Questions comparative genomics, day 4

Question 1
For the following sequence.

> Unidentified sequence 1
GGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCAC

a) Identify from which species it is derived.

b) Does its RNA have a functional RNA secondary structure? There are different ways of detecting whether an RNA structure is likely functional: compensatory base-pair changes and a low or well-defined free energy secondary structure. We will try both. First we will compare its predicted secondary structure with that of random sequences.

Find an RNA folding server (Google for RNAfold) to fold the sequence. Estimate the frequency of the Minimum Free Energy folding in the ensemble of secondary structures predicted for this sequence (the partition function, an option in RNAfold). You will have to compare that with random RNA sequences with the same nucleotide frequencies. How many standard deviations is the MFE of this secondary structure lower than that of random sequences (the so-called Z score). Fold 10 randomized sequences to obtain an average expected free energy for this sequence and the standard deviation.

Hint: Obtain random RNA sequences with the same nucleotide frequencies by shuffling the above sequence by hand (be creative, and compare your results with others). You can also use a sequence shuffler on the web like: http://www.cellbiol.com/scripts/randomizer/sequence_randomizer.html*)*

c) Is the MFE of the original sequence similar to the centroid of the ensemble? How about the MFE versus the centroid of random sequences?

d) For finding compensatory base-pair changes you will have to find homologs of your sequence that are not identical. Try it with this one:

> Unidentified sequence 2
GGUCUCUCUAGGUAGACCAGAUCUGAGCCUGGGAGCUCUCUGGCUAUCUGGGGAACCCAC

Are there compensatory base-pair changes between these two sequences? If so, which ones?

Question 2

a) Repeat question 1 for the following sequences:

>Seq_1

TTTTTAGGGAAAATTTGGCCTTCCAACAAGGGAAGGCCGGGAAATTTTCCTC

>Seq2

TTTTTAGGGAAAATTTGGCCTTCCAACAAAGGGAGGCCAGGAAACTTTCCTC

Hint: Not all base pair changes are "compensatory" sensu strictu…

b) There is an RFam database of RNA secondary structures (http://rfam.xfam.org/search). Find the sequences there. What is their function?

c) Is the structure predicted in RFAM the same as by RNAfold. Do the differences correspond to base pairs that are predicted by RNAfold with high or with low probability?

Question 3

For the following sequences:

>ssb

ATTGACCTGAATGAATATACAGTATTGGAATGCATTACCCGGAGTGTTGTGTAACAATGTCTGGCCAGGTTTGTTTCCCGG

>uvrA

TCCGGGTAATGCATTCCAATACTGTATATTCATTCAGGTCAATTTGTGTCATAATTAACCGTTTGTGATCGCCGGTAGCAC

>uvrD

TACTGCCGCATCTGGAAATTTCCCGGTTGGCATCTCTGACCTCGCTGATATAATCAGCAAATCTGTATATATACCCAGCTT

>rpsU_dnaG_rpoD

CGGTGCTTTACAAAGCAGCAGCAATTGCAGTAAAATTCCGCACCATTTTGAAATAAGCTGGCGTTGATGCCAGCGGCAAAC

>recA

TTGTGGCAACAATTTCTACAAAACACTTGATACTGTATGAGCATACAGTATAATTGCTTCAACAGAACATATTGACTATCC

>umuC

GACAAATATTGATAGCCTGAATCAGTATTGATCTGCTGGCAAGAACAGACTACTGTATATAAAAACAGTATAACTTCAGGC

>sulA

AGGCTCTTTCCGAAAATAGGGTTGATCTTTGTTGTCACTGGATGTACTGTACATCCATACAGTAACTCACAGGGGCTGGAT

>uvrB

TTACGCTGTATCAGAAATATTATGGTGATGAACTGTTTTTTTATCCAGTATAATTTGTTGGCATAATTAAGTACGACGAGT

>phrB

AGTTATCGCCGTGGCGAGCAACCACTTCTTGCGCCGCTGATGCGTATCAAACACTATATGGCGCTTTATCCTGACGCCTGG

a) Detect a shared motif with Gibbs sampling, (http://bayesweb.wadsworth.org/cgi-bin/gibbs.12.pl?data_type=DNA). Is the motif present in all the sequences? If not, can you still find some similar motifs "by eye" in the sequences in which the motif has not been reported? (you can also answer this part of the question with the MEME, see question D)

Hint: These are bacterial sequences, so check "Prokaryotic defaults". The server is a bit "fickle", do not start clicking on everything that moves. Once you get the results you have to scroll down a bit to get to the motifs.

b) Make a sequence logo of your motif (http://weblogo.berkeley.edu/logo.cgi).

Hint: Do note that you will have to give the aligned sequences to get the motif. The motif search provides you with that.

c) One of the most important aspects of bioinformatics is actually not getting the results (a computer always will give you something), but the interpretation of the results. What do these genes have in common, and why they are regulated together?

Hint:You can use the gene names provided in the FASTA sequences to search in the biomedical literature (e.g. at http://www.ncbi.nlm.nih.gov). Enter multiple gene names to find what they have in common. Review articles can be a good source of information.

d) Now try another method, MEME, which is based on so-called Expectation Maximization. How do the results compare with the results from the gibbs sampling? Do you find similar motifs?

Hint: Try: http://meme-suite.org/. *Use the default settings but make sure that it finds one motif per sequence.*

e) Examine the level of sequence conservation in the aligned sequences. Are certain positions better conserved than others? And can you explain that from the interaction of this motif with the transcription factor? For a 3D image of this interaction examine: Zhang, Pigli and Rice, Nature 2010. How does knowing the transcription factor and its function compare with your results from question 3C? What is the name of the motif?

Question 4

a) The following sequences are from the 3’ end of Hepatitis G virus the family Flaviviridae. Align the sequences with clustal Omega, detect RNA secondary structures within them that are conserved, and show evidence of compensatory base pair changes.

> Seq 1

GGCCTACGGCTCCCTCCCCCTGAGATTGCTGGTATCCCCGGGGGTTTCCCCTCTCCCCCCCCCTTTATGGGGGTGGTTC
ATCAATTGGATTTCACAAGCCAGAGGAGTCGCTGGCGGGGGTTGGGGGTCTTAGCCCTGCTCATCGTGGCCCTCTTC
GGGTGAACTAAATTCATCTGTTGCGGCAAGGTCTGGTGACTGATCATCACCGGAGGAGGTTCCCGCCCTCCCCGCCCC
AGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGGGAGGCATGGTGGTTACTAACCCCCTGGCAGGGTCAAAGC
CTGATGGTGCTAATGCACTGCCACTTCGGTGGCGGGTCGCTACCTTATAGCGTAATCCGTGACTACGGGCTGCTCGCA
GAGCCCTCCCCGGATGGGGCACAGTGCACTGTGATCTGAA

> Seq 2

AGGCCTGCGGCTTCCCCCTCCGGAGATTGCTGGTATCCCCGGGGGTTTCCCCCTTTCCCCCCCCCTATATGGGGGTGGTT
CATCAATTGGATTTCACAAGCCAGAGGAGTCGCTGGCGGTGGTTGGGGTTCTTAGCCCTGCTCATCGTAGCCCTCTTC
GGGTGAACTAAATTCATCTGTTGCGGCAAGGTCCGGTGACTGATCATCACTGGAGGAGGTTCCCGCCCTCCCCGCCCC
AGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGGGAGGCATGGTGGTTACTAACCCCCTGGCAGGGTCAAAGC
CTGATGGTGCTAATGCACTGCCACTTCGGTGGCGGGTCGCTACCTTATAGCGTAATCCGTGACTACGGGCTGCTCGCA
GAGCCCTCCCCGGATGGGGCACAGTGCACTGTGATCTGAAGGGGTGCACC

> Seq 3

GGGCCTGTTGTGGCATCCAGGCCTGCGGCTTCCCCCCCCCGAGATTGCTGGTATCCCGGGGGGTTTCCCCCTTTCCCCC
CCCTACATGGGGGTGGTTCATCAATTGGATTTAACAAGCCAGAGGAGTCGCTGGCGGTGGCTGGGGTTCTTAGCCCT
GCTCATCGTAGCCCTCTTCGGGTGAACTAAATTCATCTGTTGCGGCAAGGTCTGGTGACTGATCATCACCGGAGGAG
GTTCCCGCCCTCCCCGCCCCAGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGGGAGGCATGGTGGTTACTAAC
CCCCTGGCAGGGTCAAAGCCTGATGGTGCTAATGCACTGCCACTTCGGTGGCGGGTCGCTACCTTATAGCGTAATCCG
TGACTACGGGCTGCTCGCAGAGCCCTCCCCGGATGGGGCACAGTGCACTGTGATCTGAAGGGGTGCACCCCGGTAAG
AGCCCGGCCCAAAG

> Seq 4

CCAGGTCTCCGGCTCCCCCCCCCCGAGATTGCTGGTATCCCGGGGGGTTTCCCCCTGTCCCCCCCCCTACATGGGGGTGG
TTCATCAATTGGATTTCACAGCCCAGCGGAGTCGCTGGCGGTGGTTGGGGTTCTTAGCCCTGCTCATCGTAGCGCTCT
TTGGGTGAACTAAATTCATCTGTTGCGGCAAGGTCGGCTAGCTGATCACTAGCTGAGGAGGTTCCCGCCCTCCCCGCC
CCAGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGGGAGGCATGGTGGTTACTAACCCCCTGGCAGGGTTAAA
GCCTGATGGTGCTAATGCACT

> Seq 5

CGGTCGCGCGGATGGGCAGAGCTGGCTCGGGGCCTGTTGTGGCATCCAGGCCTCCGGCTCCCTCCACCCGAGATTGCT
GGAATCCCGGGTGGGTTTCCCCTGTCCCCCCCCCTACATGGGGGTGGTTCATCAATTGGATTTCACAGCCCAGCGGAGT
CGCTGGCGGTGGTTGGGGTTCTTAGCCCTGCTCATCGTAGCACTCTTTGGGTGAACTAAATTCATCTGTTGCGGCAA
GGTTGGGTGACTGATCATCACCTGAGGAGGTTCCCGCCCTCCCCGCCCCAGGGGTCTCCCCGCTGGGTAAAAAGGGCC
CGGCCTTGGGAGGCATGGTGGTTACTAAC

> Seq 6

GGATGGGCGGAGCTGGCTCGGGGCCTGTTGTGGCATCCTGGCCTCCGGTTCCCCCCCCCCGAGATTGCTGGTATCCCG
GGGGGTTTCCCCCTGTCCCCCCCCCTATATGGGGGTGGTTCATCAATTGGATTTTACAGCCCAGCGGAGTCGCTGGCGG
TGGCTGGGGTTCTTAGCCCTGTTCATCGTAGCGCTCTTTGGGTGAACTAAATTTATCTGTTGCGGCAAGGTCGGCTA
GCTGATCACTAGCTGAGGAGGTTCCCGCCCTCCCCGCCCCAGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGG
GAGGCATGGTGGTTACTAACCCCCTGGCAGGGTTAAAGCCTGATGGTGCTAATGCACTGCCGTTCGCGGCGGGTCGC

TACCTTATAGCGTAATCCGTGACTACGGGCTGCTCGCAGAGCCCTCCCCGGATGGGGCACAGTGCACTGTGATCTGAA
GGGGTGCACCCCGGTAAGAGCTCGGCCCAAAGGCCGGGTTCTACT

> Seq 7
GGTTGGGCGGAACTGGCTCCGGGCCTGTTGTGGCATCCAAGGCTCCGGCTCCCCCCCCCCGAGATTGCTGGTATCCCG
GGGGGTTTCCCCCTGTCCCCCCCCTACATGGGGGTGGTTCATCAATTGGATTTCACAACCCAGCGGAGTCGCTGGCGG
TGGTTGGGGTTCTTAGCCCTGCTCATCGTGGCGCTCTTTGGGTGAACTAAATTCATCTGTTGCGGCAAGGTCGGCCG
ACTGATCATCGGCTGAGGAGGTTCCCGCCCTCCCCGCCCCAGGGGTCTCCCCGCTGGGTAAAAAGGGCCCGGCCTTGG
GGGGCATGGTGGTTACTAACCCCCTGGCAGGGTCATCGCCTGATGGTGCTAATGCACTGCCGTTT

*Hint: You can use the tool ALI fold (*http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi *). Just use the default settings.*

b) What is the largest number of alternative base pairs for an interacting pair of nucleotides that has been found in this alignment?

c) The most reliable parts of the secondary structure prediction have the “steepest slopes” in the mountain range plot. What are the most reliable parts of the predicted secondary structure? Do they all show evidence of compensatory base pair changes? How can a secondary structure have a high confidence score (steep slope) even when there are no compensatory base pair changes?
Which hairpin is the most reliable?

Hint: The most reliable parts of the secondary structure prediction have the “steepest slopes” in the mountain range plot.

Question 5

Using one of the secondary structure prediction servers, predict the secondary structure of the following sequences:

>seq1

CGGGGUGAGGUAGUAGGUUGUGUGGUUUCAGGGCAGUGAUGUUGCCCCUCGGAAGAUAACUAUACAACCUACUGCCUUCCCUG

>seq2

GCAGGGUGAGGUAGUAGGUUGUGUGGUUUCAGGGCAGUGAUGUUGCCCCUCCGAAGAUAACUAUACAACCUACUGCCUUCCCUGA

>seq3

GACAGUGCAGUCACCCAUAAAGUAGAAAGCACUACUAACAGCACUGGAGGGUGUAGUGUUUCCUACUUUAUGGAUGAGUGUACUGUG

>seq4

ACCCAUAAAGUAGAAAGCACUACUAACAGCACUGGAGGGUGUAGUGUUUCCUACUUUAUGGAUG

>seq5

AGUAUAACUAGCUAAACCGCAGUACUCUAGGGCAUUCAUUUGUUACAUAGAAUACUGAGGCCUAGCUGAUUAUACU

a) Using the miRNA prediction server http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi , to predict miRNA precursors in the sequences above (the server is a Support Vector Machine (SVM). In an SVM the scores above 0 are generally “good”)

Hint: You can use multiple sequences as input.

b) Using publicly accessible sequence databases, find out what these sequences are. Which ones are miRNAs ? (Hint: search the miRNA registry at http://www.mirbase.org/). How do you explain the inconsistencies between the prediction results and the “true” miRNAs?

Hint: Nobody is perfect.