Lab 8 – Multiple Sequence Alignments and Gene Variants

This week we're going to continue investigating the gene you're studying for your final project. We'll start by generating a multiple sequence alignment of nucleotide sequences, and of amino acid sequences homologous to your gene. We'll also explore the codon bias of the organism you study is focusing on, estimate the type of selective pressure your gene is under by calculating the ratio of synonymous and non- synonymous mutation, and note sequence and epigenetic modifications found in your gene.

Multiple Sequence Alignment

First, gather at least 10 sequences which you consider to be homologus to your sequence (perform a BLAST search or download from NCBI's HomoloGene if there is an entry for your gene). http://www.ncbi.nlm.nih.gov/homologene

Using Ugene, perform a multiple sequence alignment using the homologous nucleotide sequences. Experiment with adjusting the parameters. Summarize the parameters you chose and justify your choices in a short paragraph (submit in your lab write-up). Save your best nucleotide alignment as a .aln file. View the alignment in Ugene, experiment with viewing the alignment under different color scenarios. Save several images of your alignment.

Repeat the same steps for your amino acid sequences.

Include this information in your lab write-up:

- Compare the nucleotide and protein alignments. Are there regions of your nucleotide alignment that appear to be conserved? Note the position of these regions. Are there regions of your amino acid alignments that appear to be conserved? Note the position of these regions. Do the regions of conservation in the amino acid alignment correspond to known conserved domains?

Codon Bias

Remember that the degeneracy of the genetic code means that several different codons can encode the same amino acid. The codons for a given amino acid may not be used with equal frequency, and the frequency with which particular codons are used vary between different species. The codon usage of different species and different coding regions (CDS's) within those species are summarized in a Codon Usage Database (accessible at http://www.kazusa.or.jp/codon/).

Search for the organism you are most interested in, using the organism's scientific name. Note that this site is very sensitive to misspelling and case. Copy the species' codon frequency table to a text document.

Include in your lab write-up:

- A copy of the codon frequency table for the species
- The number of CDS's and codons used to generate the species' codon frequency table

From the species codon table page, search for the CDS encoding your protein. Try several variants of your gene's full name. The program may find many CDS's for your gene, if so, click on the "Sum up codon usage" button and copy the CDS specific codon usage table into a text document. If you cannot find an entry for your gene search Homo sapiens for the prion protein.

Include in your lab write-up:

- A copy of the codon frequency table for your gene's coding region
- The number of CDS's and codons used to generate the codon frequency table for your gene's coding region
- Are there differences between the general codon frequency table for the species and the codon frequency table for your gene's coding region? For which amino acids?

Selective Pressure

Mutations in a coding nucleotide sequence which do not result in a change in the amino acid encoded are called synonymous mutations (Ks). Mutations in coding nucleotide sequence which result in a change in the amino acid encoded are non-synonymous mutations (Ka). Differences in the number of synonymous and non-synonymous mutations can be used to estimate the selective pressure on a gene. You may recall from lecture that this is measured as a ratio referred to as the Ka/Ks or dN/dS ratio. Refresh your understanding of the following terms and define them in terms of the value of the Ka/Ks ratio (use these terms and their definitions in your lab write-up):

- Positive selection
- Negative selection
- Neutral selection

Obtain the coding sequences of your homologous sequences in FASTA. Upload coding sequences to the online Ka/Ks Calculation Tool (available at http://services.cbu.uib.no/tools/kaks). Save the phylogenetic tree generated (compare it to the trees generated in the lecture homework) and the Ka/Ks values calculated for each branch of the tree (include and interpret these results in your lab write-up).

Sequence Variants

Visit dbSNP and search for interesting variants of your gene. If your gene doesn't have a dbSNP entry, use the prion related protein (PRNP) for the remainder of this exercise.

Open a variant entry and explore the record - many sections may contain information relevant to your final project (e.g. 3D structure mapping under NCBI Resource Links, Population Diversity).

Look at your gene model in the Genome Viewer. The GenomeViewer is organized in layers - sometimes called "tracks". You many notice your gene model is made up of thick and thin sections. What do these represent? Several of the other tracks there may be

colored boxes with numbers in them (if your gene doesn't have many, look at your teammate's genes). What do the colored boxes with numbers in them represent? Look for Association Results, Clinical Channel, and Cited Variants. Are the boxes distributed across the gene? Are they mainly within introns or exons? Are the boxes clustered in one region of the gene? (explain your findings in your lab write-up, pay especially close attention to clustering of variant in particular regions of your gene).

Find the Gene Models section. Copy the information in the gene model table into an excel sheet.

Are there other dbSNP entries for your gene (or PRNP)? If so repeat the steps above for another entry.

Include in your lab write-up:

- A copy of the gene model table
- Describe the types of mutations found in your gene. Did you find any mutations associated with particular phenotypes, disease susceptibility or pathogenicity? Note where these mutations occur relative to your nucleotide multiple sequence alignment.

Epigenetic Features and Variants

Open NCBI's Epigenomics viewer. NCBI only supports genome viewer data for the human genome, the genome of the model plant Arabidopsis thaliana, the genome of the nematode Caenorhabditis elegans, the genome of the fruit fly Drosophila melanogaster, and the mouse genome. If your gene or a homolog thereof is not found in these genomes, search the human genome for the retinobasltoma protein (pRb).

Explore the default tracks, noting the position of any CpG islands and epigenetics features. Then look at the Configure Page section, turn an interesting Variation track – for example try the SNP and Clinical Channels. Zoom in on the chromosome and note the position of any CpG islands and clusters of epigenetic modifications.

Include in your lab write-up:

- Are there several tissue types or experiments listed among the epigenomic features? If so note variation in the location of the epigenomic feature and any differences by tissue type or disease state .
- Are there CpG islands in the proximity of your gene? Note the position of these features, and the portion of the gene they are located in or near.

Summary

Summarize you findings by considering the position of the various sequence and epigenetic variants relative to the regions of your gene's nucleotide and protein multiple sequence alignments. Be sure to save all of this information as it will be very useful for your final paper and for upcoming exercises.