How to Use Biology Workbench v.3.2

Tutorial Contents

Overview:

Getting into it:

Taking it to the next level: *COMING SOON*

About the Biology Workbench

Biology Workbench Summary

The Biology Workbench is a computational interface and environment that permits anyone with a Web browser to readily use bioinformatics, for research, teaching, or learning. It consists of a set of scripts that links the user's Browser to a collection of information sources (databases) and application programs. The scripts are specialized for the interface of each program and information source. Functionally they transform the interface for each object, whether database or application program, into a common Web-based form that permits them to be seamlessly interconnected. The user is then able to compile a customized search-and-analyze computational strategy to answer complicated questions about the contents of the databases. By reducing all the formats to a common point-and-click Web interface, the users are freed from the necessity of knowing details of the object formats. Also the scripts work through the interfaces very rapidly, so various operations can be done quickly. Reasonable default parameters for the various operations are built into the Web interface. However the knowledgeable user can easily adjust parameters for search and analysis via Web menus. The present version of the Biology Workbench contains a large array of databases and computational tools that are most useful to the molecular biology community for understanding sequence relationships among proteins and nucleic acids. All databases and computer programs are freely accessible to anybody in the world with a networked computer, via Silicon Graphics servers at the National Center for Supercomputing Applications and the San Diego Supercomputer Center. Reflecting the revolutionary architecture of the Workbench, NCSA and SDSC have taken the revolutionary step of making their supercomputers and the Workbench freely available to all without prior arrangement. The Workbench has been publicly available since June 1996. It has steadily grown in the number of users, and the amount of use. Presently there are approximately 11,000 registered users who use the Workbench for about 150,000 computing sessions a month.

See **Appendix A** for more information regarding the growing field of bioinformatics.

Education Project Summary

The goal of this project is to promote the use of molecular data in the identification and exploration of biological problems with an evolutionary perspective throughout undergraduate biology curricula. This will be accomplished by providing undergraduate faculty and students access to the powerful state-of-the-art bioinformatics tools currently in use by the research community transformed to afford a cognitively supportive environment. These products will enable undergraduate students to investigate current problems in biology using molecular biology tools and skills. The objectives of the project are:

- Create the Biology Student Workbench, an educational front-end to the powerful suite of tools comprising the Biology Workbench for use by undergraduate students and instructors
- Develop, test, evaluate and disseminate supportive inquiry-based curricular materials for diverse areas of biology nationwide, and

- Establish a community of inquiry, with scientists, educators and biology students using advanced science concepts and Workbench tools, which will sustain, support and further disseminate the use of research tools within education.

The Biology Workbench (http://workbench.sdsc.edu/) is widely recognized as a significant bioinformatics resource because it provides a suite of interactive tools which draw on a host of biology databases and allows users to compare molecular sequences using high performance computing facilities, visualize and manipulate molecular structures, and generate phylogenetic hypotheses. The Biology Student Workbench (http://bioweb.ncsa.uiuc.edu/educwb/) will bring the advanced computational infrastructure used by today's scientists to any student desktop machine with a web browser. This access to a multiplicity of research analysis tools and data sources will provide a rich environment for promoting student inquiry.

Preparing to enter the Biology Workbench

About the tutorial:

This tutorial is designed for version 3.2 of the Biology Workbench, which is the current release. We recommend going through this tutorial with the Biology Workbench open in a second window. Alternatively, we recommend that you print the tutorial and point your browser to the Biology Workbench. Very soon, this and other tutorials will be available in PDF format (for use with the Adobe Acrobat Reader), which facilitates quality printing of documents.

Account Set-Up:

If you have never before used the Biology Workbench, you must first register a free account. This merely serves to allocate hard disk space for the sessions that you will run. Please select the "Set-up a free account" link in the Biology Workbench window that you just opened, and follow the instructions. Be assured that the username and password you provide are held in strictest confidence, and you will only be sent e-mail if you require account maintenance. If you have used the Biology Workbench previously, and already have an account (NOTE: If you have not used the Workbench since 3.0, and that is your only account, you will have to reregister for a 3.2 account. Please contact bwhelp@sdsc.edu for help transferring your sessions.), feel free to proceed.

Differences between versions 3.0 and 3.2

The current release version of the Biology Workbench is 3.2. If you have used the Biology Workbench previously (before February 2000), you might know that the last version to see extensive use was 3.0, and you will be interested in the changes outlined here. If you are new to the Workbench, this page will not be very interesting to you, and you should proceed in good conscience. This is not a comprehensive list, but it does give most of the major changes.

- 3.0 sessions cannot be imported to 3.2. You need to contact the Workbench administrators (bwbhelp@sdsc.edu) to do this. If you have been automatically moved over from 3.0 at the old Biology Workbench site, this conversion will have been done automatically – please check your account to see if the conversion was successful.

- Ndjinn Multiple Database Search has replaced the SRS Database search. Ndjinn is a text-based database engine, and brief instructions on its use are provided on the setup page and in the help file.

- The MSASHADE alignment-coloring program has been replaced by BOXSHADE. BOXSHADE is a more complete implementation, and many additional options are available in this module.

- Many programs have replaced CLUSTAL_W, the module that drew rudimentary phylogenetic trees from ClustalW generated tree files. DRAWTREE draws unrooted phylogenetic trees from those ClustalW tree files. DRAWGRAM draws "rooted" trees (the root is inferred) from the ClustalW tree files. CLUSTALDIST draws ClustalW-generated distance matrices, and CLUSTALTREE gives a text, "Clustal Format" tree output.

- DNADIST and DNAPARS have been added for nucleic alignments, to allow one to do the same analyses that PROTDIST and PROTPARS do on protein alignments

- MOTIFGREPDB and MOTIFGREP have been replaced by PATTERNMATCH and PATTERNMATCHDB

- The View tool now allows the user to view the sequences or alignments in various formats. In addition, it allows the user to download a sequence (or all the sequences) in the format being viewed (the icons on the title bar are for downloading – see the help file for View for more information).

- The Add tool now can handle multiple-sequence files in the Protein and Nucleic modes, and has a few additional options, like color-coding of non-standard amino acid or nucleic codes. Read the help file on Add and Edit for more information on the changes in these tools.

- Sessions are now handled somewhat differently. There is no longer a "Default Mode" in which the data gets lost after exiting the browser. Instead, one is placed in a "Default

Session" when they enter the Biology Workbench. All data in the default session is saved, unless the user specifically removes it. The user is allowed to rename the default session, which then creates a new, empty default sessions, and the data that was formerly in the default session will be a session with the new name that was specified. The default session can also be copied, but the users will probably find it easiest just to create a new session and work from there.

- Multiple database selection is now available within all the BLAST tools, the FASTA and SSEARCH tools, and PATTERNMATCHDB.

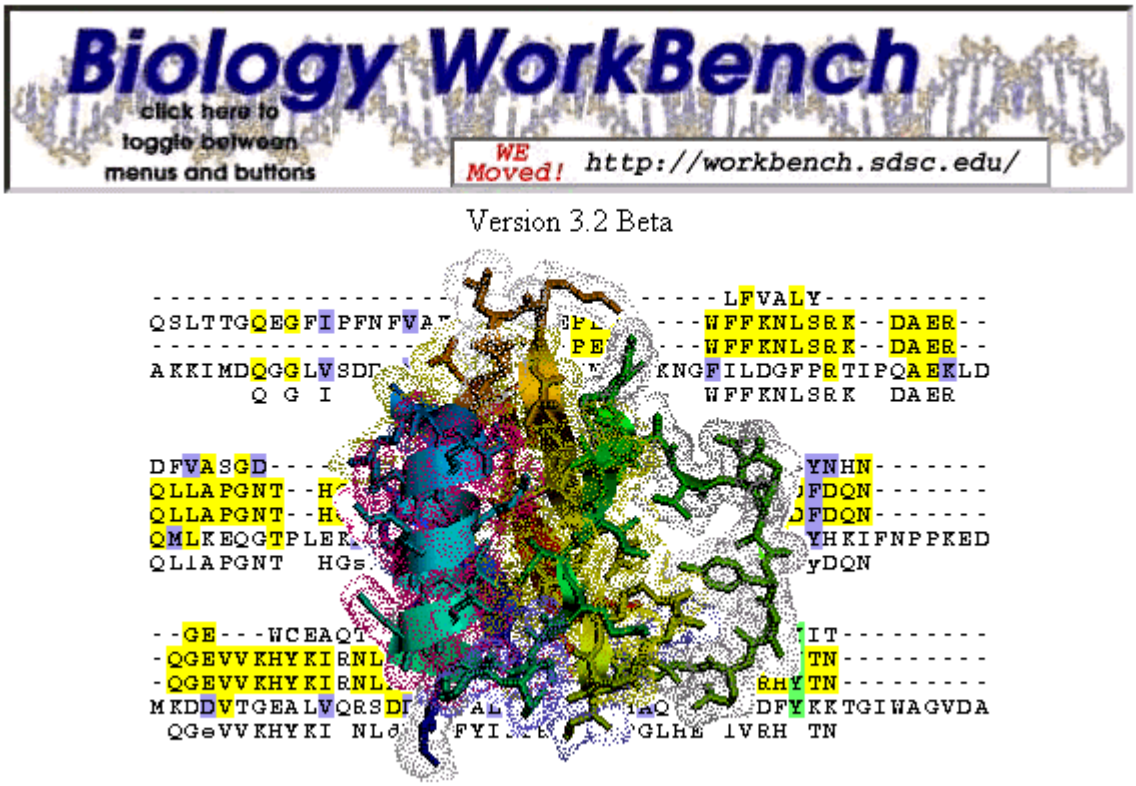
- PSIBLAST has been added in a limited fashion. The ability to use user-defined position-specific matrices is not yet available, but otherwise this should be a fairly useful implementation of PSIBLAST.

- Minor updates and interface improvements have been made in a large number of programs.

Entering the Biology Workbench

Once you have registered an account, please click on the link that reads, "Enter the Biology Workbench 3.2". Note: you will be prompted for the username and password you specified in setting up your account.

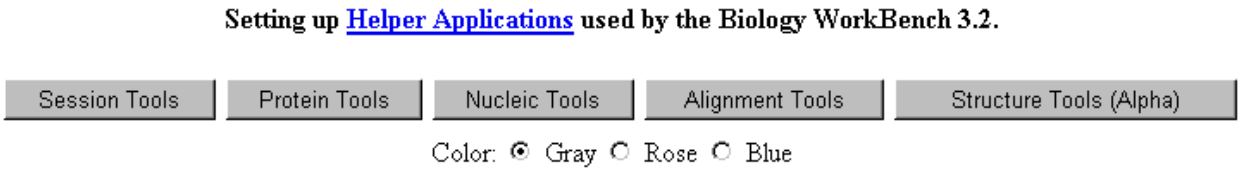
If you have correctly entered your logon and password, you should arrive at a page showing the following image:



On this home page, you must scroll down for information about the Workbench (version differences, FAQs, updates, etc.), and for the toolbar to lead you into the Workbench.

Overview of Tools in the Biology Workbench

The tools in the Biology Workbench are divided into five categories, displayed prominently as a row of buttons:



Helper Applications
Select this link in the Biology Workbench to learn more about helper applications and plug-ins that can be used with the Biology Workbench.

Session Tools
This is the place to start using the Biology Workbench. Session Tools allow you to save your work for future reference. Each time you use Version 3.2 of the Biology Workbench, you must either begin a new session or resume a previously created session.

Protein Tools
Protein Tools contain applications for predicting the secondary structure of a protein, given its amino acid sequence, applications for analyzing the amino acid composition of a protein, and general sequence analysis applications.

Nucleic Tools
Nucleic Tools combine general sequence analysis applications, with applications that are specifically for analyzing DNA or RNA.

Alignment Tools
Alignment Tools allow you to view and analyze aligned sequences.

Structure Tools (Alpha)
Structural Tools allow you to view, analyze, and manipulate sequence structures.

To hopefully clear up any questions thus far, here is a simple illustration. The tools, while displayed nicely side-by-side, are really divided up into three categories or levels. Someone who has never used the Workbench before would start with Session tools, move to either Protein or Nucleic tools (depending on the biological problem to be solved) to find or upload sequences, and then move to Alignment tools.

1) Session tools
2) Protein tools or Nucleic tools

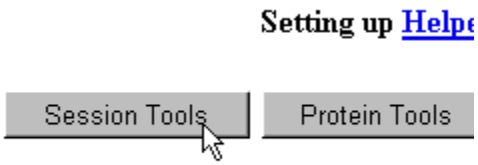
3) Alignment tools

Those who have used the Workbench to create a session may already have the desired sequences uploaded, and, after activating the appropriate session, might wish to go directly to Alignment tools to do their work.

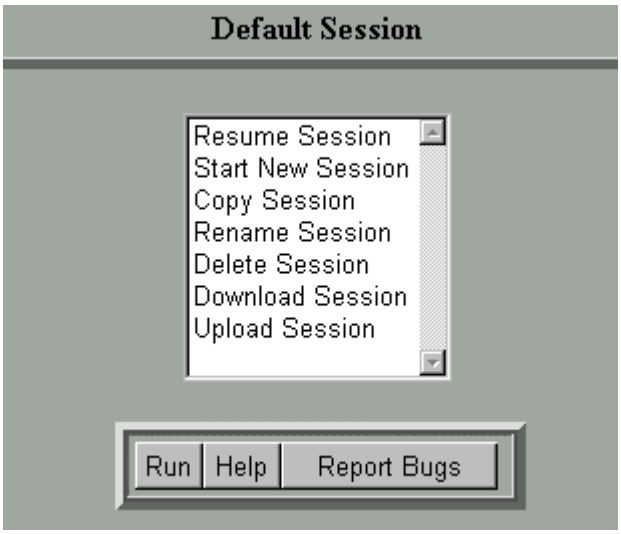
Session Tools

Each time you use the Biology Workbench, you must either start a new session or resume a previously saved session. Sessions function like folders; you can store the sequences you are interested in here. These sessions are saved on-line so they are available to be used again later. When first entering the Workbench through Session Tools, the Default session will be selected. This is a fully functional session, but if you would like to keep your Workbench sessions separate, we would recommend opening new and separate sessions.

1) Please select "Session Tools" in the Biology Workbench.



2) You will be presented will a list of session options.



Here's a short description of each of the options:

Session Tools	**Description**
Resume Session	Activates a previously existing session, allowing you access to those sequences
Start New Session	Allows you to create a new session
Copy Session	Creates a new session with all of the stored information of a desired session
Rename Session	Allows you to change the name of a session
Delete Session	Removes a session permanently
Download Session	Saves a copy of the session to your computer
Upload Session	You can restore a session from a saved copy on your computer

Highlight "Start New Session" and then select the "Run" button.

3) Next you will be presented with a text box in which you may describe or name your session.

4) After entering a session description, select "Start New Session".

5) At the top of the screen should see the name/description of the session that you just created. Your work will now be saved and can be accessed again later.

Protein Tools

These are the protein sequence tools in list form. More information regarding the tools is provided in the Biology Workbench. Select the tool for which you would like more info, click the "Help" button on the page, and a new window will open with the information you need. The most popular tools are shown in **bold** face and denoted with an arrow.

Select All Sequences
Deselect All Sequences
→ **Ndjinn - Multiple Database Search (pronounced "engine")**
Retrieve BATCH Output
Add New Protein Sequence
Edit Protein Sequence(s)
Delete Protein Sequence(s)
Copy Protein Sequence(s)
View Protein Sequence(s)
Download Protein Sequence(s)
View Database Records of Imported Sequences
View Available Scoring Matrices
BL2SEQ - Compare proteins to a reference protein with BLAST
→ **BLASTP - Compare a PS (protein sequence) to a PS DB (database)**
TBLASTN - Compare a PS to a translated DB
→ **PSIBLAST - Position Specific Iterative BLAST**
FASTA - Heuristic Sequence Similarity Search (PS Or DB)
TFASTA - Compare a PS to a NS (nucleic sequence) -> PS DB
TFASTX - Comp PS to Trans DNA (NS Or DB)
LALIGN - Calculate N-Best Local PS Alignments
→ **CLUSTALW - Multiple Sequence Alignment**
MSA - Multiple Sequence Alignment (Sum-of-Pairs Criterion)
LFASTA - Local Alignment of Two PS
ROBUST - Global alignment of Two PS (Show Robust Pairs)
ALIGN - Optimal Global Alignment of Two PS
SSEARCH - Smith-Waterman Local Alignment of Proteins
SIM - N-Best Local Similarities Using Affine Weights
BESTSCOR - Calculate the Best Self-Comparison Score
PRSS - Compare a PS to a Shuffled PS
SAPS - Statistical Analysis of PS
AASTATS - Statistics Based on Amino Acid Abundance
CTREE - Align protein sequences with confidence estimates
GREASE - Kyte-Doolittle Hydropathy Profile
PROSEARCH - Search Prosite DB for Patterns in a PS
PFSCAN - Sequence Search Against a Set of Profiles
PROSITE - Search Prosite DB for Patterns in a PS
PATTERNMATCHDB - Search for Regular Expressions in a protein sequence DB
PATTERNMATCH - Search for Regular Expressions in a protein sequence
GOR4 - Predict Secondary Structure of PS

RANDSEQ - Randomize a Sequence
CHOFAS - Predict Secondary Structure of PS(s) (Chou-Fasman)
HTH - Predict HTH Motifs in Protein Chains
PELE - Protein Structure Prediction
DSSP - Secondary Structure/Solvent Exposure of PDB Proteins

Nucleic Tools

These are the nucleic sequence tools in list form. More information regarding the tools is provided in the Biology Workbench. Select the tool for which you would like more info, click the "Help" button on the page, and a new window will open with the information you need. The most popular tools are shown in **bold** face and denoted with an arrow.

Select All Sequences
Deselect All Sequences
→ **Ndjinn - Multiple Database Search (pronounced "engine")**
Retrieve BATCH Output
Add New Nucleic Sequence
Edit Nucleic Sequence(s)
Delete Nucleic Sequence(s)
Copy Nucleic Sequence(s)
View Nucleic Sequence(s)
Download Nucleic Sequence(s)
View Database Records of Imported Sequences
BL2SEQ - Compare nucleotides to a reference nucleotide with BLAST
→ **BLASTN- Compare a NS (nucleic sequence) to a NS DB (database)**
BLASTX - Compare a PS-Derived-from-NS to a PS DB
TBLASTX - Compare a translated NS to a translated DB
FASTA - Nucleic Acid Sequence Comparisons (NS or DB)
FASTX - Compare Translated NS to PS DB
SSEARCH - Smith-Waterman Local Alignment
SENSEI - Search NS DB with large NS
→ **CLUSTALW - Multiple Sequence Alignment**
LFASTA - Calculate Local Sequence Alignments (Heuristic)
ALIGN - Optimal Global Sequence Alignment
LALIGN - Calculate Optimal Local Sequence Alignments
PATTERNMATCHDB - Search for Regular Expressions in a nucleic sequence DB
PATTERNMATCH - Search for Regular Expressions in a nucleic sequence
TACG - Analyze a NS for Restriction Enzyme Sites
PRIMER3 - Design Primer Pairs and Probes
NASTATS - Nucleic Acid Statistics
BESTSCOR - Calculate the Best Self-Comparison Score
PFSCAN - Sequence Search Against a Set of Profiles
SIXFRAME - Generate & Import 6 Frame Translations on a NS
REVCOMP - Generate Reverse Complement of NS
RANDSEQ - Randomize a Sequence

Alignment Tools

These are the sequence alignment tools in list form. More information regarding the tools is provided in the Biology Workbench. Select the tool for which you would like more info, click the "Help" button on the page, and a new window will open with the information you need. The most popular tools are shown in **bold** face and denoted with an arrow.

Select All Alignments
Deselect All Alignments
→ **Ndjinn - Multiple Database Search (pronounced "engine")**
Add New Aligned Sequence
Retrieve BATCH Output
Edit Aligned Sequence(s)
Delete Aligned Sequence(s)
Copy Aligned Sequence(s)
View Aligned Sequence(s)
Split Alignment Into Component Sequences
Split Alignment Into Component Sequences and Remove Gap Characters (-)
Download Aligned Sequence(s)
→ **BOXSHADE- Color-Coded Plots of Pre-Aligned Sequences**
TMAP - Prediction of Transmembrane Segments
→ **DRAWTREE- Draw Unrooted Phylogenetic Tree from Alignment**
→ **DRAWGRAM- Draw Rooted Phylogenetic Tree from Alignment**
→ **CLUSTALDIST- Generate Distance Matrix with Clustal W**
→ **CLUSTALTREE- Phylogenetic Analysis with Clustal W**
→ **DNADIST - Compute Evolutionary Distance Matrix from NS (nucleic sequence) Alignment**
→ **PROTDIST - Compute Evolutionary Distance Matrix from PS (protein sequence) Alignment**
DNAPARS - Infer an Unrooted Phylogeny from NS Alignment
PROTPARS - Infer an Unrooted Phylogeny from PS Alignment

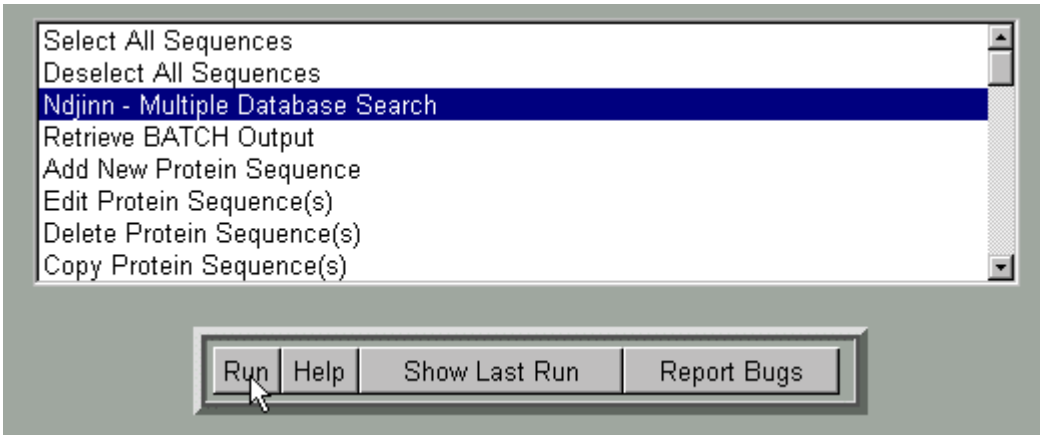
Structure Tools (Alpha)

These are the structure tools currently available. More information regarding the tools is provided in the Biology Workbench. Select the tool for which you would like more info, click the "Help" button on the page, and a new window will open with the information you need.

CONVERT - File format conversion utility
TNT - Macromolecular Refinement Package

Ndjinn Multiple Database Search

After you have started a new session or resumed an existing session, you may continue by selecting "Protein Tools", "Nucleic Tools", "Alignment Tools", or "Structural Tools", all of which have been previously discussed. If you are just beginning a session in the Workbench, you more than likely have not imported sequences from any of the available sequence databases. Therefore, many of the tools are not yet useful to you, and the categories of "Alignment Tools" and "Structural Tools" are not applicable. Selecting "Protein Tools" or "Nucleic Tools" will allow you to select sequences pertinent to your investigation. After selecting a tool, you will be presented with a scrollable list of databases. If you selected "Protein Tools" in the Biology Workbench, you would see a list very similar to the one below.

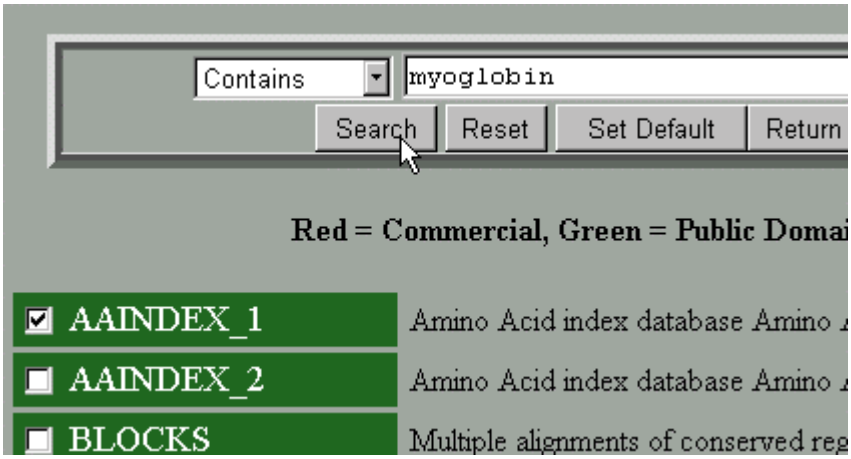


Often the first step in using the Biology Workbench is to use the "Ndjinn Multiple Database Search". The "Ndjinn Multiple Database Search" allows you to search as many or as few of the databases in the Biology Workbench as desired for the text that you specify. By highlighting "Ndjinn Multiple Database Search" and clicking "Run" in any of the three main environments (Protein Tools, Nucleic Tools, or Alignment Tools), you will be presented with a check-box list of databases from which to choose, and a text line which to construct a query. This text line is designed to permit the construction of queries that are as precise or as general as you desire, by utilizing Boolean strings. This will become clearer as you go through the next sections on "Constructing a Query."

See **Appendix B** for a collection of information about the databases that can be used with the Ndjinn search.

Constructing a Query

Part I: Selecting Databases



This image is of the database selection/query form in the Workbench. Any search will be performed with the databases that the user selects with the checkboxes on the left; one may search as many databases simultaneously as are desired. One caveat - the more databases that are selected for a query, the longer the query will take to return results.

Part II: Database Fields

The search engine performs a full text search of all databases selected for the string in the text line. Choices for searching the text consist of whether the database "Contains", "Begins With", "Ends With", or is an "Exact Match" of the string.

Part III: Boolean Operators

For simple searches, one might choose to search one or several databases for a single word.

Example (as seen above): *myoglobin*

For constructing complex/limiting searches, Boolean operators may be used within the text line. Simply place the words "AND", "NOT", or "OR" between the words of your query, and the search engine will take that into account during the full text search of the chosen database(s). If you would like to set the order of searching, place parentheses around the words to be sought and any associated operators.

Example: *(myoglobin AND human) OR orangutan*

Practice Constructing Queries

Part I: The Beginning

To see how the query process works, here is a beginning exercise:

0) Please create an account and begin a new session, if you have not already done so.

1) In the Biology Workbench, select "Protein Tools". You should arrive at a new page that contains a scroll menu.

2) Highlight "Ndjinn Multiple Database Search" from the scroll menu and then press the "Run" button. You should then be presented with a query field and a number of sequence databases, as well as genome databases.

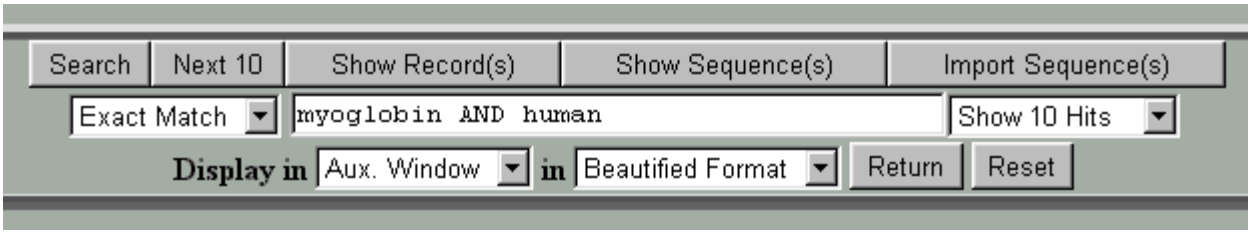
3) Select the check buttons for the "PIR1" and "SWISSPROT" databases.

4) Type "myoglobin" in the object field and press the "search" button.

Part II: Narrowing Your Search

You will observe that many objects were retrieved (204 at this writing).

Perhaps this includes some information you don't want, and you would like to automatically filter that information out, rather than inspecting it and discarding it manually. For example, suppose you only wanted to look at ape and human myoglobin, perhaps to discover which of the apes are our closest relatives. Now you can construct the search for "myoglobin AND (gorilla OR chimpanzee OR orangutan OR human)". Now when you press the "search" button, you will retrieve only a few sequences (28 at this writing).



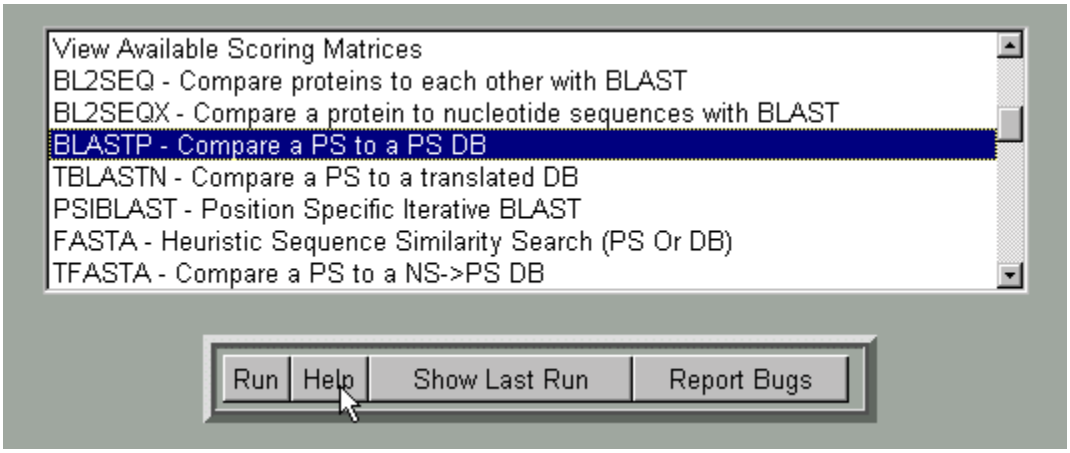
Part III: More Advanced Searches

Perhaps now you decide that basing a phylogenetic analysis on only one protein might be shaky, so you want to look at a second protein. Say you chose hemoglobin. You want to download data that will let you construct a phylogenetic tree among the apes with both hemoglobin and myoglobin and see if you get the same answer. This time construct the query as "(hemoglobin or myoglobin) and (chimpanzee or gorilla or human or orangutan)." Now when you press the "submit" button you retrieve 90 entries (at this writing), all the hemoglobin or myoglobin entries for all the ape species that you asked for.

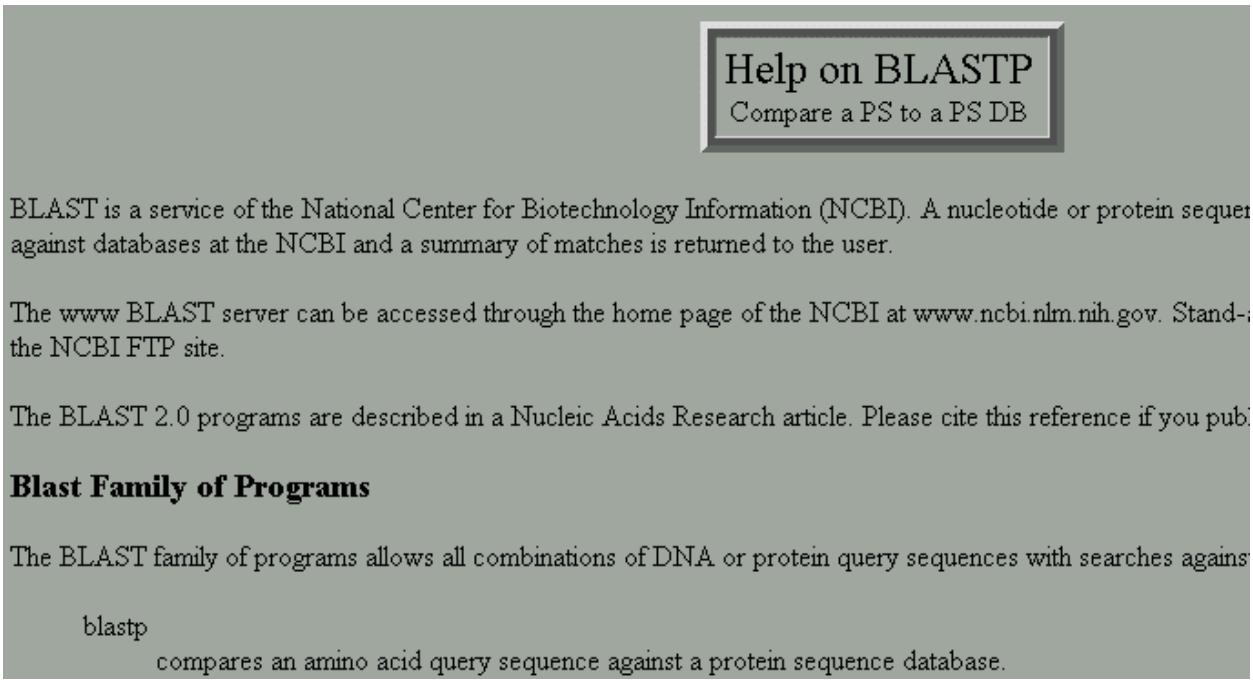
Importing Sequences to the Biology Workbench

After you have found the sequences you are looking for, you must bring them into the Biology Workbench in order to utilize further tools. From the point you chose to run an Ndjinn search, you have been communicating with the supercomputers, and they have diligently been retrieving the information you requested. In order for the retrieved sequences to be of any use to you, you must bring them from the databases into your account. This can be accomplished by checking the box associated with the sequence(s) you wish to keep and clicking the "Import Sequence(s)" button. You will be returned to the main page (your session), and your sequence(s) will be visible to you. If you have imported more than one sequence, the "Alignment Tools" will now be useful. Also interesting are the "Show Record(s)" and "Show Sequence(s)" buttons, which can aid you in choosing and understanding the sequences through links to primary research and even textbooks.

The following sections (not yet written) - sequence processing, alignment processing, and post processing - are fairly subjective, and largely dependent upon the task which you are undertaking. Perhaps the most useful tools are BLAST and CLUSTALW, along with the phylogenetic tools BOXSHADE, DRAWTREE, etc., but obviously that is not certain. The Biology Workbench gives excellent help on the tools. For example, say you wanted information about BLASTP. In the scroll box of tools, highlight BLASTP and click on HELP, as seen below.



This brings up the HELP documentation for BLAST in another window (as seen below), which includes a link to the NCBI BLAST web page.



So, now you have the tools you need to begin using the Biology Workbench. Good luck!

Sequence Processing

- Sequence Editing and Manipulation
- Refining a Search Using BLAST
- Multiple Sequence Alignment Using CLUSTALW

Alignment Processing

- Sequence Similarity Shading Using BOXSHADE
- Distance Matrix Analysis Using CLUSTALDIST
- Create Phylogenetic Trees Using DRAWTREE

Post-Processing

APPENDIX A: Bioinformatics

A leading application in which search, analysis, and visualization are intimately linked into a comprehensive computational environment under a Web browser is the Biology Workbench. The Biology Workbench provides a point-and-click access to all the leading biology molecular sequence and structure databases integrated with analysis and visualization tools. The Workbench architecture is a new way of dealing with information on the Web and was so recognized by a patent awarded on January 12, 1999.

To understand the educational significance of the Workbench, it is necessary to consider the scientific context. The raw material of biology is information, the results of experiments in the laboratory and observations in the field. The science of biology is all about constructing meaning from the information. In the last couple of decades, an enormous new category of information has become available about living systems. An array of technical revolutions in molecular biology, still ongoing, have made it possible to get enormous amounts of information about the sequences of amino acids in proteins and the sequences of bases in nucleic acids. To construct meaning from the sequence information, the array of computer techniques called bioinformatics has been developed. To understand the basic idea of bioinformatics, one might think of a written language. The text you are reading consists of a series of letters, words, sentences, and paragraphs. If you did not know the meanings of the words and the rules of the language, this page would just be a collection of meaningless symbols. Similarly, the first time scientists saw gene and protein sequences, they saw a string of symbols with no clear meaning in terms of biological function. But now, bioinformatics is showing us many things about what sequences mean. Using bioinformatics, sequences are being used to reveal relationships among different life forms that we could not find out any other way. Bioinformatics is revealing the rules and meaning of a language that is new to human beings but in fact is a billion years old - the Language of Life.

Bioinformatics is not a separate area of study in biology. Rather, the importance of bioinformatics to biology is that it adds value and new dimensions to everything else that biologists do. Consider for example structural biology - the determination of protein structures by x-ray crystallography or nuclear magnetic resonance spectroscopy. Because of bioinformatics analysis of sequences, when one determines the structure of one protein, one has a very good idea of the structure of many related molecules. For example, the first structure of a class of proteins called potassium channels was determined last year. These proteins selectively permit the passage of potassium ions across cell membranes for modulation of electrical signals and maintenance of appropriate ionic environments inside and outside cells. Because of the ability bioinformatics gives us to compare sequences from various cells, tissues, and organisms, knowledge of the one known structure can be used to provide insights into the structural correlates of ionic selectivity, permeability regulation, toxin sensitivity, etc., of potassium channels from all forms of life. Potassium channels are just one example; bioinformatics multiplies many-fold the insights obtainable from any biomolecular structure determination. This knowledge is not only useful for basic knowledge in biology. It also pertains to such practical considerations as drug design, understanding the mechanisms of action of environmental pollutants, and understanding the mechanisms of a plethora of diseases such as cancer, sickle-cell disease, cystic fibrosis, etc. There is always value to understanding the

molecular basis of disease, and bioinformatics always multiplies the insights obtainable from any particular molecular biology experimental information. Today, the vast majority of research papers published in molecular biology have a bioinformatics component in which the sequence of the molecule being studied is compared and contrasted with the sequences of related molecules to extend and generalize the insights directly obtainable from experiment.

APPENDIX B: Database Information

AAINDEX - Amino Acid Index
http://www.genome.ad.jp/dbget/

- AAINDEX_1 - Amino Acid Indices
- AAINDEX_2 - Amino Acid Mutation Matrices

BLOCKS - Multiple alignments of conserved regions of protein families
http://www.blocks.fhcrc.org

DBCAT - DBCat Public Catalog of Databases
http://www.infobiogen.fr/services/dbcat/

DOGS - Database of Genome Sizes
http://www.cbs.dtu.dk/databases/DOGS/

ECDC - E.Coli Database Collection
http://susi.bio.uni-giessen.de/ecdc/ecdc.html

ENZYME - Repository of information relative to the nomenclature of enzymes
http://www.expasy.ch/enzyme/

EPD - Eukaryotic Promoter Database
http://www.epd.isb-sib.ch/

FlyBase - A comprehensive database for information on the genetics and molecular biology of Drosophila
http://flybase.bio.indiana.edu:82/

GenBank
http://www.ncbi.nlm.nih.gov/Genbank/

- All
- Bacterial sequences
- Expressed sequence tags
- Genome survey sequence entries
- High throughput genomic sequencing entries
- Invertebrate sequences
- Mammalian sequences
- Updates
- Patent sequences
- Phage sequences
- Plant sequences (incl. fungi and algae)
- Primate sequences
- Rodent sequences
- Sequence tagged site entries

- Synthetic and chimeric sequences
- Unannotated sequence entries
- Viral sequences
- Misc. vertebrate sequences
- Gene products database

IMGT - ImMunoGeneTics
http://www.ebi.ac.uk/imgt/

KEGG
http://www.genome.ad.jp/kegg/
- Genes database
- Pathway database

Ligand
http://www.genome.ad.jp/dbget/ligand.html
- Compound database
- Enzyme database

MHC binding peptides
http://wehih.wehi.edu.au/mhcpep/

NRL 3-D - Sequence-structure database derived from the 3-dimensional structure of proteins deposited with the Brookhaven National Laboratory's Protein Data Bank
http://www-nbrf.georgetown.edu/pirwww/search/textnrl3d.html

OMIM - Online Mendelian Inheritance in Man
http://www.ncbi.nlm.nih.gov/Omim/

PDBFINDER - Sequences derived from the PDB, DSSP, and HSSP databases
http://www.biochemtech.uni-halle.de/info_protein/overview.html/

PIR - Protein Information Resource
http://www.psc.edu/general/software/packages/nbrf-pir/nbrf.html
- PIR1 - Fully classified entries
- PIR2 - Verified and classified entries
- PIR3 - Unverified entries
- PIR4 - Unencoded or untranslated entries

PRINTS - Protein Motif Fingerprint Database
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html

PRODOM - The Protein Domain Database
http://protein.toulouse.inra.fr/prodom.html

PROSITE
http://www.expasy.ch/prosite/

- PROSITE_DOC

REBASE - Restriction Enzyme Database
http://rebase.neb.com/rebase/rebase.html

SDSCNR - Non-Redundant sequence database made from all protein databases within the Biology Workbench

SEQANAL - Sequence Analysis Bibliographic Reference Data Bank
http://www.expasy.ch/seqanalref/

SWISSPROT
http://www.expasy.ch/sprot/

- SWISSnew - Updates

TRANSFAC - Transcription Factor Database
http://transfac.gbf-braunschweig.de/TRANSFAC/

- TRANSFAC_CELL - Explanaitons of cellular sources that interact with sites in Transfac SITE
- TRANSFAC_CLASS - Explanations of DNA binding domains of transcription factor classes
- TRANSFAC_FACTOR - Tukaryotic cis-acting regulatory DNA elements and trans-acting factors
- TRANSFAC_GENE
- TRANSFAC_SITE - Putative regulatory protein binding sites

TrEMBL
http://www.expasy.ch/sprot/sprot-top.html

- TrEMBLnew - Updates