

Bioinformatics 1 -- Homework 6 -- Fall 2011

Due Dec 5

(1) Gene finding, ontology, splicing:

In NCBI find the nucleotide sequence: NG_011507.1

Homo sapiens glutamate receptor, ionotropic, N-methyl D-aspartate 1 (GRIN1) on chromosome 9.

(a) Use NCBI gene viewer to look at the intron/exon structure of this gene.

Report the following

- Start / end points of exons 5 and 6,
- Frame of the short intron between exons 5 and 6,
- Amino acid sequence of the spliced exon 5,6 fragment,
- Sequence of same fragment if splicing is silenced.

(b) Use **amigo.geneontology.org** to find this gene. Display the GO terms for this gene, only in homo sapiens, and only those relating to cellular component, in graphical format.

(c) Using only the terms in GO (all three sub-ontologies), write a very short essay describing the cellular location, molecular function and biological process of the NMDA glutamate receptor.

(d) Using the MFOLD server (<http://mfold.bioinfo.rpi.edu>) predict the RNA structure of the 5,6 intron. Report the stability (ΔG). Paste the computed structure. Find and label the most likely branch site A.

(2) Satellite sequences.

Find the nucleotide sequence with accession number AI253829. "Giant tiger prawn eyestalk cDNA library *Penaeus monodon*." In UGENE, find the repeat sequence region using "Compare/Dotplot."

(a) Where is the repeat region? What is the size of the large repeating unit?

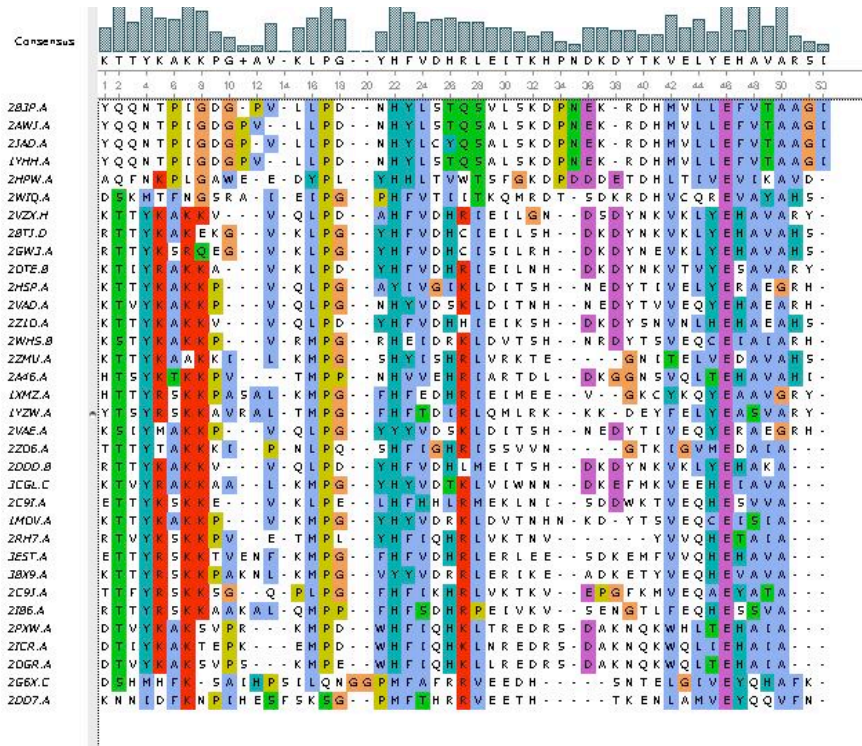
(b) Make copies of the sequence and aligning repeats by hand in UGENE. Trim the edges. Print the aligned repeats and include it in the homework.

(c) Write a Prosite-style expression for the repeat. This should be a highly selective model.

(d) Draw a HMM that models the low complexity and microsatellite content of the repeat. Use a start and end state (non-emitting). Each state may emit of one or more nucleotides, or one nucleotide from a probability distribution. Use as few states as possible. This should be a more sensitive model than the one in (c), but still somewhat selective.

(3) Profile HMMs

Given the following MSA, design and draw a profile HMM, adhering strictly to the conventions as follows: 3 states per position -- match, deletion, insertion only have arrows to states in the next position. insertion state has arrows to states in the same position.



Write a sequence that would have a high Viterbi score but is not one of the sequences used to generate the HMM.

(4) Expectation/maximization

Write the **maximum likelihood transition matrix** for the HMM below (Fig 1), given a very small set (one sequence, length=10) of training data. The sequence itself is not shown. Instead, for your convenience, the sequence data has already been converted to **Forward** (Table 1) and **Backward** (Table 2) values. [For example, $\alpha_5(M)$ is the pre-calculated value for the sum of all paths through the HMM up to $t=5$ in the sequence.]

(a) Write the formula used to sum the maximum likelihood values for all four transitions $X \rightarrow M$, $M \rightarrow M$, $X \rightarrow X$, and $M \rightarrow X$, starting with forward/backward data. Define each term in the formula. Specify how the sums are normalized.

(b) Calculate the maximum likelihood transitions, write them in Table 3, below. Show your work.

Fig 1.

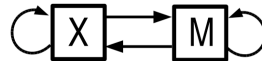


Table 1 Forward values

t=	1	2	3	4	5	6	7	8	9	10
$\alpha_t(M)$	0.00	0.00	0.00	0.30	0.40	0.80	0.90	1.00	1.00	1.00
$\alpha_t(X)$	1.00	1.00	1.00	0.70	0.60	0.20	0.10	0.00	0.00	0.00

Table 2 Backward values

t=	1	2	3	4	5	6	7	8	9	10
$\beta_t(M)$	0.00	0.00	0.10	0.20	0.50	0.60	1.00	0.80	0.90	1.00
$\beta_t(X)$	1.00	1.00	0.90	0.80	0.50	0.40	0.00	0.20	0.10	0.00

Table 3. Transition probabilities (current values)

to	M	X
from		
M	0.9	0.1
X	0.2	0.8

Table 4. Fill in the **maximum likelihood transitions**

to from	M	X
M		
X		