NBIS-UtilityCode Manual

On github: https://github.com/NBISweden/NBIS-UtilityCode

Purpose

The NBIS-UtilityCode repository contains a collection of utilities and tools, ranging from simple file manipulation to statistics and machine learning libraries. The code is written in C++ and compiles on the Linux and Mac OSX operating systems. The Makefile system supports easy prototyping and testing by automatically scanning for main routines and compiling the programs (for more details, see the README on the github site). The NBIS-UtilityCode code base is distributed under the GPL 3.0 license.

General

Input files are automatically uncompressed when presented in gzip (.gz) format. Sequence files can either be in fasta or fastq format and are automatically detected. To print the available parameters and get a brief description, run any program without arguments.

Executables

Simple tools

FastaByLength: filters a fasta or fastq file by sequence length.
FastaN50: prints the N50 of a sequence file.
FastaSizes: prints the size of each entry in a sequence file and/or the total sequence size.
FastaToProteins: translates nucleotides into amino acids in the current frame.
FindORFs: finds open reading frames.
GetColumns: manipulates text files.
SubsetFasta: generates a subset of paired-end read fastq files.
FishersExcactTest: Fisher’s exact test for enrichment.
PearsonCorr: Pearson’s rho with significance p-value.
SpearmanCorr: Spearman’s rho with significance p-value.

Workflow related

Grapevine: a double-abstract meta-scripting processor based on user-configurable syntax. For details, see (doc/GrapevineManual.pdf).
RunGrapevineFlow: a job submission system, currently only supporting SLURM, which can be used to execute grapevine workflows.

Libraries

util/

SysTime.h – get system time and date.
FindProcess.h – reads the list of active processes (Linux only!)
SComm.h – sender and receiver for TCP/IP and UDP/IP sockets.
StreamComm.h – stream communication based on UDP/IP.

SyncConn.h – ‘handshake’ synchronization for two-way communication via TCP/IP or UDP/IP.

base/

ThreadHandler.h – thread objects, mutexes etc. for multithreading.
CommandLineParser.h – parser command line arguments in a type-safe way.
FileParser.h – file parser for ASCII files.
RandomStuff.h – random number generator wrapper.
StringUtil.h – string utilities.
SVector.h – vector utilities (sorting, binary search etc.).
StreamParser.h – reads data from a stringstream like a file.

visual/

This programmatic graphics library is based on the Whiteboard class, for more documentation see https://academic.oup.com/bioinformatics/article/31/12/2054/214244).

ml/

NNet.h – a multi-layered Self Organizing Map organized in a hypertorus. Supports both unsupervised and supervised analyses.
DLNet.h – A backpropagation neural network (currently lacks parallelization).
NPCIO.h – a wrapper around input and output for neural networks.

nl/

This natural language processing library is based on the open source version of the Voxado code base.