There are many different algorithms for searching sequence databases, but BLAST algorithms are some of the most popular, because of their speed. As you will see below, the key to BLAST’s speed is its use of local alignments that serve as seeds for more extensive alignments. In fact, BLAST is an acronym for Basic Local Alignment Search Tool (Altschul et al., 1990). A set of BLAST tools for searching nucleotide and protein sequences is available for use at the NCBI site. You have already used the BLASTN algorithm to search for nucleotide matches between PCR primers and genomic DNA (Chapter 7). In this lab, you will use the BLASTP algorithm to search for homologs of \textit{S. cerevisiae} Met proteins in other organisms.

BLAST searches begin with a query sequence that will be matched against sequence databases specified by the user. As the algorithms work through the data, they compute the probability that each potential match may have arisen by chance alone, which would not be consistent with an evolutionary relationship. BLAST algorithms begin by breaking down the query sequence into a series of short overlapping “words” and assigning numerical values to the words. Words above a threshold value for statistical significance are then used to search databases. The default word size for BLASTN is 28 nucleotides. Because there are only four possible nucleotides in DNA, a sequence of this length would be expected to occur randomly once in every 428, or 1017, nucleotides, which is far longer than any genome. The default word size for BLASTP is three amino acids. Because proteins contain 20 different amino acids, a tripeptide sequence would be expected to arise randomly once in every 8000 tripeptides, which is longer than any protein. The figure below outlines the basic strategy used by the BLAST algorithms.
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.

In this lab, we will use the BLASTP algorithm, which is more useful than BLASTN for studying protein evolution. Unlike BLASTN, BLASTP overlooks synonymous gene mutations that do not change an amino acid. Synonymous substitutions do not affect the function of a protein and would therefore not be selected against during evolution. BLASTP uses a weighted scoring matrix, BLOSUM 62 (Henikoff & Henikoff, 1999), that factors in the frequencies with which particular amino acid substitutions have taken place during protein evolution.

We will return to this discussion of BLASTP after an introduction and chance to work with the BLOSUM62 matrix.