CHARITÉ / HUMBOLDT UNIVERSITÄT ZU BERLIN
INSTITUTE FOR THEORETICAL BIOLOGY

Prof. Hanspeter Herzel
Dr. Grigory Bordyugov
Institute for Theoretical Biology
D-10115 Berlin

email: h.herzel@biologie.hu-berlin.de
email: Grigory.Bordyugov@hu-berlin.de
Tel: +49 30 2903 9106 / 6044
Fax: +49 30 2903 8801

MODULE IV - BIOINFORMATICS: ASSIGNMENT 2

Please email a printable, A4-sized PDF file with your solution to Grigory.Bordyugov@hu-berlin.de **with subject "mmm2012" until May 3, 24:00 or return your solution as a hard copy at the beginning of a lecture.**

It seems to be of advantage to use the web resources listed in this home assignment in the morning (or, equivalently, in the night USA time) in order to reduce processing time due to a lower server load.

1. Sequence statistics

What is the probability to find the motif `ANRCTGSC` in a Bernoulli sequence ($p_i = 1/4$)? How many such motifs are expected in 100 kb? What is the standard deviation?

2. Exons and Introns

1. Could the sequence `AAGCCTGGAACTACG` be part of an open reading frame? If it could, which amino acids are encoded? What codon table did you use?

2. Can you find this sequence in human using BLASTp? How many entries do you find? BLAST web page: `http://blast.ncbi.nlm.nih.gov/Blast.cgi`?

3. How many times this sequence is expected in a Bernoulli sequence ($p_i = 1/4$) of length 10^9 base pairs (1 Gb)?

4. Determine the nucleotide frequency (`A`,`C`,`G`,`T`) for the following sequences within all three positions of a putative reading frame (Hint: use R):

   (a) Sequence A, follow the link: `http://j.mp/ifNNaE`

   (b) Sequence B, follow the link: `http://j.mp/gXjItZ`

5. Determine the percentage of relative nucleotide frequency $p_i, i =$ `A`,`C`,`G`,`T` of sequence B. Normalize the 3×4 table for sequence B to 100% per position. Which are the three highest deviations to the corresponding p_i? Which of these deviations to an equal distribution are significant? Which sequence might be protein coding?

6. Create and run an R script to determine the position asymmetry (PA) for sequences A and B.

2. Alignment on the Internet

Use SIM `http://web.expasy.org/sim/` - an alignment tool for protein sequences on the ExPASy server `http://www.expasy.ch/tools` to produce a local alignment of sequencies

1. Sequence 1:
   `QTSYREIVLSYFSPNSNLNQSIDNFVNMAFFADVPVTKVVEIHMELMDEFAKKLRVE`
   Sequence 2:
   `IDAVIFILALFPLPIASSALFAASITFVEIHMDLIDAFWQQFRLE`

1. Use PAM40 matrix and gap open penalty GOP=10 and gap extension penalty GEP=3.

2. How is the score and the second best alignment changing, if for GOP=10 and GEP=3 the scoring matrix is changed from PAM40 to PAM250 to PAM400? What is changing if GOP<10?

4. Exact matching search

1. How many entries can be found for the sequences `PEPTIDE` and `SEVERAL` in UniProtKB database located at PIR DB `http://www-nbrf.georgetown.edu/pirwww/`? Use the tool “peptide search”.

2. Is there a membrane protein, which contains the sequence `CHANNEL` or `KANAL` (e.g. a transporter)?

5. BLAST Search

During sequencing the following sequences were detected:
`MSSEAETQQPPAAPPAAPALSAADTKPGTTGSGAGSGGPGGLTS`
and
`AGCAGACATTTTATGCACCAAAAGAGAACTGCAATGTTTCAGGACCCACAGGAGCGACCC`

1. From where these sequences might originate? Go to the NCBI site and run a basic BLAST sequence search `http://blast.ncbi.nlm.nih.gov/Blast.cgi`.

2. How the E-value and the score associated with each hit can be interpreted?

3. What does tBLASTx mean?

4. To which phylum the sequences belong to?

6. FASTA Search

Search for homologs to the following endonuclease by using the FASTA3 tool `http://www.ebi.ac.uk/fasta33/` on EBI. The endonuclease sequence in FASTA format is available at `http://j.mp/e7lFYV` The sequence has a length of 279 amino acids and can be aligned by using FASTA3:

1. against SwissProt DB,
2. against USPTO Patents DB.
3. Write down the gene names of the given and the homologous endocucleases, EC numbers, origin / organism as well as restriction sites.

7. Information on IUPAC and IUBMB

What organisation is responsible for nomenclature for chemistry and which for biochemistry `http://www.chem.qmul.ac.uk/iupac/jcbn/`?
What do the letters `B, W, X, Y, Z` stand for (IUB one letter code for amino acids `http://www.chem.qmul.ac.uk/iupac/AminoAcid/A2021.html#AA21`)?

8. PubMed

In recent years, evidences arise for hereditary predisposition of the Parkinson's disease. Find out with the help of PubMed `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed` search (keywords: `parkinson risk genetic`), what noticeable problems and correlations were described (negative risk factor). What proteins seem to be mainly involved?

9. CLUSTALW

Calculate by using the CLUSTALW server in Japan `http://www.genome.jp/tools/clustalw/` a multiple alignment and a phylogenetic tree for the sequences. Take the following five sequences for the alignment:

```
> seq1        > seq2         > seq3         > seq4         > seq5
armerhase     hasenbraten    arsenbraten    arsenhase      rasenhase
```

1. What kind of data format is used above?
2. Which are the two most related sequences in terms of scoring?
3. How could one comment an entry out of the CLUSTALW output `http://align.genome.jp/clustalw/clustalw_help.html#output_format`?
4. Is it a local or global alignment?
5. Add the sixth sequence

   ```
   > seq6
   armerdummerhase
   ```

   Why now is the left alignment is favored over the right one, which does not look worse:

   ```
   seq1 ------AR-MER-HASE-            seq1 ARMER--------HASE-
   seq6 ARMERDUM-MER-HASE-            seq6 ARMERDUM-MER-HASE-
   ```

6. How one might interpret both outputs and how the data format is called?

```
((s1:0.19444,s4:0.02778):0.11111,
(s2:0.09091,s3:0.09091):0.24242, s5:0.22222);
```

and

```
(((s1:0.06987, s6:0.15236) :0.14205, s4:0.08018) :0.05871,
(s2:0.09091, s3:0.09091) :0.24558, s5:0.21907);
```

10. Sequence Retrieval System (SRS)

1. Create a dataset which contains all entries for MAP2K1 and MAP2K2 at Uniprot DB by using SRS `http://www.ebi.ac.uk/uniprot/search/SearchTools.html`. Search using the "Gene Name" option. Note the numbers of entries for MAP2K1 and MAP2K2.
2. Save the results for MAP2K1 as a text file on hard disc (choose FastaSeq format).
3. Take the mammalian entries for MAP2K1 and save them into a second text file (organism: mammalia).
4. Save the results for MAP2K1 and MAP2K2 together in a third text file.

11. Multiple alignments and phylogenetic trees

1. Construct by using data of 10.2 a multiple alignment with CLUSTALW. Take default parameters for the alignment.
2. Select and execute N-J-tree in the Selection menu at the end of the CLUSTALW output for the construction of a phylogenetic tree.
3. Repeat first two steps with the data from 10.3.
4. Repeat first two steps with the data from 10.4.
5. Describe the differences in the trees shortly.
6. Display (and refine) the trees graphically by using iTOL (`http://itol.embl.de`).