**Outline: Orthologs Project**

**PART A:** The algorithm needs to be designed so that all of the human accession numbers can be sent to multiple different nodes using UNIX. Each node will BLAST a different species. Then, each node will implement the code as described below in PART A with the output being a separate ortholog file for each species.

1. **Gathering Accession Numbers**
   1. **Create files that will be accessed from within the home directory. Inside of an active directory (subdirectory of home directory) create subdirectories for each gene, where BLAST reports will be stored. Store Master file in Home Directory**
      1. File with 2 columns: gene name and accession number for Homo sapiens
         1. Homo\_sapiens\_Accession.csv
         2. Parse file (BLAST input, gene for BLAST report file names, and BLAST report parsing) and make a list that will be used to store the best orthologs in Master\_file one row at a time
      2. File with 2 columns: gene name and accession number for Macaca Mulatta
         1. Macaca\_mulatta\_Accession.csv
      3. File with 1 column: organism name for ortholog search (Genus and species)
         1. Organisms.csv
         2. Use file to make a list that will be used for parsing (BLAST input, BLAST report file names, BLAST report parsing, and Master\_file column header)
      4. Make sure to keep up with the proper directory for each task.
   2. **BLAST Human accession numbers (**NCBI servers? Ole Miss SC?/Local?)
      1. *result\_handle1 = NCBIWWW.qblast("blastn","refseq\_rna",Accession[1], entrez\_query = str(Org) +'[Organism]', hitlist\_size=10)*
         1. *blastn algorithms search nucleotide databases using a nucleotide query*
         2. *refseq\_rna database*
         3. *Accession[1] – Homo sapiens accession number*
         4. *Entrez\_query – use field to BLAST HS accession number against a particular organism*
         5. *Hitlist\_size – only get a list of 10 orthologous genes*
   3. **Handle the BLAST report**
      1. Save the BLAST report as an XML file and begin parsing
      2. Using the SearchIO feature (experimental), save the accession number for the hit with the best score and the accession that is not a pseudogene (XR). Also make sure the gene name is the same as the HS gene name. If they are not, then flag the accession number (flag is lower case).
      3. Append the accession number to the gene\_list. After all the organisms have been BLASTED, append the gene\_list to the Master\_File. This will add one (gene) row to the file.

**PART B**: The algorithm needs to be designed so that all the accession numbers of each gene are sent out to different nodes for alignment and analysis using UNIX. The different nodes will be using the code described in PART B. In order to do this an intermittent step will need to be included in which a Master Ortholog file is built so that the accession numbers of each gene can be accessed from a central location. Separate alignment files will also be created.

1. **Extracting features from genbank files and aligning those features**
   1. Search the non-human orthologs of each gene in NCBI
   2. Download in Genbank format.
   3. Use the Homo sapiens Genbank file to check against the other orthologs to see if they are the right gene.
      1. Remove the wrong genes
   4. Use CDS feature to locate the coding sequence
      1. Other features: *SeqFeature.type*
      2. Location: *SeqFeature.location*
   5. Build a separate file for each ortholog (for every sequence)
      1. Naming convention: *‘gene’\_‘species’\_‘Refseq#’.txt*
   6. Translate into protein (amino acid) sequence
   7. Align the CDS features (nucleotide and amino acid sequences) using Clustal Omega
      1. Create a profile to align against
         1. Homo sapiens should be the profile
      2. Align other species against the profile alignment using multiple iterations
      3. Save alignment files in a format to be used later
         1. .phylip formatted files can be used with the Phylip and PAML programs
         2. .aln is a clustal format
         3. .fasta can also be the output format
   8. Other important fragments to align
      1. Intracellular regions
      2. Extracellular regions
      3. Ignore the transmembrane
         1. They are mostly the same
      4. N terminus and ?IC2?
      5. C Terminus and IC3
         1. Signaling
2. **Phylogenetic Analysis**
   * 1. PHYLIP
        1. Input a phylip formatted alignment file
        2. Perform maximum likelihood analysis and generate distance matrices
        3. Output will be in the form of a tree or newick format
     2. PAML
3. **Polymorphisms**
   1. Find SNP data in humans and compare to available Rhesus SNP data
   2. Dr. Vallender has several Rhesus genomes that he has sequenced in order to get evolutionary data from.