

Name:

Gene:

Task 3: *Gene Finding*

Your next assignment (“*if you agree to accept it*” – although we will certainly disavow all knowledge of you in the future!) is to analyse the yeast genome surrounding your gene for putative transcriptional units and compare this with the published map.

Our suggestion is that you start by using the *GeneMark* algorithm which was developed initially for finding bacterial genes, but has been modified with a “learning set” to also search for genes in many other organisms including yeast. The program can be found in a number of sites, e.g.

<http://opal.biology.gatech.edu/GeneMark/genemark24.cgi>

<http://www.ebi.ac.uk/genemark/>

Step 1: Recover 1 kb of DNA sequence on either side of your gene ORF using the pull-down menu option on the Yeast Genome Database after entering your gene ID number:

[http:// www.yeastgenome.org](http://www.yeastgenome.org)

Step 2: Convert the sequence to FASTA format and paste it into the GeneMark program interface.

Step 3: Depending on which web site above you choose, the options are represented slightly differently. Make sure you select yeast or *S.cerevisiae* for the species/organism. Try various options and outputs until you find one that you can readily interpret.

Step 4: Evaluate your output - in particular paying attention to whether or not the algorithm correctly identifies your gene as well as those reported as flanking it on either the Watson or Crick strands.

Step 5: Depending upon the distance of adjacent genes from your gene as indicated on the YGD genomic browser, you may want to increase the number of nucleotides recovered in association with your gene (2kb or more). You need to ensure that the sequence includes these ORFs and their putative promoter regions.

Scan your data output for any or all of the following features that characterize a eukaryotic transcriptional unit:

- promoter elements such as the TATA box
- transcription initiation site
- 5' UTR
- translational initiation site conforming to the Kozak consensus sequence (nnRnnAUGG)
- 3' UTR with poly(A) addition signal sequence (AATAAA) - recall that the RNA primary transcript is cleaved ~30 bases downstream to form the pre-mRNA

if your gene contains intron(s) (few yeast genes do), the donor and acceptor splice sites, the branch A site and their associated consensus sequences

Your report should include a comparison of your data with that presented in the yeast database genome browser. Please **provide an annotated drawing** of the region of the yeast genome surrounding your gene (remember BOTH strands) with the ORFs and any other pertinent gene identifiers or signatures clearly marked.

To what extent does the GeneMark algorithm validate the published gene arrangement on the yeast chromosome surrounding your gene?

Are any putative genes missed or novel ones proposed by Gene Mark?

What is the effect of adjusting the parameters of the algorithm (step size, window size, detection threshold) from the default settings?

Now retrieve the DNA sequence for the *E.coli* (or other prokaryote) homologue of your gene that you discovered during your Milestone 2. This may require a little creativity and searching on your part using *Entrez* from NCBI or one of the other Sequence Retrieval Systems available:

<http://www.ncbi.nlm.nih.gov/Entrez/>

<http://www.sanger.ac.uk/>

<http://www.genome.ad.jp/dbget/>

or alternatively, going directly to the *E.coli* database, where you can also retrieve DNA sequence both upstream and downstream of the ORF for your yeast gene homologue:

<http://www.genome.wisc.edu/>

Once you have retrieved the bacterial genomic sequence, paste it into the GeneMark program interface as specified in the Steps above for your yeast gene. Change the pull-down organism profile accordingly and then run the algorithm.

See if GeneMark detects your gene homologue and whether you can recognize in the sequence any of the following features specific to prokaryotic transcriptional units:

-35 sequence

-10 sequence (TATA region)

transcriptional termination sequence (complementary GC rich sequence followed by a run of A residues)

ribosome binding site in the 5'UTR (Shine-Dalgarno sequence) 5-15 bases before the ATG initiator methionine codon

If you do not find a putative promoter associated with your prokaryote homologue, you may need to search further upstream or downstream of the gene, since it may be part of an operon transcribed from a distant promoter as part of a polycistronic mRNA.

If your prokaryote gene is part of an operon, then you should search for and report the chromosomal location of at least one of the associated genes in yeast. Use the SGD to find the yeast gene equivalent(s) from the prokaryote operon.

Due date: Wednesday, March 2nd 2005

Please e-mail to:

btjaden@wellesley.edu
dwebb@wellesley.edu