

---

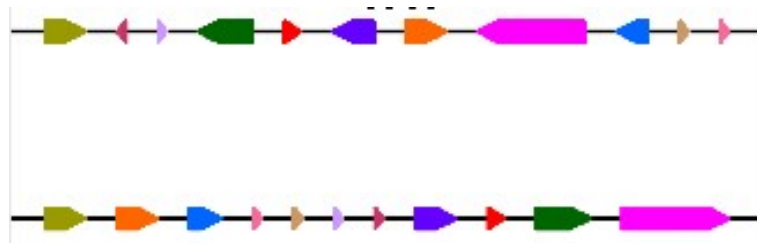
# CS/ST 480

## HOMEWORK 5 (DUE MAY 3RD)

---

### 1. Genome Rearrangements [20 pts].

Comparison of the mouse and human X-chromosomes yields 11 blocks that are conserved between the two species. The orders of these blocks in the two genomes are:



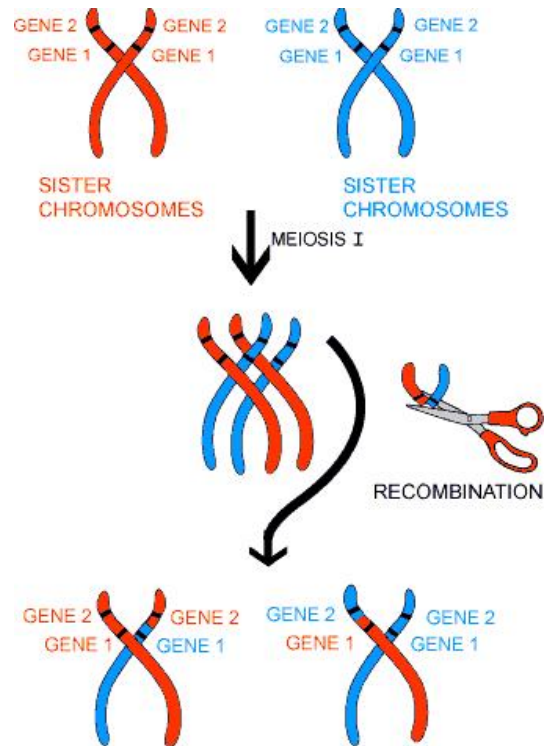
Blocks of the same color have an established correspondence between them. This figure can be represented by following signed permutations:

Mouse	1	-7	6	-10	9	-8	2	-11	-3	5	4
Human	1	2	3	4	5	6	7	8	9	10	11

Ignore the sign of the permutations, and find a series of reversals that transforms the mouse X-chromosome into the human X-chromosome using the SimpleReversalSort and ImprovedBreakpointReversalSort methods presented in class. Find a bound on the number of reversals using the cycle decomposition method. Finally, what can you say about the reversal distance between these two permutations on the basis of these results?

### 2. Evolution by recombination [20 pts].

Every person has two copies of each chromosome (with the exception of the sex chromosomes): one member of each pair is inherited from the father and one from the mother. The members of a pair of chromosomes (known as homologous chromosomes) contain corresponding sets of genes, but each pair of genes may not be identical. During the formation of sex cells (egg and sperm), homologous chromosomes exchange segments by a process called *recombination*. In this exercise we will address the question of reconstructing the history of a population of sequences generated by this process.



Let  $s, t$  be two sequences of length  $n$ . An *equal cross-over* of  $s$  and  $t$  at position  $i$  can generate either  $s[0 : i]t[i : n]$  or  $t[0 : i]s[i : n]$  (the notation follows python notation:  $s[i : j]$  is the concatenation of  $s[i], \dots, s[j - 1]$ ).

We assume a model of evolution by equal cross-over events. Initially the population consists of two sequences that generate a third sequence by cross-over. At each step, two sequences out of the population generate a new sequence by cross over. We are given a population of  $m$  sequences generated by this process. Propose an algorithm that discovers the two ancestor sequences in the population. Assume you have at your disposal an algorithm that given sequences  $s, t$  and  $u$  determines whether  $u$  was generated from  $s$  and  $t$  by equal cross-over. What is the running time of your algorithm?

**3. UPGMA and NJ: part 1 [10 pts].**

A phylogeny is to be constructed for 5 species: A, B, C, D, E. The following dissimilarity matrix has been constructed by the scientist based on nucleotide sequence information:

	A	B	C	D	E
A	0	2	16	16	16
B		0	16	16	16
C			0	12	12
D				0	6
E					0

- (a) Construct a phylogenetic tree for these species using UPGMA. Calculate the edge lengths and display them on your tree diagram. Show the root node on your tree diagram.
- (b) Construct an (unrooted) phylogenetic tree for these species using Neighbor Joining. Calculate the edge lengths and display them on your tree diagram.
- (c) Does the tree constructed using NJ have the same topology as the tree obtained by UPGMA if you view the UPGMA tree as an unrooted tree (remove the root node and redisplay the unrooted tree)? If the two unrooted trees agree then you are done. If they don't agree, explain which one of the trees is more likely to be the true tree and how you arrived at your conclusion.

**4. UPGMA and NJ: part 2 [10 pts].**

As in part 1, you are given the following dissimilarity matrix:

	A	B	C	D	E
A	0	9	5	8	9
B		0	12	15	16
C			0	5	6
D				0	3
E					0

- (a) Construct a phylogenetic tree for these species using UPGMA. Calculate the edge lengths and display them on your tree diagram.
- (b) Construct an (unrooted) phylogenetic tree for these species using Neighbor Joining. Calculate the edge lengths and display them on your tree diagram.
- (c) Does the tree constructed from NJ equal the tree obtained by UPGMA if you view the UPGMA tree as an unrooted tree (remove the root node and redisplay the unrooted tree)? If the unrooted trees agree then you are done. If they don't agree, explain which one of the trees is more likely to be the true tree and how you arrived at your conclusion.

**5. UPGMA and NJ: part 3 [10 pts].**

Consider the following dissimilarity matrix for a group of 4 species: A, B, C, D.

	A	B	C	D
A	0	120	51	52
B		0	53	54
C			0	55
D				0

- (a) Construct a phylogenetic tree for these species using UPGMA. Calculate the edge lengths and display them on your tree diagram.

- (b) If possible, construct an (unrooted) phylogenetic tree for these species using Neighbor Joining. Calculate the edge lengths and display them on your tree diagram. Are you surprised by your results? Under what circumstances can you guarantee that such surprising results will not occur?
- (c) Would you have confidence that the UPGMA method has actually reconstructed the true tree? Explain why?

6. **Snake Phylogeny [30 pts].**

Portions of two mitochondrial genes (12S and 16S ribosomal genes) for thirty-six species of snakes, representing nearly all extant families, are available for download (see the course webpage or WebCT). The 12S and 16S sequences have been concatenated, in that order, so that there is a single sequence for each snake species. Three additional sequences (two lizards and one tuatara) are also included in the sequence file. A phylogenetic tree is to be constructed for these species. The non-snake sequences are included to help with rooting the constructed phylogenetic tree. The reference below provides additional details.

Obtain a phylogenetic tree for these species. The following steps give one way of doing this.

- (a) Use the publicly available program CLUSTALW (or a more user-friendly version CLUSTALX) to perform a multiple alignment of the given sequences.
- (b) Use the publicly available program PHYLIP (this is actually a suite of programs) to compute a pairwise distances matrix using the default distance measure. (We will briefly discuss the distance measure choices in class.)
- (c) Use the UPGMA option of the NEIGHBOR-JOINING program from PHYLIP to obtain a phylogenetic tree. You can use DRAWGRAM or DRAWTREE programs from PHYLIP to produce a high resolution tree diagram for your result. Alternatively, you can also use the publicly available program TREEVIEW. The advantage of TREEVIEW is that it has built-in facilities for editing the tree (rearranging the different branch positions so that the tree is more easily interpretable).
- (d) Repeat the previous step, but now use NEIGHBOR JOINING option of the neighbor joining program from PHYLIP. Compare the phylogenetic trees you have constructed with each other and also with the tree given in the journal article. Add relevant biological comments regarding your results.

**Reference:** *“Higher-Level Snake Phylogeny Inferred from Mitochondrial DNA Sequences of 12s rRNA and 16s rRNA Genes”* by Heise et al (available for download from WebCT).

**Software:** you can either run CLUSTALW on your own machine (software available at: <ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/>, <ftp://ftp.ebi.ac.uk/>

## HOMEWORK 5

---

pub/software/mac/clustalw/, or <ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw/>, depending on your OS), or use one of many CLUSTALW servers (e.g. <http://www.ebi.ac.uk/clustalw/>). Phylip can be downloaded from <http://evolution.genetics.washington.edu/phylip/getme.html>.