**MCB 5472 Assignment #5: RBH Orthologs and PSI-BLAST**

**February 17, 2014**

Today’s assignment will expand on last week’s work by using BLAST to count ortholog families between two genomes using the Reverse BLAST Hit method. Besides being generally useful, the purpose of this experiment is to help you further understand hashes and how useful they are. The second part is a brief exercise comparing blastp, psi-blast and tblastn for sequence detection. The purpose of this part is to highlight how these methods ask slightly different questions of the data and to get you to think about why the results differ between them.

**Before the start of next week’s class:** Tabulate the results of each exercise and send them to me ([jonathan.klassen@uconn.edu](mailto:jonathan.klassen@uconn.edu)) along with a short 1-2 sentence interpretation of then, plus the exact copies of representative terminal commands that you used to answer the question, i.e., perl and/or BLAST+, sufficient for me to be able to reproduce your results. For question 2, please send your input sequence as well and a list of the genomes that you queried it against.

**Question 1: [10 marks]** Determine the number of orthologous proteins shared between our two test complete genomes, E.coli O104:H4 2009EL 2050 and E.coli O104:H4 2009EL 2071 using the RBH method, and compare this value to the number of proteins unique to each genome. You should consider the appropriate alignment and sequence similarity or E-value thresholds to use to identify orthologs between two strains belonging to the same species. This will require that you perform two BLASTp searches using all proteins in each genome where the query genome for one search becomes the reference genome in the other. You will use hashes to parse these BLAST results and to identify and count orthologs.

**Question 2: [5 marks]** Download one or more complete genome sequence(s) of your choice and their proteins (this can be our E.coli) and search it/them for sequences belonging to a molecular parasite (e.g., transposon, integrase, phage coat protein; just make sure that it infects the type of genome that you are searching). Compare the following search methods: (i) BLASTp; (ii) PSI-BLAST using a pssm that you constructed using NCBI’s nr database; (iii) tBLASTn using that same pssm.

*Note 1: Although I did not cover this in Monday’s lecture, tblastn is a way to search nucleotide sequences using a pssm. Note that this only works for pssms built from protein sequences, unlike HMMs that can also handle models built from nucleotide sequences. Syntax:* tblastn -in\_pssm integrase1.pssm -db all.fna -evalue 1e-5 -num\_threads 2 -out psitblastn.out -outfmt 6 -comp\_based\_stats 1

*Note 2: Note the* –num\_threads *flag in the above example. This parallelizes the BLAST algorithm to distribute its computations over multiple nodes, very useful in practice.*

*Note 3: Also note the* –outfmt 6 *above, which gives you the same table as* –outfmt 7 *like we used last week but without the table headers and comment lines. This lets you count hit directly from the terminal using the “*wc*” (word count) command. Using “*wc –l*” just prints the number of lines in your file.*