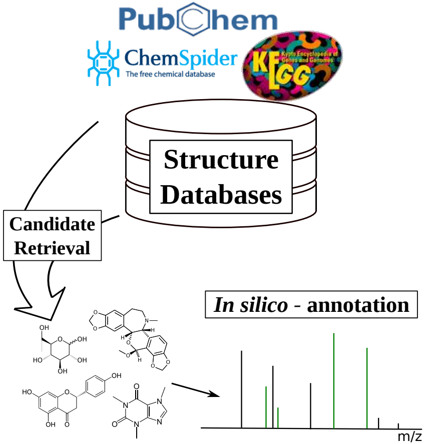
Non-Target Screening: *In silico* fragmentation and reference information

Christoph Ruttkies

Leibniz Institute of Plant Biochemistry
Department of Stress and Developmental Biology
Bioinformatics & Mass Spectrometry

16-9-2014

Workflow

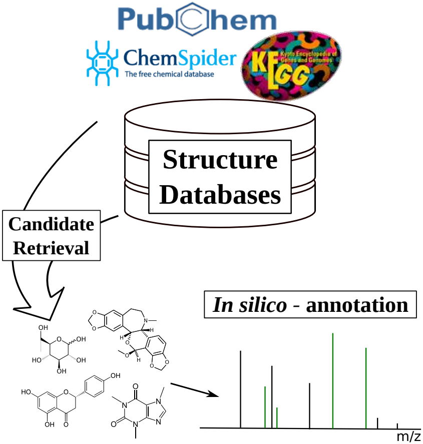


Structure Databases

Candidate Retrieval

In silico - annotation

Workflow



Structure Databases

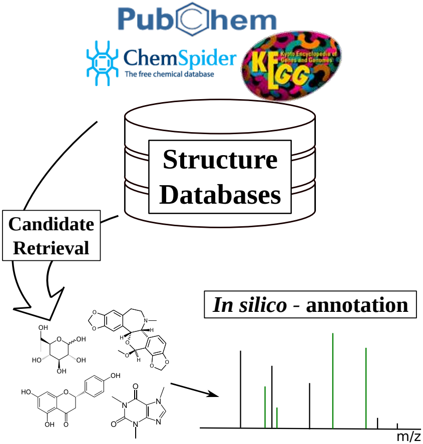
Candidate Retrieval

In silico - annotation

PubChem
ChemSpider
The free chemical database
KEGG
Kyoto Encyclopedia of Genes and Genomes
m/z

Candidate selection

Workflow


PubChem
ChemSpider
The free chemical database
KEGG
Kyoto Encyclopedia of Genes and Genomes
m/z

Structure Databases

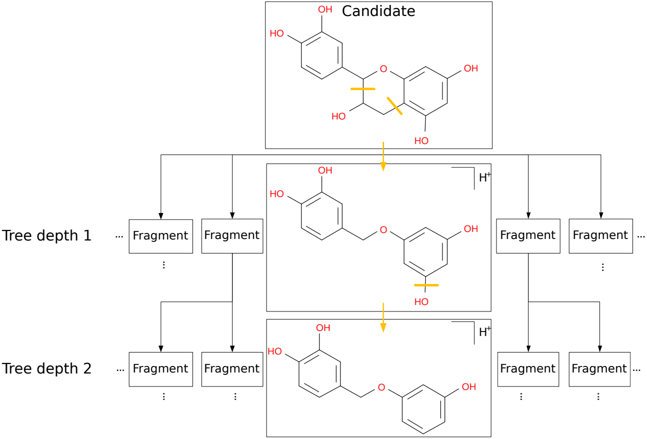
Candidate Retrieval

In silico - annotation

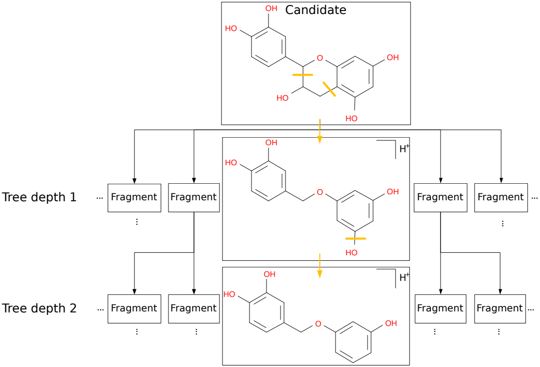
Candidate selection

Fragmentation

In silico fragmentation



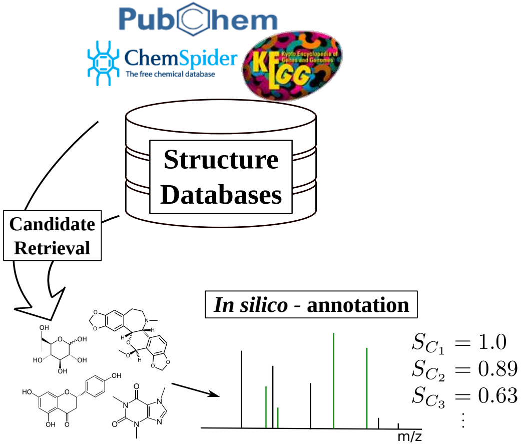
In silico fragmentation



- here: combinatorial
- also rule based fragmentation approaches[1]

[1] Kerber, A. *et al.* (2013): Mathematical Chemistry and Chemoinformatics: Structure Generation, Elucidation and Quantitative Structure-Property Relationships

Workflow

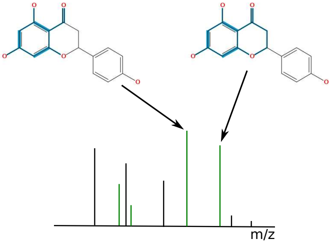


Candidate selection

Fragmentation

Scoring/ Ranking

Scoring



- each candidate for a spectrum is scored separately

→ mass of matched peak is considered

→ intensity of peak is considered

→ strength of bonds broken are considered

Where are we? What do we want to achieve?

Current state:

- HR-MS/MS data of molecules from environmental reference standards
- benchmark data of 466 merged spectra (+, -)
- MetFrag ranks 12 % correctly in top spot (PubChem + mol. formula)

Aim:

improve HR-MS/MS based identification (MetFrag) with additional information from the experimental context

references in literature (PubMed)

Where are we? What do we want to achieve?

Current state:

- HR-MS/MS data of molecules from environmental reference standards
- benchmark data of 466 merged spectra (+, -)

- MetFrag ranks 12 % correctly in top spot (PubChem + mol. formula)

Aim:

improve HR-MS/MS based identification (MetFrag) with additional information from the experimental context

references in literature (PubMed)

Where are we? What do we want to achieve?

Current state:

- HR-MS/MS data of molecules from environmental reference standards
- benchmark data of 466 merged spectra (+, -)

- MetFrag ranks 12 % correctly in top spot (PubChem + mol. formula)

Aim:

improve HR-MS/MS based identification (MetFrag) with additional information from the experimental context

references in literature (PubMed) ???

Preliminary work

Little, J.L. *et al.* (2011): **Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider**[1]

"Known Unknowns" - unknown to the investigator but known to the chemical literature, reference database, internet resource

Approach:

- selecting compounds by accurate mass search (ChemSpider, PubChem)
- sorting them (descending) by number of references
- most useful results at the top

The more an environmental compound is used the more likely it is to be found in a sample.

[1] J. Am. Soc. Mass Spectrom. 23, 179-185

Preliminary work

Little, J.L. *et al.* (2011): **Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider**[1]

"Known Unknowns" - unknown to the investigator but known to the chemical literature, reference database, internet resource

Approach:

- selecting compounds by accurate mass search (ChemSpider, PubChem)
- sorting them (descending) by number of references
- most useful results at the top

The more an environmental compound is used the more likely it is to be found in a sample.

[1] J. Am. Soc. Mass Spectrom. 23, 179-185

Preliminary work

Little, J.L. *et al.* (2011): **Identification of “Known Unknowns” Utilizing Accurate Mass Data and ChemSpider**[1]

“Known Unknowns” - unknown to the investigator but known to the chemical literature, reference database, internet resource

Approach:

- selecting compounds by accurate mass searc[illegible]der, PubChem)
- sorting them (descending) by number of referen[illegible]
- most useful results at the top

no HR-MS/MS info included

The more an environmental compound is used the more likely it is to be found in a sample.

[1] J. Am. Soc. Mass Spectrom. 23, 179-185

Reference resources for chemicals

Scopus

PubMed

Google scholar

SciFinder®
Essential content. Proven results.™

ChemSpider
The free chemical database

WEB OF SCIENCE

Why we use PubMed? → **Content**

- technical reasons
- open access
- PubChem (direct link)

- 23M citations (incl. MEDLINE)
- biomedical, chemical, biological, environmental[1]

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Reference resources for chemicals

Scopus

PubMed

Google scholar

SciFinder® Essential content. Proven results.™

ChemSpider The free chemical database

WEB OF SCIENCE

Why we use PubMed? → **Content**

- technical reasons
- open access
- PubChem (direct link)

- 23M citations (incl. MEDLINE)
- biomedical, chemical, biological, environmental[1]

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Reference resources for chemicals

Scopus

PubMed

Google scholar

SciFinder® Essential content. Proven results.™

ChemSpider The free chemical database

WEB OF SCIENCE

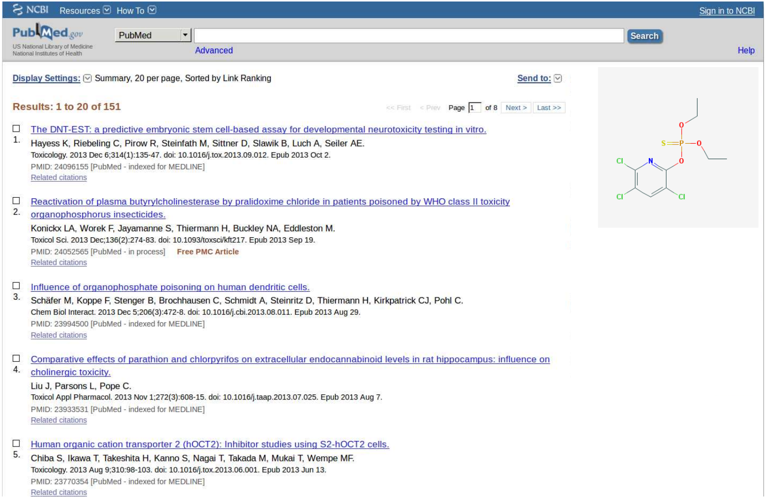
Why we use PubMed? → **Content**

- technical reasons
- open access
- PubChem (direct link)

- 23M citations (incl. MEDLINE)
- biomedical, chemical, biological, environmental[1]

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Example: Chlorpyriphos's PubMed References



Display Settings: Summary, 20 per page, Sorted by Link Ranking

Send to:

Results: 1 to 20 of 151

<< First < Prev Page 1 of 8 Next > Last >>

□ 1. The DNT-EST: a predictive embryonic stem cell-based assay for developmental neurotoxicity testing in vitro.
Hayess K, Riebeling C, Pirow R, Steinfath M, Sittner D, Slawik B, Luch A, Seiler AE.
Toxicology. 2013 Dec 6;314(1):135-47. doi: 10.1016/j.tox.2013.09.012. Epub 2013 Oct 2.
PMID: 24096155 [PubMed - indexed for MEDLINE]
Related citations

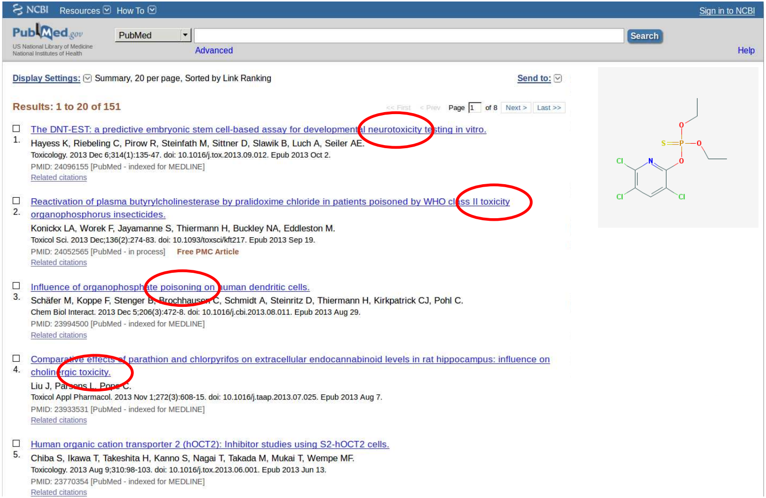
□ 2. Reactivation of plasma butyrylcholinesterase by pralidoxime chloride in patients poisoned by WHO class II toxicity organophosphorus insecticides.
Konickx LA, Worek F, Jayamanne S, Thiermann H, Buckley NA, Eddleston M.
Toxicol Sci. 2013 Dec;136(2):274-83. doi: 10.1093/toxsci/kft217. Epub 2013 Sep 19.
PMID: 24052565 [PubMed - in process] Free PMC Article
Related citations

□ 3. Influence of organophosphate poisoning on human dendritic cells.
Schäfer M, Koppe F, Stenger B, Brochhausen C, Schmidt A, Steinritz D, Thiermann H, Kirkpatrick CJ, Pohl C.
Chem Biol Interact. 2013 Dec 5;206(3):472-8. doi: 10.1016/j.cbi.2013.08.011. Epub 2013 Aug 29.
PMID: 23994500 [PubMed - indexed for MEDLINE]
Related citations

□ 4. Comparative effects of parathion and chlorpyrifos on extracellular endocannabinoid levels in rat hippocampus: influence on cholinergic toxicity.
Liu J, Parsons L, Pope C.
Toxicol Appl Pharmacol. 2013 Nov 1;272(3):608-15. doi: 10.1016/j.taap.2013.07.025. Epub 2013 Aug 7.
PMID: 23933531 [PubMed - indexed for MEDLINE]
Related citations

□ 5. Human organic cation transporter 2 (hOCT2): Inhibitor studies using S2-hOCT2 cells.
Chiba S, Ikawa T, Takeshita H, Kanno S, Nagai T, Takada M, Mukai T, Wempe MF.
Toxicology. 2013 Aug 9;310:98-103. doi: 10.1016/j.tox.2013.06.001. Epub 2013 Jun 13.
PMID: 23770354 [PubMed - indexed for MEDLINE]
Related citations

Example: Chlorpyriphos's PubMed References



Display Settings: Summary, 20 per page, Sorted by Link Ranking

Results: 1 to 20 of 151

□ 1. The DNT-EST: a predictive embryonic stem cell-based assay for developmental neurotoxicity testing in vitro.
Hayess K, Riebeling C, Pirow R, Steinfath M, Sittner D, Slawik B, Luch A, Seiler AE.
Toxicology. 2013 Dec 6;314(1):135-47. doi: 10.1016/j.tox.2013.09.012. Epub 2013 Oct 2.
PMID: 24096155 [PubMed - indexed for MEDLINE]
Related citations

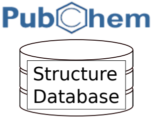
□ 2. Reactivation of plasma butyrylcholinesterase by pralidoxime chloride in patients poisoned by WHO class II toxicity organophosphorus insecticides.
Konickx LA, Worek F, Jayamanne S, Thiermann H, Buckley NA, Eddleston M.
Toxicol Sci. 2013 Dec;136(2):274-83. doi: 10.1093/toxsci/kft217. Epub 2013 Sep 19.
PMID: 24052565 [PubMed - in process] Free PMC Article
Related citations

□ 3. Influence of organophosphate poisoning on human dendritic cells.
Schäfer M, Koppe F, Stenger B, Brochhausen C, Schmidt A, Steinritz D, Thiermann H, Kirkpatrick CJ, Pohl C.
Chem Biol Interact. 2013 Dec 5;206(3):472-8. doi: 10.1016/j.cbi.2013.08.011. Epub 2013 Aug 29.
PMID: 23994500 [PubMed - indexed for MEDLINE]
Related citations

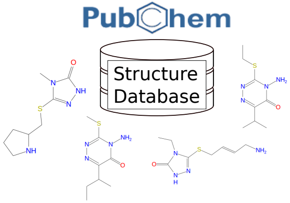
□ 4. Comparative effects of parathion and chlorpyrifos on extracellular endocannabinoid levels in rat hippocampus: influence on cholinergic toxicity.
Liu J, Parsons L, Pope C.
Toxicol Appl Pharmacol. 2013 Nov 1;272(3):608-15. doi: 10.1016/j.taap.2013.07.025. Epub 2013 Aug 7.
PMID: 23933531 [PubMed - indexed for MEDLINE]
Related citations

□ 5. Human organic cation transporter 2 (hOCT2): Inhibitor studies using S2-hOCT2 cells.
Chiba S, Ikawa T, Takeshita H, Kanno S, Nagai T, Takada M, Mukai T, Wempe MF.
Toxicology. 2013 Aug 9;310:98-103. doi: 10.1016/j.tox.2013.06.001. Epub 2013 Jun 13.
PMID: 23770354 [PubMed - indexed for MEDLINE]
Related citations

Combining MetFrag with References



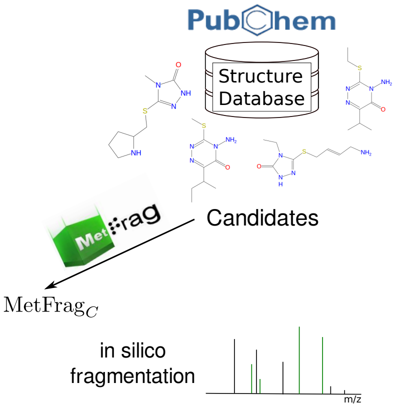
Combining MetFrag with References



Structure Database

Candidates

Combining MetFrag with References



PubChem

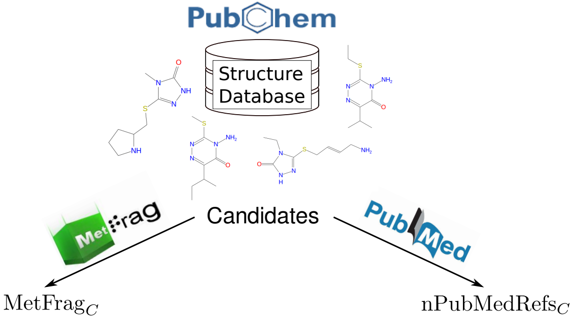
Structure Database

Candidates

MetFrag$_C$

in silico fragmentation

Combining MetFrag with References



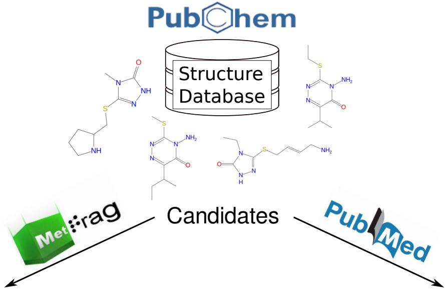
Structure Database

Candidates

MetFrag_C

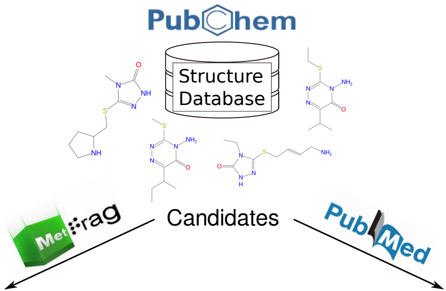
nPubMedRefs_C

Combining MetFrag with References



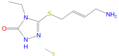
$\alpha \cdot \mathrm{MetFrag}_C \quad + \quad \beta \cdot \mathrm{nPubMedRefs}_C$

Combining MetFrag with References

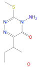


$$\alpha \cdot \mathrm{MetFrag}_C \quad + \quad \beta \cdot \mathrm{nPubMedRefs}_C$$

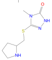
$$S_{C_1} = 1.0$$



$$S_{C_2} = 0.97$$



$$S_{C_3} = 0.71$$



$$\vdots$$

Example: Chlorpyriphos

HR-MS/MS data of Challenge 9 (Chlorpyrifos)
CASMI contest 2013[1]

[1] http://casmi-contest.org/2013/challenges.shtml

Example: Chlorpyriphos

HR-MS/MS data of Challenge 9 (Chlorpyrifos)
CASMI contest 2013[1]

→ PubChem + 348.9264 Da = 114 candidates

→ MetFrag alone: Rank 3

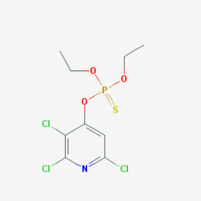
[1] http://casmi-contest.org/2013/challenges.shtml

Example: Chlorpyriphos

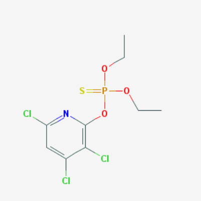
HR-MS/MS data of Challenge 9 (Chlorpyrifos)
CASMI contest 2013[1]

→ PubChem + 348.9264 Da = 114 candidates

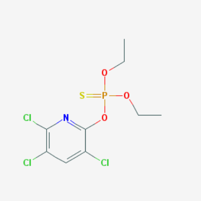
→ MetFrag alone: Rank 3



(1) CID: 57354037



(2) CID: 13274485



(3) Chlorpyrifos

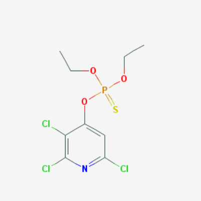
[1] http://casmi-contest.org/2013/challenges.shtml

Example: Chlorpyriphos

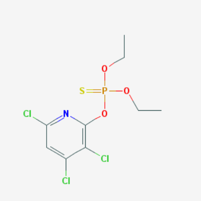
HR-MS/MS data of Challenge 9 (Chlorpyrifos)
CASMI contest 2013[1]

→ PubChem + 348.9264 Da = 114 candidates

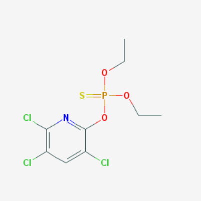
→ MetFrag alone: Rank 3



(1) CID: 57354037
No. Refs: 0



(2) CID: 13274485
No. Refs: 0



(3) Chlorpyrifos
No. Refs: 151

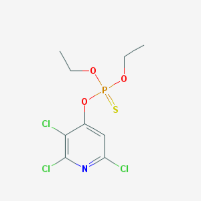
[1] http://casmi-contest.org/2013/challenges.shtml

Example: Chlorpyriphos

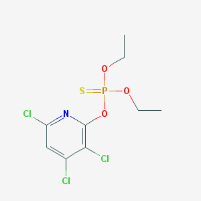
HR-MS/MS data of Challenge 9 (Chlorpyrifos)
CASMI contest 2013[1]

→ PubChem + 348.9264 Da = 114 candidates

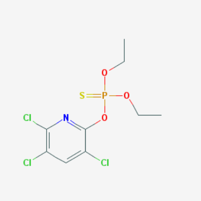
→ MetFrag + PubMed Refs: Rank 1



(1) CID: 57354037
No. Refs: 0



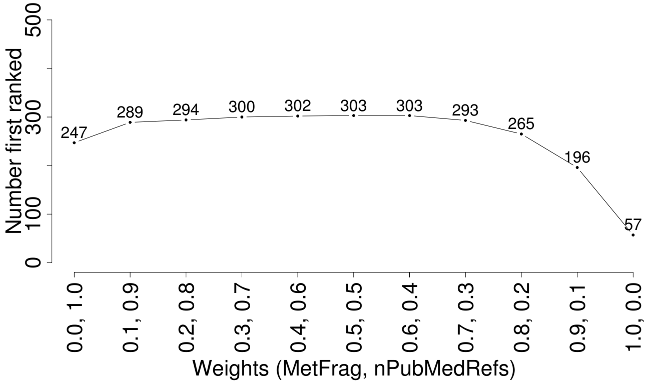
(2) CID: 13274485
No. Refs: 0



(3) Chlorpyrifos
No. Refs: 151

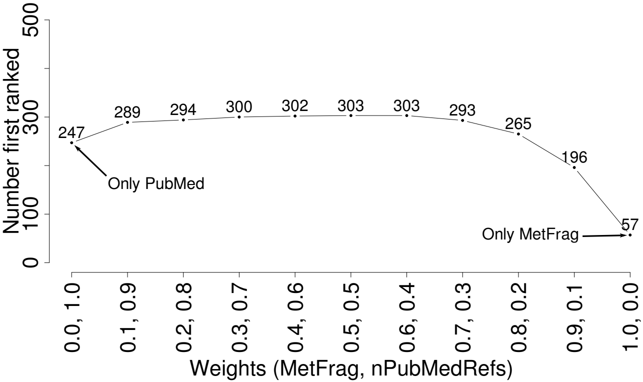
[1] http://casmi-contest.org/2013/challenges.shtml

Number of ranked first correctly (466 spectra)



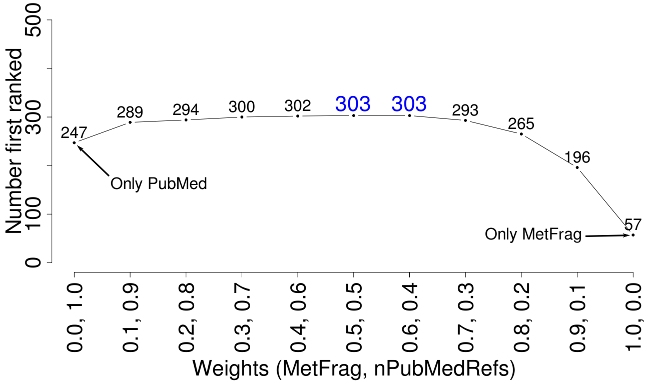
$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{nPubMedRefs}_C$$

Number of ranked first correctly (466 spectra)



$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{nPubMedRefs}_C$$

Number of ranked first correctly (466 spectra)



$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{nPubMedRefs}_C$$

Aim: Further improvement

- include further parameters to improve identification
- information weighted based on (user-)defined priorities

$$\begin{aligned} S_C = \; & \omega_{\mathbf{1}} \cdot (\textbf{MetFrag})_C \\ & + \omega_{\mathbf{2}} \cdot (\textbf{PubMedRefs})_C \end{aligned}$$

Aim: Further improvement

- include further parameters to improve identification
- information weighted based on (user-)defined priorities

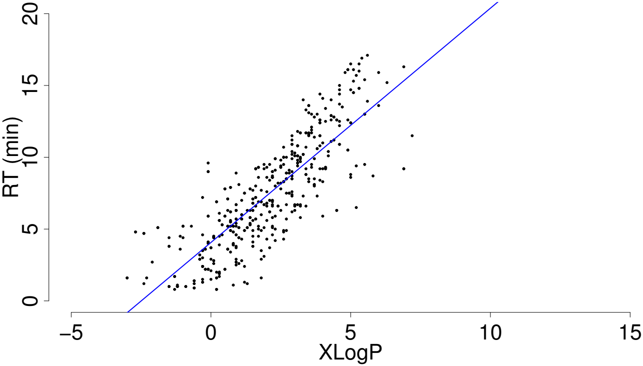
$$\begin{aligned} S_C = \; & \omega_{\mathbf{1}} \cdot (\textbf{MetFrag})_C & & \text{first ranked} \\ & & & \sim 12\ \% \\ & + \omega_{\mathbf{2}} \cdot (\textbf{PubMedRefs})_C & & \sim 65\ \% \end{aligned}$$

Aim: Further improvement

- include further parameters to improve identification
- information weighted based on (user-)defined priorities

$$\begin{aligned} S_C = \ & \boldsymbol{\omega_1} \cdot (\textbf{MetFrag})_C & & \text{first ranked} \\ & & & \sim 12\ \% \\ & + \boldsymbol{\omega_2} \cdot (\textbf{PubMedRefs})_C & & \sim 65\ \% \\ & + \boldsymbol{\omega_3} \cdot \text{Retention time prediction} & & \end{aligned}$$

Retention time modelling



$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{nPubMedRefs}_C + \gamma \cdot \mathrm{RT}_C$$

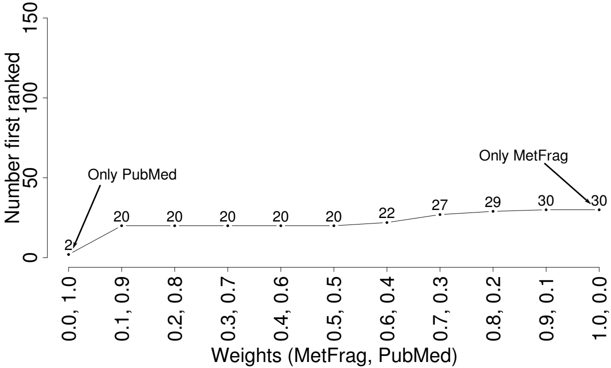
Aim: Further improvement

- include further parameters to improve identification
- information weighted based on (user-)defined priorities

$$\begin{aligned} & & \text{first ranked} \\ S_C = \; & \omega_\mathbf{1} \cdot (\textbf{MetFrag})_C & \sim 12\ \% \\ & + \omega_\mathbf{2} \cdot (\textbf{PubMedRefs})_C & \sim 65\ \% \\ & + \omega_\mathbf{3} \cdot \text{Retention time prediction} & \sim 66\ \% \end{aligned}$$

Cases with no references

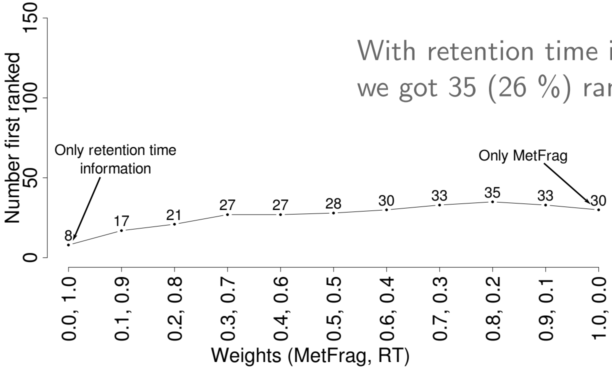
For **134** of 466 spectra the correct candidate has no PubMed reference.



$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{nPubMedRefs}_C$$

Cases with no references

For **134** of 466 spectra the correct candidate has no PubMed reference.



With retention time information we got 35 (26 %) ranked first.

$$S_C = \alpha \cdot \mathrm{MetFrag}_C + \beta \cdot \mathrm{RT}_C$$

Aim: Further improvement

- include further parameters to improve identification
- information weighted based on (user-)defined priorities

$$
\begin{array}{rlc}
 & & \text{first ranked} \\
S_C = & \boldsymbol{\omega_1} \cdot (\textbf{MetFrag})_C & \sim 12\ \% \\
 & + \boldsymbol{\omega_2} \cdot (\textbf{PubMedRefs})_C & \sim 65\ \% \\
 & + \boldsymbol{\omega_3} \cdot \text{Retention time prediction} & \sim 66\ \% \\
 & + \boldsymbol{\omega_4} \cdot \text{Spectral information} & ? \\
 & + \boldsymbol{\omega_5} \cdot \text{Substructure (non}-\text{)presence} & ? \\
 & + \boldsymbol{\omega_6} \cdot \text{Effect prediction} & ? \\
 & + \boldsymbol{\omega_i} \cdot \ldots &
\end{array}
$$

Future aim: create an interface that allows user-defined scoring terms

Acknowledgement

Dierk Scheel	IPB Halle
Steffen Neumann	IPB Halle
Emma Schymanski	Eawag Zurich
Carsten Kuhl	IPB Halle
Michael Gerlich	IPB Halle
Susann Mönchgesang	IPB Halle
Heinz Singer	Eawag Zurich
Michael Stravs	Eawag Zurich
Juliane Hollender	Eawag Zurich

MetFrag MetFusion

MassBank CASMI

Funding:

solutions

DFG Deutsche Forschungsgemeinschaft

Acknowledgement

Dierk Scheel	IPB Halle
Steffen Neumann	IPB Halle
Emma Schymanski	Eawag Zurich
Carsten Kuhl	IPB Halle
Michael Gerlich	IPB Halle
Susann Mönchgesang	IPB Halle
Heinz Singer	Eawag Zurich
Michael Stravs	Eawag Zurich
Juliane Hollender	Eawag Zurich

MetFrag MetFusion

MassBank CASMI

Funding:

solutions

DFG Deutsche Forschungsgemeinschaft

Thank you for attention!

Q&A?

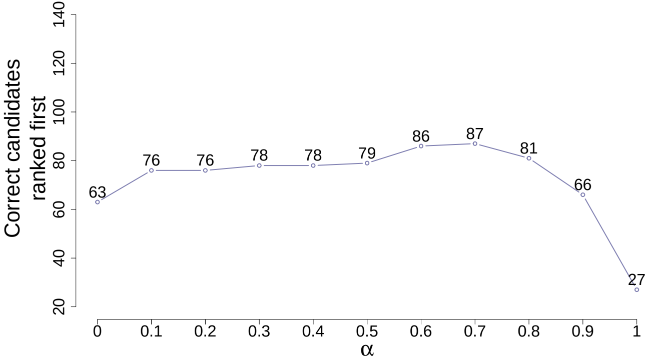
Appendix

MetFrag with References: Results

- 333 of 421 MS/MS spectra have candidates with PubMed entries
- **Question:** Do we need MetFrag when there is a PubMed entry?
- **Answer:** Show results on the 333 spectra where PubMed entries are present in the candidate set.

MetFrag with References: Results

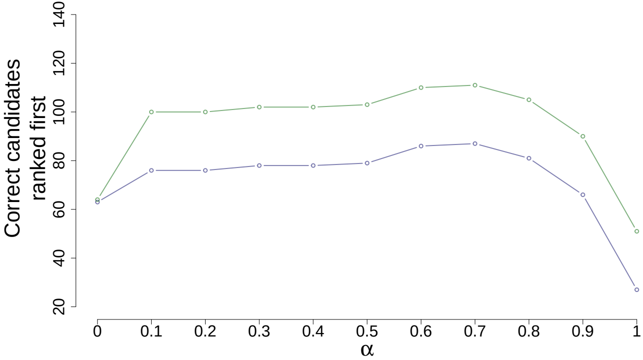
MS/MS dataset of 333 merged spectra



$$S_C = \alpha \cdot \mathrm{MetFrag}_C + (1-\alpha) \cdot \mathrm{nPubMedRefs}_C$$

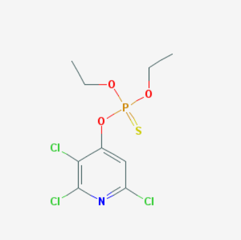
MetFrag with References: Results

MS/MS dataset of 333 merged spectra

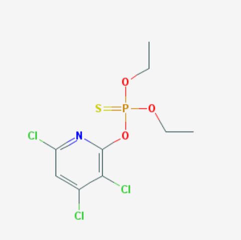


$$S_C = \alpha \cdot \mathrm{MetFrag}_C + (1-\alpha) \cdot \mathrm{nPubMedRefs}_C$$

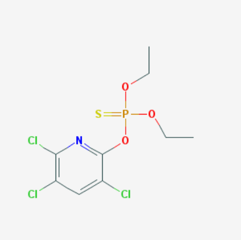
CASMI 2013 - Challenge 9



(a) 57354037



(b) 13274485



(c) Chlorpyrifos

User-defined scoring terms

- user provides a list of structures with determined features

CID	Toxicity	Effect x_1	Effect x_2	Pred.Ret.time	...
321	1	1	-1	5.63	...
26881	1	-1	-1	10.22	...
9752	-1	-1	1	11.33	...
909	1	1	-1	9.87	...
87665	1	-1	-1	9.41	...

- identification of compounds relies on the given information with user-defined priorities

CASMI 2013: Challenge 9

CASMI 2013

Critical Assessment of Small Molecule Identification

- MS data for 16 challenges provided
- measured compounds are unknown to the investigator
- use your desired workflow and identification method

- we tackled **Challenge 9** originating from an environmental sample together with **PubMed**

Normalised PubMed reference score

$$\mathrm{nPubMedRefs}_{C} = \frac{\mathrm{PubMedRefs}_{C}}{max(\mathrm{PubMedRefs}_{C})}$$

MetFrag score

score of one candidate c:

$$S_c = \sum_{\mathrm{f} \in \mathrm{F_c}} \frac{\left(10.0 \cdot \frac{\mathrm{Mass_f}}{\mathrm{Mass_c}}\right)^{\alpha} \cdot (\mathrm{RelInt_f})^{\beta}}{\left(\sum_{\mathrm{b} \in \mathrm{f}} \mathrm{BDE_b}\right)^{\gamma}}$$

MetFrag score

score of one candidate c:

$$S_c = \sum_{\mathrm{f} \in \mathrm{F_c}} \frac{\left(10.0 \cdot \frac{\mathrm{Mass_f}}{\mathrm{Mass_c}}\right)^{\alpha} \cdot (\mathrm{RelInt_f})^{\beta}}{\left(\sum_{\mathrm{b} \in \mathrm{f}} \mathrm{BDE_b}\right)^{\gamma}}$$

- all matched fragments f of candidate c

MetFrag score

score of one candidate c:

$$S_c = \sum_{\mathrm{f} \in \mathrm{F_c}} \frac{\left(10.0 \cdot \frac{\mathrm{Mass_f}}{\mathrm{Mass_c}}\right)^{\alpha} \cdot (\mathrm{RelInt_f})^{\beta}}{\left(\sum_{\mathrm{b} \in \mathrm{f}} \mathrm{BDE_b}\right)^{\gamma}}$$

- all matched fragments f of candidate c
- relative mass of matched fragment f and the candidate c

MetFrag score

score of one candidate c:

$$S_c = \sum_{\mathrm{f} \in \mathrm{F_c}} \frac{\left(10.0 \cdot \frac{\mathrm{Mass_f}}{\mathrm{Mass_c}}\right)^{\alpha} \cdot (\mathrm{RelInt_f})^{\beta}}{\left(\sum_{\mathrm{b} \in \mathrm{f}} \mathrm{BDE_b}\right)^{\gamma}}$$

- all matched fragments f of candidate c
- relative mass of matched fragment f and the candidate c
- relative intensity of explained peak

MetFrag score

score of one candidate c:

$$S_c = \sum_{\mathrm{f} \in \mathrm{F_c}} \frac{\left(10.0 \cdot \frac{\mathrm{Mass_f}}{\mathrm{Mass_c}}\right)^{\alpha} \cdot (\mathrm{RelInt_f})^{\beta}}{\left(\sum_{\mathrm{b} \in \mathrm{f}} \mathrm{BDE_b}\right)^{\gamma}}$$

- all matched fragments f of candidate c
- relative mass of matched fragment f and the candidate c
- relative intensity of explained peak
- sum of Bond Dissociation Energies (BDEs) of all broken bonds b