# CS 3824
# Project Description

**Given:** February 5, 2010 **Due:** May 4, 2010

**Introduction.** This is a project to estimate the mutation rates of proteins in yeast species. The project is to be accomplished as a team (teams will be assigned soon) with everyone on each team working together towards the overall goal. Each of you will write up a report on the project separately for grading, due on May 4, 2010. The steps in the project are detailed below.

**Yeast Phylogeny.** In the paper

> Estimating the tempo and mode of gene family evolution from comparative genomic data. Matthew W. Hahn, Tijl De Bie, Jason E. Stajich, Chi Nguyen, and Nello Cristianini. Genome Research 15, pp. 1153–1160, 2005,

there is a phylogenetic tree for five contemporary yeast species — *Saccharomyces cerevisiae, Saccharomyces paradoxus, Saccharomyces mikatae, Saccharomyces kudriavzevii,* and *Saccharomyces bayanus* — that includes the branch lengths in millions of years. Use this tree when you estimate mutation rates below.

**Yeast Proteins.** Use the Entrez Protein interface

`http://www.ncbi.nlm.nih.gov/sites/entrez?db=Protein&itool=toolbar`

at NCBI[1] to find a random protein in one of the yeast species. For example, accession AAT42142 is a protein in *Saccharomyces bayanus.* Use the Protein BLAST interface

`http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome`

to verify whether that protein in the other four yeast species is present in the NCBI protein database. If so, you will get BLAST reports with very small E-values and very good alignments. If not, try a different random protein.

**Alignments.** You will need a multiple sequence alignment (MSA) for the five protein sequences you have. This can be done with BLAST. More conveniently, it can be done with an MSA tool like MUSCLE

`http://www.drive5.com/muscle/`.

What you can see in the multiple sequence alignment is those positions (columns) where mutations have occurred and the pattern of those mutations.

**Estimating Mutation Rates.** Use the pattern of mutations that you found to estimate the mutation rates (amino acid changes per million years) on each branch of the yeast phylogenetic tree. Also, estimate an average mutation rate for this protein.

[1] NCBI databases can also be accessed through a programming interface called eUtils: `http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html`.

**Iterating.** Now that you have the hang of it, automate the entire process of selecting a random protein, verifying that it is in all five species, generating a multiple sequence alignment, and estimating mutation rates. Use your software to select 25 more random proteins and to estimate mutation rates for each one.

**Report.** Write a report (preferably in LaTeX) detailing what you did in the project and what results you obtained. Be certain to clarify what part of the work of your team you actually did. Include a table or graph of the average mutation rates you obtained. There is likely to be a large spread of mutation rates. Hypothesize why that might be so (you are free to use the literature to help you here). The report should be 4–5 pages long to be comprehensive. Feel free to add any software you wrote as an appendix. Submit the report by 4:00 PM on May 4, 2010.