HST.508 **Problem Set 2** **Due Nov 19, 2005**

1.Kimura 2-parameter model.

Consider the two-parameter Kimura model with rate of transition α and rate of transversion β.

1) Consider a position that has nucleotide A at time *t=0*. Calculate probabilities of A, T, C and G at this position at time *t>0.* Start with an equation $dP/dt=\lambda P$, where λ is the matrix of transitions. Solve this linear system to obtain sought probabilities.

2) Consider two homologous sequences of length *L* that have *P* transitions and *Q* transversions. Calculate the number of substitutions *K* that have occurred between these two sequences.

2.Number of gapped alignments.

If the distance between two sequences depends on the number of types of pairwise matches in an additive fashion, then dynamic programming provides a fast way of finding the optimal alignment. However, certain applications require more complicated distance functions (that are non-additive in the number of pairwise matches) and dynamic programming cannot be used. For these applications smart (approximate) optimization techniques must be used because evaluating the score of every possible alignment to find the optimal one is computationally prohibitive because the total number of possible alignments is very large.

-Calculate the number of possible alignments between two sequences of length *L1* and *L2*, such that the number of matches is *l.*

-By summing over all possible values of *l*, obtain the total number of possible alignments.

3.MacDonald – Kreitman (MK) test.

Use the Seattle SNPs data (your dataset for PS1 http://pga.mbt.washington.edu/) to compare polymorphisms in humans with human-chimp divergence. Pick five genes from the SNP dataset (e.g. genes with many SNPs, genes with non-synonymous SNPs, genes with interesting function; your choice). Fetch their sequences and BLAT them against chimp genome http://genome.ucsc.edu/cgi-bin/hgBlat . Make counts and fill an MK 2x2 table and perform the MK test. Rationalize your results.

4.Genomic evolution of bacterial biosynthetic pathways.

We discussed that chromosomal proximity and phylogenetic profiles can be used to infer which genes are functionally related (e.g., genes of the same biochemical pathway). Although these arguments work well on average individual pathways may have peculiar genomic organization and evolution. Such peculiarities may suggest that a pathway is not inherited and regulated as a single module. Here your goal is to explore genomic organization (i.e. chromosomal proximity, gene

order) and genomic evolution (i.e. phylogenetic co-occurrence) of any two of the following bacterial pathways: purine metabolism, biosynthesis of branched amino acids, histidine biosynthesis, cysteine biosynthesis.

For each pathway:

-Use the KEGG database to identify the pathway in E.coli

-Using KEGG and its links to genomes find the relative location of the participating genes in the E.coli genome

-Identify orthologs (genes with the same function in different organisms; KEGG has the appropriate links to this information) in other *related* genomes.

1)Do genes of a given pathway tend to be close to each other on the chromosome (you may want to calculate expected distance between a random pair of genes and bring up some statistical arguments to support your claim)?

2)Do genes cluster in other related genomes?

3)Do you see same genes close to each other, or do you see some degree of shuffling? Is gene order preserved?

4)Can you see some pattern in phylogenetic co-occurrence?

5)Most of the genes involved encode to metabolic enzymes, but you may be able to see a few genes that cluster together with the enzymes that are not enzymes. What are these genes? Why do they cluster with enzymes of a particular pathway?

6)Identify transporters (see ABC transport pathways in KEGG) that are involved in the pathway. These are genes that transport precursors or products of a pathway in and out of the cell (e.g. cysP transporter and cysteine biosynthesis). Do transporters cluster with metabolic enzymes? Do they show a pattern of phylogenetic co-occurrence?

You are welcome to use other web recourses if you find them useful.