH	KLRVDPVNFK
110	120	130	140	
LLSHCLLVTL	AAHLPAEFTP	AVHASLDKFL	ASVSTVLTSK	YR

## Human Hemoglobin subunit beta (HBB\_HUMAN)

10	20	30	40	50
MVHLTPEEKS	AVTALWGKVN	VDEVGGEALG	RLLVVYPWTQ	RFFESFGDLS
60	70	80	90	100
TPDAVMGNPK	VKAHGKKVLG	AFSDGLAHL D	NLKGTFATLS	ELHCDKLHVD
110	120	130	140	
PENFRLLGNV	LVCVLAHHFG	KEFTPPVQAA	YQKV VAGVAN	ALAHKYH

3)

### Globales Alignment

Beide Sequenzen werden komplett betrachtet, sodass beide in etwa gleich lang sein sollten → nur sinnvoll, wenn beide Sequenzen einander über ihre gesamte Länge hinweg ähnlich

### Lokales Alignment

Suche der am besten übereinstimmenden Teilsequenzen → sinnvoll, wenn die Sequenzen nicht über ihre ganze Länge homolog (evolutionär verwandt) sind, sondern nur in Teilen

### (1) Globales Alignment mit voreingestellten Parametern

```
#####
# Program: needle
# Rundate: Tue 10 Jul 2018 21:01:35
# Commandline: needle
# -auto
# -stdout
# -asequence emboss_needle-I20180710-210134-0282-50243963-p1m.asequence
# -bsequence emboss_needle-I20180710-210134-0282-50243963-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: 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-LSPADKTNVKAAWGKVGGAHAGEYGAEALERMFSLFPTTKTYFPHF-D      48
                  || |:|:|:|.|.|.||| | :..|.|.|||.|:|:|:|.|:|:|.|. |
EMBOSS_001      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD      48

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

EMBOSS_001      94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      142
                  |||.||:|:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
EMBOSS_001      99 VDPENFRLLGNVLVLCVLAHHFGKEFTPFVQAAYQKVVAGVANALAHKYH      147
```

## **(2) Globales Alignment mit Pam80**

```
#####
# Program: needle
# Rundate: Tue 10 Jul 2018 21:06:34
# Commandline: needle
#   -auto
#   -stdout
# -asequence emboss_needle-I20180710-210632-0909-64082032-plm.asequence
# -bsequence emboss_needle-I20180710-210632-0909-64082032-plm.bsequence
# -datafile EPAM80
# -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: EPAM80
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    87/149 (58.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 318.5
#
#
#=====

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

EMBOSS_001     49 LSH-----GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDDLHAKLR      93
                  ||.      |:.||.|||||.|.:.:.|||.|:..:..:|:|:|..||.
EMBOSS_001     49 LSTPDVAMGNPKVKAHGKKVLGAFSDGLAHLNLTGTATLSELHCDKLH        98

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

### **(3) Globales Alignment mit GAP OPEN penalty = 50**

```
#####
# Program: needle
# Rundate: Tue 10 Jul 2018 21:11:36
# Commandline: needle
#   -auto
#   -stdout
#   -asequence emboss_needle-I20180710-211134-0982-84718056-p2m.asequence
#   -bsequence emboss_needle-I20180710-211134-0982-84718056-p2m.bsequence
#   -datafile EBLOSUM62
#   -gapopen 50.0
#   -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: 50.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      61/149 (40.9%)
# Similarity:    87/149 (58.4%)
# Gaps:          9/149 ( 6.0%)
# Score: 210.0
#
#
#=====

EMBOSS_001      1 -MVLSPADKTNVKAANGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHF-- 47
                  :.:|.:|.:|.:|.:|  :..|.|.|||.|:..:|.|:..:|.|.
EMBOSS_001      1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD 48

EMBOSS_001     48 ----DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR 93
                  |...|.:.:|.|.|||||.|.:.:..:|:|:|:..:..:|:|:|..|.
EMBOSS_001     49 LSTPDAMGNPKVKAHGKKVLGAFSDGLAHLNLTGTFTLSELHCDKLH 98

EMBOSS_001     94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR 142
                  |||.||:|.|.:.:|..|.|.|||.|.|.:.:|.|:..:|.|.|.
EMBOSS_001     99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
```

#### (4) Lokales Alignment (Water) mit voreingestellten Parametern

```
#####  
# Program: water  
# Rundate: Tue 10 Jul 2018 21:18:55  
# Commandline: water  
# -auto  
# -stdout  
# -asequence emboss_water-I20180710-211852-0917-19526312-plm.asequence  
# -bsequence emboss_water-I20180710-211852-0917-19526312-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	LSPADKTNVKAANGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-	50
		: .: .: . .      :.. . .   . :~::~ . :~::~ .~.	
EMBOSS_001	4	LTPEEKSAVTALWGKKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST	51
EMBOSS_001	51	----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDP	96
		. :~::~ . .     .~::~ :~::~ :~::~ :~::~ :~::~ :~::~ :~::~	
EMBOSS_001	52	PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP	101
EMBOSS_001	97	VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY	141
		. :~::~ .~::~ :~::~ .~::~ .~::~ :~::~ :~::~ :~::~ :~::~ :~::~	
EMBOSS_001	102	ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY	146

## Multiple sequence alignment (Globin family)

```
Helix      AAAAAAAAAAAAAAAAAA      BBBBBBBBBBBBBBBBBBCCCCCCCCCCCC
HBA_HUMAN  -----VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  -----VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  -----VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP -----LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTIRSAAWAPVYS--TYETSGVDILVKFFTSTPAAQEFPFKF
LGB2_LUPLU -----GALTESQAALVKSSWEEFN--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI -----GLSAAQRQVIAATWKDIAGADNGAGVGKDKLIKFLSAHPQMAAVFG-F
Consensus  Ls.... v a W kv . . g . L.. f . P . F F

Helix      DDDDDDDDEEEEEEEEEEEEEEEEEEEEEEE      FFFFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNVAHV--D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL--D--NLKGTFFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK---K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVS KG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVVRHKGYGN
Consensus  . t . . . v..Hg kv. a a...l d . a l. l H .

Helix      FFGGGGGGGGGGGGGGGGGGGGG      HHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLP AEFTPAVHASLDKFLASVSTVLTSKYR-----
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVAVAGVANALAHKYH-----
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVS YMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-----
GLB5_PETMA -QVDPQYFKVLA AVIADTV AAG-----DAGFEKLMSMICILLRSAY-----
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAKDAWAAAYADISGALISGLQS-----
Consensus  v. f l . . . . . f . aa. k . . l sky
```

Das Alignment aus der Vorlesung betrachtet mehrere Sequenzen der Globin-Familie, sodass, um ein stimmiges Gesamtbild zu erhalten, anders aligniert wurde. Beispielsweise ist aufgrund der Sequenz von GLB5\_PETMA notwendig bei allen anderen Sequenzen Lücken zu Beginn einzufügen.

### Erläuterung der verwendeten Parameter

#### PAM80-Matrix

= Point Accepted Mutation Matrix

Substitutionsmatrix, die anhand statistischer Sequenzunterschiede die Wahrscheinlichkeiten einer Veränderung einer Sequenz ausgibt. Die Zahl gibt die Wahrscheinlichkeit der Umwandlung einer Aminosäure in die andere an (80 % Wahrscheinlichkeit der Umwandlung, 20 % Wahrscheinlichkeit der Nicht-Umwandlung).

### BLOSUM62-Matrix

= **B**LOcks **S**Ubstitution **M**atrix

Substitutionsmatrix, die mittels lokalem Alignment, Scores zwischen evolutionär divergierenden Proteinsequenzen. Sie wird bei hochkonservierten Regionen von Proteinfamilien verwendet. Die Zahl gibt, dass die Sequenzen, die benutzt wurden um diese Matrix zu erstellen zu ca. 62 % identisch waren. Allen BLOSUM-Matrices ist gemeinsam, dass sie auf beobachteten Alignments basieren und nicht wie PAM-Matrices aus Vergleichen nahe verwandter Proteine extrapolarisiert werden.

### Gap Open penalty

Gibt den Wert der Bestrafung für das Einfügen einer einzelnen Lücke in die alignnten Sequenzen an. Wenn der Wert hoch ist, ist die Bestrafung bezüglich des finalen Scores stärker.

### Gap Extend penalty

Gibt den Wert der Bestrafung für das Einfügen mehrerer aufeinander folgender Lücken in die alignnten Sequenzen an.

Je höher der Score, desto besser das Alignment. Zur Berechnung des Scores werden Gaps mit negativen Werten gewichtet, wohingegen Matches positive Werte erhalten. Missmatches verrechnen sich mit größeren negativen Zahlen.