Programming Lab II
Handout 6

Antonio de la Vega de León, Bijun Zhang, Thomas Blaschke, Dr. Martin Vogt
martin.vogt@bit.uni-bonn.de

Jun 9,2016

Assignments for classes on Jun 21 and Jun 23

Sequence alignments

Theory. Sequence alignment is a widely used method for the comparison of two or more macromolecular sequences (e.g. DNA, RNA, proteins). Similar residues of the sequences are aligned, and gaps are introduced into the sequences where necessary. The quality of such a sequence alignment is assessed by means of a score. First, the score for every position of the alignment is assessed and then summed up to yield the overall alignment score. The score for one position reflects the similarity of the aligned residues at this position; the more similar the residues, the higher the score. Hence, the overall alignment score measures the similarity of the aligned sequences. Very simple scoring schemes assign a positive score to a position if the aligned residues are of the same type and a constant negative score if they are distinct. This approach is mostly used for nucleotide sequences. Substitution matrices present a more elaborate scoring approach: for every possible pair of residues, an individual score is determined. The following shows the BLOSUM62 substitution matrix.

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
```

The BLOSUM matrices (blocks substitution matrices) are a family of substitution matrices widely used for protein sequence alignment. The score for the alignment of two residues (amino acids) is derived from their evolutionary relatedness. The degree of relatedness is determined from a database of known related protein sequences. These related protein sequences are aligned to each other in gapless multiple sequence alignments called "blocks". Then, for each pair of amino

acids, the frequency of their occurrence in the same position of a block is counted. If two amino acids are aligned frequently in the same position, they are considered to be highly similar.

BLOSUM scoring scheme. The BLOSUM score compares for each amino acid pair how often they actually co-occur in the same position of the blocks database with the statistical expectation for this frequency. The final score s_{ab} for two amino acids a and b is called the log odds ratio and is defined as follows:

$$s_{ab} = 2 \cdot \log_2 \left(\frac{p_{ab}}{e_{ab}} \right) = 2 \cdot \log_2 \left(\frac{\text{observed frequency of pair } ab}{\text{expected probability of pair } ab} \right) \tag{1}$$

Here, p_{ab} is the *observed frequency* for amino acids a and b to occur in the same position of the alignment database, whereas e_{ab} denotes the *expected frequency* for a and b to be aligned in the same position. If two amino acids occur in the same position more frequently than expected in the database of related protein sequences, the two amino acids are likely to be similar. Then, the log odds score is positive, and aligning these amino acids gives a positive contribution to the overall score of the alignment. However, if the substitution of two amino acids is observed less frequently than expected, the corresponding log odds ratio will be negative. Aligning such an amino acid pair will decrease the overall score of an alignment.

Calculation of the score. First of all, the absolute frequencies of each single amino acid (f_a) and each pair of aligned amino acids (f_{ab}) in the alignment database are counted and stored in a frequency table. Here, one does not distinguish between f_{ab} and f_{ba}. Then, the relative frequencies can be determined from that:

$$p_a = \frac{f_a}{\sum_i f_i} \tag{2}$$

where the sum is taken over all amino acids and

$$p_{ab} = \frac{f_{ab}}{\sum_{\{i,j\}} f_{ij}} = \frac{\text{frequency of pair } ab}{\text{sum of frequencies of all pairs}} \tag{3}$$

where the sum is taken over all unordered pairs of amino acids.

The next step is the computation of the expected occurrence of an aligned amino acid pair. This is computed as the statistical probability of co-occurrence of the two amino acids from their relative frequency in the database:

$$e_{ab} = p_a \cdot p_b + p_b \cdot p_a = 2 \cdot p_a \cdot p_b \text{ for two distinct amino acids} \tag{4}$$

and

$$e_{aa} = p_a \cdot p_a \text{ for an amino acid pair consisting of the same amino acid.} \tag{5}$$

From these values, the log odds score s_{ab} is computed. The score is given as an integer number. So the score s_{ab} should be **rounded to the nearest integer number**.

Ex.0 To test your understanding, consider this small example of a "blocks alignment database" consisting of 5 sequences of length 12 each. From these to numbers one can determinate the denominators in eqs. (2) and (3). What is $\sum_i f_i$ and $\sum_{\{i,j\}} f_{ij}$?

```
TSVKTYAKFVTH
TSVKTYAKFSTH
TSVKTYAKFVTH
LSVKKYPKYVVQ
SSVKKYPKYSVL
```

Count the frequencies f_a for all amino acids in the alignment and f_{ab} for all amino acid pairs occurring in the same column of the alignment. (For the pairs, do not consider the order of the amino acids: do not distinguish between `VS` and `SV`, for example). From these values, calculate the relative frequencies p_a for each occurring amino acid and p_{ab} for each occurring

amino acid pair. Finally, calculate the expected probability and the score for each amino acid pair. Fill your results into the given tables. You can compare your results to the file `blosum_ex0-solution.txt`,which will be uploaded to the course folder at some point.

	T	L	S	V	K	Y	A	P	F	H	Q
f_a											
p_a											

	TT	TL	TS	LS	SS	VV	KK	TK	YY	AA
f_{ab}										
p_{ab}										
e_{ab}										
s_{ab}										

	AP	PP	FF	FY	VS	TV	HH	HQ	HL	QL
f_{ab}										
p_{ab}										
e_{ab}										
s_{ab}										

Ex.1 Write a program that computes a scoring matrix score from a given block alignment database. You should be able to call your problem like `python blosum.py alignment.dat blosum_matrix.out` from the command line. The input file `alignment.dat` (found in the course folder) contains a small alignment "database". It contains a number of aligned sequences of equal length with no gaps with one sequence per line.Your program should compute the substitution matrix from this alignment. The output file `blosum_matrix.out` should contain an output of the matrix like in the example of the BLOSUM matrix given above. (The original BLOSUM62 matrix is provided as `blosum62.txt` in the group folder. Are the scores of the matrix you have calculated for `alignment.dat` similar?)

Your program will need to.

(a) read in the alignment data in an appropriate data structure,

(b) determine the log-odds scores for each possible alignment of amino acids,

(c) produce a (nicely formatted) output of the resulting scoring matrix.

Note: The "blocks alignment database" given isn't very large and it might happen that alignments between some pairs of amino acids do not occur at all, i.e., they will have a count of 0. You will run into numerical difficulties if you try to take the logarithm as this would yield a score of $-\infty$. To avoid this, set the score to a large negative number (e.g. -99). However it is even better to start counting all the frequencies from 1 and not from 0 thereby pretending that there is one additional alignment of each amino acid pair by default. This or a similar strategy is frequently used when calculating frequencies/probabilities from a limited amount of data. The additional counts are also known as (Laplace) pseudocounts.

Further reading. The original publication of the BLOSUM matrix method by Henikoff & Henikoff might be helpful for understanding the theory about substitution matrices. It is deposited as `blosum_paper.pdf` in the course folder. Another good introduction is presented in the Nature Biotechnology Primer on BLOSUM matrices, which you will find at the same location as `blosum_primer.pdf`.