

PIDC – User Guide ver. 0.92

University of Helsinki, Department of Computer Science,
FDK Research Unit, Project ICOMIC
<http://www.cs.helsinki.fi/research/icomica>

Contents

1. Introduction to PIDC
2. Installation
3. Running PIDC
4. Format of input files
5. Format of output files
6. Editing the Initialization file

1. Introduction to PIDC

This document gives an introduction to PIDC (Positional Isotopomer Distribution Calculator), a software package that computes a positional isotopomer distribution of the selected atom type of the compound when mass spectra of the compound and its fragments are given. PIDC can analyze mass spectrometric data originating from GC-MS and LC-MS, when either full scanning or daughter ion scanning is used in the fragmentation stage. A more verbose description to the computational methods applied in the PIDC is given in the manuscript “Computing Positional Isotopomer Distributions from Tandem Mass Spectrometric Data” [1].

PIDC-application runs in the MATLAB 6® -environment (<http://www.mathworks.com/products/matlab>). PIDC might also work with the previous versions of MATLAB, but this can not be guaranteed or supported. PIDC is tested in the Windows® and Linux operating systems.

PIDC is distributed as a collection of MATLAB’s M-source files. These source files or part of them can freely be used in other freely distributed applications. However, it is prohibited to redistribute edited source files under name PIDC without permission. PIDC is a research prototype, not intended for commercial use. It is distributed “as is” and authors do not take any responsibility of any harm or damage caused by PIDC. Notifications of the erroneous function of PIDC as well as suggestions for improvements can be send to the e-mail address ari.rantanen@cs.helsinki.fi.

2. Installation

After downloading a zip-compressed installation package of PIDC (PIDC.zip) extract files to a new directory. The extraction can be done using popular compression utilites such as Winzip or pkunzip in Windows or unzip in Linux. After extraction the M-files needed by PIDC can be found from the created directory together with one example input file, *example_input.txt*.

After the extraction of the installation package start MATLAB. There exists a drop down list in the main window of the MATLAB titled “*current directory*”. Input there a

complete path to the directory where the M-files were extracted. Now PIDC is ready to run.

3. Running PIDC

PIDC is started by typing “*PIDC*” to a command window of MATLAB. After a short while the main window of PIDC shows up. From this window you can specify the name of the input and output files, start the computation of the positional isotopomers and close the application. Names of the input and output files are given by typing the full pathnames of the files in the corresponding edit fields or, alternatively, by clicking the “*Browse*”-button and the selecting the file from the file dialog.

Note that the file extensions “*_positional.txt*” and “*_mass.txt*” are automatically added to respective output filenames. Therefore, you should not type them in by hand.

When the input and output files are specified, press the button “*Calculate*” to compute a positional isotopomer distribution of the *labeled atom type* of the compound. The labeled atom type defaults to carbon, see below about specifying the atom type.

When the calculation is started “*Ready for Computation*” text in the main windows turns into text “*Computing, wait*”. When the “*Ready for Computation*” is visible again the computation is finished and the results can be read from the output files. Output files are standard text files that use tabulators as separators, so the files can be easily imported, for example, to a spreadsheet application.

4. Format of input files

The input file describing the mass spectra information is in standard text format where different values in the same row are separated with tabulator characters. Every row in the input file corresponds to a one peak in a one spectrum. Information about the peaks originating from the same spectrum have to be listed in consecutive rows of the input file. Also, information about the mass spectrum of the whole molecule has to listed before information of its fragments.

For every peak the following data has to be provided in the given order:

- the weight of the mass isotopomer (m/z) corresponding that peak (an integer),
- the intensity of the peak,
- atomic formula of the molecule or its fragment corresponding the peak (e.g., $C_3H_8NO_2$), and
- identification of the positions of the labeled atoms that belong to the molecule or fragment.

The positions of the labeled atoms are given as relative to the original molecule: if, for example, a fragment contains carbons numbered 2 and 3 of in original molecule, the positions of labeled atoms are given as “*C2C3*”. Character is ‘*C*’ is a separator character in the string specifying the atom positions regardless the atom type actually used as labels. Positions of the atoms can be defined using any characters or strings that do not

include characters ‘C’, ‘c’ or any special characters. It suffices that same definitions to the same atoms are used in the molecule and its fragments. For example, strings such as “C3’C3” or “C α C β C γ ” are legal identifier strings.

If the daughter ion scanning is used, one also has to specify for each daughter peak the mass of the parent ion that the daughter peak originated from. This is done by adding an integer to the corresponding row of the input file: 0 means that the daughter peak originated from a parent ion with the molecular weight (i.e. not containing any higher isotopes of atoms), 1 denotes that that the parent peak whose mass is (base mass + 1) was fragmented (one extra neutron in some atom), etc. Information about the fragment spectra originating from full and daughter ion scanning can be concatenated to the same input file: if the mass of the parent ion is not specified PIDC interprets corresponding fragment to be born as a result of full scanning.

If the GC-MS with derivatization of the molecules is used, labeled atoms resulting from the derivatization have to be identified separately from the labels in the rest of the molecule. This is done by denoting carbons resulting from derivatization with letter ‘D’ in the atom formulas in the input file. If other labels than carbon are used, M-file “*initialize.m*” has to be modified for this purpose (see below).

Here is an example of the input file that contains the information of the mass spectra of the alanine ion (C3H8NO2) and its fragments C2H6N, C2H3 and CH3N. Daughter ion scanning is executed twice, first to the +1 peak of the spectrum of the original molecule, then to the peak +2.

90	743899136	C3H8NO2	C1C2C3	
91	113999872	C3H8NO2	C1C2C3	
92	6172928	C3H8NO2	C1C2C3	
44	123844	C2H6N	C2C3	1
45	174496	C2H6N	C2C3	1
27	8794	C2H3	C2C3	1
28	13993	C2H3	C2C3	1
29	14401	CH3N	C2	1
44	50044	C2H6N	C2C3	2
45	36667	C2H6N	C2C3	2
46	21638	C2H6N	C2C3	2
47	13342	C2H6N	C2C3	2

The three first rows correspond to the parent ion spectra, i.e. the alanine molecule as a whole. The next four rows correspond to the fragment without the carboxyl group (C2H6N), originated from alanine with one extra neutron in some atom. The next row corresponds to a fragment consisting of the middle carbon of alanine with the amino group (CH3N), also originated with alanine having one extra neutron. The last four rows represent the spectra of the fragment without the carboxyl group, this time from alanine with two extra neutrons.

5. Format of output files

Output having the suffix “*_positional.txt*” contains the information about (possibly partial) positional isotopomer distribution computed by PIDC. At the beginning of the file the creation date of the file is shown.

The positional isotopomers in the distribution are represented as bit vector where *1* means that label was present in the corresponding location of the molecule and *0* that label was absent in that location. For example, vector $[1\ 0\ 0]$ means that there was a label in the first atom position of the molecule but not in the second and third positions. The mapping between positions of the bit vector and naming of atom positions of the molecule given in the input file is shown below distribution (*‘Order of labels’*). If the positional isotopomer distribution was not completely discovered, the distribution shown contains sums of two or more positional isotopomers. For example, if the distribution contains a component $[1\ 0\ 0] - [0\ 0\ 1]$, PIDC was able to compute only the relative amount of the sum (isotopomer $[1\ 0\ 0]$ – isotopomer $[0\ 0\ 1]$) in the compound, not the relative amounts of those isotopomers separately.

Output file with the suffix “*_mass.txt*” contains the mass isotopomer distributions of the labeled atom type of the molecule and its fragments. Compared to the input data the effect of the natural abundance of the heavier isotopes of the non-labeled atom types is removed from these distributions.

6. Editing the Initialization file

M-file *initialize.m* contains parameters that specify the exact operation of the PIDC. You have to edit these parameters by hand in order to change the default behavior of the software. The parameters are:

- *labelType*; Type of the atom used in labeling. Default type is *C*.
- *epsilon*; A minimum distance from zero that a value in the coefficient matrix must have in order to be counted as non-zero during inversion. Also positional isotopomers whose coefficients do not differ from zero more than $|epsilon|$ are discarded from the output.

In addition, *initialize.m* contains information about the natural abundance of the heavier isotopes of atoms contained in amino acids. You can add new atoms to this list, if needed. Note that atoms of the same type as labeled ones but resulting from derivatization have to be named differently than labeled ones in the atom formulas of the compound and its fragments as well as in the *initialize.m*. For example, in the initialization file the letter ‘*C*’ is reserved for labeled carbons and ‘*D*’ to the carbons resulting from derivatization. These letters have to be used in the atomic formulas of the input files to separate amounts of carbons resulting from different sources.

- [1] A. Rantanen, J. Rousu, J. T. Kokkonen, V. Tarkiainen, and R. Ketola: Computing Positional Isotopomer Distributions from Tandem Mass Spectrometric Data, To appear in *Metabolic Engineering*.

