CHARITÉ / HUMBOLDT UNIVERSITÄT ZU BERL

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 9101 / 9121
Fax: +49 30 2903 8801

MODULE IV - BIOINFORMATICS: ASSIGNMENT 2

Please your return solution to G Bordyugov (office #1324 in the ground floor of the ITB) in hard copy until Monday May 5th, 15:00.

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 amino acid motif `ANRCTGSC` in a Bernoulli sequence if you know the statistics of the DNA nucleotids $p_i = 1/4$, i = A,C,G,T? How many such motifs are expected in 100 kilo base pairs? What is the standard deviation of the number of found motifs?

2. Sequence statistics and BLAST search

1. Could the sequence
`ATTCTTTTACTCAAGAATGCATGGAGGAGAAATCTTTCTTTTGCCGTGTCAG`
be part of an open reading frame? If it could, which amino acids are encoded?

2. Can you find a transcript of this sequence in the human using BLASTp? How many entries do you find? BLAST web page: `http://blast.ncbi.nlm.nih.gov/Blast.cgi`? What does the E-value stand for?

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

3. Exons and Introns

1. Determine the frequency of the four nucleotides A,C,G,T in the following sequences within all three positions for:

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

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

Represent your result as a 3 × 4 table.

2. 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?

3. Write an R script for computation of the position asymmetry (PA) and use it for sequences A and B.

4. Which sequence might be protein coding and why?

4. Alignment on the Internet

Use SIM `http://web.expasy.org/sim/` – an alignment tool for protein sequences – to produce a local alignment of sequences:

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?

5. 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)?

6. BLAST Search

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

1. Where can these sequences originate from? 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?

7. FASTA Search

Please be advised that FASTA online search can take some time.
An endonuclease sequence is given at `http://j.mp/e71FYV`. Search for homologs to the endonuclease using the FASTA3 tool `http://www.ebi.ac.uk/Tools/sss/fasta/` on EBI. The sequence has a length of 277 amino acids and can be aligned by using FASTA3 against SwissProt DB. Write down the gene names of the given and the homologous endocucleases, EC numbers, origin / organism as well as restriction sites.

8. Literature search on 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. Is it a local or global alignment?

4. 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-
   ```

5. How one might interpret those both exemplary 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 `www.uniprot.org` Search using the "Advanced Search" and "Gene Name" options. 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 (taxonomy: 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. Display (and refine) the trees graphically by using iTOL (`http://itol.embl.de`).