metabomatching

Rico Rueedi

February 15th, 2017

Updates on:

www.unil.ch/cbg/index.php?title=metabomatching

Please address any questions, comments, or bugs to

metabomatching@unil.ch

For bugs, please mention the version of metabomatching that produced the error. If possible, include the relevant `ps.<description>` subdirectories.

Contents

1 Quick Start **2**
1.1 Docker . . . 2
1.2 Matlab/octave . . . 2

2 Complete Instructions **3**
2.1 Pseudospectra and Scores . . . 4
2.1.1 Pseudospectrum file format, one per file. . . . 4
2.1.2 Pseudospectrum file format, multiple per file. . . . 5
2.2 Settings . . . 5
2.3 Reference Spectral Databases . . . 6
2.3.1 HMDB . . . 7
2.3.2 BMRB . . . 8
2.3.3 BMRB . . . 9
2.3.4 Chemical Abstracts Service Registry Number (CASRN) . . . 9
2.3.5 Providing a Custom Spectral Reference Database . . . 9

3 Metabomatching Figures **10**

1 Quick Start

1.1 Docker

The simplest method to run metabomatching is to use the metabomatching docker image. With a working docker installation, run

```
docker pull metabomatching/metabomatching-pre
```

to download the metabomatching container. The download may take some time, as the container includes a full Linux operating system, octave to run metabomatching, and inkscape to convert metabomatching figures from SVG to PDF.

Minimal Run Run

```
docker run -i -v <directory>:/mm-ps metabomatching/metabomatching-pre
```

with `<directory>` the absolute path to the local directory to which you want metabomatching to write its results. This runs metabomatching with default settings on a default pseudospectrum.

1.2 Matlab/octave

The metabomatching Matlab code, which is compatible with octave, can be obtained from GitHub. The download includes: code files to run metabomatching and produce SVG images of the results, spectral reference databases built from HMDB and BMRB, a CASRN to metabolite name conversion list, and a test pseudospectrum.

Minimal Run In Matlab/octave, set as working directory the directory into which metabomatching was downloaded, and run the file

```
metabomatching.m
```

This will apply metabomatching with default settings to the test pseudospectrum located in the subdirectory `ps.test`. The metabomatching scores for each metabolite are stored in the 4th column of the tab-separated file `ps.text/test.scores.tsv` and presented graphically in the file `ps.test/test.svg`.

2 Complete Instructions

Metabomatching operates on the contents of the directory in which it is installed. Specifically, for each subdirectory of the installation directory named

```
ps.<description>
```

metabomatching will be run over all pseudospectrum files named

```
<tag>.pseudospectrum.tsv
```

within. Metabomatching settings are fixed for each subdirectory, and all pseudospectra within a subdirectory must be defined on the same chemical shift binning. Details on the content and format of pseudospectra files are given in section 2.1.

Example. Suppose we would like to apply metabomatching to the pseudospectra of two SNPs, rs1 and rs2, obtained in one association study (study1), and to the pseudospectra of two SNPs, rs1 and rs3, obtained in another study (study2). Using the SNP names as tags for the pseudospectra and the study names as description for the subdirectories, the folder structure in the installation should then be

```
<installation directory>/
|-- data/
    |-- ...
|-- documentation/
    |-- ...
|-- fos/
    |-- ...
|-- func/
    |-- ...
|-- ps.study1/
|   |-- rs1.pseudospectrum.tsv
|   |-- rs2.pseudospectrum.tsv
|-- ps.study2/
|   |-- rs1.pseudospectrum.tsv
|   |-- rs3.pseudospectrum.tsv
|-- metabomatching.m
```

Metabomatching Results. For each pseudospectrum file

```
<tag>.pseudospectrum.tsv
```

in each subdirectory, metabomatching outputs one scores file

```
<tag>.scores.tsv
```

and one SVG file

```
<tag>.svg
```

showing the main metabomatching results. If running the metabomatching docker the SVG file is also converted to PDF, resulting in the additional PDF file

`<tag>.pdf`

Details on the content and format of scores files are given in section 2.1. Details for the SVG/PDF files are given in section 3. For each subdirectory, metabomatching also writes files related to the settings and reference data used. These are described in section 2.2.

Spectral Reference Data. Metabomatching compares pseudospectra to NMR spectra listed in a reference database. The `data` subdirectory in the installation directory contains nine databases. Six contain reference spectra from HMDB, three from BMRB. Details as to their content and format are given in section 2.3.

2.1 Pseudospectra and Scores

For an untargeted metabolome composed of N features, the pseudospectrum of a certain variable is the collection of N association p-values, N effect sizes, and N standard errors resulting from the associations of the variable with the N metabolome features.

2.1.1 Pseudospectrum file format, one per file.

A pseudospectrum is provided to metabomatching as a tab-separated file named

`<tag>.pseudospectrum.tsv`

The file contains four columns. The first row must be a header line containing the four labels `shift`, `beta` (effect size), `se` (standard error), and `p`. The subsequent N rows contain the association data for the N metabolome features, indexed by their chemical shift. Columns and data rows may appear in any order. The tag may be composed of letters, numbers, hyphens, and underscores.

Example. Pseudospectrum file with the header row and three data rows.

```
shift   beta    se      p
1.11    0.03041 0.08991 0.2647
1.12    -0.5301 0.8771  6.1451E-11
1.13    -0.7171 0.1871  0.003201
```

Scores file format. For every pseudospectrum, metabomatching computes M scores, with M the number of metabolite reference spectra provided, and writes the corresponding scores file, named

`<tag>.scores.tsv`

The file contains three columns, listing the metabolite CASRN, a spectrum identification number (see 2.3), and the metabomatching score.

Example. Scores file for three metabolites.

```
70-47-3 10003021     3.8131
56-84-8 10003111     0.0002
56-88-2 10003311     0.0002
```

2.1.2 Pseudospectrum file format, multiple per file.

Multiple pseudospectra can be provided to metabomatching in a single tab-separated file, named

```
<study_tag>.pseudospectrum.tsv
```

For K pseudospectra, the file contains $1 + 3K$ colums. The first row must be a header line containing the label `shift`, and the set of labels `beta/<tag>` (effect size), `se/<tag>` (standard error), and `p/<tag>` for each of the K pseudospectra.

Example. Pseudospectrum file with the header row and three data rows, for two pseudospectra, labeled `rs18` and `ex27`

```
shift   beta/rs18      se/rs189 p/rs189 beta/ex72  se/ex27    p/ex27
1.11    0.109   0.0871  0.207   0.556   0.0964  1.15E-08
1.12    -2.66E-06       0.0872  0.999   -0.0154 0.0979  0.874
1.13    -0.0634 0.0871  0.466   0.491   0.0964  4.37E-07
```

Scores file format. For every pseudospectrum, metabomatching computes M scores, with M the number of metabolite reference spectra provided, and writes the corresponding scores file, named

```
<study_tag>.scores.tsv
```

The file contains $2 + K$ columns, listing the metabolite CASRN, a spectrum identification number (see 2.3), and the metabomatching score for each of the K pseudospectra.

Example. Scores file for three metabolites and two pseudospectra labeled `rs189` and `ex27`

```
cas     id      score/rs189     score/ex27
70-47-3 10003021        3.81    0.715
56-84-8 10003111        0.002   1.21
56-88-2 10003311        0.002   267
```

2.2 Settings

The settings used by metabomatching in each subdirectory can be specified with the optional file

```
parameters.in.tsv
```

There are seven settings that control the metabomatching algorithm, and two common graphical settings described below. Additional graphical settings are described in section 3. For any non-specified setting, metabomatching uses a default value. On completion, metabomatching writes all settings it used to the file

```
parameters.out.tsv
```

key	description	values	default
variant	Select among the 1-compound, 2-compound, ±-, and ±-2-compound-variants of metabomatching. Note that 2-compound and ±-2-compound metabomatching are computationally demanding, and may exceed the capabilities of standard personal computers.	1c 2c pm pm2c	1c
mode	Controls the spectrum description type used by metabomatching, and determines, in conjunction with setting reference, the spectral reference file metabomatching tries to load.	peak multiplet	peak
dsh	Controls the neighborhood parameter (δ if in peak mode, γ if in multiplet mode).	$(0, 1]$	0.025 (δ) 0.01 (γ)
scoring	Selects the type of scoring used in the match sum	chisq z	chisq
decorr_lambda	Sets the shrinkage parameter λ used when applying decorrelation. $\lambda = 1$ implies no decorrelation.	$[0, 1]$	1
p_pm	Sets the p-value threshold below which pseudospectrum features are considered to carry signal (rather than noise) for the purpose of ±-metabomatching (variant = pm or pm2c). When running metabomatching for one effect direction, the effect size for features of association p-value below p_suggestive and of opposite effect direction is set to 0.	$[0, 0.05]$	1E-4
reference	Selects which of the four standard reference sets metabomatching uses. Metabomatching copies the corresponding reference file to each subdirectory. Metabomatching uses the specified reference set in combination with any slo(m|p) files already present in the subdirectory. If reference is set to existing, metabomatching uses only reference file(s) already present in the subdirectory. This applies to both slo(m|p) files and casname files.	hmdb bmrb umdb umrb smdb smrb	hmdb
n_show	Sets the number of candidate metabolites shown in the results SVG file. Due to the graphical layout, only multiples of 4 are allowed as values for n_show.	4, 8, 12, ...	8
p_significant	Sets the p-value threshold below which pseudospectrum features are considered significant. This has no effect on metabomatching, and serves only to label features in the SVG file. Significant features are clustered (distance below 2δ if mode = peak or $2\gamma + 0.3$ if mode = multiplet). High values for p_significant may tag many features significant and break the layout of the SVG file.	$(0, 0.05]$	5E-8

independently of whether the settings were provided or used by default. For any other value in $[0, 1)$ the feature-feature correlation matrix must be available in the subdirectory, as a comma-separated file named

```
correlation.csv
```

The file has no header or row descriptors, and should contain N columns and N rows, with N the number of non header rows in the pseudospectra.

2.3 Reference Spectral Databases

The spectral reference databases included with the metabomatching software were obtained from HMDB (Human Metabolome Database, DS Wishart et al., *Nucleic Acids Res.*, 41:D801-7) and BMRB (Biomagnetic Resonance Database, EL Ulrich et al., *Nucleic Acids Res.*, 36:D402-8). Resulting from different experiments, the spectra for

individual metabolites are usually different in the two databases, sometimes significantly so.

2.3.1 HMDB

In HMDB, the spectrum for a metabolite is characterized twice. Once as a list of multiplets, and once as a list of peaks. For most spectra, the two descriptions are different enough to result in differences in metabomatching rankings. The differences can be significant. HMDB contains a proton NMR spectrum for 835 metabolites. Every metabolite in the database is identified by a metabolite ID of the form `HMDB00001`.

Peak Descriptions File. The full set of peak description metabolite spectra is listed in the reference file

```
hmdb.slop
```

This file is a four column tab-separated file, with each row containing the CASRN (as a string), the peak position in ppm (real number), the peak height (real number), and the spectrum ID (integer), for one peak of the peak description spectrum of one metabolite. The spectrum ID is defined as `1<xxxxxx>1`, with `<xxxxxx>` the HMDB ID stripped of the characters `HMDB` but keeping leading zeroes. Peak heights are used in the SVG image, but not by metabomatching.

Example. Spectra of dimethylglycine (CASRN 1118-68-9, HMDB00092) and glycine (CASRN 56-40-6, HMDB000123)

```
1118-68-9   2.91    1.000   1000921
1118-68-9   3.71    0.277   1000921
56-40-6     3.54    1.000   1001231
```

Multiplet Descriptions File. The full set of multiplet description metabolite spectra is listed in the reference file

```
hmdb.slom
```

This file is a four column tab-separated file, with each row containing the CASRN (as a string), the left edge of the multiplet range (real number), the right edge of the multiplet range (real number), the proton count (integer), and the spectrum ID (integer), for one multiplet range of the multiplet description spectrum of one metabolite. The spectrum ID is defined as `1<xxxxxx>1`, with `<xxxxxx>` the HMDB ID stripped of the characters `HMDB` but keeping leading zeroes. Proton counts are used in the SVG image, but not by metabomatching.

Example. Spectra of dimethylglycine (CASRN 1118-68-9, HMDB00092) and glycine (CASRN 56-40-6, HMDB000123)

```
1118-68-9   2.90    2.93    6   1000921
1118-68-9   3.70    3.72    2   1000921
56-40-6     3.46    3.60    2   1001231
```

Subsets. Two additional reference files are available containing the spectra of metabolites known to be contained in urine samples. This subset was defined by Bouatra et al. (*PLOS One*, 8(9):e73076), and named UMDB (Urine Metabolome Database). It lists the spectra for 180 metabolites. The files are named

```
umdb.slop
umdb.slom
```

for the peak and multiplet description spectra, respectively. Similarly, two additional reference files are available containing the spectra of metabolites known to be contained in serum samples. This subset was defined by Gowda et al. (*Anal. Chem.*, 86(11):5433-5440). It lists the spectra for 180 metabolites. The files are named

```
smdb.slop
smdb.slom
```

2.3.2 BMRB

In BMRB, the spectrum for a metabolite is characterized as a list of peaks. Each metabolite can have several spectra, obtained from different experiments. For most spectra, the descriptions are different enough to affect, slightly, metabomatching rankings. BMRB contains 975 proton NMR spectra for 740 metabolites. Every metabolite in the database is identified by a metabolite ID of the form BMRB00001.

Peak Descriptions File. The full set of peak description metabolite spectra is listed in the reference file

```
bmrb.slop
```

This file is a four column tab-separated file, with each row containing the CASRN (as a string), the peak position in ppm (real number), the peak height (real number), and the spectrum ID (integer) for one peak of the peak description spectrum of one metabolite. The spectrum ID is defined as `2<xxxxxx><y>`, with `<xxxxxx>` the BMRB ID stripped of the characters `BMRB` but keeping leading zeroes, and `<y>` indexes different spectra for the same BMRB metabolite. Peak height is used in the SVG image, but not by metabomatching.

Example. Spectra of dimethylglycine (CASRN 1118-68-9, BMRB02411) and glycine (CASRN 56-40-6). Glycine appears in BMRB with two different BMRB IDs (BMRB00089 and BMRB00977), and has four listed spectra (3 for BMRB00089, and 1 for BMRB00977).

```
1118-68-9   2.916   1.000   2002411
1118-68-9   3.714   0.336   2002411
56-40-6     3.541   1.000   2000891
56-40-6     3.542   1.000   2000892
56-40-6     3.544   1.000   2000893
56-40-6     3.545   1.000   2009771
```

Subsets. An additional reference file is available containing the spectra of metabolites known to be contained in urine samples. It lists the spectra for 180 metabolites. The file is named

```
umrb.slop
umrb.slom
```

for the peak and multiplet description spectra, respectively. Similarly, an additional reference file is available containing the spectra of metabolites known to be contained in serum samples. It lists the spectra for 180 metabolites. The file is named

```
smrb.slop
```

2.3.3 BMRB

In BMRB, the spectrum for a metabolite is characterized as a list of peaks. Each metabolite can have several spectra, obtained from different experiments. For most spectra, the descriptions are different enough to affect, slightly, metabomatching rankings. BMRB contains 975 proton NMR spectra for 740 metabolites. Every metabolite in the database is identified by a metabolite ID of the form BMRB00001.

2.3.4 Chemical Abstracts Service Registry Number (CASRN)

Metabomatching uses CASRN (Chemical Abstracts Service Registry Number) to identify metabolites. Metabolite names are used only in the SVG image. The file

```
all.casname
```

provides the CASRN to metabolite name table, as a two column tab-separated file.

Example. Entries for dimethylglycine and glycine

```
1118-68-9   dimethylglycine
56-40-6     glycine
```

2.3.5 Providing a Custom Spectral Reference Database

Additional spectra can be provided to metabomatching as a custom spectral reference database. The database itself needs a name and a number (3 and above, as 1 and 2 are already assigned to HMDB and BMRB respectively). Each spectrum in the database needs a CASRN identifying the metabolite, and a 6 digit ID number indexing the metabolite in the database as well as a 1 digit number indexing different spectra for the same metabolite.

Example. To provide two new spectra for glycine and two new spectra for the new metabolite *customine* in a custom database called MYDB, we would create the database file

```
mydb.slop
```

of content

```
56-40-6     3.541   1.000   4000011
56-40-6     3.542   1.000   4000012
12345-67-8  1.234   1.000   4000021
12345-67-8  1.236   1.000   4000022
```

Because customine is not known in `all.casname`, we would also create the CASRN-name table

```
mydb.casname
```

of content

```
12345-67-8  customine
```

Both `mydb.slop` and `mydb.casname` need to be placed in the `ps.<description>` subdirectories for which metabomatching should use MYDB, and can be used exclusively, or in combination with a standard database (see `reference` setting).

3 Metabomatching Figures

For every `<tag>.pseudospectrum.tsv` file in every `ps.<description>` subdirectory, metabomatching creates an SVG image that shows the pseudospectrum and the spectra of the `n_show` top-ranked candidate metabolites. The image is designed with a width of 180mm. It is optimized for use with the `Open Sans` font, freely available from fonts.google.com.

The SVG image for the test pseudospectrum included in the metabomatching download is shown in 1. The SVG image for 2-compound metabomatching (2) has some slight differences.

Titles. The titles for the metabomatching figures are provided either with an additional `description.tsv` file, or directly in the `parameters.in.tsv` file.

The `description.tsv` file is a two-column tab-separated file, in which the first column contains the pseudospectrum tag and the second a description. The pseudospectrum panel title will then read

```
Pseudospectrum of <description>
```

in bold face. In the case where pseudospectra result from an mGWAS, the description file may be formatted as a three-column file, in which the first column contains the pseudospectrum tag, the second the SNP ID, and the third the gene name. The title will then read

```
Pseudospectrum of <SNPID> in <GENE>
```

in bold face and with the gene name in italics.

Example. For the test pseudospectrum, the description file reads

```
test    rs37369 AGXT2
```

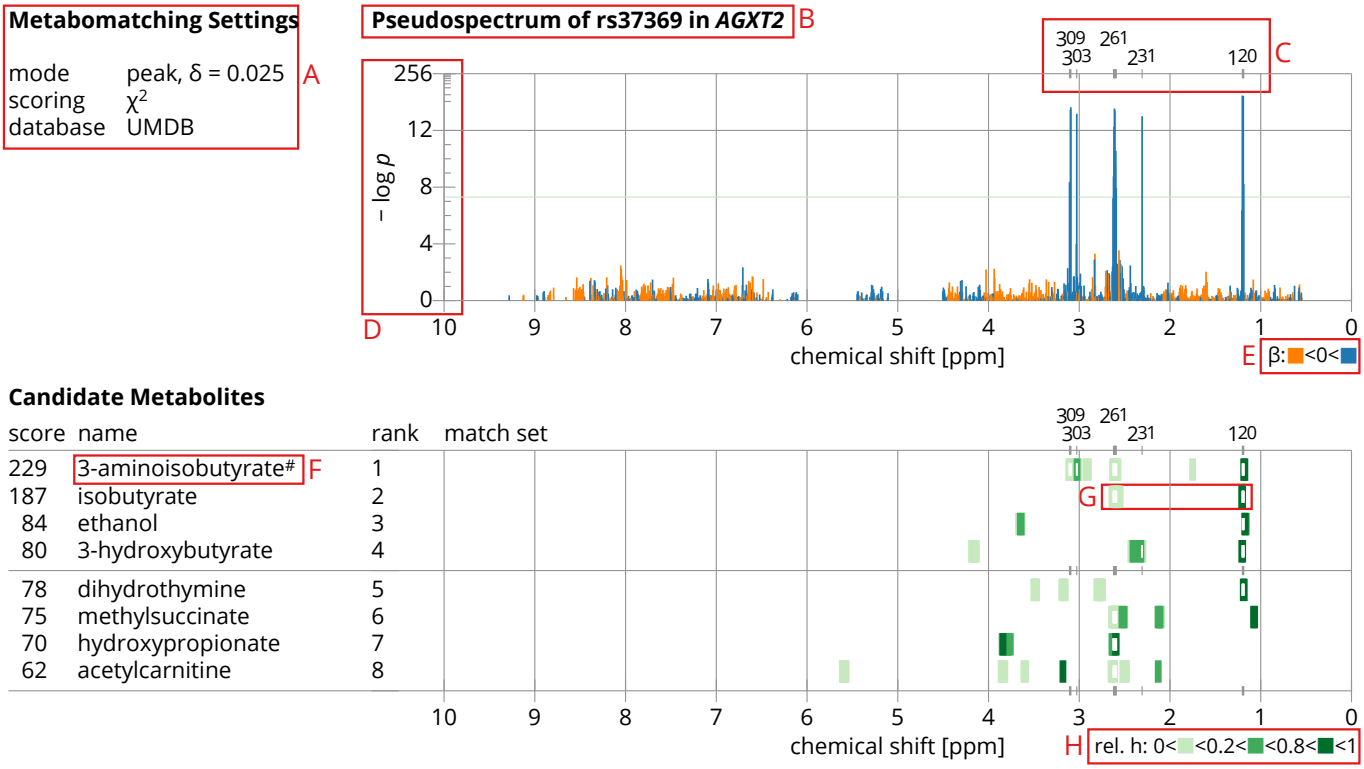


Figure 1: Metabomatching results for the pseudospectrum of SNP rs37369 in *AGXT2*. The figure contains three panels: the settings panel (*Metabomatching Settings*, top left), the pseudospectrum panel (*Pseudospectrum of . . .*, top right), and the spectrum panel (*Candidate Metabolites*, bottom).

A. The settings panel lists the main settings used in the metabomatching run.

B. The title of the pseudospectrum panel can be provided by the user (see below).

C. Features in the pseudospectrum with p-value below the significance threshold defined by `p_significant` are highlighted. The position of each significant feature is marked by a tick on the upper edge of the pseudospectrum box. In addition, significant peaks are clustered (according to δ or $\gamma + 0.015$), and the position of the peak with the strongest association in the cluster is labeled. Marks are repeated in the spectrum panel.

D. The scale used for the pseudospectrum y-axis is linear for values of $-\log p$ between 0 and 12, logarithmic for values above 12, and capped at 256.

E. Features are colored orange if their association effect size β is negative, blue if it is positive.

F. A list of metabolites that are shown regardless of their ranks, can be provided by the user (see below). Such metabolites are marked by a "#" next to their name.

G, H. The match sets are shown on the same chemical shift axis as the pseudospectrum, with one match set per row. For each match set element, the relative height of the peak that produced the element is color-coded. Elements are plotted in order of increasing relative height. In multiplet mode, relative proton counts are used instead of relative heights. The positions of significantly associated features are marked by white ticks in the match set rows to highlight match set elements that include these features.

In the `parameters.in.tsv` file, descriptions are provided as

```
description_<tag> <description>
```

Control Metabolites. Control metabolites are provided to metabomatching either with an additional `cascontrol.tsv` file, or directly in the `parameters.in.tsv` file.

The `cascontrol.tsv` file is a two column tab-separated file, in which the first column contains the tag of the pseudospectrum and the second the CASRN of the desired control metabolite. More than one control metabolite may be provided for any pseudospectrum, each in a new line.

Example. For the test pseudospectrum, the `cascontrol.csv` file reads

```
test    144-90-1
```

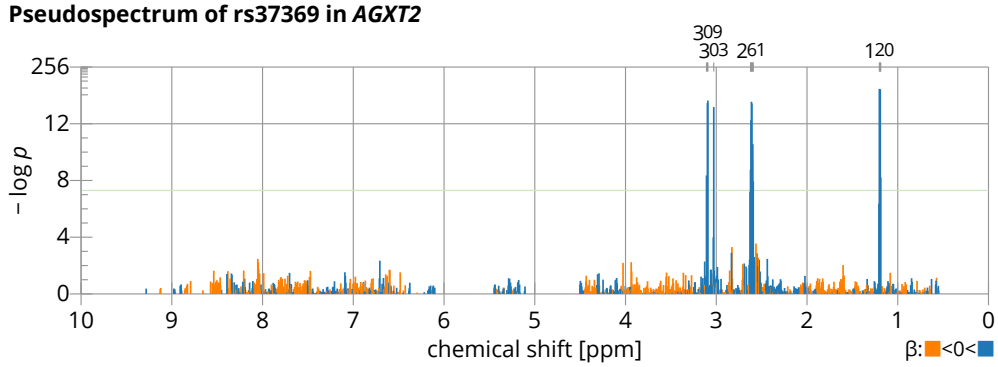
In the `parameters.in.tsv`, control metabolites are provided as

```
cas_control_<tag>_<#>   <CASRN>
```

Metabomatching Settings

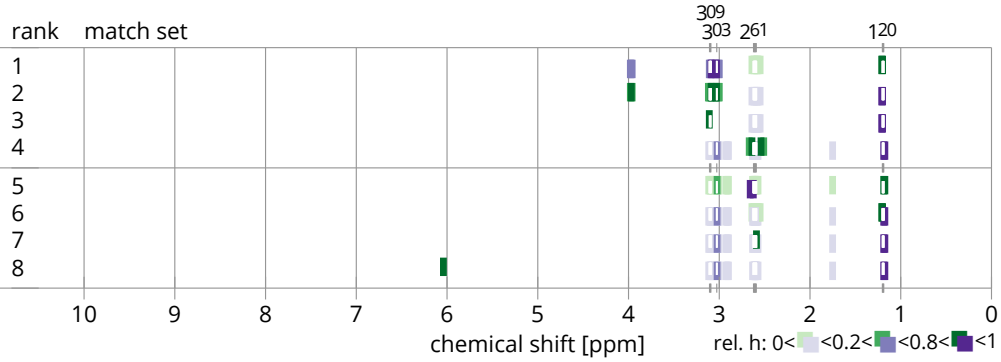
variant	2-compound
mode	peak, δ = 0.025
scoring	χ^2
database	HMDB

Pseudospectrum of rs37369 in *AGXT2*



Candidate Metabolites

score	name (■&■)	rank	match set
239	isobutyrate & D-cysteine	1	
238	cysteine & isobutyrate	2	
232	malonate & isobutyrate	3	
231	citrate & 144-90-1	4	
230	144-90-1 & 4408-78-0	5	
229	isobutyrate & 144-90-1	6	
229	methylamine & 144-90-1	7	
229	maleate & 144-90-1	8	



144-90-1: 3-aminoisobutyrate

4408-78-0: phosphonoacetate

Figure 2: 2-compound metabomatching results. When running 2-compound metabomatching, pairs of candidate metabolites are scored and ranked. As a result the *Candidate Metabolites* panel lists the names and spectra of two metabolites, with the spectra colored green for one metabolite, and purple for the other. Where metabolite names would overrun the *name* column, CASRN are listed instead, and a CASRN to name table is printed at the bottom of the figure.