

1001001100001
0100001110100
0100001110100

Homework One

Please hand in the solutions to the following problems on Thursday, February 14, 2008. Hand in a hard copy and a disk containing your solutions.

Problem One

In a typical description of a eukaryotic gene, several sequence elements in the promoter of the gene are described as important cis-acting elements involved as positioning/initiation and upstream elements used by RNA polymerase II during transcription initiation of the gene.

You have recently defined two new cis-acting elements in a gene you are characterizing. These elements have the sequences ATGCGCTTA and CCATGGTATTAA and are located at -80 to -88 and -40 to -51, respectively, with respect to the transcription initiation site of the gene.

Characterization of the mRNA generated from this gene indicates that the AUG for translation initiation of the mRNA is located at +40.

- a) What are **cis-acting elements** (also known as cis-elements or cis-acting factors)?
- b) Label the double-stranded DNA diagram illustrating the site of transcription initiation and the location and sequence of the two upstream **cis-acting elements**.
 - 1) Include the sequences for both DNA strands when labeling the positions and providing the sequences of the elements.
 - 2) Label the location and provide the sequence for the **translation initiation site** for both DNA strands.
- c) Label each DNA strand with respect to their respective 5' and 3' orientation clearly showing:
 - 1) the template or anti-sense DNA strand, and
 - 2) the coding, RNA, or message sense strand.

Use the following diagram format for your answers:

5' -----> 3'

3' <----- 5'

Problem Two

Obtain the E.coli lactose operon sequence from NCBI. The accession number is: J01636 and the GI is: 146575.

Choose two gene-prediction packages that you can find on the internet.

- A) Discuss what algorithmic technique each package uses.
- B) Use each package to predict the start and stop positions of the genes in the E.coli sequence.
- C) Compare the output of both packages and summarize the differences and similarities. Use a maximum of two paragraphs.

Problem Three

“A vision for the future of genomics research” by Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer, was published in Nature in April 2003. The article can be found at:

<http://www.nature.com/nature/journal/v422/n6934/full/nature01626.html>

The authors formulate the vision for the future of genomics research into:

- a) Three major themes: genomics to biology, genomics to health, and genomics to society. For each theme, they present a series of grand challenges which are intended to be bold, ambitious research targets for the scientific community.
- b) Six crosscutting elements:

1) resources	2) technology development
3) computational biology	4) training
5) ethical, legal and social implications	6) education.

- 1) Choose one grand challenge from one of the three major themes and write a one-page essay explaining:
 - what the challenge is about
 - the role of bioinformatics in the challenge
- 2) Choose an element from the six crosscutting elements and write a one-page essay explaining:
 - What the element is about
 - the role of bioinformatics in the challenge

In both cases, do not copy what is written in the article. Use the references as starting points for your literature search.

Problem Four

“Initial sequence of the chimpanzee genome and comparison with the human genome” by The Chimpanzee Sequencing and Analysis Consortium was published in Nature in September 2005. The article can be found at:

<http://www.nature.com/nature/journal/v437/n7055/pdf/nature04072.pdf>

- a) Choose one of the nine main findings mentioned and described on pages 69 and 70, and write a one-page essay explaining it.
- b) The article has 13 figures. Choose one of them and write a one-page essay explaining it.

c) The article has 8 tables. Choose one of them and write a one-page essay explaining it.

In all three cases, do not copy what is written in the article. Use the references as starting points for your literature search. Explain the role of bioinformatics in all three questions above.

Problem Five

The gene DCC (deleted in colorectal cancer), located on human chromosome 18q21.3, encodes for a tumor suppressor protein; expression of the gene is reduced significantly in most colorectal carcinomas. The protein sequence of human DCC can be found by searching NCBI Entrez for RefSeq accession number NP_005206.

a) Perform a BLASTP search, using this sequence as the query and Swiss-Prot as the target database. Limit the search to mammalian species only, and use BLOSUM62 as the scoring matrix.

i) The DCC protein from human is most closely related to the DCC protein from what other mammal?

ii) What percent identity do they share?

iii) What is the percent similarity?

iv) What is the length of the alignment? Were both proteins aligned along their entire length?

b) Does the DCC protein contain any low-complexity regions that have been masked-out by BLASTP? If so, where?

c) Using BLAST2Sequences and BLOSUM62, what percent identity and similarity do the first two hits from mouse in the BLASTP hit share with one another? How much of the alignment is accounted for by gaps? Does changing the matrix to BLOSUM90 significantly change these numbers?

d) Perform a FASTA search, using the DCC sequence as the query and Swiss-Prot as the target database. Again, use BLOSUM62 as the scoring matrix, and use $k_{\text{tup}} = 1$ for the word size. One of the returned hits is for the Wiskott-Aldrich syndrome protein (sp|Q92558). Does the protein share significant similarity with human DCC?

e) Based on the BLASTP and FASTA results, can any general observations be made regarding the putative function or cellular role of DCC? Describe what possible functions of this protein may be in the cell, based on all of the significant hits in the BLASTP and FASTA results. [BO2005]

Problem Six

The article “Gene Regulation in Hematopoiesis: New Lessons from Thalassemia” by Douglas R. Higgs*, appeared in the Hematology in 2004.

Please read the article available at

<http://www.cs.sjsu.edu/faculty/khuri/CS123B/higgs2004.pdf>

and explain in your own words the results presented in

- a) Figure 1, make sure to explain what every box on the chromosomes represents, including LCR, HS
- b) Figure 2, make sure to explain each part separately (a, b, and c) including the meaning of hematopoiesis
- c) Figure 3, make sure to explain the meaning of synteny, cis-elements, SRO, and PAC
- d) Table 1. The table has 15 items under Location. Choose 5 and fully explain what they represent in the table.