UNIVERSITY OF Southampton
School of Biological Sciences

PICSI: Proteomics Identification from Cross-Species Inference

Richard J. Edwards © 2010.

Contents

Tables

1. Introduction

This manual gives an overview of PICSI, a utility for simple assembly and annotation of EST data using ungapped alignment and BLAST homology searches or for searching EST libraries for query proteins of interest. General details about Command-line options can be found in the RJE Appendices document included with this download. Details of command-line options specific to PICSI can be found in the distributed readme.txt and readme.html files.

Like the software itself, this manual is a 'work in progress' to some degree. If the version you are now reading does not make sense, then it may be worth checking the website to see if a more recent version is available, as indicated by the Version section of the manual. Check the readme on the website for up-to-date options etc. In particular, default values for options are subject to change and should be checked in the readme.

Good luck.

Rich Edwards, 2010.

1.1. Version

This manual is designed to accompany PICSI version 1.0.

The manual was last edited on 15 January 2010.

1.2. Using this Manual

As much as possible, I shall try to make a clear distinction between explanatory text (this) and text to be typed at the command-prompt etc. Command prompt text will be `written in Courier New` to make the distinction clearer. Program options, also called 'command-line parameters', will be **`written in bold Courier New`** (and coloured **`red`** for fixed portions or **`dark`** for user-defined portions, such as file names etc.). Command-line examples will be given in (green) *`italicised Courier New`*. Optional parameters will (if I remember) be `[`in square brackets`]`. Names of files will be marked in coloured normal text.

1.3. Why use PICSI?

PICSI is a program for cross-species protein identifications using searches against NCBInr, for example. MASCOT results files are processed using BUDAPEST. Hits are then converted into peptides for redundancy removal. Hits from a given (known) query species are preferentially kept and any peptides belonging to those hits are purged from hits returned by other species.

All protein hits are then classified:

- UNIQUE = Contains 2+ peptides, including 1+ unique peptides
- NR = Contains 2+ peptides; None unique but 1+ peptides only found in other "NR" proteins
- REDUNDANT = Contains 2+ peptides but all found in proteins also containing UNIQUE peptides
- REJECT = Identified by <2 peptides once query-species peptides subtracted

Hits that are not rejected are clustered by BLAST. Protein classifications are output along with the individual peptide classifications.

1.4. Getting Help

Much of the information here is also contained in the documentation of the Python modules themselves. A full list of command-line parameters can be printed to screen using the **`help`** option, with short descriptions for each one:

```
python PICSI.py help
```

General details about Command-line options can be found in the PEAT Appendices document included with this download. Details of command-line options specific to PICSI can be found in the distributed readme.txt and readme.html files.

If still stuck, then please e-mail me (r.edwards@southampton.ac.uk) whatever question you have. If it is the results of an error message, then please send me that and/or the log file (see Chapter 2) too.

1.4.1. Something Missing?

As much as possible, the important parts of the software are described in detail in this manual. If something is not covered, it is generally not very important and/or still under development, and can therefore be safely ignored. If, however, curiosity gets the better of you, and/or you think that something important is missing (or badly explained), please contact me.

1.5. Citing PICSI

PICSI is part of a manuscript in preparation (Jones *et al.* in prep.). Until published, please cite the PICSI Website.

1.6. Availability and Local Installation

PICSI is distributed as a number of open source Python modules. It should therefore work on any system with Python installed without any extra setup required. If you do not have Python, you can download it free from www.python.org at http://www.python.org/download/. The modules are written in Python 2.5. The Python website has good information about how to download and install Python but if you have any problems, please get in touch and I will help if I can.

All the required files should have been provided in the download zip file. Details can be found at http://www.southampton.ac.uk/~re1u06/software/ and the accompanying PEAT Appendices document. The Python Modules are open source and may be changed if desired, although please give me credit for any useful bits you pillage. I cannot accept any responsibility if you make changes and the program stops working, however!

Note that the organisation of the modules and the complexity of some of the classes is due to the fact that most of them are designed to be used in a number of different tools. As a result, not all the options listed in the __doc__() (`help`) will be of relevance. If you want some help understanding the way the modules and classes are set up so you can edit them, just contact me.

2. Fundamentals

2.1. Running PICSI

2.1.1. The Basics

If you have python installed on your system, you should be able to run PICSI directly from the command line in the form:

```
python PICSI.py seqin=FILENAME
```

To run with default settings, no other commands are needed. Otherwise, see the relevant sections of this manual.

IMPORTANT: If filenames contain spaces, they should be enclosed in double quotes: *data="example file"*. That said, it is recommended that files do not contain spaces as function cannot be guaranteed if they do.

2.1.2. Options

Command-line options are suggested in the following sections. General details about Command-line options can be found in the RJE Appendices document included with this download. Details of command-line options specific to PICSI can be found in the distributed readme.txt and readme.html files. These may be given after the run command, as above, or loaded from one or more *.ini files (see RJE Appendices for details).

2.1.3. Running in Windows

If running in Windows, you can just double-click the PICSI.py file. It is recommended to use the **win32=T** option. (Place this command in a file called PICSI.ini.)

2.2. Input

PICSI requires a set of MASCOT results files and the relevant protein hits downloaded from NCBI or UniProtKB in fasta format. For full functionality, the UniProtKB (Bairoch, et al., 2005) species code for the query organism (*i.e.* the organism on which the experiments were performed) and a delimited text file of UniProtKB species code mappings (generated by RJE_UNIPROT).

2.3. Output

Primary output for PICSI is a delimited text file [picsi.clean.tdt] containing the processed protein hits (Table 2.1).

In addition, each MASCOT search has a peptide table produced [picsi.SEARCH.peptide.tdt] that lists all the peptide sequences used for identifications, which hits they identified and whether they are UNIQUE or common to a CLUSTER of BLAST-related (Altschul, et al., 1990) proteins (or COMMON across clusters, which is unlikely).

2.4. Commandline Options

A full list of commandline options can be found in the readme file or by running:

```
python picsi.py help
```

Table 2.1. Fields for main PICSI output file.

Field	Description
search	MASCOT search ID
hit	Protein hit number from MASCOT
class	PICSI classification of protein (see 1.3)
cluster	BLAST-based clustering ID for non-REJECT proteins (independent numbering for each search)
accnum	Accession number of protein
spec	Species code for protein
species	Full species (if possible) for protein
desc	Protein description
pepcount	Initial MASCOT peptide count, including PTMs etc.
pep_con	No. of different peptides "converted" by PICSI - PTMs ignored and MS ambiguities (e.g. Ile vs Leu) considered
pep_rem	No. of peptides removed due to being found in hits from query species
pep_uniq	No. unique peptides, found only in this hit
pep_nr	No. peptides found in 2+ "NR" proteins (see 1.3)
pep_red	No. redundant peptides also found in a UNIQUE protein (see 1.3)
peplist	Original list of peptides
conpep	Converted PISCI peptide list

3. Appendices

3.1. Troubleshooting & FAQ

There are currently no specific Troubleshooting issues arising with PICSI. Please see general items in the PEAT Appendices document and contact me if you experience any problems not covered.

3.2. References

- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. J Mol Biol, 215: 403-410.
- Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N and Yeh LS (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res., 33: D154-159.