# Overview

The software package SiLiX implements an **ultra-efficient algorithm for the clustering of homologous sequences**, based on single transitive links (single linkage) with alignment coverage constraints.

SiLiX is now incorporated in hundreds of projects in the world (in particular in France) and is
widely cited [↗]
.

SiLiX adopts a graph-theoretical framework to interpret similarity pairs as edges of a network. A very efficient algorithm, based on the Disjoint Sets Data Structure, allows the computation of sequence families with **low time and space requirements**.
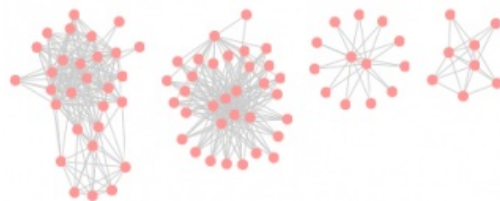
A parallel version of SiLiX, based on MPI, is also available in this package and has been proved to be scalable, so that its allows the study of **very large datasets**.

SiLiX is already included in the analysis pipeline for HOGENOM.

BMC
Bioinformatics

LICENCE

SiLiX is licensed under the
General Public License ↗
.

DOWNLOAD

You can **download the latest version**
HERE ↗
.

Alternatively, you can **clone our git repository**
HERE ↗
.

SYSTEM REQUIREMENTS & DEPENDENCIES

SiLiX is written in ANSI C++ and has been tested on Linux and MacOsX.

Necessary :

* The C++
Boost:program_options ↗
package (include files AND shared library) must be installed. This is free and easy to install on every systems with one of the
following procedures :

> Getting the Debian package on Ubuntu/Debian Linux :

```
sudo apt-get install libboost-dev
sudo apt-get install libboost-program-options-dev
```

> Building and installing from source on other Linux or MacOSX : download Boost and follow this guideline (replace xx_x by

current version number) :

```
tar zxvf boost_1_xx_x.tar.gz
cd boost_1_xx_x/
./bootstrap.sh —with-libraries=program_options
sudo ./b2 install
```

> Installing Boost on MacOSX with MacPorts

Optional :

> *To maximize performance, the C++ Boost:unordered_map class must be installed*

> *For the unit tests performed during the checking, CppUnit must be installed.*

> *To enable parallelism, the MPI library must be installed (tested with openMPI).*

INSTALLATION

Compilation and installation are compliant with **the GNU standard procedure**

URL de la page : https://lbbe-web.univ-
lyon1.fr/fr/SiLix

```
tar zxvf silix-1.x.x.tar.gz
cd silix-1.x.x
./configure
make
make check
make install
```

but additional **optional configure options :**

> enabling the use of MPI library and switch to the parallel version of SiLiX

```
enable-mpi
```

> enabling the use of Boost:unordered_map class

```
enable-hash
```

> specifying a path where the programs must be installed

```
prefix=install_path
```

> verbose mode (not important). Enabling the output of % identity and % coverage information in the file created with —net
option

```
enable-verbose
```

PROGRAMS USE (important changes since versions 1.2.x)

Two programs with man pages are available :

> **Filtering+Clustering** : The user provides a fasta file and the result file(s) of a all-against-all BLAST search in tabular format (
> -outfmt 6 option in blastall, i.e. query id, subject id, percent identity, alignment length, number of mismatches, number of
> gap openings, query start, query end, subject start, subject end, Expect value, HSP bit score)

```
silix [OPTIONS] <FASTAFILE> <BLASTFILE>084D4F
```

To get information or help :

```
man silix
silix --help
```

> **Clustering** : The user provides an input a list of pairs of sequence IDs

```
silixx <NB> <FILE>
```

To get information or help :

```
man silixx
silixx --help
```

Running the parallel version of silix

First, the user must have a collection of N blast result files to be processed in parallel.

After having use ./configure with the option —enable-mpi, the user must adopt the classical way to run a program using MPI :

```
mpirun -np NP silix [OPTIONS] <FASTAFILE> <MULTIBLASTFILE>
```

with NP the chosen number of processors (in practise, NP<=N)

WARNING

Many users contact us because they got the following error :

```
Error in Fam : unable to open file xxxx
```

where "xxxx" is the id of the first sequence in the blast output file.

What does this error mean ? If you configure silix with the option "—enable-mpi" (see previous section), then the command line interface is slightly different : a MULTIBLASTFILE is expected, not a BLASTFILE !

CLASSICAL SKETCH (important changes since versions 1.2.x)

In the following, we use auxiliary programs that are in the utils/ directory of the package, but not installed.

> Blasting all versus all

```
formatdb -i seq.fasta -n seq.db
blastp -db seq.db -query seq.fasta -outfm 6 -out blastall.out
```

or for older versions of Blast :

```
blastall -p blastp -d seq.db -i seq.fasta -m 8 -o blastall.out
```

> Running silix. Requires fasta files. The options are the filtering parameters, with the following default values (see

Penel et al, BMC Bioinformatics, 2009 ↗

) :

two sequences in a pair are included in the same family if remaining HSPs (Homologous Segment Pairs) cover at least 80% of the
protein length and if their similarity is over 35% ; a partial sequence is included if its length is ≥100 amino-acids or ≥50% of the
length of the complete protein.

```
silix  seq.fasta blastall.out -f FAM > seq.fnodes
```

Here, we specified a prefix "FAM" for the family ids.

Nota Bene: For running the parallel version, the user displays the list of blast results files to be processed in parallel

```
mpirun -np 4 silix  seq.fasta filenames.txt -f FAM > seq.fnodes
```

where "filenames.txt" is

```
blastall1.out
blastall2.out
blastall3.out
blastall4.out
```

> Retrieving family sizes

```
utils/silix-fsize seq.fnodes > seq.fsize
```

> Splitting sequences in multiple fasta files

```
silix-split seq.fasta seq.fnodes
```

> With —net option, a file blastall.net is created and contains all the pairs taken into account after filtering. A message is
sent to STDERR :

```
Inflating blastall.net
```

Naming Conventions

".net" are the extension for files of format

```
SEQID1 SEQID2
SEQID3 SEQID4
```

".fnodes" are the extension for files of format

```
FAMID1 SEQID1
FAMID1 SEQID2
FAMID2 SEQID3
FAMID2 SEQID4
```

References

SiLiX is developed by :
- Laurent Duret
- Vincent Miele
- Simon Penel

If you use SiLiX in a published work, please cite the following reference :

Miele,V., Penel, S. and Duret,L., Ultra-fast sequence clustering from similarity networks with SiLiX, BMC Bioinformatics

URL de la page : https://lbbe-web.univ-lyon1.fr/fr/SiLix

Contact

For any bugs, information or feedback, please contact :

Vincent Miele

Contact