



ANGIS

**Current list of Programs and
Databases on
BIOMANAGER**

OVERVIEW

The following pages contain information about the programs currently available on BioManager. At the present time there are in excess of 200 programs available on this interface, many are under-utilized.

You will need a login/username to access the programs. If you have problems accessing the interfaces, or wish to enquire about subscribing, contact us at help@angis.org.au or call 1800 728 028.

BioManager (formerly BioNavigator) is the product of the now defunct company Entigen. ANGIS has one of the few copies of this interface in the world.

The following is a list of the programs that you have access to as ANGIS subscribers. Only a brief outline of each program is included to be used as a guide in deciding which program is suitable for your problem. Additional information including references for citation and limitations in the program can be located by clicking on the links within the programs

The following pages list all the programs and databases that are available on BioManager. In some cases there are multiple programs that perform the same or similar function.

BIOMANAGER DATABASES	2
SEQUENCE DATABASE SEARCHING	5
CONTIG ASSEMBLY	7
TWO SEQUENCE ANALYSIS	8
MULTIPLE SEQUENCE ANALYSIS	10
MOLECULAR EVOLUTION	12
PROTEIN STRUCTURE	14
SEQUENCE FAMILY	16
MOLECULAR MODELLING	17
MOTIF ANALYSIS	18
GENE DETECTION	20
RESTRICTION MAPPING/PCR	22
DNA/RNA	24
STATISTICAL ANALYSIS	26
FILE MANAGEMENT	27
OTHER PROGRAMS	30

BIOMANAGER DATABASES

Non Redundant databases:

A non-redundant database contains entries compiled from a variety of similar databases e.g. GenBank and EMBL or PIR, SWISS-PROT and GenPep). The purpose of a non-redundant database is to minimize the repetition (or redundancy) of sequence data and thereby enable a faster and more efficient search. Sequence databases are merged and identical sequences from the less informative databases are discarded. The NR databases are only available for searching in Blast1. Blast2 uses the suite of GenBank databases.

Expressed Sequence Tags (ESTs)

ESTs are sequences of cDNA that have been reverse-transcribed from mRNA and their function is not necessarily known. They have applications in the discovery of new genes, mapping of various genomes, and identification of coding regions in genomic sequences.

High Throughput Genome Sequences (HTGs)

The HTG division contains 'unfinished' DNA sequences generated by the high-throughput sequencing centres. Unfinished HTG sequences containing contigs greater than 2 kb are assigned an accession number and deposited in the HTG division. A typical HTG record might consist of all the first pass sequence data generated from a single cosmid, BAC, YAC, or P1 clone which together comprise more than 2 kb and contain one or more gaps. A single accession number is assigned to this collection of sequences and each record includes a clear indication of the status (phase 1 or 2) plus a prominent warning that the sequence data is "unfinished" and may contain errors.

Genome Survey Sequences (GSSs)

The GSSs are similar in nature to ESTs, except that the sequences are genomic in origin, rather than cDNA. The GSSs may be (but are not limited to) the following types of data: random "single pass read" genome survey sequences single pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be), exon trapped genomic sequences Alu PCR sequences

Sequence Tagged Sites (STS):

STS are short DNA segments with a single location in the genome. This feature of STS makes them useful tags for mapping.

Blocks +

Blocks are multiple aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks Database are made automatically by looking for the most highly conserved regions in families of proteins documented in the PROSITE Database. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the chance distribution of matches. It is these calibrated blocks that make up the Blocks Database.

Course DNA

Course DNA is a subset of GenBank and is used strictly for ANGIS courses only. This database is not recommended for any other purposes.

Course Protein

Course Protein is a subset of SWISS-PROT and is used strictly for ANGIS courses only. This database is not recommended for any other purposes.

Enzyme

This is a database of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and

Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided.

Database	Text Search	Sequence Search
Blocks+	Y	Y
Course DNA	Y	Y
Course Protein	Y	Y
Enzyme	Y	N
GenBank	Y	N
GenBank ESTs	Y	Y
GenBank GSSs	Y	Y
GenBank HTG	Y	Y
GenBank Main	Y	Y
GenBank Patents	Y	Y
GenBank STSs	Y	Y
PDB	Y	Y
NR Database	N	Y Blast 1 only
Pfam	Y	N
Prosite Documents	Y	N
Prosite	Y	N
StackDB	N	Y
SWISSPOT+SP-TrEMBL	Y	N
SP-TrEMBL	Y	Y
SWISS-PROT	Y	Y

GenBank

GenBank is a comprehensive database of publicly available, annotated nucleotide sequences, including ESTs, HTGs, GSSs and STSs. GenBank also contains all new or revised entries submitted to GenBank since the last full release. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is collaboration between the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

GenBank ESTs

GenBank ESTs is an expressed sequence tag (EST) database created by extracting the EST entries from the GenBank database. Expressed Sequence Tags (ESTs) are sequences of cDNA that have been reverse-transcribed from mRNA and their function is not necessarily known. They have applications in the discovery of new genes, mapping of various genomes, and identification of coding regions in genomic sequences.

GenBank GSSs

GenBank GSSs is a genome survey sequence (GSS) database that has been created by extracting the GSS entries from the GenBank database.

GenBank HTGs

GenBank HTGs is a high throughput genome sequences (HTGs) database that has been created by extracting the HTGs entries from the GenBank database.

GenBank Main

The GenBank database in BioManager has been split into six subsets to enable faster searching. GenBank Main contains all the entries from GenBank which are not contained in the subsets GenBank ESTs, GenBank HTGs, GenBank GSSs, GenBank Patents or GenBank STSs

GenBank Patents

GenBank Patents is a nucleotide sequence database that has been created by combining the relevant subsection of the GenBank database.

GenBank STSs

GenBank STSs is a sequence tagged site (STS) database that has been created by extracting the STS entries from the GenBank database.

PDB

PDB (Protein Data Bank) is the single international repository for the processing and distribution of 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR.

Pfam

Pfam is a database of multiple sequence alignments and Hidden Markov Models (HMMs) for many common protein families and domains.

PROSITE

Protein signatures are used to assign newly sequenced proteins to a specific family of proteins and thus to formulate hypotheses about its function. PROSITE currently contains signatures specific for about a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins.

PROSITE Documents

For each entry in PROSITE, a corresponding entry containing information about the protein signature (each of which specifies the structure and function of a protein family or domain) can be found here.

StackDB

This is a database of sequences expressed in the human genome. The STACK (Sequence Tag Alignment and Consensus Knowledge) project aims to generate comprehensive representation of the sequence of each of the expressed genes in the human genome.

SWISS-PROT + SP-TrEMBL

The SWISS-PROT + SP-TrEMBL database contains all the entries from SWISS-PROT and SP-TrEMBL.

SP-TrEMBL

Sp-TrEMBL is a subdivision of the TrEMBL database (computer-annotated database of translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT). It contains TrEMBL entries that have been assigned SWISS-PROT accession numbers and will eventually be incorporated into the SWISS-PROT database. Entries in SP-TrEMBL are removed when they are incorporated into the SWISS-PROT database. TrEMBL supplements SWISS-PROT in order to speed up the release of sequence data without jeopardizing the quality standards of SWISS-PROT. Thus, SP-TrEMBL is effectively a preliminary section of SWISS-PROT.

SWISS-PROT

SWISS-PROT is a comprehensive, annotated protein sequence database, containing the latest full release of SWISS-PROT and all new or revised entries submitted to SWISS-PROT since the last full release. SWISS-PROT is a curated database. It is the most extensively annotated of all the sequence databases and is very useful for this reason.

SEQUENCE DATABASE SEARCHING

BlastN

BlastN compares a nucleotide query sequence to a nucleotide sequence database.

BlastP

BlastP compares a protein query sequence to a protein sequence database.

BlastX

BlastX compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database

BlockMaker

BlockMaker finds blocks in a group of related protein sequences. Blocks are short multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. Typically, a group of proteins has more than one region in common and their relationship is represented as a series of blocks separated by unaligned regions.

BlockSearcher

Compares a protein sequence to the Blocks+ database of protein blocks.

FastA

Compares a protein sequence to a protein sequence database or a nucleotide sequence to a nucleotide sequence database using the FASTA algorithm. The search speed and selectivity are controlled with the word size option.

FastX

Compares a nucleotide sequence translated in three frames to a protein sequence database, allowing frameshifts only between codons.

FastY

Compares a nucleotide sequence translated in three frames to a protein sequence database, allowing frameshifts only within codons.

HMMpfam

Searches a hidden Markov model database for matches to a query sequence.

HMMsearch

Search a sequence database for matches to a hidden Markov model.

PSI-Blast

PSI-BLAST (Position-Specific Iterated BLAST) compares an input protein sequence queries with protein databases.

Overlap

Compares two sets of DNA sequences to each other in both orientations, using a WordSearch style comparison. Has a very high limit on total sequence length for genome scale sequence analysis but it is too larger for general use on most systems.

Ssearch

Ssearch compares a protein or DNA sequence to all of the entries in a sequence database using the rigorous Smith-Waterman algorithm (Smith and Waterman (1983)). This may be the most sensitive method available for similarity searches. It is 10-50 times slower than BLAST and FastA.

tBlastN

Compares a protein sequence to a nucleotide sequence database. Each sequence in the nucleotide database is translated into each of six frames and the protein sequences are matched to the query protein sequence. Uses the BLAST algorithm.

tBlastX

Compares a nucleotide sequence to a nucleotide sequence database. Both the query sequence and the nucleotide sequence database are translated into each of the six frames, and the proposed protein sequences are compared to identify significant matches.

tFastA

Uses a Person and Lipman search for similarity between a query peptide sequence and any group of nucleotide sequences. The sequence is translated into all six reading frames before the comparison.

tFastX

Compares a protein sequence to a nucleotide sequence database. A protein sequence is compared against only two sequences (the forward and reverse orientation) from each nucleotide sequence. For a given forward or reverse orientation, the similarity score for alignments that allow frameshifts between codons are calculated, thus considering all possible reading frames.

TFastY

Compares a protein sequence to a nucleotide sequence database. A protein sequence is compared against only two sequences (the forward and reverse orientation) from each nucleotide sequence. For a given forward or reverse orientation, the similarity score for alignments that allow frameshifts within codons are calculated, thus considering all possible reading frames.

WordSearch

Identifies sequences similar to a query sequence using a Wilbur and Lipman search.

CONTIG ASSEMBLY

CodonCode Assembler

This is a program for assembling shotgun DNA sequence data. Allows use of entire read (not just trimmed high quality part). Uses a combination of user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats. Constructs contig sequence as a mosaic of the highest quality parts of reads and is able to handle very large datasets.

CodonCode BaseCaller

This is a base-calling program for DNA sequence traces. This program calls and assigns quality values to the bases of DNA trace sequence data files and writes the base calls and quality values to output files. Can read trace data from SCF files and ABI model 373 and 377 DNA sequencer chromatography files.

CodonCode Matcher

This is used for comparing any two sets of (long or short) DNA sequences.

TWO SEQUENCE ANALYSIS

BestFit

Makes an alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the algorithm of Smith and Waterman.

Compare

Compares two protein or nucleotide sequences and creates a file of the points of similarity between them for plotting with DotPlot. Compare finds the points using either a window/stringency or a word match criterion. Compare makes a file with the coordinates of each point where two sequences are similar.

Diffseq

Diffseq takes two overlapping, nearly identical sequences and reports the differences between them, together with any features that overlap with these regions

Dotmatcher

Displays a dotplot of two sequences

Dotpath

Displays a non-overlapping wordmatch dotplot of two sequences

DotPlot

Makes a dot-plot with the output file from Compare or Stemloop. Dot-plotting shows all of the structures in common between two sequences or all of the repeated or inverted repeated structures in one sequence.

Dottup

Displays a wordmatch dotplot of two sequences

Est2Genome

Align EST and genomic DNA sequences

ESTwise

Can be used to compare a protein sequence or a protein profile HMM (Hidden Markov Models) to an EST or cDNA sequence. This comparison of DNA sequence at the level of its protein translation allows the simultaneous prediction of gene structure with homology based alignment.

FrameAlign

Creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in all possible reading frames on a single strand of a nucleotide sequence.

Gap

Uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps. A scoring matrix is used to assign values for symbol matches. In addition, a gap creation penalty and a gap extension penalty are required to limit the insertion of gaps into the alignment.

GeneWise

Compares a protein sequence or a protein profile HMM (hidden Markov model) to a genomic DNA sequence. This comparison of DNA sequence at the level of its protein translation allows the simultaneous prediction of gene structure with homology based alignment.

Global Pair Alignment

Performs a global alignment of two sequences. It introduces a certain number of gaps into either pairwise aligned sequences or groups of sequences to find a minimal global distance.

HMMalign

Aligns sequences to an existing hidden Markov model.

Local Pair Alignment

Compares two protein or nucleotide sequences to identify regions of similarity. Will report several alignments by identifying and displaying several similar regions and similarities due to internal repeats.

Matcher

Matcher compares two sequences looking for local sequence similarities using a rigorous algorithm. Matcher is based on Bill Pearson's 'lalign' application, version 2.0u4 Feb. 1996.

Megamerger

Megamerger takes two overlapping sequences and merges them into one sequence. It could thus be regarded as the opposite of what splitter does

Merger

This joins two overlapping nucleic acid sequences into one merged sequence.

Needle

This program uses the Needleman-Wunsch global alignment algorithm to find the optimum alignment (including gaps) of two sequences when considering their entire length.

Stretcher

Stretcher calculates a global alignment of two sequences using a modification of the classic dynamic programming algorithm which uses linear space.

Supermatcher

This is a rough and ready local alignment program for large sequences.

Water

Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment.

Wordmatch

Finds all exact matches of a given minimum size between 2 sequences displaying the start points in each sequence and the match length.

MULTIPLE SEQUENCE ALIGNMENT

ClustalW (accurate)

The ClustalW (accurate) pair-wise similarity scores are calculated from accurate, global alignments using a dynamic programming method. The parameters are used to give initial alignments that are then re-scored to give percent identity scores. The scores are converted to distances for the trees that are used to guide the final alignment. This program is used for aligning short sequences and is slow and accurate. It will be VERY SLOW for many (e.g. >20) long (e.g. >1000 residues) sequences.

ClustalW (fast)

The pair-wise similarity scores are calculated from fast, global alignments using the Wilbur and Lipman method. This method involves two techniques, firstly, the exactly matching fragments (word size) are considered and secondly, the 'best' diagonals (the ones with most word size matches) are used. This program is used for aligning long nucleotide or protein sequences (e.g. >1000 residues).

ClustalW (profiles)

This program is used to align two existing alignments (either of which may consist of just one sequence) or to add a series of new sequences to an existing alignment. Often, just a few sequences cause misalignments in the progressive algorithm and these can be removed from the process and then added at the end by profile alignment. A second use is where one has a high quality reference alignment and wishes to keep it fixed while adding new sequences automatically.

Cons

Calculates a consensus sequence from a multiple sequence alignment.

Edit (Jalview)

Edit (Jalview) allows editing of multiple sequence alignments. This program is accessed by viewing the multiple sequence alignment and choosing the Edit option.

NoOverlap

This program determines if there are regions where a group of nucleotide sequences do not share any common sub-sequences. Hybridization probes specific enough to detect individual members of a gene family can be prepared if a region 100 bases or longer can be found that does not have a perfect match of nine or more bases with any other member of the family. NoOverlap is designed to find such regions.

Overlap

Overlap accepts a set of sequences as input and uses the algorithm of Wilbur and Lipman (1983) to compare each sequence with each other in both orientations. Unlike WordSearch, Overlap looks for overlaps between sequences rather than simply regions of similarity. An overlap is a highly similar region between two sequences that run the entire length of a register of comparison.

OldDistances

OldDistances writes a matrix of the pair wise similarities between up to 50 different sequences in a multiple sequence alignment. The similarity value is the number of "matches" between each sequence pair divided by the sequence length.

PileUp

PileUp creates a multiple sequence alignment using a simplification of the progressive alignment method of Feng and Doolittle. The multiple alignment procedure begins with the pair wise alignment of the two most similar sequences, producing a cluster of two aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pair wise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pair wise alignments that include increasingly dissimilar sequences and clusters, until all sequences have been included in the final pair wise alignment. A dendrogram, or tree representation of clustering relationships, is produced in the initial stages. PileUp can align up to 500

sequences, with any single sequence in the final alignment restricted to a maximum length of 7,000 characters (including gap characters inserted into the sequence to create the alignment).

Plotcon

Displays a graphical representation of the similarity along a set of aligned sequences.

Plotorf

A graphical representation of where the open reading frames are in all 6 reading frames is shown.

PlotSimilarity

PlotSimilarity calculates the average similarity among all members of a group of aligned sequences at each position in the alignment, using a user-specified sliding window of comparison. The window of comparison is moved along all sequences, one position at a time, and the average similarity over the entire window is plotted at the middle position of the window. The average similarity across the entire alignment is plotted as a dotted line.

Pretty

Pretty prints sequences with their columns aligned. Can display a consensus for the alignment, allowing you to look at relationships among the sequences. This program can be used for aligned sequences in an MSF or RSF file, or for separate sequences that have had gaps added to make them all align.

PrettyBox

Produces a PostScript file containing an alignment with residues shaded on the basis of agreement to a calculated consensus sequence, allowing you to identify relationships among the sequences.

PrettyPlot

PrettyPlot displays multiple sequence alignments and calculates a consensus sequence.

MOLECULAR EVOLUTION

Consense

Computes consensus trees by the majority-rule consensus tree method, which also allows one to easily find the strict consensus. eConsense is a modified version with command line control added.

Diverge

Diverge measures the percent divergence of two protein coding sequences using the method of Perler and Efstratiadis. eDiverge is a version of Diverge with command line control.

DNAcomp

Estimates phylogenies from nucleic acid sequence data using the compatibility criterion, which searches for the largest number of sites which could have all states (nucleotides) uniquely evolved on the same tree. Compatibility is particularly appropriate when sites vary greatly in their rates of evolution, but we do not know in advance which are the less reliable ones.

DNAdist

Computes four different distances between species from nucleic acid sequences. The distances can then be used in the distance matrix programs.

- A. the Jukes-Cantor formula,
- B. one based on Kimura's 2- parameter method,
- C. Jin and Nei's distance which allows for rate variation from site to site,
- D. maximum likelihood method using the model employed in DNAML which can be very slow.

DNAinvar

Nucleic acid sequence data from four species. Computes Lake's and Cavender's phylogenetic invariants, which test alternative tree topologies. The program also tabulates the frequencies of occurrence of the different nucleotide patterns. Lake's invariants are the method that he calls "evolutionary parsimony".

DNAml

Estimates phylogenies from nucleotide sequences by maximum likelihood. The model employed allows for unequal expected frequencies of the four nucleotides, for unequal rates of transitions and transversions, and for different (pre-specified) rates of change in different categories of sites, with the program inferring which sites have which rates.

DNAmlk

Estimates phylogenies from nucleotide sequences by maximum likelihood but assumes a molecular clock.

DNApars

Estimates phylogenies by the parsimony method using nucleotide sequences. This program uses the IUB ambiguity codes, and estimates ancestral nucleotide states. Gaps are treated as a fifth nucleotide state.

DNApenny

Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search. This may not be practical for more than 10 or 11 species.

DrawGram

DrawGram interactively plots a cladogram or phenogram-like rooted tree diagram, with many options including orientation of tree and branches, style of tree, label sizes and angles, tree depth, margin sizes, stem lengths, and placement of nodes in the tree.

DrawTree

Interactively plots an unrooted tree diagram.

FastDNAmI

Estimates phylogenies from nucleotide sequences by maximum likelihood (a 'faster' version of DNAML).

Fitsch

Estimates phylogenies from distance matrix data under the "additive tree model" (the distances are expected to equal the sums of branch lengths between the species). Uses the Fitch-Margoliash criterion and some related least squares' criteria. Does not assume an evolutionary clock.

Kitch

Estimates phylogenies from distance matrix data under the "ultrametric" model (the additive tree model with an evolutionary clock assumed). The Fitch-Margoliash criterion and other least squares' criteria are assumed. This program will be useful with distances computed from DNA sequences, with DNA hybridization measurements, and with genetic distances computed from gene frequencies.

Neighbor

Neighbor Joining is a distance matrix method producing an unrooted tree without the assumption of a clock. Branch lengths are not optimized by the least squares' criterion. Fast and for larger data sets.

Protdist

Computes a distance measure for protein sequences using maximum likelihood estimates based on the Dayhoff PAM matrix, Kimura's 1983 approximation to it, or a model based on the genetic code plus a constraint on changing to a different category of amino acid.

Protml

Estimate phylogenies from protein sequences by maximum likelihood.

ProtPars

Estimates phylogenies from protein sequences (input using the standard one-letter code for amino acids) using the parsimony method, in a variant which counts only those nucleotide changes that change the amino acid, on the assumption that silent changes are more easily accomplished.

ReTree

Reads in a tree and allows you to re-root the tree, to flip branches, to change species names and branch lengths, and then write the result out. Can be used to convert between rooted and unrooted trees.

SeqBoot

Produces multiple data sets from a molecular sequence data set by bootstrap, jack-knife, or permutation re-sampling. This can be used together with the consensus tree program CONSENSE to do bootstraps.

PROTEIN STRUCTURE

Antigenic

Antigenic predicts potentially antigenic regions of a protein sequence, using the method of Kolaskar and Tongaonkar.

CoilScan

CoilScan finds coiled-coil segments in protein sequences by comparing each residue in the sequence to a weight matrix tabulated from known coiled-coil protein segments. This prediction method works only for solvent exposed coiled coils, particularly for parallel and anti-parallel two-stranded coiled coils and for parallel three-stranded coiled coils.

Grease

Plots the hydrophobicity profile of a protein using Kyte-Doolittle hydrophathy plot.

HelicalWheel

Plots a peptide sequence as a helical wheel to help you recognize amphiphilic regions.

HelixTurnHelix

HelixTurnHelix uses the Dodd/Egan matrix to test for the presence of a helix-turn-helix DNA-binding motif in a protein sequence.

HTHscan

Scans protein sequences for the presence of helix-turn-helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.

Iep

This calculates the isoelectric point of a protein from its amino acid composition assuming that no electrostatic interactions change the propensity for ionization.

Isoelectric

Isoelectric calculates the isoelectric point of a protein from its amino acid composition assuming no electrostatic interactions occur that perturb ionization.

Moment

Moment plots the height of the hydrophobic moment calculated for all possible angles of rotation for a window that you specify. The contours should help identify peaks of hydrophobic moment at particular positions and angles of rotation.

Octanol

Displays protein hydrophathy

PepCoil

PepCoil identifies potential coiled-coil regions of protein sequences using the algorithm of Lupas, van Dyke & Stock.

PepNet

PepNet is a program to view the two-dimensional helical representation of protein sequences.

PepPlot

PepPlot shows several common measures of protein secondary structure together on one coordinated plot. Most of the curves are the average, sum, or product of some residue-specific attribute within a window. Throughout the plot, the blue curves are for beta-sheets and the red curves are for alpha-helices; black is used for turns and hydrophathy.

PepStats

PepStats gives a short statistical summary on the composition of a protein sequence and gives the molecular weight and isoelectric point.

PeptideStructure

Predicts secondary structure for a peptide sequence. Predictions include (in addition to alpha, beta, coil, and turn) measures for antigenicity, flexibility, hydrophobicity, and surface probability.

PepWindow

Plots measures of protein hydrophathy according to the method of Kyte & Doolittle.

PepWheel

Displays peptide sequences in a helical representation.

PlotStructure

Plots the measures of protein secondary structure in the output file from PeptideStructure. The measures can be shown on parallel panels of a graph or with a two-dimensional "squiggly" representation.

ProfileScan

ProfileScan uses a database of profiles to find structural and sequence motifs in protein sequences.

SPScan

SPScan scans protein sequences for the presence of secretory signal peptides (SPs).

tMap

This program predicts transmembrane segments in proteins,

SEQUENCE FAMILY

HMMbuild

This program reads a multiple protein sequence alignment file to create a Hidden Markov model (HMM).

HMMcalibrate

Correct statistics in a newly created Hidden Markov Model. Takes a HMM and empirically determines parameters that are used to make searches more sensitive, by calculating more accurate expectation value scores.

HMMemit

HMMemit reads a Hidden Markov model (HMM) file and generates from it either a number of sequences or a single majority-rule consensus.

ProfileGap

ProfileGap uses the method of Gribskov, et al. 1987 and Smith and Waterman (1981) to make an optimal alignment between a profile and one or more sequences.

ProfileSearch

ProfileSearch accepts a profile from ProfileMake and uses it to search a database (or any set of sequences you specify) for sequences that are similar to the aligned probe sequences used to create the profile. The output list can be displayed as optimal alignments with ProfileSegments.

ProfileSegments

ProfileSegments creates optimal alignments showing the segments of similarity found by ProfileSearch.

MOLECULAR MODELLING

Determine Hydrogen Bonds

This program determines the optimal hydrogen bonds between the hydrogen of the donor atom and the lone pair of the acceptor atom. It uses four parameters.

1) Distance between the donor and acceptor atom, 2) Distance between the (calculated) hydrogen position, and the acceptor atom, 3) Angle from donor atom over the hydrogen to the acceptor atom, 4) Angle from the hydrogen over the acceptor to a 'virtual' atom.

Determine Interatomic Contacts

This program allows you to list all the contacts between two ranges that you select.

Determine Salt Bridges

Determine Salt Bridges will search for salt-bridges between two ranges. A salt-bridge is defined as a basic nitrogen and an acidic oxygen having their atomic centers within 4.5 Angstroms of each other:

Mutate Residue

This program allows you to mutate a single or multiple residue/s from your PDB file into another type of amino acid. A new PDB file will be generated which can then be used as an input file for any of the other molecular modelling programs.

Overlay

Overlays 2 PDB files. This program is designed for overlaying a PDB file containing 1 to 10 mutations on the original template PDB file.

Suggest Cysteine mutation

This program will try to mutate all residues temporarily into cysteine to find out where a CYS-CYS bridge i.e. a disulfide bond could be formed.

Suggest Mutation

This program will allow you to determine candidates for possible mutation. It will require you to input 2 ranges. For each selected residue you get one number as output. This number is the "happiness" factor. A negative score means that the residue is not optimally "happy" in its environment and may be destabilizing the protein. A positive score indicates that the residue is "happy". Positive scores do not forbid you to make a mutation, but negative scores are just a better bet. Asterisks in your result file indicate where to look first.

Suggest X to Pro Mutation

This program suggests which residues would be good candidates to mutate to Proline. You would want to mutate a residue X to a proline in order to stabilize the protein. Mutating a residue X to a Proline will reduce the number of degrees of freedom that previously existed when the protein folded. This reduction in the number of degrees of freedom (entropy in this instance) which the protein needs to "consider" when folding up is referred to as entropic stabilization. Another example of entropic stabilization is the mutation of glycine residues to alanines.

MOTIF ANALYSIS

Clover

Clover is a program for identifying functional sites in DNA sequences. If you give it a set of DNA sequences that share a common function, it will compare them to a library of sequence motifs (e.g. transcription factor binding patterns), and identify which if any of the motifs are statistically overrepresented in the sequence set.

Dreg

Regular expression search of a nucleotide sequence

FindPatterns

Locates short sequence patterns. FindPatterns can recognize patterns with some symbols mismatched but not with gaps. It supports the IUPAC-IUB nucleotide ambiguity codes for searching through nucleotide sequences.

Fuzznuc

Fuzznuc uses PROSITE style patterns to search nucleotide sequences. Patterns are specifications of a (typically short) length of sequence to be found.

Isochores

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches (>300kb) of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of GC levels (GC is the molar ratio of guanine + cytosine in DNA), which is narrow in cold-blooded vertebrates, but broad in warm-blooded vertebrates. This application plots GC content over a sequence.

MEME

MEME uses the method of Bailey and Elkan (Bailey and Elkan, 1994) to identify likely motifs within the input set of sequences. You may specify a range of motif widths to target, as well as the number of unique motifs to search for. Uses Bayesian probability to incorporate prior knowledge of the similarities among amino acids, into its predictions of likely motifs. The resulting motifs are output as profiles. A profile is a log-odds matrix used to judge how well an unknown sequence segment matches the motif.

Motifs

Looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns. Motifs can display an abstract of the current literature on each of the motifs it finds.

MotifSearch

Uses a set of profiles (representing similarities within a family of sequences) as a query to either a) search a database for new sequences similar to the original family, or b) annotate the members of the original family with details of the matches between the profiles and each of the members. Normally, the profiles are created with the program MEME.

OddComp

Finds protein sequence regions with a biased composition

Patmatdb

Takes a protein motif and compares it to a set of protein sequences. It returns the number of matches there were between the motif and each matched sequence, length of match, start and end positions of match, and writes out an alignment

Patmatmotifs

Compares a protein sequence to the PROSITE motif database.

Preg

Regular expression search of a protein sequence

ProfileMake

Creates a position-specific scoring table, called a profile, which quantitatively represents the information from a group of aligned sequences.

Profit

profit takes a simple frequency matrix produced by prophecy and searches with this to find matches in the input sequence(s) you are searching.

Prophecy

Prophecy produces a simple frequency matrix for use by profit or a position specific weighted profile using either the Gribskov or Henikoff method for use by prophet.

tfScan

Scans DNA sequences for transcription factors using the TRANSFAC database

GENE DETECTION

Backtranslate

Backtranslates an amino acid sequence into a nucleotide sequence. The output helps you identify areas with fewer ambiguities that might be candidates for synthetic probes.

Backtranseq

Translates a Protein sequence into a DNA sequence

Checktrans

Reports STOP codons and ORF statistics of a protein

Chips

Codon usage statistics

CodonPreference

A frame-specific gene finder that tries to recognise protein coding sequences by virtue of the similarity of their codon usage to a codon frequency table or by the bias of their composition (usually GC) in the third position of each codon.

Cpgport

Scans a nucleotide sequence for regions with higher than expected frequencies of the di-nucleotide CG.

Cusp

Reads one or more coding sequences and calculates a codon frequency table. The output file can be used as a codon usage table in other applications.

Flip6Frames

Displays all six frame translations of a given nucleotide sequence.

FlipORF

Finds and translates open reading frames in a nucleotide sequence.

Frames

Shows open reading frames for the six translation frames of a DNA sequence. Frames can superimpose the pattern of rare codon choices if you provide it with codon frequency tables.

GenScan

GenScan is a general purpose gene identification program which analyses genomic DNA sequences from a variety of organisms including human, other vertebrates, invertebrates and plants.

GetORF

Finds and extracts open reading frames (ORFs)

Preg

This searches for matches of a regular expression to a protein sequence.

Quicktandem

Scans for potential tandem repeats in a nucleotide sequence.

Showorf

Showorf displays a nucleic acid sequence with its protein translation in a style suitable for publication. The translation can be done in any frame or combination of frames.

Sigcleave

Sigcleave locates signal sequences and identifies the cleavage site. The method is 95% accurate in resolving signal sequences from non-signal sequences with a cut-off score of 3.5 and 75-80% accurate in identifying the cleavage site.

Syco

Synonymous codon usage Gribkov statistic plot

Terminator

Searches for prokaryotic factor-independent RNA polymerase terminators using the method of Brendel and Trifnov.

TestCode

Identifies potential protein coding regions in nucleic acid sequences by plotting a measure of the non-randomness of the composition at every third base. The statistic does not require a codon frequency table.

Translate

Translates nucleotide sequences into protein. There are two versions of Translate in BioManager.

Wobble

Wobble plots the third position variability as an indicator of a potential coding region.

RESTRICTION MAPPING/PCR

CodeHop

The CodeHop program designs a pool of primers containing all possible 11- or 12-mers for the 3' degenerate core region and having the most probable nucleotide predicted for each position in the 5' non-degenerate clamp region.

Digest

Digest finds the positions where a specified proteolytic enzyme cuts a peptide sequence. It will list the positions where the agent cuts, together with the peptides produced.

Fingerprint

Fingerprint cuts any subrange of a nucleotide sequence at Gs (as if it were digested with T1 ribonuclease) and arranges the fragments in order of their U (or T) content. Within families of U (or T) content, the fragments are arranged by A content and then by C content. Fingerprint shows how the fragments would be labelled if the original molecule had been labelled with any single alpha-(32)P-triphosphate by creating a table of the labelled nucleotides that would be found from an alkaline hydrolysis of each fragment. The labels, after hydrolysis, remain on the 3' side of the nearest neighbour 5' to every nucleotide of the kind labelled.

Map

Map displays a sequence that is being assembled or analysed intensively. Map asks you to select the enzymes whose restriction sites should be marked individually by typing their names. You can choose to have your sequence translated in any or all of the six possible translation frames. You can also choose to have only the open reading frames translated.

MapPlot

MapPlot uses colour to distinguish the types of overhang left after digestion (5' overhangs are green, 3' overhangs are blue, blunt ends are black, and undetermined overhangs are red). The site, cut position, and total number of cuts are also shown for each enzyme. The enzymes that do not cut are listed below the plot. You may choose to plot only enzymes that have six base recognition sites or enzymes that cut the molecule only once.

MapSort

MapSort predicts how the fragments of an enzyme digest will look on a gel. You can concatenate your sequence with its vector before running MapSort to see if a single step isolation is possible and you can examine the pattern of fragments from a multi-enzyme digest. Enzymes that cut the sequence, as well as those that do not, are shown at the bottom of the output. The output contains a complete list of all the enzymes considered. You can see the cut sites graphically with MapPlot or PlasmidMap.

PeptideMap

PeptideMap marks a peptide sequence at every position where a known proteolytic enzyme or reagent might cut it.

PeptideSort

Shows the peptide fragments from a digest of an amino acid sequence. It sorts the peptides by position, putative molecular weight, and relative HPLC retention at pH 2.1, and shows the composition of each peptide. It also prints a summary of the composition of the whole protein.

PlasmidMap

Draws a circular plot of a plasmid construct. It can display restriction patterns, inserts, and known genetic elements. The plot is suitable for publication, record keeping, or analysis.

Prime

Selects oligonucleotide primers for a template DNA sequence. The primers may be useful for the polymerase chain reaction (PCR) or for DNA sequencing. You can allow Prime to choose primers from the whole template or limit the choices to a particular set of primers listed in a file.

Primer3

Primer3 picks primers for PCR reactions, considering as criteria:

- A. oligonucleotide melting temperature, size, GC content, and primer-dimer possibilities,
- B. PCR product size,
- C. positional constraints within the source sequence, and
- D. miscellaneous other constraints.

All of these criteria are user-specifiable as constraints, and some are specifiable as terms in an objective function that characterizes an optimal primer pair.

TACG

TACG searches a nucleotide sequence for matches based on the descriptions stored in a database of restriction enzyme recognition sites (REBASE). It accepts IUPAC degeneracy (yrmkwsbdhv) and performs all possible operations on that sequence.

RNA/DNA

Banana

Bending and curvature plot in B-DNA

Btwisted

Calculates the twisting in a B-DNA sequence

Circles

The circular graph plotted by Circles represents the sequence as a segment of a circle. You can set the radius and the angular width of one base so plots of different secondary structures are strictly comparable. All of the bases that participate in stems are shown with arcs or chords connecting them; hairpin, bulge, interior, and bifurcation loops are easily seen.

Dan

Calculates DNA RNA/DNA melting temperature

Domes

Domes represent a folded RNA sequence as a line with elliptical arcs connecting the bonded bases. This representation has the property that length of the arcs is proportional to the distance (along the primary structure) between the bases; hairpin, bulge, interior, and bifurcation loops are easily seen.

eInverted

eInverted looks for inverted repeats (stem loops) in a nucleotide sequence. It will find inverted repeats that include a proportion of mismatches and gaps (bulges in the stem loop).

eQuicktandem

eQuicktandem scans a sequence for potential tandem repeats up to a specified size.

eTandem

This program is usually used after eQuicktandem has been run to identify potential repeat sizes. It calculates a consensus for the repeat region and gives a score for how many matches there are to the consensus - the number of mismatches.

MFold

Predicts optimal and sub-optimal secondary structures for an RNA or DNA molecule using the most recent energy minimization method of Zuker. MFold calculates energy matrices that determine all optimal and sub-optimal secondary structures for an RNA or DNA molecule. The program writes these energy matrices to an output file.

Mountains

Mountains make a graph that looks like a mountain range. Horizontal striations upon a particular peak are bonds between bases and vertical links between the horizontal striations represent stems.

Palindrome

Palindrome searches for inverted repeats in a DNA or RNA sequences by using a running window which is fixed at the 3' end of the sequence and shifts towards the 5' end by one.

PlotFold

PlotFold displays the optimal and sub-optimal secondary structures for an RNA molecule predicted by MFold.

Squiggles

Squiggles create a representation similar to what you might draw by hand; that is, bonds formed between bases are drawn as chords. The entire structure looks like an airport or intersecting railroad tracks. The spider-like graph plotted by Squiggles represents the sequence as bases connected by chords. Bases are shown participating in stems, as well as in hairpin, bulge, interior and bifurcation loops. These structures are easily seen in Squiggles. To make your graph more readable, you can label the bases, show sequence numbers at set intervals, and pivot up to nine stems.

StemLoop

Locates inverted repeats (stems) within a sequence. You specify the minimum stem length, minimum and maximum loop sizes, and the minimum number of bonds per stem. All stems or only the best stems can be displayed on your screen or written into a file.

STATISTICAL ANALYSIS

CodonFrequency

Tabulates codon usage from sequences and/or existing codon usage tables. The output is correctly formatted for input into the CodonPreference and Frames programs.

Composition

Determines the composition of sequences, including di-nucleotide and tri-nucleotide content.

Cpgplot

Plots the frequency of the occurrence of CpG di-nucleotides and C and G percentage relative to their position in a sequence.

Cpgreport

This program reads in one or more sequences and finds regions where there is a high absolute frequency of CpG dimers as well as a high proportion of CpG compared to GpC.

Freak

Freak takes one or more sequences as input and a set of bases or residues to search for. It then calculates the frequency of these bases/residues in a window as it moves along the sequence.

Geecee

This calculates the fraction of G+C bases of the input nucleic acid sequence(s).

PRSS

Evaluates the significance of pair-wise similarity scores using a Monte Carlo analysis. The Smith-Waterman algorithm is used to calculate similarity scores for the two sequences, and then the second sequence is shuffled and compared with the first sequence, 200-to 1000 times. If the similarity between the two sequences is real then the similarity score for the un-shuffled sequences will be a lot higher than the shuffled score. The score is more informative than just percent similarity, because it takes into account length of match and gaps inserted.

SEG

Replaces low complexity regions in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search, the regions with X characters are ignored.

StatPlot

Plots a set of parallel curves from a table of numbers like the table written by the Windows program. The statistics in each column of the table are associated with a position in the sequence.

WordCount

Counts the commonest words in a sequence and reports them in order of frequency and sequence.

XNU

Replaces statistically significant tandem repeats in protein sequences with X characters. If a resulting protein sequence is used as a query for a BLAST search the regions with X characters are ignored.

FILE MANAGEMENT

Basepairplot

Plots the percentage occurrence and the observed over expected frequency of a di-nucleotide pair relative to their position in a sequence.

Consensus

Extracts the consensus sequence from a multiple sequence alignment, producing either a nucleotide or a protein sequence file.

Corrupt

Uses a random number generator to add errors to nucleotide sequences.

Edit

This program allows you to modify a selected input file e.g. sequence, or alignment file.

Extract Hit Sequences

Extract sequences which match your criteria from blast results.

Extract Query Sequence

Extract the query sequence(s) of the output(s) from database searching programs(s) - e.g. BLAST, FASTA.

Feature Extract

Extracts features from annotated nucleotide sequences.

Filter Hits

Reduce blast files by deleting irrelevant hits.

Mask for Repeat

Screens DNA sequences for low complexity DNA sequences and interspersed repeats. The output of the program is a modified version of the query sequence in which all the sequences found to be similar to the repeat sequence(s) have been masked (replaced by Ns).

Mask for Vector

Screens DNA sequences for vector sequence(s). The output of the program is a modified version of the query sequence in which all the sequences found to be similar to the vector sequence(s) have been masked (replaced by Ns).

Msbar

Mutate sequence beyond all recognition

Newcpgseek

Newcpgseek reports CpG rich regions of a sequence as candidate CpG islands.

Reverse

Reverse reverses and/or complements the symbols in a sequence. The complements of all of the supported IUPAC-IUB nucleic acid symbols are listed.

Shuffle

Shuffle uses a random number generator to scramble the positions of the symbols in a sequence.

Slice

Slice allows you to excise specified regions of a sequence or sequences and concatenate the fragments into a single assembly.

Splitter

This simple editing program allows you to split a long sequence into smaller, optionally overlapping, subsequences.

OTHER PROGRAMS

BoxShade

BoxShade is a program for creating “publish quality” printouts from multiple aligned protein or DNA sequences. The program does no alignment by itself, it has to take as input a file pre-processed by a multiple alignment program or a multiple file editor e.g. ClustalW, PileUp (GCG). In the standard BoxShade output, identical and similar residues in the multiple-alignment chart are represented by different colours or shadings. There are many options concerning the kind of shading to be applied, sequence numbering, consensus output and so on.

Create Set

Create Set allows you to merge any number of files of the same file type to create a set.

FindKm

Find Km and Vmax for an enzyme reaction by a Hanes/Woolf plot

Intersection

Intersection allows you to create a new set from the intersection of two or more sets of the same type. The new set contains the items that were common to all sets.

Union

Union allows you to create a new set by merging the contents of two or more sets of the same type. The new set will contain all of the files that were members of both (all) input sets.