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 1

Please your return solution to G Bordyugov (office #1324 in the ground floor of the ITB) in hard copy until April 4th, 15:00.

1. Finding useful resources on the Internet

Do an Internet research and write a *short* description for the following databases:

1. Federated search engines: Entrez, Pubmed
2. Genes: GENBANK, Ensembl, MGC (Mammalian Gene Collection), Gene Ontology (GO)
3. Proteins: UniProt, PDB, PFAM
4. Protein-Protein Interactions Databases: UniHI, HPRD, MIPS

For each database, please describe what data are collected, what you think distinguishes the database from the similar ones, and what other databases the given one is connected to. Hint: An updated list for databases can be found here:
`http://nar.oxfordjournals.org/content/41/D1/D1.short`

2. DB entries #1

Find and compare the entries for human period circadian protein homolog 2 (PER2) at UniProt including entry name (protein identifier), gene ontology (GO identifiers), length and molecular weight, and nuclear export signal.

3. 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.

4. Entropy and the Number Guessing Game

Sometimes entropy is defined as a number of questions that must be asked in order to define some object uniquely, for example, to define an integer number from a certain range. This idea is illustrated by a simple game where the computer tries to guess the number the user has picked by asking a sequence of questions.

1. Suppose that user picked an integer number between 0 and N. A computer program can apply the so-called *binary search* to guess this number in the following way:

   (a) The range of possible numbers is described by variables `min` and `max`, which in the beginning are set to 0 and N, respectively.

   (b) Computer calculates the integer midpoint `mid` between `min` and `max`.

   (c) Computer makes a guess by asking user if `mid` is larger than, smaller than, or equal to the number picked by user.

   (d) If `mid` is equal to the picked number, the algorithm stops and prints the found number. Otherwise, if `mid` is larger than the picked number, `max` is set equal to `mid`. Otherwise, `min` is set equal to `mid`.

   (e) The process repeats from (b) with values of `min` and `max` updated as described in (d).

2. For three different numbers between 0 and 100, provide a sequence of computer guesses according to the algorithm above together with the values of `min`, `mid` and `max` variables. What is the maximum number of guesses that are needed in order to find a number between 0 and 100?

3. Write an R function `nggame(N)` that plays the Number Guessing Game with user. Do a little research on the Internet on how to make R ask for a user input.

4. Make computer play the game against itself by having it pick a random number between 0 and N and trying to find it by the binary search. To do that, write an R function `pick_n_find(N)` which takes N as a argument. This function should pick a random number between 0 and N and perform the binary search to find the number, returning the number of guesses it took.

5. Run the function `find_n_pick(N)` several times for different N. Plot the number of guesses against N. Scale N exponentially, I suggest $N = 2^n$ for $n = 4..20$. For each N, average the number of guesses over 100 games with different picked numbers. Can you explain the scaling of the number of guesses as N increases?

5. Enzyme Nomenclature

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

6. PPI database search

Find interactions partners for proteins CRY1, BMAL1, PER1 in UniHI and HPRD databases.

What differences between those two databases can you figure out? Are there differences between the results for the interaction partners between the databases?

7. Sequence statistics

1. Consider a plasmid with 10000 base pairs and assume no sequential correlation of nucleotide probabilities. One probability is known: $p_T = 0.2$. Determine the probabilities p_A, p_C, and p_G. How many restrictions sites `AACGTT` and `CGTACG` can be expected in the given plasmid on average? What is the standard deviation of the number of restriction sites? 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?

8. 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

This table is also available as a text file at `http://itb.biologie.hu-berlin.de/~bordyugov/tut/mmm2014/table.txt`

1. Calculate the Position Count Matrix (PCM) from the given data.

2. Calculate the position weight matrix (PWM) with elements W_{ij} derived from the PCM assuming $p_T = 0.3$.

Hint: *Before* calculating the PCM, add 1 pseudo-count in each position of the PCM to avoid $\log_2 0$ (please also explain why should this be avoided?). 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?

4. Suggest a possible consensus sequence.

5. Calculate the scores for the following putative binding sites:

   (a) `TTGGATG`

   (b) `AATGAGG`

   (c) `AGTGGAG`

   What cut-off value would you stick to and why? Which of the three sequences would be therefore predicted as a potential new binding site?

6. Suggest a hypothetical high-score binding site.

7. Create a sequence logo for the weight matrix. Use the WebLogo `http://weblogo.threeplusone.com/`

9. A little bit of analysis

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

$$f(x) = xe^{-x}.$$

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

2. What are the zeroes of the function:

$$g(t) = \frac{\sin t}{t}.$$

   Draw the function in the interval $(0, 6\pi)$.

Please don't hesitate to contact us if you have difficulties with the home assignments!