CHARITÉ / HUMBOLDT UNIVERSITÄT ZU BERL
INSTITUTE FOR THEORETICAL BIOLOGY

Prof. Hanspeter Herzel — h.herzel@biologie.hu-berlin.de
Dr. Grigory Bordyugov — Grigory.Bordyugov@hu-berlin.de
Institute for Theoretical Biology — Tel: +49 30 2903 9101 / 9121
D-10115 Berlin — Fax: +49 30 2903 8801

MODULE IV - BIOINFORMATICS: ASSIGNMENT 2

Please return your solution to G Bordyugov (office #201 in Haus 4, Charité Campus) in hard copy until Thursday April 21st, 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 the processing time due to lower server load.

1. Sequence statistics

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

2. Exons and Introns

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

   (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 $p_i, i = \mathtt{A,C,G,T}$ 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?

3. Online alignment tools

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 best 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. 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?

6. FASTA Search

Please be advised that FASTA online search can take some time.
An endonuclease sequence is given at `http://j.mp/e7lFYV`. 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.

7. 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. How one might interpret the following exemplary CLUSTALW outputs and what data format is this?

```
((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);
```

8. Statistical testing of synthetic data

1. Implement a function `drawfromgaussian(mu,sigma,N)` that generates and returns a vector of `N` random numbers drawn from a Gaussian distribution with the mean `mu` and standard deviation `sigma`.
2. For each N from $N = 3, 10, 30, 300, 1000, 3000, 10000$, generate five pairs of vectors consisting of N normally distributed numbers (with means and standard deviations of your choice) such that the means are a) 0.1Σ, b) 0.5Σ, c) Σ, d) 3Σ, and e) 6Σ apart, where Σ is the sum of the standard deviations of both distributions in a pair.
3. Apply the t-test to each of the pairs to test the hypothesis that random numbers in both vectors are drawn from distributions with the same mean. For each pair, specify the calculated p-value.