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 1

Please email a printable, A4-sized PDF file with your solution to Grigory.Bordyugov@hu-berlin.de **with subject "mmm2012" until April 19, 24:00 or return your solution as a hard copy at the beginning of a lecture.**

1. Finding useful resources on the Internet

The following terms for exemplary and/or commonly used databases (DB) may be entered in the window of a search engine to obtain a current Internet link to that resource, if available. Thereby, inform yourself about the provided information. Note the Internet link and just **a few** (!!!) words for at least 8 DBs: Which is the responsible institution and where is it located? What kinds of data are collected? (Try to order the information in a useful way.)

DB search terms

1. NCBI: GENBANK, Entrez, Pubmed
2. EBI: EMBL, UniProt, MSD, ArrayExpress, Ensembl, IntAct
3. EXPASY: PROSITE, UniProt
4. GDB HUMAN, Gene Ontology (GO), KEGG, MGC (Mammalian Gene Collection), MGI, GSF, eGENOME
5. PDB, PFAM, PIR, RFAM, SMART, TIGR
6. Protein-Protein Interactions Databases: UniHI, HPRD, MIPS

Hint: An updated, more complete list for 2012 might be found here:
`http://nar.oxfordjournals.org/content/40/D1/D1.abstract`

2. Searching keywords in DBs

How many entries can be found for period circadian protein homolog 2 (PER2) at:

1. GenBank?

2. UniProt?
3. PubMed?

3. DB entries #1

Find and compare the entries for human period circadian protein homolog 2 (PER2) at UniProt including:

1. Entry name (protein identifier)
2. Gene Ontology (GO identifiers)
3. Length and molecular weight
4. Nuclear export signal

4. DB entries #2

Find the entries for human Aryl hydrocarbon receptor nuclear translocator-like (ARNTL) at MGC (Mammalian Gene Collection):

1. How many entries are listed? What are the differences?
2. Calculate the distances between the putative transcription initiation and translation start sites as well as between the translation stop and the transcription termination sites.

5. Prediction of protease activity, calculation of peptide mass

How many fragments of molecular weight > 500 Dalton might result by cleaving human PER2 using:

1. Trypsin?
2. Proteinase K?
3. What are the cleavage rules for Trypsin and for Proteinase K used by the PEPTIDE MASS program (`http://www.expasy.ch/tools/peptide-mass.html`, see "Instructions")?
4. Find out the theoretical and monoisotopic mass of human PER1 using the PEPTIDE MASS program and compare the results.

6. Definition of key patterns

In many cases, the function of a new sequence can be defined based on a single motif (key pattern), which can be found in the PROSITE DB (`http://www.expasy.org/prosite/`). Specify the consensus pattern for the human PER2.

7. Enzyme Nomenclature

What does EC number classify and who is responsible for administration and definition of EC numbers? Which EC number does deoxyribodipyrimidine photo-lyase possess and what do the different numbers mean? Hint: use `http://www.brenda-enzymes.org`.

8. PPI database search

Find interactions partners for proteins PER2, BMAL, CRY2 in the following databases.

1. UniHI, HPRD.
2. What is the difference between these two PPI databases?
3. How many interactions partners did you find for search proteins in each database?

9. Sequence statistics

1. We consider a plasmid with 10000 base pairs and independent nucleotides with $p_T = p_A = 0.2$ and $p_C = p_G = 0.3$. How many restrictions sites `AACGTT` and `CGTACG` can be expected on average? What is the standard deviation? Calculate the probability that no site `CGTACG` is found.
2. Let us assume $p_i = 1/4$ independence of nucleotides and subsequent triplets. What is the probability to observe for $N = 100$ triplets $k = 0,1,2,\ldots,6$ stop codons (`TGA`, `TAG`, `TAA`)? How many random ORFs with 100 codons can be expected in a mammalian genome?

10. Positional weight matrix

Transcription factor binding sites are usually slightly variable in their sequences. Positional weight matrix summarizes information about binding sites sequence alignment. It also allows predicting the occurrence of new sites and estimating their binding energy for transcription factor. Here is an example of binding sites sequence alignment:

	alignment position						
site	1	2	3	4	5	6	7
1	T	T	C	T	T	C	T
2	C	T	A	T	A	A	C
3	T	C	G	G	A	G	G
4	C	T	G	A	A	T	G
5	T	T	G	G	A	C	G
6	T	C	G	T	G	C	G
7	T	T	G	G	A	G	C
8	T	T	G	T	A	A	G
9	T	A	C	C	A	A	G
10	T	G	C	A	A	A	G
11	A	T	G	A	T	C	T
12	A	T	G	A	A	T	G
13	T	C	A	T	T	G	G
14	T	A	G	A	T	G	T
15	A	G	G	C	A	T	A

1. Calculate a matrix that presents how many times nucleotide i was observed in position j of the alignment (position count matrix, PCM).
2. Calculate the position weight matrix (PWM) with elements W_{ij} derived from the matrix above. In this example assume $p_T = p_A = 0.3$ and $p_C = p_G = 0.2$ (overall frequency of the letters within Drosophila melanogaster genome).

Hint: *Before* calculating f_{ij}, add 1 pseudo-count in each position of the PCM to avoid $\log_2 0$. Do not forget to modify N accordingly.

3. What does a positive W_{ij} mean for a letter i at position j in the matrix? How the weight of a new binding site might be interpreted biologically? In which case the information content of a matrix would become increasing?

4. Give a possible consensus sequence.

5. Weight matrix can be used to evaluate the resemblance of any $L = 7$ bp DNA sequence to the training set of binding sites. The score for this sequence is calculated as the sum of the values that each base of the sequence has in the weight matrix. Any sequence with score that is higher then the predefined cut-off is a potential new binding site. Calculate the weights for the following putative binding sites:

   (a) `TTGGATG`

   (b) `GATGAGA`

   (c) `GGTGGAA`

   Which cut-off would you define (why?), and which of the three sequences would be therefore predicted as a potential new binding site?

6. Invent an imaginary binding site the same factor might likely bind to.

7. Create a sequence logo of the given weight matrix. Use the WebLogo `http://weblogo.berkeley.edu/`.

11. Functions

1. Determine zeroes, extrema and asymptotic behaviour for the function:

$$f(x) = xe^{-x}, \quad \text{with} \quad x \in (0,\infty).$$

   Draw the function in the interval $(0,5)$.

2. What are the zeroes of the function:

$$g(t) = \frac{\sin t}{t}, \quad \text{with} \quad t \in (0,\infty).$$

   Draw the function in the interval $(0,4\pi)$.