Bioinformatik Assignment 6- Shelly Harel

2) *Proteinsequenz von „Human Hemoglobin subunit alpha “:*
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

Proteinsequenz von „Human Hemoglobin subunit beta “:
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH

3) *Zentrale Unterschiede zwischen globalen und lokalen Alignment zweier Sequenzen:*
Beim globalen Alignment werden alle Zeichen beider Strings miteinander verglichen. Sie werden also verwendet, wenn die Sequenzen ähnlich lang sind und starke Übereinstimmungen erwartet werden.

Beim lokalen Alignment hingegen wird nur ein Teil der ersten Sequenz mit einem Teil der zweiten Sequenz verglichen. Hier können die Strings unterschiedlich lang sein.

4) (a) *(1) Globales Alignment mit voreingestellten Parametern: (BLOSUM62)*

```
#    -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-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
EMBOSS_001         1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

EMBOSS_001        49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
EMBOSS_001        49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

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

(2) Globales Alignment mit einer anderen Substitution MATRIX: (PAM10)

```
#    -datafile EPAM10
#    -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: EPAM10
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 203
# Identity:      61/203 (30.0%)
# Similarity:    61/203 (30.0%)
# Gaps:         117/203 (57.6%)
# Score: 136.0
#
#
#=======================================

EMBOSS_001         1 MV-LSPADKTNVKAAWGKV------GAHAGEYGAEALERM---------F     34
                     || |.|..|..|.|.||||      |        |||.|.         |
EMBOSS_001         1 MVHLTPEEKSAVTALWGKVNVDEVGG--------EALGRLLVVYPWTQRF     42

EMBOSS_001        35 LSFPTTKTYFPHF----DLSHGSAQ----------VKGHGKKV--A--DA     66
                             |    |||               ||.|||||  |  |.
EMBOSS_001        43 ------------FESFGDLS-----TPDAVMGNPKVKAHGKKVLGAFSDG     75

EMBOSS_001        67 LTNAVAHVDDMPN-----ALSALSDLHAHKLRVDPVNFKLLSH---CLLV    108
                        |    ||.|   |      |  .||.||..||.|||.||.||..   |.|
EMBOSS_001        76 L----AHLD---NLKGTFA--TLSELHCDKLHVDPENFRLLGNVLVCVL-    115

EMBOSS_001       109 TLAAHLPA----EFTPAVHASLDKFLASVSTVLTSKYR------------    142
                        ||       ||||.|.|      |          |.
EMBOSS_001       116 ---AH---HFGKEFTPPVQA------A---------YQKVVAGVANALAH    144

EMBOSS_001       143 ---    142

EMBOSS_001       145 KYH    147
```

(3) Globales Alignment mit einer anderen GAP OPEN penalty: (BLOSUM62 mit veränderter GAP OPEN)

```
#    -datafile EBLOSUM62
#    -gapopen 100.0
#    -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: 100.0
# Extend_penalty: 0.5
#
# Length: 147
# Identity:      44/147 (29.9%)
# Similarity:    62/147 (42.2%)
# Gaps:           5/147 ( 3.4%)
# Score: 146.0
#
#
#=======================================

EMBOSS_001          1 -----MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFP     45
                            .....|.....................:......|..:.....
EMBOSS_001          1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS     50

EMBOSS_001         46 HFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD     95
                      ..|...|:.:||.|||||..|.:::.:||:|::....:.||:||..||.||
EMBOSS_001         51 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    100

EMBOSS_001         96 PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                      |.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
EMBOSS_001        101 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

(4) Lokales Alignment mit voreingestellten Parametern: (BLOSUM62)

```
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 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: 145
# Identity:      63/145 (43.4%)
# Similarity:    88/145 (60.7%)
# Gaps:           8/145 ( 5.5%)
# Score: 293.5
#
#
#=======================================

EMBOSS_001         3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-     50
                     |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |||
EMBOSS_001         4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST     51

EMBOSS_001        51 ----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDP     96
                         .|:.:||.|||||..|.::.:||:|::....:.||:||..||.|||
EMBOSS_001        52 PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP    101

EMBOSS_001        97 VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY    141
                     .||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||
EMBOSS_001       102 ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY    146
```

Vergleich zur Vorlesung:

```
HBA_HUMAN   ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN   --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
HBA_HUMAN   -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN   GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
HBA_HUMAN   -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN   -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
```

In der Vorlesung wurden die beiden Sequenzen nicht nur miteinander, sondern auch mit anderen Proteinsequenzen der Globinfamilie verglichen. Dies führte dazu, dass manche Bereiche also anders aliniert wurden, damit das Gesamtbild stimmt. Beispielsweise wurde in der Vorlesung eine Consensus Sequenz angegeben, die Leucin und Serin nach einander beinhaltet, weshalb die Valine hier verschoben sind und nicht matchen, wie im ersten Beispiel (1).

(b) In (1) wurde die BLOSUM62 Matrix verwendet. BLOSUM = Block substitution matrices.
Blocks sind multiple Sequenz Alignments ohne Gaps, die zur am meisten konservierten Region der involvierten Proteine korrespondieren. In jeden solchen multiplen Alignment wurden die Sequenzen, die ähnliche % Identität gezeigt haben in Gruppen geclustert und gemittelt. Mithilfe dieser Gruppen wurde die Substitutionshäufigkeit aller Aminosäuren Paare berechnet und eine Matrix erstellt. BLOSUM62 bedeutet also, dass die Sequenzen die benutzt wurden, um diese Matrix zu erstellen zu ca. 62% identisch waren. Deswegen ist es sinnvoll diese Matrix bei Proteinen die weniger als 62% Identität zeigen zu benutzen.

Die hier benutzte Gap Penalty war 10.0, was bedeutet, dass 10 Punkte für ein Gap abgezogen wurden.

In (2) wurde die PAM10 Matrix verwendet. PAM = Point accepted mutation.
Diese Matrizen beruhen auf den Austausch einzelner Aminosäuren in der Primärstruktur der Proteine. PAM Matrizen entstanden aus Beobachtungen von Mutationen in Phylogenetischen Stammbäumen von nah verwandten Proteinen. Diese wurden gewählt, da sie eine große Ähnlichkeit zu ihren Vorgängern hatten. Die benutzten Alignments mussten also mindesten 85% Identität aufweisen. So konnte man davon ausgehen, dass die vorherrschenden Mutationen, aus einer und nicht mehrerer Mutationen an derselben Stelle hervorgerufen wurde. Bei einer PAM10 Matrix werden bloß 10 Mutationen pro 100 Aminosäuren vermutet, weshalb nur sehr ähnliche Sequenzen sinnvolle Alignments bringen.

Hier wurde ebenfalls eine Gap Penalty von 10.0 benutzt.

Bei (3) wurde ein BLOSUM62 Matrix benutzt, aber eine Gap Penalty von 100.0 gewählt. Sowohl die Identität als auch die Ähnlichkeit ist im Vergleich zur (1) gesunken. Da weniger Lücken eingefügt werden dürfen und somit mehr Mismatches entstehen.

Bei der (4) wurde im Vergleich zu den vorherigen Fällen eine lokalen Alignment statt ein globales durchgeführt. Es wurde also nicht die gesamte Sequenz verglichen, sondern nur bestimmte Bereiche. Es wurde jedoch wieder eine BLOSUM62 Matrix mit einer Gap Penalty von 10.0 verwendet.