Sequence Similarity Searching

Take a Class

This guide supports the Galter Library class called **Sequence Similarity Searching**. See our [Classes schedule](#) for the next available offering. If this class is not on our upcoming schedule, it is still available to you or your group by request.

Background

Sequence similarity searches can be done in multiple ways on multiple platforms. BLAST (Basic Local Alignment Search Tool) aligns a sequence to a database of other sequences or can align two sequences to each other. BLAST can be done with nucleotide sequences or with amino acid protein sequences. There are a number of special BLAST variations that allow searching among more dissimilar sequences. Multiple sequence alignment (done by CLUSTALW, COBALT, MUSCLE, ProbCons and others) consists of aligning several sequences to each other.

Why use BLAST?

Specialized BLAST tools are available for many purposes; choosing the right BLAST tool and utilizing it properly can help you:

- Find statistically significant matches, based on sequence similarity, to a protein or nucleotide sequence of interest.
- Obtain information on the inferred function of a gene or protein.
- Help determine whether an implied homology between two sequences is justified
- Find conserved domains in your sequence that are common to many sequences
- Search for sequence motifs or patterns that are similar to a sequence of interest in a particular region
- Compare known sequences from different taxonomic groups
- Limit a search to particular segments of the database such as a particular species' genome
- Search for a protein sequence of interest using a nucleotide sequence as the query and vice versa
- Discover suspected cloning vector sequences in your sequence

The Basis of Sequence Similarity Searches

Most sequence similarity search tools are based on foundations created by seminal pattern matching algorithms.

**Needleman-Wunsch**

This algorithm finds the best *global* alignment between any 2 sequences.
CPU and time-intensive
Often misses domain or motif alignments in sequences, since it disfavors local alignment of highly similar regions

In this example of a Needleman-Wunsch matrix:

Match = +1
Mismatch = -1
Gap = -1

Scores are filled in beginning in the upper left corner at the beginning of both sequences. When matrix is complete, the best value is found in the right-most column, and a path is tracked backwards to the beginning of the two sequences to find the best alignment.

Smith-Waterman

Smith-Waterman is an extension of Needleman-Wunsch that compares segments of all possible lengths (thus creating local alignments) between two sequences to maximize alignment

- Very sensitive search
- CPU and time-intensive
Above is an example of a S/W matrix

Match = + 1
MisMatch = 0.3
Gap = \(1 + 0.3k^*\)
\(^*\) where \(k = \text{number of residues included in gap}\)

All cell values start at zero and are not allowed to fall below zero (so a new alignment path can begin at any point).
Values in cells, like in Needleman-Wunsch, are based upon value of cell plus the highest value in row, column or direct diagonal using gap penalties. FASTA and BLAST are based on this type of comparison matrix.

**FASTA**

FASTA generates local alignments. The algorithm uses lookup tables (hash tables) to increase speed. Sensitivity and speed are determined by the size of the "word" used for the initial lookup table. FASTA builds diagonals in conjunction with the results of the lookup tables.

- Sensitive search
- Fast


**BLAST**

BLAST generates local alignments. It begins a search by indexing all character strings of a certain wordsize within the query sequence by their starting position in the query. BLAST then scans the target database looking for matches between the "words" indexed in the query and strings found within the database sequences. When a word match is found (two nearby words in the case of protein searches), BLAST attempts to extend both forward and backward from the match to produce an alignment. BLAST will continue this extension as long as the alignment score continues to increase or until it drops by a critical amount owing to the negative scores given by mismatches.

- Fairly sensitive search
- Very fast

The Basis of BLAST

The initial search is done for a word of length "W" that scores at least "T" when compared to the query using a substitution matrix.

PAM and BLOSUM - The Matrices of Choice (For Protein sequences only)

**PAM:** Percent Accepted Mutation. A unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit is the amount of evolution that will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have a certain amount (x) of evolutionary divergence. [Taken from the NCBI Glossary - PAM definition].

How PAM matrices were derived:

- Comparison of 71 groups of closely related proteins yielding 1,572 changes. (>85% identity)
- Different PAM matrices are derived from the PAM 1 matrix by matrix multiplication
- The matrices are converted to log odds matrices. (Frequency of change divided by probability of chance alignment converted to log base 2.)
- A PAM 250 matrix has 250 point changes per 100 amino acids. It is similar in stringency to a BLOSUM45 matrix


**BLOSUM:** Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid over-weighting closely related family members. [Taken from the NCBI Glossary - BLOSUM definition]

How BLOSUM matrices were derived:

- Based on comparisons of Blocks of sequences derived from the Blocks database
- The Blocks database contains a multiple alignment of ungapped segments corresponding to the most highly conserved regions of proteins. - (Thus emphasizes local alignment versus global alignment)


BLOSUM62 performed better than all versions of PAM matrices in finding more distant relationships between protein sequences in published comparisons. This is why it is usually the default choice of matrix for most BLAST programs.

---

The Statistics of BLAST

- What's an "E value"?
  - **Expectation value.** The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.
- What's "S"?
  - **Raw Score.** The score of an alignment, S, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (see PAM, BLOSUM). Gap scores are typically calculated as the sum of G, the gap opening penalty and L, the gap extension penalty. For a gap of length n, the gap cost would be \( G + L \). The choice of gap costs, G and L is empirical, but it is customary to choose a high value for G (10-15).
What's a "Bit score"?

The value S is derived from the raw score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

What about Nucleotide Searching?

Whenever possible, it's usually best to BLAST with amino acid sequences using BLASTp. There are 2 reasons for this:

1. BLASTn for nucleotide sequences assumes that all substitutions in base pairs are equal when this is not the truth. The rate of transition mutations (purine to purine or pyrimidine to pyrimidine) is approximately 1.5-5X that of transversion mutations (purine to pyrimidine or vice versa) in all genomes where it has been measured (see Wakely, Mol Biol Evol 11(3):436-42, 1994).
2. Code Degeneracy. Some amino acids are coded by more than one codon (eg. serine is coded by UCU or AGC). This leads to great variation in how the BLAST algorithm may interpret a nucleotide sequence.

However, it's still useful to run BLAST on nucleotide sequences. Treat it like an experiment: try blastn, megablast and blastx or tblastx.

The BLAST Interface

Go to http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

or you can get there from the NCBI home page by clicking on "BLAST" in the popular links menu.

Use MyNCBI to Save BLAST Searches

You can use MyNCBI to save your BLAST results. Click on the MyNCBI link in the upper right corner of any NCBI page to log into MyNCBI where you can manage your searches and collections.
Choosing the Right BLAST

- Choose your BLAST platform from the BLAST main page: protein, nucleotide or translated BLAST, or choose a specialized BLAST
  - Specialized BLASTs have been created to help you find vector contamination, create primers, search specifically for SNPs and many more choices
- When performing a nucleotide BLAST, try experimenting with blastx and tblastn, as well as nucleotide blast (blastn)
- After selecting your BLAST tool, enter your sequence information in the Query Sequence area

Entering Query Sequences

On the BLAST query page, you may notice a number of small question marks (?). If you click on any of these, a line of text will open that explains the field and gives you links to more detail about the section in the BLAST tutorial pages on the NCBI website.

There is a "mystery sequence" you can use for this class. Copy the whole sequence, along with the line that begins with the ">", and paste it into the box on the blastn page. Try it once without changing ANY of the settings (use Megablast),
then try it again with blastn instead of Megablast.

When entering sequences in the query area

- You can enter a copied sequence in FASTA format: characterized by the carat `>` followed by a descriptive line of text.
- You can also use just the gi or accession number for a sequence you find in Entrez Protein or Entrez Nucleotide. The gi number is often consistent across all international databases, but not always, so if you find a sequence outside of NCBI's databases (Entrez), enter the FASTA format in the search box on the BLAST page.
- You can also use the Browse button to search for a FASTA formatted file on your computer to upload.

Choosing a Database to Search (Nucleotide BLAST)

For nucleotide BLAST, you have many genomic and nucleotide databases to choose from.

- Choose genomic and transcript databases for mouse or human
- Choose the non-redundant nucleotide databases (nr/nt) for a search of ALL available nucleotides in ALL species
- Choose a subset of nucleotide databases such as reference sequences, expressed sequence tags, high-throughput sequence data and many more

![Choose Search Set](image)

- You can use the Organism region to select a specific species to search. Just start typing a species in the box and the "smart index" will suggest species to choose from. You can also choose to exclude species using this region by checking the Exclude checkbox.
- Use the Exclude region to exclude predicted model sequences or unfiltered environmental samples
- Use the Entrez Query region to define fields or molecule types to search using square brackets and boolean operators (click the ? and select more in the pop-up to get more information on constructing queries.
- For our example query, just choose the nucleotide collection (nr/nt)

Choosing the Type of BLAST Program (Nucleotide BLAST)

- Megablast is good for searching for genes that have high similarity between species, and is the default for blastn
- Discontiguous megablast is good for VERY dissimilar sequences
- Blastn is the original nucleotide BLAST and a good place to start when you are not sure if you are going to find high or low similarity

Algorithm Settings (Nucleotide BLAST)

Click on the Algorithm parameters text link (below the BLAST button). This will expand a menu of algorithm settings that are specific to the type of BLAST you've chosen in the section above. Generally, the default settings have been shown to be optimal for each type of BLAST, but you may want to try adjusting these settings to see how your results will be affected.
Max target sequences - set at 100. You can re-set this number for more or fewer results.

Short queries - checked by default to automatically adjust for short query sequences. If your sequence isn't short, you can uncheck this box, but generally it does not provide you any greater benefit to uncheck it.

Expect threshold - this is the number of sequences you can expect to be matched purely by chance. The default is 10, but for a more stringent search, set this number lower or set it higher for a more relaxed search.

Word size - the number of nucleotides that are used to start a match (called "seeding"). Default is 28 for megablast, and 11 for the less stringent discontiguous megablast and blastn.

Match, mismatch scores - set at 2 and -3 for blastn or discontiguous megablast but matches are lower and the mismatches higher (e.g. 1,-2) for megablast. This reflects a ratio: a ratio of 0.5 (1,-2) is best for 95% conserved sequences, so is used for megablast, while the ratio of 0.66 (2,-3) is used for the less conserved searches in blastn and discontiguous megablast.

Gap costs - linear for megablast and based on the match and mismatch scores, so gap costs will compound as the gaps are found and extended. For all other nucleotide BLASTs, the gap opening cost is 5, and the cost to extend a gap is 2. If you are doing a search across a number of species, you may want to reduce the existence and extension costs, to assign less penalty to gaps. If you want your sequences to be as similar as possible, select higher numbers for both existence and extension gap costs.

Filters and masking - When to filter or mask?
- Filter for regions in genomic sequences which you suspect have low complexity regions, such as SINES, LINES or virus-inserted repeats
- You can select filtering species-specific repeats
- Masking features can be useful for regions that may code for structural features that don't have functional significance and you want to de-emphasize these regions when searching for matches (this is especially pertinent for protein BLASTs).
  - Masking for lookup table only will mask features when the initial lookup search is done, and thus speed up the search, but will unmask them for the final similarity score assignment
  - To mask features for lookup and scoring, choose the mask lower case letters option. Then, in a Word document or other text program, highlight the region and change its case to lower case.

Hit the BLAST Button!

It is generally useful to check the box Show results in a new window. This way you can always come back to the BLAST setup page, change a few parameters and try the search again.

---

Setting up a Protein BLAST (blastp)

There is an Entrez Protein record you can use for protein BLAST in this section. It is in a public MyNCBI record at:

http://www.ncbi.nlm.nih.gov/sites/myncbi/collections/public/1BixP7hJvx64q1YgMhDqWs/

To BLAST from this record, just click on the BLAST link on the right side of the Entrez Protein record.

For protein BLASTS, you have fewer options for databases to search, but more algorithm and matrix choices. You can search:

- All "non-redundant" protein sequences
  - Be warned, "non-redundant" is not literally true. There will be many instances of the same protein represented many times in this database choice. The reason for this is that there are often many sequence submissions for any particular protein, with small differences in length, mutation of amino acids or different isoforms
- Reference proteins (refseq), which is the subset of proteins in the databases that have been curated and verified by experimentation and validation by the NCBI
- Swissprot protein sequences - searches only Swissprot sequences. This search returns a slightly smaller set of proteins
Patented protein sequences
Protein Data Bank sequences - this search is especially useful if you want to find whether your protein sequence has high similarity with known structures
Environmental samples

Algorithm Parameters

The general guidelines for algorithm selection described above for nucleotide BLAST apply also to protein BLAST, except you have slightly different gap costs and you now have the choice of scoring matrices.

PAM or BLOSUM?

- The lower the number following the word PAM, the more stringent the criteria for the search
- The lower the number following the word BLOSUM, the less stringent the criteria for the search
- BLOSUM62 is the default and performs with the best combination of specificity and precision
- Overall, BLOSUM is more efficient and more accurate in finding matches, but PAM can be useful when searching for more weakly related sequences
- Try the same search with different matrices and compare results

PSI-BLAST and PHI-BLAST

PSI-BLAST is Position-Specific Iterated BLAST. It is especially useful if you suspect weak sequence similarity due to evolutionary change, but still suspect overall conservation (such as a specific protein fold or function). PSI-BLAST creates a positional matrix—a PSSM: Position Specific Scoring Matrix—afer your first BLAST iteration. It finds regions in this matrix of highest correlation between your query and the database matches, then weights these regions more heavily on successive iterations of the BLAST.

- You can use the matrix created by PSI-BLAST to apply to different databases
- When results fall below the E-value cutoff, they should generally be excluded in subsequent iterations, unless they are interesting or expected
  - Including these results may be useful when attempting to find all members of a protein family, or finding the most diverse members of the family
- PSI-BLAST is good for finding distantly-related protein sequences for protein phylogenetic discovery
- Use good judgment: think about how realistic distant matches may be

PHI-BLAST (Pattern Hit Initiated BLAST) is a subset of protein BLAST in which you can input a specific pattern that you want to find in all protein matches. It is useful when you are looking for a region of functional importance in proteins (such as specific binding sites). The pattern syntax is from PROSITE.

An example: [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]
This translates to mean: any one of the amino acids LIVMF followed by G followed by E followed by any single character followed by any one of GAS followed by any one of LIVM followed by any 5 to 11 characters followed by R followed by one of STAQ followed by A followed by any single character followed by one of LIVMA followed by any single character followed by one of STACV.

For more information on PHI-BLAST syntax (along with the above example), see the [NCBI PHI-BLAST rules page](#).

**After You Hit the BLAST Button**

In protein BLAST, you will see an intermediate page that will show you if you have any regions in your protein that have matches in the NCBI Conserved Domains Database (CDD).

---

**Formatting and Saving BLAST Results**

On your BLAST results page, you will see a number of options at the top of the page that let you:

- Edit and resubmit your query
- Save your search strategies
- Change the formatting of the BLAST output
- Download the BLAST in XML format

You can also select specific sequences from your results by clicking the checkboxes next to their sequence identifiers, and change their format (to FASTA, for example), then save them to your Collections in MyNCBI, download them, or view them in your browser window.
At the top or the bottom of this section of the BLAST results, click the text "Get selected sequences".

An Entrez Protein page with your selected sequences will open.

On the Entrez Protein page, change your display to FASTA. Click Apply.
In the **Send to** pull-down window, select your destination

- Use Collections to save to MyNCBI
- Use Clipboard to save the results *temporarily* to a clipboard space in NCBI
- Use File to save to your computer, so you can perform a multiple sequence alignment later on the sequences.
  - Choose **FASTA** format and save the file
  - Files will be saved as a .fasta file. This format is compatible with all multiple sequence alignment programs.

Now you can use these sequences for multiple sequence alignment.
Multiple Sequence Alignment

BLAST is the preferred platform for sequence similarity searching, and most users employ BLAST at the NCBI site. When aligning multiple sequences, however, many options are available.

For years, most researchers used CLUSTALW for multiple alignment of sequences, and this program worked well with good speed. However, there are a number of newer multiple alignment programs that perform with better accuracy and speed than CLUSTALW (including the direct replacement: Clustal Omega).

You can choose from several multiple alignment tools at the European Bioinformatics Institute's website or at the Max-Planck Institute's Bioinformatics toolkit website (click on the "Alignment" tab at the top of the page) or at the online programs page of Phylogeny.fr. Max-Plank has a good selection of multiple alignment tools in addition to having a number of other bioinformatics tools, but you cannot launch your multiple alignments directly in Jalview upon completion of the alignment. You can, however, save your file and use it in other applications, such as phylogenetic analysis.

EBI and Phylogeny.fr servers allow you to open your multiple alignment in Jalview (a Java-based multiple alignment tool), which is useful for visualization and manual editing of multiple alignments.

The sequence alignment tools at EBI are located under either the "Proteins" or the "DNA & RNA" sections on the EBI's Services page:

http://www.ebi.ac.uk/services

Not all tools listed on these pages are for multiple alignment, but ones that are useful are Clustal Omega, Kalign, Lalign, MAFFT, and T-Coffee. MUSCLE performs well for most protein alignments and leads to less need for manual editing of sequences than is necessary with some other programs.

Click on the MUSCLE link.
You can paste your sequences in FASTA format in the box, or upload your .fasta file. You can change the output format as well as other features of the output. Most results are returned quickly using the interactive format, unless you have many sequences of great length.

On the results page:

- You can "View the alignment file" - it's an HTML text format
- You can scroll down the page to see the alignment
- Or you can open the alignment with EBI's ClustalW_Phylogeny tree builder.

Other Multiple Alignment Algorithms

- COBALT - from the NCBI BLAST platform - pairwise construction of multiple alignments
- MultiAlin - while some authors still use this for DNA sequence multiple alignment, it is not as fast or accurate as more modern multiple alignment programs
- ProbCons - best used at the Phylogeny.fr platform, so you can use Jalview to edit and view alignments. ProbCons performs better than most other multiple alignment programs for most alignments

Resources and Help

Sequence similarity searching is a basic but powerful tool that can be used to discover related molecules to your molecule of interest. There are many sources available online that will provide information and instruction on BLAST.
- NCBI's BLAST help page, with links to the NCBI Handbook on BLAST
- NCBI's BLAST tutorials page
- Help links from each of the sequence analysis tools at the EBI sequence tools page
- Contact the Biosciences & Bioinformatics Librarian for help with BLASTing and multiple sequence alignment