Chapter 9  
Protein conservation

The budding yeast, *Saccharomyces cerevisiae*, and the fission yeast, *Schizosaccharomyces pombe*, are separated by close to a billion years of evolution. Our goal for the semester is to determine if the proteins involved in methionine and cysteine biosynthesis have been conserved between these evolutionary divergent species. In this lab, you will use bioinformatics tools to study the evolution of these protein sequences.

Objectives

- Learn the one letter code for amino acids

- Understand how amino acid side chain chemistry is reflected in the BLOSUM62 matrix

- Use the BLASTP algorithm to compare protein sequences

- Prepare a multiple sequence alignment to identify conserved regions in a protein
As species evolve, their proteins change. The rate at which an individual protein sequence changes varies widely, reflecting the evolutionary pressures that organisms experience and the physiological role of the protein. Our goal this semester is to determine if the proteins involved in Met and Cys biosynthesis have been functionally conserved between S. pombe and S. cerevisiae, species that are separated by close to a billion years of evolution. In this lab, you will search databases for homologs of S. cerevisiae sequences in several species, including S. pombe. Homologs are similar DNA sequences that are descended from a common gene. When homologs are found in different species, they are referred to as orthologs.

Homologs within the same genome are referred to as paralogs. Paralogs arise by gene duplication, but diversify over time and assume distinct functions. Although a whole genome duplication occurred during the evolution of S. cerevisiae (Kellis et al., 2004), only a few genes in the methionine superpathway have paralogs. Interestingly, MET25 is paralogous to three genes involved in sulfur transfer: STR1 (CYS3), STR2 and STR4, reflecting multiple gene duplications. The presence of these four distinct enzymes confers unusual flexibility to S. cerevisiae in its use of sulfur sources. The SAM1 and SAM2 genes are also paralogs, but their sequences have remained almost identical, providing functional redundancy.

Protein function is intimately related to its structure. You’ll recall that the final folded form of a protein is determined by its primary sequence, the sequence of amino acids. Protein functionality changes less rapidly during evolution when the amino acid substitutions are conservative. Conservative substitutions occur when the size and chemistry of a new amino acid side chain is similar to the one it is replacing. In this lab, we will begin with a discussion of amino acid side chains. You will then use the BLASTP algorithm to identify orthologs in several model organisms. You will perform a multiple sequence alignment that will distinguish regions which are more highly conserved than others.

As you work through the exercises, you will note that protein sequences in databases are written in the 1-letter code. Familiarity with the 1-letter code is an essential skill for molecular biologists.

Amino acid R groups have distinct chemistries

Each of the 20 amino acids commonly found in proteins has an R group with its own distinctive chemistry. R groups differ in their size, polarity, charge and bonding potentials. When thinking about evolutionary changes in proteins, it is helpful to group the amino acids by their chemistry in a Venn diagram, shown on the opposite page. In general, replacing one amino acid with a second amino acid from the same sector can be considered a conservative change. Size is also important. R groups vary considerably in the bulkiness of their chains. Substitution of a large R group for a small one can significantly alter the function of a protein.
Exercise 1 - The 1-letter code for amino acids


1. Under the amino sequence below, write the same sequence using the 1-letter code.

   Met-Glu-Asn-Asp-Glu-Leu-Pro-Ile-Cys-Lys-Glu-Asp-Pro-Glu-Cys-Lys-Glu-Asp

2. What is the net charge of this peptide? (Assign -1 for each acidic amino acid and +1 for each basic amino acid. Add up the total charges.)

3. How many hydrophobic amino acids are found in this peptide?

4. Write the name of a music group that you enjoy. Then transpose the name into an amino acid sequence written with the 3-letter code. Pass the amino acid sequence to a friend and have him/her decode it. (Note: the 1-letter code uses all of the alphabet, except B, J, O, U, X and Z).
Chapter 9

**BLAST algorithms are used to search databases**

There are many different algorithms for searching databases, but BLAST algorithms are some of the most popular. BLAST is an acronym for Basic Local Alignment Search Tool, which aptly describes its function. The original BLAST algorithms were developed and published by a group of investigators in 1990 (Altschul et al., 1990). Since that time, investigators have refined and diversified the original BLAST programs so that a whole family of BLAST programs is currently available. BLAST programs are widely used because they are able to quickly search through large amounts of sequence data. The key to BLAST’s speed is its use of local alignments that serve as seeds for more extensive alignments. Other algorithms perform global sequence alignments that match the entire sequence of a query sequence against database information. Global sequence alignments are more sensitive than BLAST in finding related sequences, but they require considerably more processing time. BLAST searches are quite adequate for our purposes, and we will use the online BLAST tools available at the NCBI site throughout this course.

The BLASTN and BLASTP algorithms are designed for searching nucleotide and protein databases, respectively, and they use different scoring matrices and search parameters. BLAST algorithms begin by breaking down query sequence into short “words” and assigning numerical values to the words. A scoring matrix is used to assign numerical values to the words. Words and synonyms above a threshold value are then used to search databases. The default word size for BLASTN is 28 nucleotides, while the default word size for BLASTP is 3 amino acids. This difference in word size is because nucleic acid sequences are written in 4 letters (ACGT), while protein

1. The **query** sequence is broken into “**words**” that will act as seeds in alignments

   Query
   
   Words

2. BLAST searches for matches (or synonyms) in **target** entries in the database

   Target
   
   Word match
   
   Word match

3. If a **target** entry has two or more matches to “**words**” from the query, the alignment is extended in both directions looking for additional similarity

   Target
   
   Word match
   
   Word match
   
   Word match

**Overview of the strategy used in BLAST algorithms**
BLASTN and BLASTP use a rolling window to break down a query sequence into words and word synonyms that form a search set. At least two words or synonyms in the search set must match a target sequence in the database, for that sequence to be reported in the results.
Protein conservation

sequence is written with 20 different letters. The probability of obtaining random, but irrelevant, matches are much greater with nucleotide sequences. A stretch of 3 amino acids should arise randomly once in every 8000 tripeptides, compared to once in every 64 trinucleotides. In this lab, we will use the BLASTP algorithm, which is more useful than BLASTN for studying protein evolution, because it overlooks synonymous gene mutations that do not change an amino acid.

In BLASTP, a sequence is broken into all possible 3-letter words using a moving window. A numerical score for each word is assigned by adding up values for amino acids from a scoring matrix. Using the BLOSUM62 matrix (below), BLASTP first finds word with a score of 12 or more in a query sequence. BLASTP then identifies potential synonyms that differ from the word at one position, and synonyms above a defined threshold value are included in the search set. NCBI BLASTP uses a default threshold of 10 for synonyms, which can adjusted by the user. Using this search set, BLAST rapidly scans a database and identifies protein sequences that contain at two words/synonyms from the search set. These sequences are set aside for the next part of the BLASTP process, when these short matches serve as seeds for more extended alignments in both directions from the original match. BLAST keeps a running raw score as it extends the matches. Each new amino acid either increases or decreases the raw score. Penalties are assigned for mismatches and for gaps between the two alignments. Again, users can adjust the gap penalties, which are usually quite severe. In the NCBI default settings, the presence of a

Overview of the BLASTP process.
The query sequence EAGLES into broken into three-letter words or synonyms that are used as a search set against records in a protein or translated nucleotide database. See the text for additional details.
Chapter 9

gap brings a penalty of 11, which increases by 1 for each missing amino acid. Once the score falls below a set level, the alignment ceases. Raw scores are then converted into bit scores by correcting for the scoring matrix used in the search and the size of the database search space.

The output data from BLASTP includes a table with the bit scores for each alignment as well as its E-value, or “expect score”. The E-value indicates the number of alignments with that particular bit score that would be expected to occur solely by chance in the search space. Alignments with the highest bit scores (and lowest E-values) are listed at the top of the table. For perfect or nearly perfect matches, the E-value is reported as zero - there is essentially no possibility that the match occurs randomly. The E-value takes into account both the length of the match and the size of the database that was surveyed. The longer the alignment, and/or the larger the database search space, the less likely that a particular alignment occurs strictly by chance. An E-value of 1 does not necessarily imply, however, that the match has no biological relevance. More nuanced analysis may simply be needed to evaluate the match.

**BLOSUM62 scoring matrix**

The results obtained in a BLASTP search depend on the scoring matrix used to assign numerical values to different words. A variety of matrices are available, whose utility depends on whether the user is comparing more highly divergent or less divergent sequences. The BLOSUM62 matrix is used as the default scoring matrix for BLASTP. The BLOSUM62 matrix was developed by analyzing a large number of protein sequences that were more than 62% identical to each other. Investigators computationally determined the frequency of all 210 possible amino acid substitutions that had occurred in these conserved blocks of proteins, and they used this data to construct the BLOSUM62 scoring matrix for amino acid substitutions. The BLOSUM62 score for a particular substitution is a log-odds score that provides a measure of the biological probability of a substitution relative to the chance probability of the substitution. For a substitution of amino acid \( i \) for amino acid \( j \), the score is expressed:

\[
S_{ij} = \frac{1}{\lambda} \log \left( \frac{p_{ij}}{q_i q_j} \right)
\]

where \( p_{ij} \) is the frequency of the substitution in homologous proteins, and \( q_i \) and \( q_j \) are the frequencies of amino acids \( i \) and \( j \) in the database. The term \( (1/\lambda) \) is a scaling factor used to generate integral values in the matrix.

The BLOSUM62 matrix (below) is consistent with strong evolutionary pressure to conserve protein function. As expected, the most common substitution for any amino acid is itself. Overall, positive scores (shaded) are less common than negative scores, suggesting that most substitutions negatively affect protein function. The most highly conserved amino acids are cysteine, tryptophan and histidine, which have the highest scores. Interestingly, these latter amino acids have unique chemistries and often play important structural or catalytic roles in proteins.
Exercise 2 - The BLOSUM62 matrix

1. Use the Venn diagram on p. 87 to hypothesize why relatively high values (2 and above) occur for the substitutions:
   
   Asp to Asn
   Asp to Glu
   His to Tyr
   Ile to Val
   Lys to Glu

2. Calculate the word score for the following tripeptides. Which would you expect to occur most frequently in a protein database?

   Leu - Cys - Pro
   Gln - Ala - Met
   Asp - Glu - His
In this exercise, you will use BLASTP to find a homolog for your protein in *S. pombe*. Direct your browser to the BLAST tools at NCBI:


- Enter the NP_ number for the *S. cerevisiae* Met protein (p. 45) in the query box
- For the database, select reference proteins
- Enter *S. pombe* for the organism
- Click BLAST

- On the results page, note the graphic summary at the top which gives you an instant idea about the extent and strength of the match with *S. pombe* sequences.

  Does the match extend for the entire length of the protein?

  How many homologs are present in the *S. pombe* genome?

**Cursor down to the alignment statistics.**

Record in the table on p. 94:

- NP_ record number of the *S. pombe* match
- total score
- how much of the query sequence was aligned (coverage)
- E-value for the alignment

**Cursor down to see the actual alignment of the *S. cerevisiae* and *S. pombe* sequences.**

- Take note of the center row in the alignment which summarizes the homology between the protein sequences. If an amino acid is conserved between the two species, its 1-letter code name is shown in that row.

  What do the pluses in the center row indicate?

  Are some regions of the sequences more conserved than others?

  Did BLASTP add gaps to one or both of the sequences in the alignment? If so, are they more common in less conserved or highly conserved regions?

Click on the link to the NP_ record for the *S. pombe* ortholog. Record the EC number for the protein. Is it identical to that for the *S. cerevisiae* enzyme?
Exercise 4 - Multiple sequence alignments

BLASTP gives a pairwise alignment of sequences that is very useful for identifying homologs. Multiple sequence alignments compare a larger number of sequences simultaneously. By comparing a larger number of sequences over a wider evolutionary range, multiple sequence alignments allow researchers to identify regions of a protein that are most highly conserved, and therefore, more likely to be important for the function of a protein. In this exercise, we will study conservation of protein sequences in a number of model organisms that are widely used in genetic studies. Model organisms are easily maintained in large numbers in the laboratory. The genomes for model organisms have been sequenced, and techniques for genetic analysis are well-developed. In addition, database and clone resources are available to support research with model organisms. The organisms below have been selected because they represent important branches of evolution and because they are potential candidates for future research in this course.

**Bacteria** - these represent two major divisions of the bacteria
- *Escherichia coli* strain K-12 (gram negative; K-12 is the standard laboratory strain)
- *Bacillus subtilis* strain 168 (gram positive reference strain)

**Eukaryotes** - model organisms
- *Saccharomyces cerevisiae* - needs to be included in trees and alignments!
- *Schizosaccharomyces pombe*
- *Arabidopsis thaliana* - thale crress; model organism for flowering plants
- *Caenorhabditis elegans* - nematode model organism used in developmental studies
- *Mus musculus* - laboratory mouse

1. Collect the sequence and BLAST data

The first step in a multiple sequence alignment is to collect the sequence data and analyze the BLASTP data that compare the sequences with the *S. cerevisiae* sequence. We will be using the reference sequences for the organisms, which begin with a NP___ number. Since you already know how to find NP____ records and set up BLASTP, we will take some shortcuts to finding the remaining numbers and BLASTP statistics. The accession numbers for the bacterial species will be available on the class website. We will use the Homologene database at NCBI (Sayers et al., 2012) to find the accession numbers for the eukaryotes. Homologene is a database that automatically detects homologs in 20 sequenced eukaryotic genomes.

**Access Homologene at:** http://www.ncbi.nlm.nih.gov/homologene

Note the species that have been searched for homologs. Enter the name of your gene into the search box. This brings up the various Homologene groups that have a gene with that name. If search brings you to a page with more than one Homologene group list, click on the Homologene group that contains the *S. cerevisiae* gene.

**Record the accession number for the Homologene group:**

93
Chapter 9

At the top of the record, Homologene describes the distribution of homologs in eukaryotes. A narrowly conserved protein might only be found in the Ascomycota, while a widely-distributed protein would be found in the Eukaryota.

- What phylogenetic divisions have homologs of your gene?

The records page has links on the left to gene summaries, which you may want to investigate. Links on the right side of the page bring you to the NP___ record. The Homologene output always provides a graphical depiction of protein domains in each of the homologs.

- How many domains are found in your protein?

- Are the domains equally well-conserved between species?

Depending on the phylogenetic distribution of your gene and the possibility of paralogs in a species, you will have different numbers of entries in the table. Record the NP___ numbers for the species listed on the Homologene page. Note: Not all the species listed on the previous page may have homologs! Add the NP_ numbers for E. coli and B. subtilis from the posted data sheet. If you have less than five entries, e.g. the protein is narrowly restricted to Ascomycota, add two additional species of your choice from the Homologene group that contains your gene.

<table>
<thead>
<tr>
<th>Species</th>
<th>NP Accession #</th>
<th>Total score</th>
<th>Coverage</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. pombe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next, perform a pairwise BLASTP alignment for each sequence against the S. cerevisiae sequence. Collecting BLASTP data is easy with Homologene: Use the grey box on the lower hand side of the page to set up each BLASTP comparison. Record the total score, % coverage and E-value for each match.

In the next step, you will prepare a multiple sequence alignment using the sequence information in the NP___ records. Using the BLASTP data, it may be possible to exclude some sequences from further study. The best matches will have high total scores and % coverage (fraction of the two proteins that are aligned) and low E-values. For the rest of this assignment, exclude sequences where the total score is less than 100 and E-values are higher than 1E-10.
2. Prepare the multiple sequence alignment.

Examine the BLASTP results that you recorded for your gene. You will next construct a multiple sequence alignment and phylogenetic tree, using the Phylogeny suite of programs. Phylogeny describes itself as providing “Robust Phylogenetic Analysis for the Non-Specialist.” You will be working with material at two different sites, so you need two operational browser pages. One browser tab should remain at NCBI, where you will retrieve records.

- Direct a browser window or tab to www.phylogeny.fr. Click the “Your workspace” tab and set up an account to store your work.
- Under the heading Phylogeny analysis tab, select One Click. Your sequences will be automatically brought through multiple alignment and phylogenetic tree building algorithms. The advanced option on this page would allow you to adjust the parameters associated with each program, but we will ignore them. We’ll let Phylogeny make the decisions for us!
- Construct a FASTA file containing all the sequences that you would like to compare. You will be pasting files from NCBI records directly into the Phylogeny data entry box. The title line of a FASTA file begins with a “>” character and end with a hard return. The title lines of NCBI FASTA files contain a GenBank reference number (gi_____), the RefSeq accession number (NP_xxxxxx), protein name and the species name in brackets. Edit out all the characters except the NP_xxxxxx name. (You will see why later!) IMPORTANT: Do NOT use a text editor or Word to work with sequences. These programs introduce hidden punctuation that will interfere with Phylogeny programs.
- The first file in our comparisons will be the S. cerevisiae protein sequence. Find the record in the NCBI Protein Database, using the NP_ number as a search term. When the record comes up, click the FASTA link at the upper left side of the record. (Alternatively, you can click to the NP_ record from the Homologene page.)
- Copy the title line (begins with >) and the entire amino acid sequence
- Paste the FASTA record into the Phylogeny data entry box
- Repeat step 3 with each of the sequences that you would like to compare.
- When you are finished, give your project a title, enter your email address (the analysis can take a little time) and click the Submit button. Your results will be posted on a web page.

3. Export the multiple sequence alignment

- Click on the alignment tab to view the multiple sequence alignment.
- Ask for the output in ClustalW format. The output appears on a new web page.
- Right-click on the page and download the Clustal alignment with a new filename that makes sense to you. The page will download as an ASCI text file that you will be able to open with Word or a text editor.
- Open the file in a word processor. Adjust the font size so sequences are properly aligned and fit on the page. Use a non-proportional font such as Courier so all characters line up.
3. **Construct a phylogenetic tree.**

- Click the tree rendering tab to access your phylogenetic tree.
- Use the editing tools to alter the appearance of your tree as you see fit. Pay particular attention to the legends in the “leaves” of the tree, which may have been altered during the rendering processing. For example, if there are no paralogs for the protein, you may want to replace the NP_ numbers with legends that make the figure more comprehensible to a reader.
- Download the file in a format of your choice.

**Discussion questions**
Which homolog would you test after the *S. pombe* homolog? Why?

Which gene has the strongest homolog in *E. coli*, *B. subtilis*, *A. thaliana*, etc.?

Look at the methionine superpathway on p. 57. Which parts of the pathway are the most widely distributed? Which reactions are more highly conserved in bacteria? in plants? in animals?

**References**


Sayers, E.W., Barrett, T., Benson, DA *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucl Acids Res* **40**: D1-D25. (Note: this is an online publication that is updated annually.)