FastaTools

FastaTools is a Graphical User Interface Perl made application that performs several operations to Fasta formatted, protein databases (DB). It has multiple features and it's been designed to be a useful tool for anyone working on the Proteomics field: an easy to use interface and a complete documentation available in this page will allow non experienced users to easily perform advanced bioinformatics tasks. FastaTools runs on almost any operating system (Linux, Windows, Mac OS) with some little requirements (described bellow). A Windows installer is also supplied with no requirements at all (not even a Perl installation is needed to run the software). To download the latest version of FastaTools, you can use the Project Page or go to the Download section of this page. If you want to learn on how to install and use FastaTools application, please go to the Documentation section of this page.

The main features of FastaTools are: **DB exploration** (the first 200 lines of the database are shown), **DB analysis** (the number of lines, proteins and peptides generated after digestion by an enzyme are reported), **Decoy DB generation** (a decoy database is generated using the specified DB), **DB Search** (a simple text or a list of terms can be introduced to be searched inside the DB) and **DB Join** (joins multiple Fasta databases).

We will greatly appreciate any comment, review or suggestion about the code, application, actions performed or new ideas about FastaTools. Feel free to visit the Project Page, or use the Contact section of this page.

Documentation

This documentation is provided in pdf format in the downloads section from the project's web page (http://code.google.com/p/fastatools/downloads/list).

Download options and install instructions

In the download section from the project's web page (http://code.google.com/p/fastatools/downloads/list), FastaTools can be downloaded as a Windows installer (FastaTools.exe) or as a single Perl script (fastatools.pl), than works over virtually every platform.

- **fastatools.pl** : This program has been tested on Linux Fedora 9.0, OpenSuse 11.0 Linux and MS Windows XP. There are two important dependencies: the Perl programming language and wxWidgets libraries (Perl bindings to the wxWindows cross-platform toolkit). For wxPerl installation, please visit the official wxPerl page: http://wxperl.sourceforge.net, where sources can be downloaded. However, pre-compiled packages of the wxPerl library are available in most of the Linux distributions. In Linux Fedora 9.0, the package perl-Wx-0.74-1, included in the main distribution, can be used to automatically install the libraries. In OpenSuse 11.0, the package perl-Wx-0.84-19 is available at the OpenSuse project page (http://software.opensuse.org/search).
Once all requirements have been satisfied, the downloaded script can be executed doing: `perl /path_to_program/fastatools.pl`

- **FastaTools.exe** : If you are under a MS Windows system, you can download and install the provided Windows installer, where no dependencies exist (nor Perl, nor Wx installed). Only double-click over the installer icon and a guided process will install the program into your computer. You can also uninstall the program as usual in Windows systems. The installer package has been built using the "Cava Packager" application (http://www.cava.co.uk/).

Graphical user interface

FastaTools has been designed with the aim to provide a very easy-to-use application. In the next figure, the main components of the user interface are described.

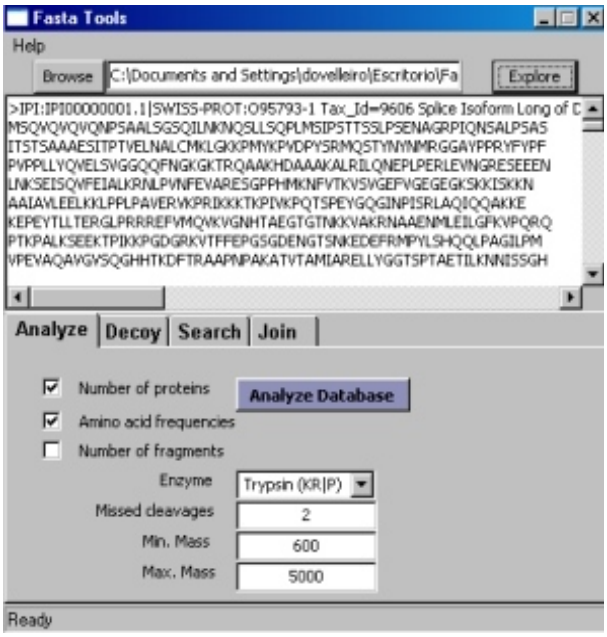


Figure 1.

'Main Fasta database' area

When the application is invoked, the **first thing to do** is choosing a Fasta formatted database to work. This is performed with the 'Browse' button in the 'Main Fasta database' area. Using the 'Explore' button to the right, the first 200 lines of the Fasta file will be displayed at the 'Information window'. It's imperative to select a Fasta formatted protein database in the 'Main Fasta database' area.

Information Window

All the information produced by FastaTools will be offered in two ways:

Using the information window: all error messages (like 'you have forgotten to choose a database'), analysis results (for example: 'This protein contains 123,456 enzymatically produced fragments') and informative remarks will be displayed through the information window.

Writing in a text file: if you choose to produce a new database, or a long list of peptide-protein pairs, FastaTools will produce a text file in the same path (at the same directory level) that the main database (displayed in the 'Main Fasta database' area).

Functions Tabs - Functions Panel

You can select four different main functions with FastaTools: Analyze, Decoy, Search and Join. Pressing the appropriate tab, a set of controls will be displayed in the Functions Panel. This will let you to perform the desired operation.

Status Bar

In the left bottom of the graphical interface, the state of the application ('Ready', or the progress of a routine) will be displayed.

FastaTools functions

The five functions provided by FastaTools will be described here. For more information and a practical approach, please download the FastaTools Tutorial available at http://code.google.com/p/fastatools/downloads/list

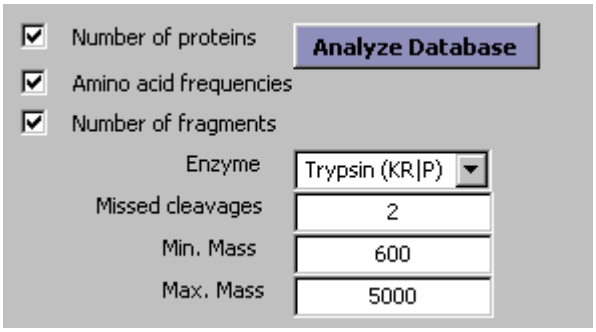
Database exploration

Once a Main database has been selected, press the “Explore” button and the first two hundred lines will be displayed.

Database analysis

The Analyze function allows the user to know three important properties of a Protein database:

1. the number of proteins contained
2. the amino acid frequencies for the 20 canonical and other forms present (for example, it's frequent to find an 'X' letter as an 'unknown amino acid'); database providers use different formats to express less common amino acid forms
3. the number of fragments produced by the chosen enzyme, in the selected mass range and taking into account a number of missed cleavages



Analyzing big databases (more than 50 MB) can be very time consuming, and it will probably need a lot of RAM from your computer: a 100 MB database will need at least 2 GB of free RAM.

Take into account that the higher number of missed cleavages and the wider mass interval used in the fragmentation pattern, longer will be the analysis time and higher the amount of memory needed.

When the database analysis has finished, the status bar shows the “Ready” message. Other tasks can be done with FastaTools then. However, it is better to restart the FastaTools application if more than one analysis operations are being performed: that will help to maintain the computer's memory at its minimum level. If big databases are analyzed, it is also advisable to turn off other applications (like browsers or spreadsheets).

Five different cleavages are allowed using three different enzymes (Trypsin-KR, Trypsin-KR|P, GluC-DE, GluC-E and Endolysin-K). The symbols after the enzyme name refer to the amino acids used to cleave (cleavage is made after this amino acid), and the | symbol indicates an amino acid that blocks the cleavage (for example: KR|P means that cleavage is made after K or R but not if the next amino acid is a P).

The two first options are selected by default. You can choose any combination of them, or all of them. Take into account, that the calculation of fragments produced by enzyme cleavage is a highly RAM intensive process. FastaTools counts the non-redundant peptides produced: it means that it has to store into memory large amounts of data. To prevent failures on the system memory, the database size is limited (only if you choose to analyze the enzymatically produced fragments) to 100 MB.
The produced output can be copied and pasted into a spreadsheet program to easily compare and analyze the results. Figure 2 displays the FastaTools produced information analyzed using LibreOffice/OpenOffice program Calc.

The amino acid composition will inform us about the presence of non typical forms of amino acids (in the Figure 2 appear four of these forms: X, Y, B and Z).
Parameters provided in the enzymatic cleavage are very useful to determine the “quality” of a database, in order to use low redundancy protein sources.

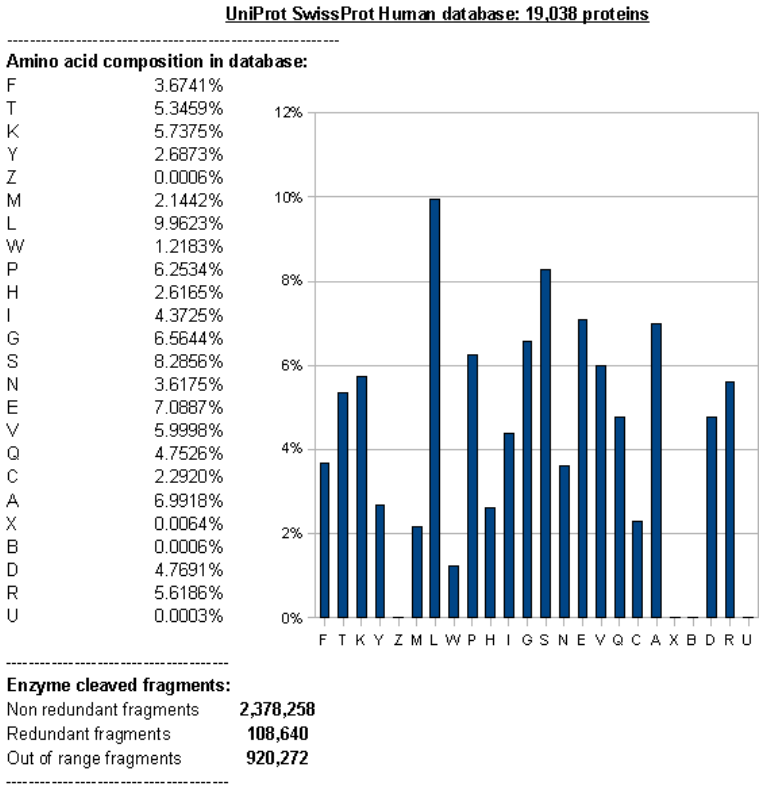


Figure 2.

Decoy databases generation

Decoy databases have become an extended practice in Proteomics. These databases are usually appended at the end of a normal database with the main purpose of obtaining an estimation of the accuracy of the peptide identifications: peptide identifications pointing to decoy sequences (in principle

not present at biological sequences), are then product of a wrong identification.

To obtain a decoy database, the original database is processed, and two processes are made: first, the header is generated using an index like 'decoy_16754'. Second, the original sequence is processed to obtain a derived sequence. FastaTools can use three different strategies to generate the decoy database: the reverse, pseudo-reverse and random strategies. There are outlined at figure 3.

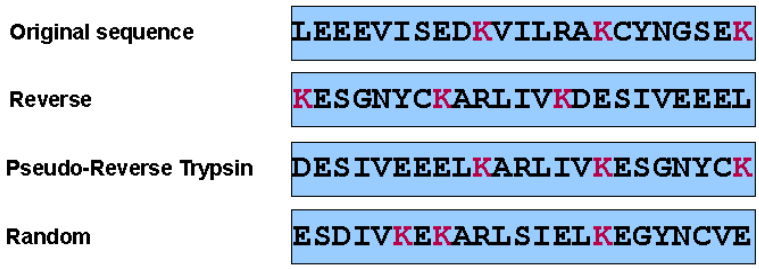


Figure 3. The three strategies implemented in FastaTools produce three different patterns in protein sequences. Reverse strategy produces only the reversion of the whole protein sequence. The pseudo-reverse strategy implies two steps: first, sequence is digested using an enzyme, the resulting peptides are individually reversed and last, they are joined together. FastaTools provides four different digestion patterns: Trypsin (KR|P), Trypsin (KR), GluC (DE) and Endolysin (K). For example: in Trypsin (KR|P), a cut is produced after the K or P amino acid symbols, except if following the K or R symbols, a P is found. The random strategy randomizes the whole protein sequence.

To obtain a decoy database, first, the user must select a “Main database”, go to the “Decoy” function selecting the appropriate tab function, select one decoy strategy (using “Decoy type” menu), and pressing the “Generate decoy” button. A new database will be generated in the “Main database” path, with the name provided in the “Out name” field. The user can append the new database to the original by selecting the “Attach to the original database” option.

Search in the database

The “Search database” function, allows a user to search in the headers or the sequences for a text, producing different outputs depending on the search type. All the outputs will be generated in the same path where the “Main Fasta database” is located. In the search field, the user can type a text to search (for example: “kinase” if a search into the headers is performed, or maybe “ELVI” if a sequence is searched), or can use a list using the “use list” button. The list must be a text file with one or more lines: each line will be used as a text to search. There are four different search types:

1- Search headers, generate sub-DB (Fasta format)

The headers of the “Main Fasta database” are searched. The produced output will be a new Fasta formatted database. When the searched text is found, the protein containing this text (the header and the sequence) will be included in the new sub-Database.

2- Search sequences, generate sub-DB (Fasta format)

The sequences of the “Main Fasta database” are searched. The produced output will be a new Fasta formatted database. When the searched sequence is found, the protein containing this sequence (the header and the sequence) will be included in the new sub-Database. Amino acids are written in capitalized letters: the user must use then capital letters if sequences are searched.

3- Search sequences, generate peptide-protein list

The sequences of the “Main Fasta database” are searched. The produced output will be a text file with two columns, separated by a Tab key character: the left field contains the sequence, and the right field contains the one protein in the “Main database” containing this peptide. If the peptide is found in more than one protein, a new line will be added, with the same peptide and another protein. If a list of

sequences is provided, each peptide contained in the list will be searched.

4- Search sequences, generate protein-peptide list

The sequences of the “Main Fasta database” are searched. The produced output will be a text file with two columns, separated by a Tab key character: the left field contains a protein name, and the right field contains the complete list of provided peptides contained in the protein, separated by a Tab key character. If a list of sequences is provided, all the proteins presenting those sequences will be listed.

Join databases

This function allows the user to join one or two databases to the “Main Fasta database”. The “File A” field stores the same database as the selected in the “Main Fasta Database” field. The fields “File B” and “File C” will provide the databases to append to the “File A”: pressing the “Browse” button, the user will select the desired database. The “File D” field contains the name of the output database, which will be created in the same path as the “Main Fasta database” file.

Tutorials

This short tutorials will describe the different functionalities of FastaTools. To follow the instructions, it is necessary that you download the following files:

- In first place, you'll need a Fasta formatted database; examples described in this tutorial use the SwissProt database (UniProtKB/Swiss-Prot Release 56.2 of 23-Sep-2008) available at ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release/ . The file you want is **uniprot_sprot.fasta.gz.** Once you have downloaded it, you'll need to decompress it: use WinZip (or a similar application) to do so.
- In second place, you'll also need the files provided in the tutorial_files file (tutorial_files.zip), available at http://code.google.com/p/fastatools/downloads/list . Once downloaded, decompress the contents using WinZip or a related application.

All examples are executed using a 2.99 GHz and 2 MB RAM computer: times will be therefore only an orientation and they will widely depend on your computer.

Tutorial 1: Generation of a Human Database

In this tutorial, a sub-database of human proteins will be generated from a general database (UniProt-SwissProt).

Select the “Main Fasta database”-> browse to the fasta file location and select the database

Press the “Explore” button-> the following output will appear in the “Information window”:

```
>sp|Q4U9M9|104K_THEAN 104 kDa microneme/rhoptry antigen OS=Theileria annulata GN=TA08425 PE=3 SV=1

MKFLVLLFNILCLFPILGADELVMSPIPTTDVQPKVTFDINSEVSSGPLYLNPVEMAGVK

YLQLQRQPGVQVHKVVEGDIVIWENEEMPLYTCAIVTQNEVPYMAYVELLEDPDLIFFLK

(...)
```

As you can see, the taxonomy identifier used is “_THEAN ”. We can assume that the human identifier will be “_HUMAN”. Using this identifier we'll generate a human sub-database

Go to the “Search” function using the appropriate tab and make sure that the “1- search headers, generate sub-DB (fasta format)” mode is selected. The, write the text “_HUMAN” in the “Search” field and press the “Begin Search” button. The following text will appear in the “Information window”:

```
The following strings are going to be searched:

_HUMAN

---- A total of 1 strings to search----

Search complete.

 A file G:\00_fastatools\sub_DB.fasta

 has been generated with the results

A total of 398181 proteins have been scanned, and 20327 matches have taken place

Total time: 58 secs
```

A file named “sub_DB.fasta” will be generated in the same path as the “Main database”. It is ten times smaller than the original file, and contains only human entries.

Using the correct taxonomy identifier is a key step in the elaboration of a sub-database on the basis of taxonomic principles. If we search for the text “Human” (instead of the correct taxonomy identifier: “_HUMAN”), a 2153 proteins database is generated, containing only human viruses.

The next table shows the human taxonomy identifiers used in some common protein databases.

Database	Human key
RefSeq	[Homo sapiens]
IPI	Tax_Id=9606
Uniprot (SwissProt/Trembl)	_HUMAN
MSDB	- human
UniRef-100	Tax=Homo sapiens
NCBI nr	[Homo sapiens]

Tutorial 2: Database Analysis

In this tutorial, the protein database will be analyzed to obtain the number of proteins contained in the database, the amino acid frequencies and the number of peptides produced by a virtual digestion with trypsin.

Select as the “Main database” the human database generated in Tutorial 1.

Go to the “Analysis” functionality using the appropriate tab and select the three fields: Number of proteins, Amino acid frequencies and Number of fragments (using Trypsin (KR|P), 2 missed cleavages and 600-5000 as the mass range). The press the “Analyze Database” and wait until the operation has finished. The following output will be generated:

```
Please, wait while the analysis is performed
Number of proteins in database: 20,327

---------------------------------------

Amino acid composition in database:

F       3.6630%          S       8.3142%

T       5.3590%          N       3.5970%

K       5.7309%          E       7.0856%

Y       2.6703%          V       5.9815%

Z       0.0006%          Q       4.7592%

M       2.1343%          C       2.2993%

L       9.9629%          A       6.9977%

W       1.2231%          X       0.0061%

P       6.2899%          B       0.0006%

H       2.6283%          D       4.7393%

I       4.3510%          R       5.6341%

G       6.5716%          U       0.0003%

---------------------------------------

Enzyme cleaved fragments:

Non redundant fragments   2,641,813

Redundant fragments   148,870

Out of range fragments   1,022,853

---------------------------------------

Total time: 242 sec
```

The most time consuming process in this analysis is the virtual fragmentation produced when the “Number of fragments” option is selected. Selecting only the “Amino acid frequencies” option will take 35 seconds to complete, and selecting only the “Number of proteins” option will take 2 seconds. Take into account that the “Number of fragments” option should be only selected using a computer with a great amount of RAM (at least 2 GB), and analyzing relatively small databases (no more than 50 MB). Not doing so will probably run your computer out of memory.

Tutorial 3: Decoy database generation

In this tutorial, a decoy database will be generated.

Select as the “Main database” the human database generated in Tutorial 1.

Go to the “Decoy” functionality using the appropriate tab and select the “Reverse” option in the “Decoy type” field. Select the “Attach to the original DB” option as marked and press the “Generate Decoy” button. A new database named “decoyDB.fasta” will appear in the “Main Database” path. Its size is double the size of the original database (usually less, because of the shortened headers size). If you open the generated database with a suitable text editor, you'll see the first half of the new database with the proteins contained in the original “Main Database”, and the second half with the reversed sequences. The headers of the decoy part will be a succession from “>decoy_1” to “>decoy_20327”.

Tutorial 4: Database search → generating sub-databases

In this tutorial we'll generate sub-databases, searching a word in the headers or looking for peptides in the protein sequences.

Select as the “Main database” the human database generated in Tutorial 1.

Go to the “Search” functionality using the appropriate tab and select the mode “1- search headers, generate sub-DB (fasta format)”. In the search field write the text “kinase”. In the “Out Name” field write “kinases.fasta”. Press the “Begin Search” button: a new fasta database named “kinases.fasta” will appear in the “Main Database” path, containing 702 proteins. Now, repeat the operation only changing the search text as “protease” and the “Out Name” as “proteases.fasta”: a database with 93 proteins will be generated. Now, press the “use list” button and browse to the file “list_1.txt” included in the “tutorial_files.zip” file. This file contains two lines, each of them containing the words “kinase” and “protease”. Change the output name as something like “kinases_or_proteases.fasta” and press the “Begin search” button: a new database will be generated, containing this time 795 proteins.

The same process could be repeated, using now sequences. Select the mode “2- search sequences, generate sub-DB (fasta format)”. Try searching first for one sequence (for example “ELVIS”, with 12 matches) and then for other sequence (for example “LIVES”, with 9 matches). Use the “list_2.txt” file, included in the “tutorial_files.zip” file. This file contains two lines, each of them containing the sequence “ELVIS” and “LIVES”. Search again: a new database will be generated, containing 21 proteins.

Tutorial 5: Database search → generating paired lists

Select as the “Main database” the human database generated in Tutorial 1. Go to the “Search” functionality using the appropriate tab and select the mode “3- search sequences, generate peptide-protein list”. Now, select the “list_3.txt” file, included in the “tutorial_files.zip” file, with the “Use list” button. This file contains nine lines, each of them containing a peptide to be searched. Type an appropriate name in the “Out Name” file (for example, “pair1.txt”) and press the “Begin Search” button: a paired list will be generated in the “Main Database” path. Repeat the whole procedure, only changing the “Out Name” file name (for example, “pair2.txt”) and the mode of operation (selecting 4-search sequences, generate protein-peptide list): a second paired list will be generated. If you open the two files, the differences will be clear:

Peptide-Protein List

LEDLVCFWEEAASAGVGPGNYSFSYQLEDEPWK	>sp|P19235|EPOR_HU...
EAASAG	>sp|Q8IWZ3|ANKH1_H...
EAASAG	>sp|P19235|EPOR_HU...
EAASAG	>sp|P51843|NR0B1_H...
EAASAG	>sp|Q96CW6|S7A6O_H...
RHFYVAK	>sp|Q9BYK8|PR285_H...
SGIPEIK	>sp|P51790|CLCN3_H...
SGIPEIK	>sp|P51793|CLCN4_H...
SGIPEIK	>sp|P51795|CLCN5_H...
EDEPWK	>sp|P19235|EPOR_HU...
DEMVEQEFNR	>sp|O60341|LSD1_HU...
LFSVAR	>sp|Q9BYK8|PR285_H...
CTDAIVSFISRHFYVAK	>sp|Q9BYK8|PR285_H...
TGAPNECR	>sp|Q9HB63|NET4_HU...

Protein-Peptide List

>sp|P51843|NR0B1_H...	EAASAG		
>sp|Q8IWZ3|ANKH1_H...	EAASAG		
>sp|P51793|CLCN4_H...	SGIPEIK		
>sp|P51790|CLCN3_H...	SGIPEIK		
>sp|P19235|EPOR_HU...	LEDLVCFWEEAASAGVGPGNYSFSYQLEDEPWK	EDEPWK	EAASAG
>sp|O60341|LSD1_HU...	DEMVEQEFNR		
>sp|Q96CW6|S7A6O_H...	EAASAG		
>sp|P51795|CLCN5_H...	SGIPEIK		
>sp|Q9HB63|NET4_HU...	TGAPNECR		
>sp|Q9BYK8|PR285_H...	LFSVAR	CTDAIVSFISRHFYVAK	RHFYVAK

Downloads

To download FastaTools, you can go to http://code.google.com/p/fastatools/downloads/list . There, you'll find:

1. Windows installer.
2. Perl script.
3. This documentation in pdf format.
4. Tutorial files.

Contact

Please, feel free to write a mail and send us any comments, suggestions or whatever.

Proteomics Laboratory at CSIC/UAB
Phone number: +34 93 581 48 55

lp.csic@uab.cat

http://proteomica.uab.cat

https://www.facebook.com/LPCSICUAB

https://twitter.com/LPCSICUAB