Formularity – User Manual

Abbreviations

CIA – Compound Identification Algorithm

FT-ICR - Fourier Transform Ion-cyclotron Resonance

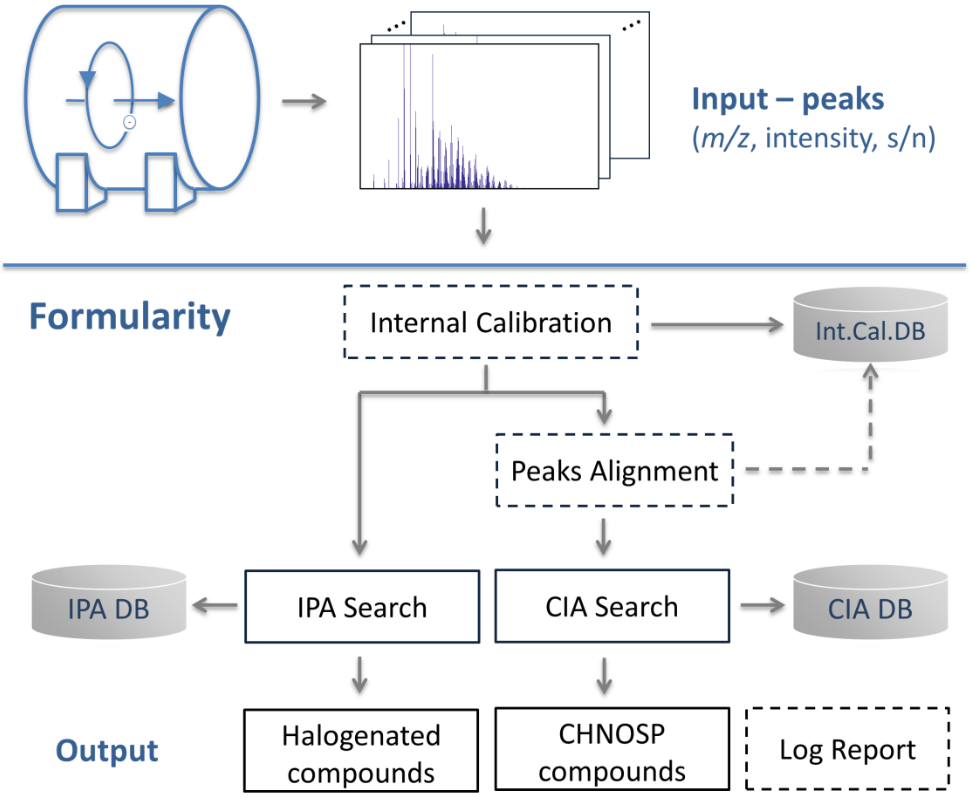
IPA – Isotopic Pattern Algorithm

ppm – parts per million

Introduction

Formularity software is tool for assignment of small-weight molecular formula from list of peaks compiled from high-resolution mass spectra. Although software could be used with any list of peaks it is targeted to FT-ICR mass spectrometers with assumed mass precision of 1 ppm or better. It consists of several independent functions used to accomplish various steps in data analysis; figure 1 shows software flowchart and dependencies. CIA search function is used for standard NOM search using universal database of C,H,N,O,S,P formula for the matching with mass spectra peaks. IPA search function is useful when looking for complex isotopic patterns for example halogenated organic matter or organometallics. Using common input and calibration functions Formularity bundles both searches into unified software interface for automated formula assignment for Hi-Res MS.

Figure 1.

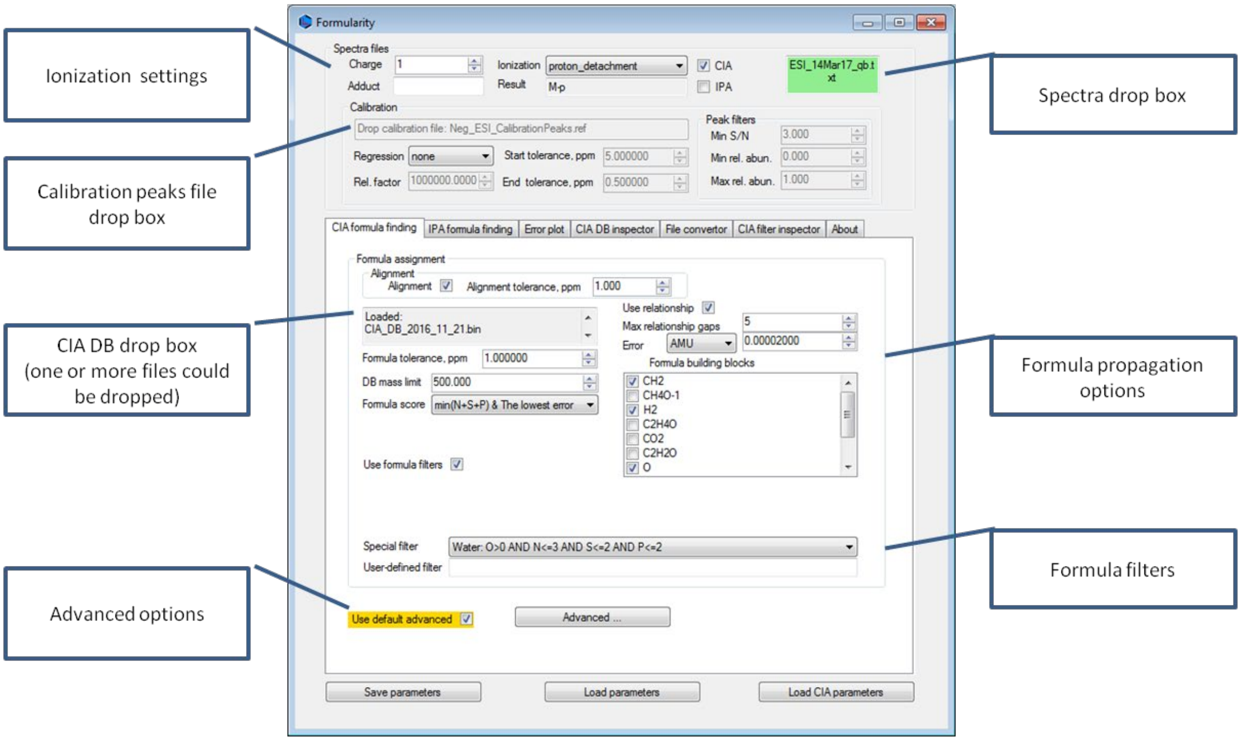


Formularity input is one or more text files with list of peaks from mass spectra. Formularity assigns peaks with molecular formula through CIA and/or IPA search functions. Each function queries its own database, for CIA database (CIA DB) contains only formula with elemental composition C, H, N, O, P, and S; IPA database (IPA DB) could contain any formula with stable isotopes elemental composition. Software features internal calibration function on predefined list of calibration peaks and alignment function allowing comparative analysis of any number of spectra.

GUI

Both search functions are activated from the same screen function once appropriate database files are loaded which is accomplished by “drag and drop” of database file to target box on “CIA formula finding” and “IPA formula finding” tabs. Once database is loaded CIA and IPA check boxes are enabled for search; dropping spectra files on designated box in “Spectra” panel will start CIA or IPA or both searches, depending on which box is checked. If requested, internal calibration is used with both search functions; list of calibration peaks is dropped on appropriate box in “Calibration” panel. Once calibration peaks file is loaded, boxes with other calibration parameters are enabled for editing. Formularity loads with parameter files saved after the last run but could be also stored and re-loaded from file. Figure 2. shows Formularity main screen with “CIA formula finding” tab active.

Figure 2.



Formularity

Spectra files

Charge 1

Adduct

Ionization proton_detachment

Result M-p

☑ CIA

□ IPA

ESI_14Mar17_qb.txt

Calibration

Drop calibration file: Neg_ESI_CalibrationPeaks.ref

Regression none

Rel. factor 1000000.0000

Start tolerance, ppm 5.000000

End tolerance, ppm 0.500000

Peak filters

Min S/N 3.000

Min rel. abun. 0.000

Max rel. abun. 1.000

Formula assignment

Alignment

Alignment ☑ Alignment tolerance, ppm 1.000

Loaded:
CIA_DB_2016_11_21.bin

Formula tolerance, ppm 1.000000

DB mass limit 500.000

Formula score min(N+S+P) & The lowest error

Use formula filters ☑

Use relationship ☑

Max relationship gaps 5

Error AMU 0.00002000

Formula building blocks

☑ CH2

□ CH4O-1

☑ H2

□ C2H4O

□ CO2

□ C2H2O

☑ O

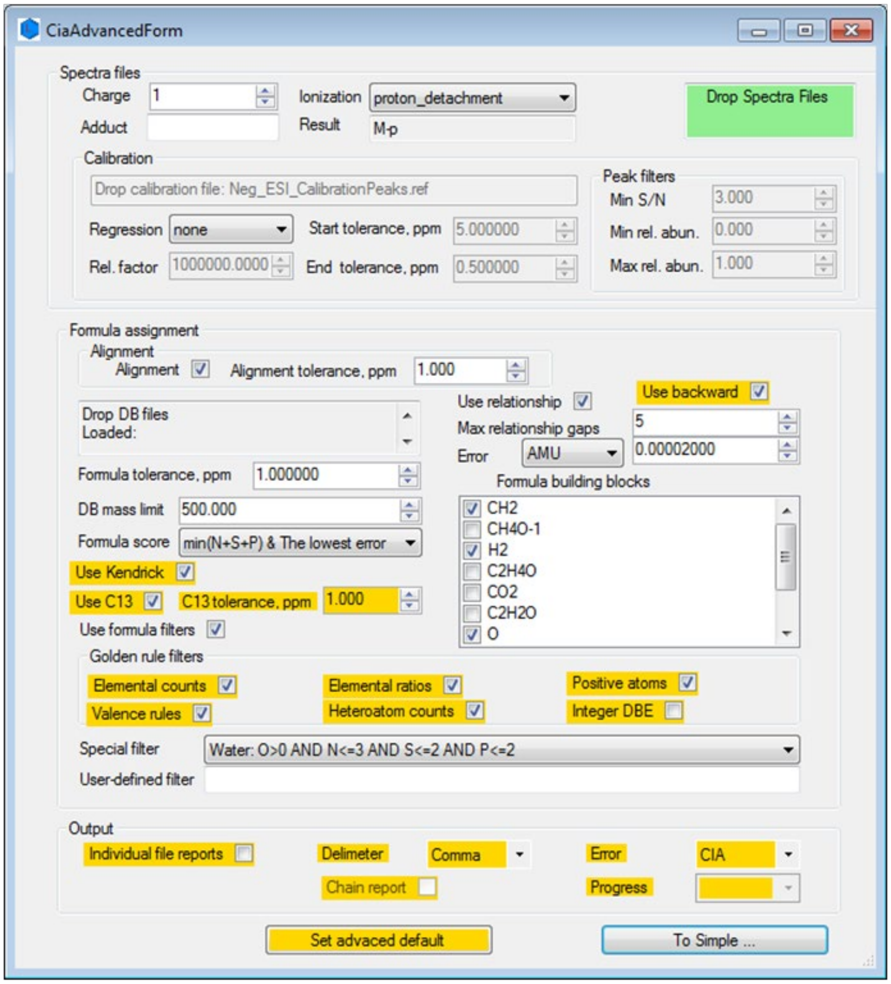
Special filter Water: O>0 AND N<=3 AND S<=2 AND P<=2

User-defined filter

Use default advanced ☑

To reduce number of options on the main search display we identified and isolated under **Advanced** tab options that are not changed very often. Advanced display has all parameters listed on "CIA formula finding" panel, advanced options are highlighted (Figure 3). Some of those are parameterized to allow for exploration of CIA behavior on changes in parameters, defaults reflects as close as possible default values in original CIA code. We recommend that with each analysis set of parameters used is saved using "Save parameters" command for future reference and results reproducibility.

Figure 3.



Spectra files

Charge 1

Ionization proton_detachment

Adduct

Result M-p

Calibration

Drop calibration file: Neg_ESI_CalibrationPeaks.ref

Regression none

Start tolerance, ppm 5.000000

Rel. factor 1000000.0000

End tolerance, ppm 0.500000

Peak filters

Min S/N 3.000

Min rel. abun. 0.000

Max rel. abun. 1.000

Formula assignment

Alignment

Alignment ☑ Alignment tolerance, ppm 1.000

Drop DB files
Loaded:

Use relationship ☑

Use backward ☑

Max relationship gaps 5

Error AMU 0.00002000

Formula building blocks

☑ CH2
□ CH4O-1
☑ H2
□ C2H4O
□ CO2
□ C2H2O
☑ O

Formula tolerance, ppm 1.000000

DB mass limit 500.000

Formula score min(N+S+P) & The lowest error

Use Kendrick ☑

Use C13 ☑ C13 tolerance, ppm 1.000

Use formula filters ☑

Golden rule filters

Elemental counts ☑ Elemental ratios ☑ Positive atoms ☑

Valence rules ☑ Heteroatom counts ☑ Integer DBE □

Special filter Water: O>0 AND N<=3 AND S<=2 AND P<=2

User-defined filter

Output

Individual file reports □

Delimeter Comma

Error CIA

Chain report □

Progress

Spectra and Calibration Panel

Spectra files panel contain set of parameters common for both search function. This includes setting appropriate assumption on type of spectra in Ionization, Adduct and Charge parameters. If different adducts have to be considered, searches have to be done separately; this software does not contain function for reconciliation of results from multiple searches. For similar reasons, spectra with different ionization method or assumed charge cannot be analyzed in the same batch! Result box displays actual calculation of ion m/z value from molecule neutral mass M for selected triplet of Ionization, Adduct, Charge settings.

Similar like peak picking, instrument manufacturer usually provides software for peaks calibration so calibration function implemented in Formularity is optional – if Regression parameter is set to “none” calibration is not used. In current implementation, the only implemented Regression model is “linear” or “quadratic” suitable to most of instrument types. To use calibration function a list of calibration peaks (formatted as Bruker Daltonics ".ref" file) is dropped on designated box and initial and target mass tolerance are specified, relative factor describes parts units; 1e6 is used for parts-per-million (ppm). Additional peaks filter (s/n and relative abundance) could be used for selection of peaks to be used in calibration. These filters have no effect on search functions. Calibration peak files list predicted charged mass (*m/z* values) so separate files for positive and negative ion modes and ionization methods have to be specified. It is up to user to provide appropriate calibration peaks file and check calibration tables as this is very sensitive step for application of both CIA and IPA functions. Ionization, Adduct and Charge settings have no influence on internal calibration since m/z values from peaks list is compared directly with values from calibration file.

Ionization – describes physics of forming ions

Adduct – can be any molecular formula, its neutral mass will be added to database formula mass during the search. Since database lists neutral (molecular) formula, peaks have to be "neutralized" prior to search. Specifying adduct serves as correction in formula calculation allowing single calculation for all spectra types.

Charge – ions measured in NOM MS measurements are usually singly charged; this setting allows search for multiply charged ions.

Result (read only) – describes how Ionization, Adduct and Charge affects calculation of ions from neutral mass M. This affects both CIA and IPA functions but has no affects on calibration peaks which are assumed to be ion m/z values. Following table lists some of commonly used ion types

Conventional notation					Formularity notation		
Ionization	Polarity	CS	Symbol	Comment	Ionization	Adduct	Charge
None				M	none		
ESI	neg	1	M-H	M-1.0072765	proton_detachment		1
ESI	neg	1	M-H	M-1.0072765	electron_attachment	-H	1
ESI	neg	1	M+Cl-	M+34.969401	electron_attachment	Cl	1
ESI	pos	1	M+H	M+1.0072765	proton_attachment		1
ESI	pos	1	M+Na	M+22.989221	electron_detachment	Na	1
ESI	pos	1	M+K	M+38.963158	electron_detachment	K	1
APPI	pos	1	M+H	M+1.0072765	electron_detachment	H	1
APPI	pos	1	M+*	M-0.0005486	electron_detachment		1
APPI	neg	1	M-H	M-1.0072765	proton_detachment		1
MALDI	pos	1	M+H	M+1.0072765	proton_attachment		1
ESI	pos	3	M+3H	M/3+1.007276	proton_attachment		3
ESI	pos	3	M+3Na	M/3+22.989218	electron_detachment	Na3	3
ESI	neg	1	M+acetate	M+59.013853	electron_attachment	C2H3O2	1
ESI Li ion	pos	1	M+Li	M+7.022399	proton_attachment	Li	1

CIA – check for CIA search (enabled only if CIA DB is loaded)

IPA – check for IPA search (enabled only if IPA DB is loaded)

Regression – calibration regression model; if “none” calibration is not used

Start tolerance, ppm – initial tolerance for matching calibration peaks with peaks of spectrum.

Rel. factor – relative factor to base calibration equation on (use 1e6 for ppm)

End tolerance, ppm – target tolerance; calibration succeeds if target tolerance is achieved.

Min S/N – signal to noise filter for calibration; only peaks with measured s/n value better than specified will be used when matching calibration peaks

Min rel. abun. and **Max rel. abun.** allow peak filter based on peak relative abundance during calibration procedure (range 0-1)

CIA Search

To perform CIA search; drag CIA DB to designated box; once database is loaded CIA check-box in Spectra panel will be enabled; check it and set other parameters on “CIA formula finding” panel. CIA formula finding can be performed with or without alignment step and with or without individual files report. To allow custom databases and also to keep main database identical to the original, Formularity software allows loading of multiple NOM databases which are merged in memory and searched simultaneously. Following is the list of relevant settings for “CIA formula finding”:

Alignment - if checked alignment of peaks between different spectra will be performed and presence absence matrix with peak intensity from each spectrum in corresponding column. Peaks within **Alignment Tolerance (ppm)** from different datasets will be declared the same and average m/z will be used as value of aligned peak; this value will also be used in database search. If internal calibration is used, alignment function acts on calibrated peaks from individual spectra. At the moment failure of calibration of any spectra will cause failure of the whole CIA search.

Formula tolerance, ppm – search tolerance for formula candidates

DB mass limit - for peaks up to specified DB mass limit formula candidates are checked directly in the database

Formula score – criteria how to resolve ambiguity when more than one formula matches peak. Even for well calibrated spectra, the lowest mass error does not guarantee correct formula assignment. Selecting simpler (lower count of heteroatoms) appears to provide better results for most of environmental samples.

Use Kendrick (Advanced) – if selected CH2 formula propagation at the end of assignment overwrites assignment with significantly worse mass error; this was not optional in original CIA code.

Use C13 and **C13 tolerance, ppm (Advanced)** – in original CIA code C13 (13C) peak was searched by default and mass tolerance was the same as **Formula tolerance, ppm.**

Use formula filters and **Golden rule filters (Advanced)** allow filtering of formula from CIA DB to be considered for assignment. Golden rules are used according to Kind & Fiehn[3] and are already enforced during compilation of CIA DB with exception of Integer DBE. Since it is possible to use CIA search with different databases these rules should be used carefully to avoid assignment of chemically impossible molecular formula. None of these filters affect IPA search or database.

Few named **Special filter**s are provided as an examples how to write **User-defined filter**s allowing user to, based on experiment prior knowledge or any other reason, limits search space based on elemental composition. For example, heuristics for CIA search of dissolved carbohydrates sample would be to write “O>0 AND O=H AND H=2C AND N+S+P=0” as user-defined filter since all inputs in CIA DB have C>0.Integer DBE filter was added to “golden rules” of formula selection; originally compiled CIA DB has not used DBE related filters; rather than adding lengthy user defined filter we provide additional check box to allow this filter and still keep original database intact.

Use relationships – database formula matched with peaks are propagated using selected **Formula building blocks**, max defined **Error** and **Max. relationship gaps**. Original CIA code performed **Use backward (Advanced)** propagation potentially re-assigning some peaks assigned directly from the database. We have found cases where this procedure caused questionable results!

Other settings in **Advanced** tab concern mostly output format of CIA search function.

CIA search algorithm and database (CIA DB) are described in original manuscript and MatLab code [1, 2] so we list here only implemented changes with note that CIA DB provided with the code is recompiled in November 2016.

For peaks with *m/z* > “DB mass limit” original MatLab code implemented as last level of ambiguous formula resolution usage of “latest found” acceptable formula; we changed this to acceptable formula with lowest mass error.

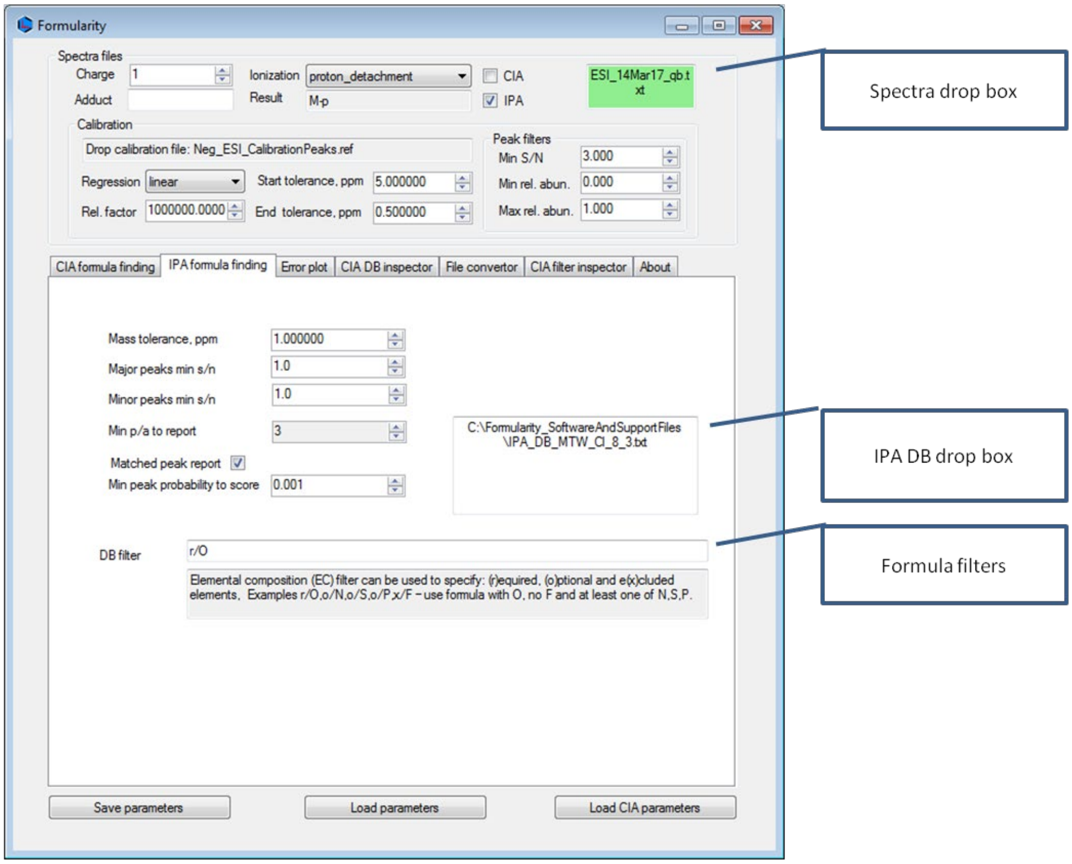
Allowing reporting of search results in individual files and signed mass measurement error allows posterior analysis of mass error and refined search based on statistical interpretation of error distribution. Original CIA reports error absolute value.

CIA DB is recompiled for better coverage with replacing strong inequalities in implementation of some golden rules (Senior rule).

IPA Search

IPA search is performed whenever IPA check box is selected; IPA database is targeted database of predicted isotopic peaks with structure. Example IPA database provided with the software could be used for search of chlorinated organic matter in drinking water. To compile IPA database we use external tools like Mercury, Deuterium or ecipex and custom made scripts for calculation end extraction of predicted isotopic peaks. To make easier for user to compile custom IPA database in current version of software IPA DB is text file with structure described below!

Figure 4.



Spectra files

Charge 1

Adduct

Ionization proton_detachment

Result M-p

☐ CIA

☑ IPA

ESI_14Mar17_qb.txt

Calibration

Drop calibration file: Neg_ESI_CalibrationPeaks.ref

Regression linear

Rel. factor 1000000.0000

Start tolerance, ppm 5.000000

End tolerance, ppm 0.500000

Peak filters

Min S/N 3.000

Min rel. abun. 0.000

Max rel. abun. 1.000

Mass tolerance, ppm 1.000000

Major peaks min s/n 1.0

Minor peaks min s/n 1.0

Min p/a to report 3

Matched peak report ☑

Min peak probability to score 0.001

C:\Formularity_SoftwareAndSupportFiles\IPA_DB_MTW_CI_8_3.txt

DB filter r/O

Elemental composition (EC) filter can be used to specify: (r)equired, (o)ptional and e(x)cluded elements. Examples r/O,o/N,o/S,o/P,x/F – use formula with O, no F and at least one of N,S,P.

Details of IPA algorithm and scoring are described in manuscript in preparation [4]; here we will just mention enough to explain IPA function parameters. To facilitate search of isotopic peaks and fine structure peaks we introduce concept of major and minor peaks; simply major peak is the most probable peak at each mass unit, all other isotopic peaks are minor. Combinatorial nature of isotopic peaks causes explosion with high number of peaks unlikely to be observed in mass spectrum. For practical and performance reasons IPA DB is compiled with same number of major and minor peaks for all formula, for example DB marked as 8x3 is made with 8 (most abundant) major peaks and 3 (most abundant) minor peaks around each major. Zeroes are filled for molecular formula with insufficient number of isotopic peaks.

Mass tolerance, ppm – search tolerance for matching of experimental and formula predicted isotopic peaks.

Major peaks min. s/n – signal to noise filter (if present in peaks list) for assignment of major peaks.

Minor peaks min. s/n – signal to noise filter (if present in peaks list) for assignment of minor peaks.

Minimum p/a to report – p/a score for molecular formula in IPA DB is the sum of 1 for each major and 0.1 for each minor peak matched with predicted isotopic peaks. This is output filter for “s_” reports reducing the output of formula with marginal matching.

Matched peak report – if checked “p_” reports will be generated listing tabulation of all matched peaks for all database formula.

Min peak probability to score – allows filtering out less likely predicted isotopic peaks without need for new database. IPA DB compilation requires time and non-trivial effort. Isotopic peaks are filtered out based on their pseudo-probability within the isotopic distributions. It makes sense to compile database with more peaks and then using this parameter refine searches eliminating peaks less likely to be observed in the spectrum. For example setting this value to 0.01 would ensure that peaks for formula in the IPA DB listed as being less likely than 1% of the most likely peak for that formula will not be used in matching (and therefore not penalized for absence). Using full database would still make great sense when peak count maximization for best formula is important parameter of evidence quality.

DB Filter – IPA DB can contain formula of any elemental composition so filter used for CIA function was not practical to implement here. Simpler version of filter is implemented for IPA DB with example of usage presented on GUI.

Input File Format

Both search function expect input file in the same format – delimited text table with first two columns assumed to be *m/z* and abundance ordered in ascending order on *m/z*. If more than 2 columns exist in the input file, third column is assumed to be “signal to noise” (s/n). If “s/n” column is present calibration function and IPDB search potentially use this column as threshold filter. MS Excel and Bruker XML formats are supported as well.

Output File Format

Output of CIA search function is table with optionally calibrated and aligned list of peaks and assigned molecular formula formatted as tabulated report for easy import to specialized post-processing software. Calibration and alignment function at the moment do not produce output without being performed with search functions. IPA search function produces pair of files “p_*” and “s_*”, listing all peaks “p_*” matched with assigned and scored IPA DB records using “mf_ind” column from database as a key in peaks table.

Columns of “s_” file:

mf_ind	- index of molecular formula in IPDB
mf	- molecular formula
ec	- elemental composition
C	- resolved C count
H	- resolved H count
O	- resolved O count
NSP	- resolved (N+S+P) count
Other	- other elements count
search_m0	- zero charge (neutral) mass of search peak (the most probable peak)
search_m	- charged mass of search peak

pa_mm_abs - isotopic peaks presence/absence score weighted with 1 for major and 0.1 for minor peak

pa_mm_max - maximum value for pa_mm score for molecular formula record

pa_mm_rel - relative pa_mm score; pa_mm_rel= pa_mm_abs/pa_mm_max

major_count - count of major peaks of formula isotopic pattern matched with observed peaks

minor_count - count of minor peaks of formula isotopic peaks (simulated isotopic distributions) matched with observed peaks

major_multi_count - reserved for future use

minor_multi_count - reserved for future use

p_d1 - d1 (taxicab) distance between IPDB formula record and normalized record of observed peak intensity matched to formula peaks

p_d2 - d2 (Euclidean) distance between IPDB formula record and normalized record of observed peak intensity matched to formula peaks

p_dinf - d2 (infinity) distance between IPDB formula record and normalized record of observed peak intensity matched to formula peaks

pa_sum_abs - isotopic peaks presence/absence score weighted with isotopic peaks pseudo-probabilities

pa_sum_max - maximum value of pa_sum for molecular formula

pa_sum_rel - relative pa_sum score; pa_sum_rel= pa_sum_abs/pa_sum_max

tmp_m_err_ppm - relative mass error of the most probable isotopic peak

Columns of “p_” file:

ind - peak index; from input file

mz - “m/z” value from input or calibrated input therefore observed “m/z”

intensity - peak intensity (from input file)

sn - peak s/n value (from input file)

rel_int - peak relative intensity (compared to base peak intensity)

mf_ind - index of molecular formula in IPDB with matching isotopic peak

mf - molecular formula from IPDB with matching isotopic peak

m0 - zero charge (neutral) mass of matching peak

search_m - charged mass of matching peak (expected “m/z”)

p - probability of observation of matched peak if mf was observed

major_ind - major index of matching peak in IPDB formula record; if matching peak is minor peak this is index of major peak minor peak belongs to

minor_ind - minor index of matching peak in IPDB formula record; -1 if major peak

err_ppm - relative mass measurement error calculated as (“mz”-“search_m”)/”search_m”*1e6

Calibration Files

Provided test calibration file illustrates structure of file; first line is comment, calibration peaks are listed as name, m/z, and optionally charge tabulation but only m/z value is used by Formularity. Provided file should successfully calibrate majority of negative mode ESI natural organic matter samples.

Log File

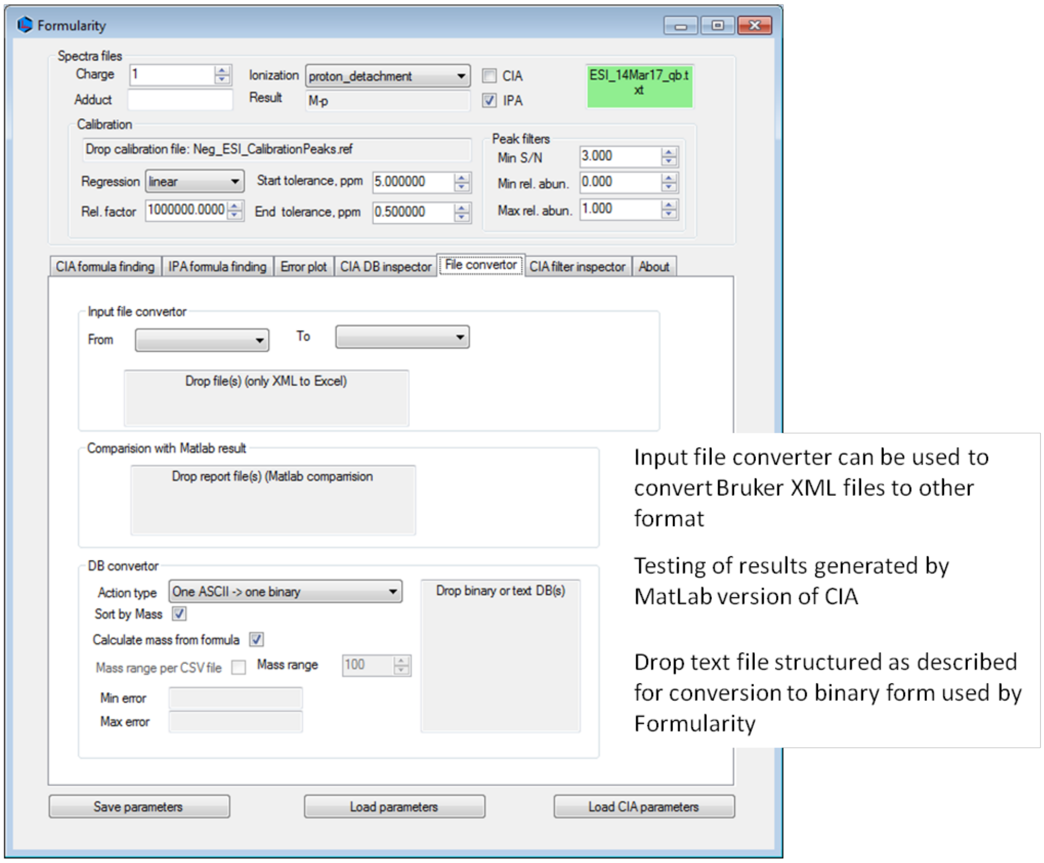
Log file is time-stamped text file used to store important messages during the processing including calibration tables if calibration is performed. Additional information could be added to log file content in future releases.

Database Structure

CIA NOM database

CIA database is binary database containing monoisotopic mass and C, H, O, N, S, P elemental counts defining molecule. To avoid any confusion CIA database should always contain neutral (molecular) formula and monoisotopic mass. Formularity is equipped with converter from tabulated to binary file allowing compilation of custom databases. Converter (Figure 5) takes as an input tabulated file without header with column in following order [ID:int][mass:double][C:int][H:int][O:int][N:int][13C:int][S:int][P:int][Na:int]. Column [13C] and [Na] could be set (with adjusted mass) but are usually 0 and are not considered in any CIA algorithm. Those two columns could be used to trick Formularity into processing formula with additional elements!

Figure 5.



Formularity

Spectra files

Charge: 1

Adduct:

Ionization: proton_detachment

Result: M-p

☐ CIA

☑ IPA

ESI_14Mar17_qb.txt

Calibration

Drop calibration file: Neg_ESI_CalibrationPeaks.ref

Regression: linear

Rel. factor: 1000000.0000

Start tolerance, ppm: 5.000000

End tolerance, ppm: 0.500000

Peak filters

Min S/N: 3.000

Min rel. abun.: 0.000

Max rel. abun.: 1.000

Input file convertor

From:

To:

Drop file(s) (only XML to Excel)

Comparision with Matlab result

Drop report file(s) (Matlab comparision)

DB convertor

Action type: One ASCII -> one binary

Sort by Mass ☑

Calculate mass from formula ☑

Mass range per CSV file ☐

Mass range: 100

Min error:

Max error:

Drop binary or text DB(s)

Input file converter can be used to convert Bruker XML files to other format

Testing of results generated by MatLab version of CIA

Drop text file structured as described for conversion to binary form used by Formularity

CIA DB does not have to be sorted because Formularity sorts all loaded databases together after loading.

IPA DB database

In the current Formularity release IPA DB is in tabulated text format to make it easier for user to decipher and compile custom database. To improve performance next releases might have it in a binary format.

First line contains number of records, number of Major(r) and minor(s) peaks; the rest of the line could have information on the database source and is ignored by the software. Each line is fixed length record of predicted (simulated) peaks for molecular formula. Record fields (columns) are [ID:int],[MF:text],[name:text],[M_im:double][M_ip:double]i=1,…,r,],[$m_{i,j}$m:double][$m_{i,j}$p:double]i=1,…,r; j=1,…,s. Database can be compiled using software like ecipex or Deuterium; please contact us if you need help compiling custom IPA database.

Future Enhancements

Spectra diagnostics based on autocorrelation spectra.

Custom compilation of calibration peaks.

Iterative formula assignments with simultaneous application of CIA and IPA functions based on feedback loop from mass error distribution.

Disclaimer

This material was prepared as an account of work sponsored by an agency of the United States Government.Neither the United States Government nor the United States Department of Energy, nor the Contractor, nor any or their employees, nor any jurisdiction or organization that has cooperated in the development of these materials, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness or any information, apparatus, product, software, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY operated by BATTELLE for the UNITED STATES DEPARTMENT OF ENERGY under Contract DE-AC05-76RL01830

Contact and Availability

Program written by Andrey Liyu and Nikola Tolic for Department of Energy (PNNL, Richland, WA) 2015-2016

E-mail: nikola.tolic@pnnl.gov or andrey.liyu@pnnl.gov

Website: https://store.pnnl.gov

Acknowledgments

Original CIA code was developed as MatLab function in Woods Hole Oceanographic Institute by Elizabeth Kujawinski and Krista Longnecker. We would like to thank Elizabeth and Krista for providing us with original MatLab code and helpful advices on code refactoring. The research was performed using EMSL, a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

References

1) Kujawinski, E. B., & Behn, M. D. (2006). Automated analysis of electrospray ionization Fourier transform ion cyclotron resonance mass spectra of natural organic matter. *Analytical chemistry*, *78*(13), 4363-4373.
2) Kujawinski, E. B., Longnecker, K., Blough, N. V., Del Vecchio, R., Finlay, L., Kitner, J. B., & Giovannoni, S. J. (2009). Identification of possible source markers in marine dissolved organic matter using ultrahigh resolution mass spectrometry. *Geochimica et Cosmochimica Acta*, *73*(15), 4384-4399.
3) Kind, T., & Fiehn, O. (2007). Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC bioinformatics, 8(1), 1.
4) Formularity: Software for Automated Formula Assignment of Natural and Derived Organic Matter from Ultra-High Resolution Mass Spectra Nikola Tolić, Yina Liu, Andrey Liyu, Yufeng Shen, Malak M. Tfaily, Elizabeth B. Kujawinski, Krista Longnecker, Li-Jung Kuo, Errol W. Robinson, Nancy J. Hess(in preparation)