A Quick Guide To MATLAB Bioinformatics Toolbox

http://www.mathworks.com

This is a Quick reference Guide for MATLAB Bioinformatics Toolbox 3. MATLAB (short for "**mat**rix **lab**oratory") is a high-performance language for technical computing, created by *The MathWorks*. It features a family of add-on application-specific solutions called toolboxes (i.e., comprehensive collections of functions) that extend the MATLAB environment to solve particular classes of problems.

Bioinformatics Toolbox offers an integrated software environment for genome and proteome analysis. In particular, it provides access to genomic and proteomic data formats, analysis techniques, and specialized visualizations for genomic and proteomic sequence and microarray analysis. The key features of the basic categories in the Bioinformatics Toolbox will be presented in the following sections.

This guide does not replace the entire documentation provided by *The MathWorks*, but can be used as a reference for those who want to explore the essential features of the Bioinformatics Toolbox. For more information on the functions syntax you can refer to:
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo

Data formats and databases

Web-based databases

getgenbank (…)	Retrieve sequence information from GenBank database
getgenpept (…)	Retrieve sequence information from GenPept database
getembl (…)	Retrieve sequence information from EMBL database
getpir (…)	Retrieve sequence data from PIR-PSD database
getpdb (…)	Retrieve protein structure from PDB database
getgeodata (…)	Get Gene Expression Omnibus (GEO) data
gethmmalignment (…)	Retrieve multiple aligned sequences from the PFAM database
gethmmprof (…)	Retrieve profile hidden Markov models from the PFAM database
gethmmtree (…)	Get phylogenetic tree data from PFAM database

Raw data

scfread (…)	Read trace data from SCF file
joinseq (…)	Join two sequences to produce the shortest supersequence
traceplot (…)	Draw nucleotide trace plots squares polynomial

Reading data formats

genbankread (…)	Read data from a GenBank file
genpeptread (…)	Read data from a GenPept file
emblread (…)	Read data from EMBL file
pirread (…)	Read data from PIR file
pdbread (…)	Read data from PDB file
fastaread (…)	Read data from FASTA report
multialignread (…)	Read multiple sequence alignment file
geosoftread (…)	Read data from a Gene Expression Omnibus (GEO) SOFT file
gprread (…)	Read microarray data from a GenePix Results (GPR) file
galread (…)	Read microarray data from a GenePix array list file
sptread (…)	Read data from a SPOT file
affyread (…)	Read microarray data from Affymetrix GeneChip file
pfamhmmread (…)	Read data from a PFAM-HMM file

Writing data formats

fastawrite (…)	Write to file with fasta format

BLAST searches

blastncbi (…)	Generate a remote BLAST request
getblast (…)	Get BLAST report from NCBI Web site
blastread (…)	Read data from NCBI BLAST report

Sequence Analysis

Sequence Statistics

aacount (…)	Count amino acids in a sequence
basecount (…)	Count the nucleotides in a sequence
codoncount (…)	Count codons in a nucleotide sequence
ntdensity (…)	Plot the density of nucleotides along a sequence
codonbias (…)	Calculate codon frequency for each amino acid in a DNA sequence
seqshoworfs (…)	Display open reading frames (ORFs) in a nucleotide sequence
randseq (…)	Generate random sequence from finite alphabet

Sequence Conversion and manipulation

nt2aa (…)	Convert a nucleotide sequence to an amino acid sequence
aa2nt (…)	Convert amino acid sequence to nucleotide sequence
dna2rna (…)	Convert DNA sequence to RNA sequence
rna2dna (…)	Convert RNA sequence to DNA sequence
seqcomplement (…)	Calculate complementary strand of nucleotide sequence
seqrcomplement (…)	Calculate reverse complement of a nucleotide sequence
atomiccomp (…)	Determine the atomic composition of a protein
aminolookup (…)	Look up the amino acid for a codon
molweight (…)	Determine the molecular weight of a protein
isoelectric (…)	Estimate isoelectric point for amino acid sequence
restrict (…)	Split nucleotide sequence at specified restriction site
cleave (…)	Cleave amino acid sequence with enzyme

Sequence Alignment

nwalign (…)	Globally align two amino acid sequences, using the Needleman-Wunsch algorithm
swalign (…)	Locally align two amino acid sequences using the Smith-Waterman algorithm
multialign (…)	Align multiple sequences using progressive method
profalign (…)	Align two profiles using Needleman-Wunsch global alignment
seqdotplot (…)	Draw a dot plot comparing two amino acid sequences
showalignment (…)	Display color-coded sequence alignment

Phylogenetic Analysis

Phylogenetic Tree Data

phytreeread (…)	Read phylogenetic tree file
phytreewrite (…)	Write phylogenetic tree object to Newick formatted file
phytree (…)	Object constructor for a phylogenetic tree object

Create a Phylogenetic Tree

dnds (…)	Estimate synonymous and nonsynonymous substitution rates
dndsml (…)	Estimate synonymous and nonsynonymous substitution rates using maximum likelihood method
seqlinkage (…)	Construct phylogenetic tree from pairwise distances

seqneighjoin (...) Neighbor-joining method for phylogenetic tree reconstruction
seqpdist (...) Calculate pairwise distance between sequences
phytreetool (...) View, edit, explore phylogenetic tree data
reroot (...) Change the root of a phylogenetic tree
seqinsertgaps (...) Insert gaps into nucleotide or amino acid sequence

Microarray data Analysis

Microarray normalization and filtering

malowess (...) Smooth microarray data using the Lowess method
manorm (...) Normalize microarray data
geneentropyfilter (...) Remove genes the profiles of which have low entropy
genevarfilter (...) Filters out genes with small variance over time
genelowvalfilter (...) Removes genes that have very low absolute expression values

Microarray Visualization

maimage (...) Display a spatial image for microarray data
maboxplot (...) Display a box plot for microarray data
maloglog (...) Create a loglog plot of microarray data
mairplot (...) Display intensity versus ratio scatter plot for microarray signals
mapcaplot (...) Create a Principal Component plot of expression profile data
clustergram (...) Create dendrogram and heat map

Microarray utility functions

probelibraryinfo (...) Extract probe set library information for probe results
probesetlookup (...) Look up gene name for probe set
probesetvalues (...) Extract probe set values from probe results
probesetlink (...) Link to NetAffx Web site
probesetplot (...) Plots values for Affymetrix CHP file probe set

Mass Spectrometry data Analysis

Reading raw data into MATLAB

jcampread (...) Read JCAMP-DX formatted files (JCAMP-DX: file format for infrared, NMR, and mass spectrometry data from the Joint Committee on Atomic and Molecular Physical Data)

Preprocessing of raw data

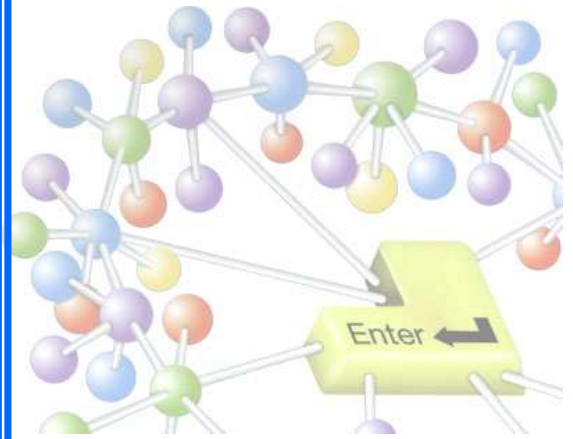
msresample (...) Resample a mass spectrometry signal
msbackadj (...) Correct the baseline of a mass spectrum
msalign (...) Align peaks in mass spectrum to reference peaks
msheatmap (...) Display color image for set of spectra
msnorm (...) Normalize set of mass spectra
mslowess (...) Smooth mass spectrum using non-parametric method
mssgolay (...) Smooth mass spectrum with least-squares polynomial

Spectrum analysis

msviewer (...) Explore MS spectrum or set of spectra with GUI

Statistical Learning

classperf (...) Evaluate performance of classifier
crossvalind (...) Generate cross-validation indices
knnclassify (...) Classify data using nearest neighbor method
knnimpute (...) Impute missing data using nearest-neighbor method
optimalleaforder (...) Determine optimal leaf ordering for hierarchical binary cluster tree
randfeatures (...) Generate randomized subset of features
rankfeatures (...) Rank key features by class separability criteria
samplealign (...) Align two data sets containing sequential observations by introducing gaps
svmclassify (...) Classify data using support vector machine
svmsmoset (...) Create or edit Sequential Minimal Optimization (SMO) options structure
svmtrain (...) Train support vector machine classifier



A Quick Guide
MATLAB Bioinformatics Toolbox