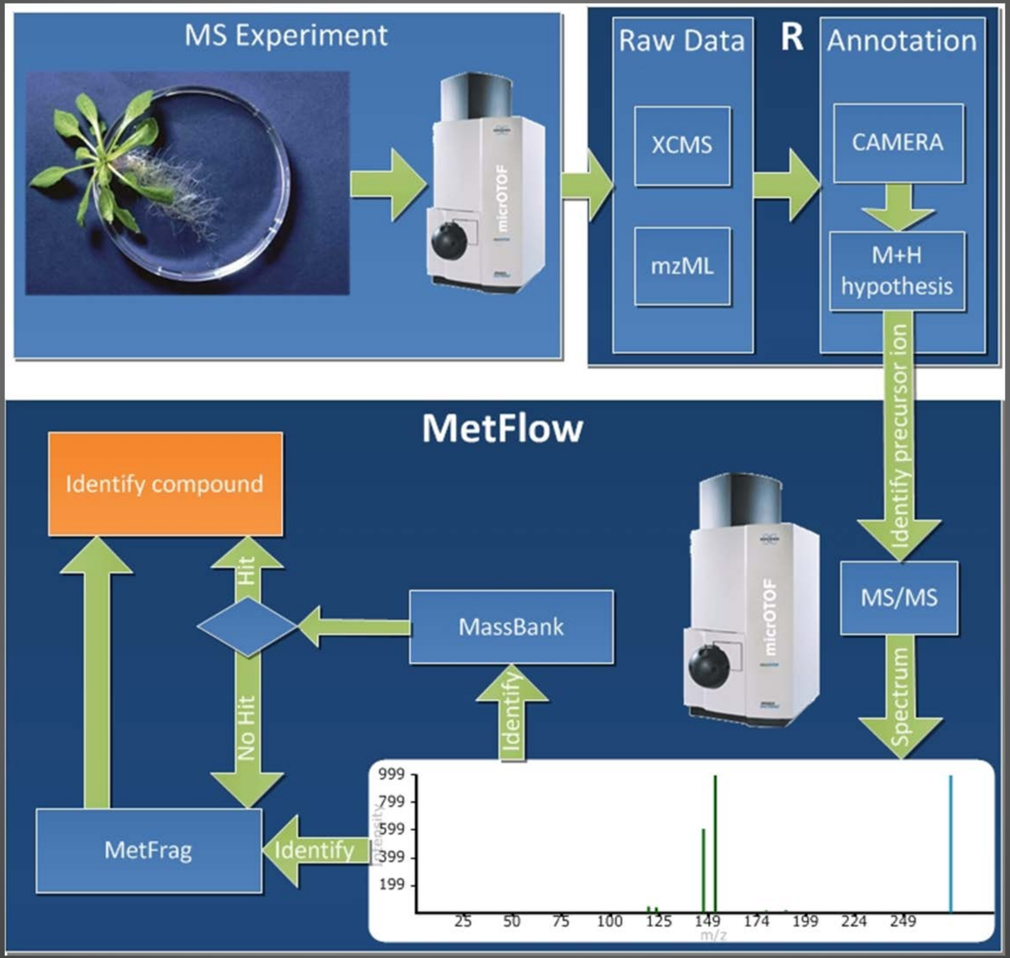
MetFusion: integration of compound identification strategies

MassBank Workshop 27.11.2012

Amsterdam, The Netherlands

Michael Gerlich, Department of Stress and Developmental Biology, IPB Halle

Introduction

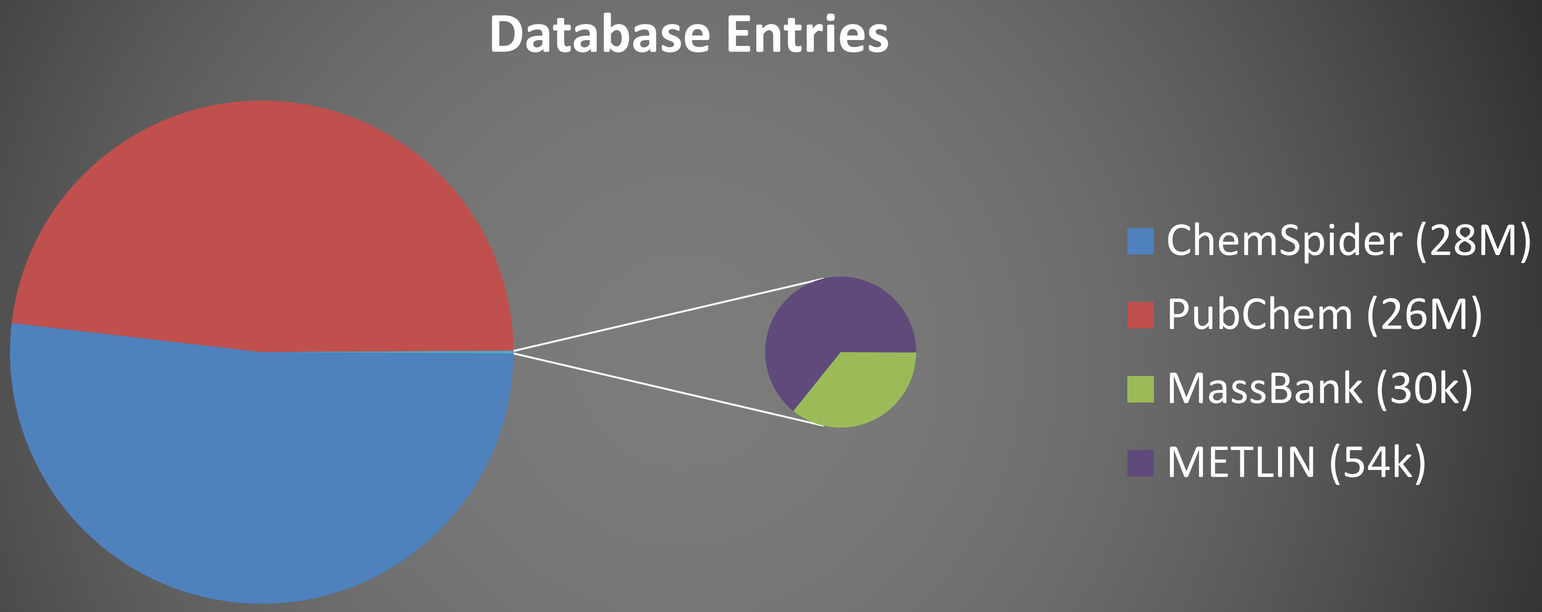


Michael Gerlich, Department of Stress and Developmental Biology, IPB Halle

Structures & Spectra

- Target list of interesting MS^2 spectra
- Requires expert knowledge
  - Time-consuming
  - Impossible to keep track of high-throughput
- Only small fraction of compounds has associated reference spectrum
  - Required for speed-up in re-identification
  - Difficult to use for *de novo* identification

Structures vs. Spectra Imbalance



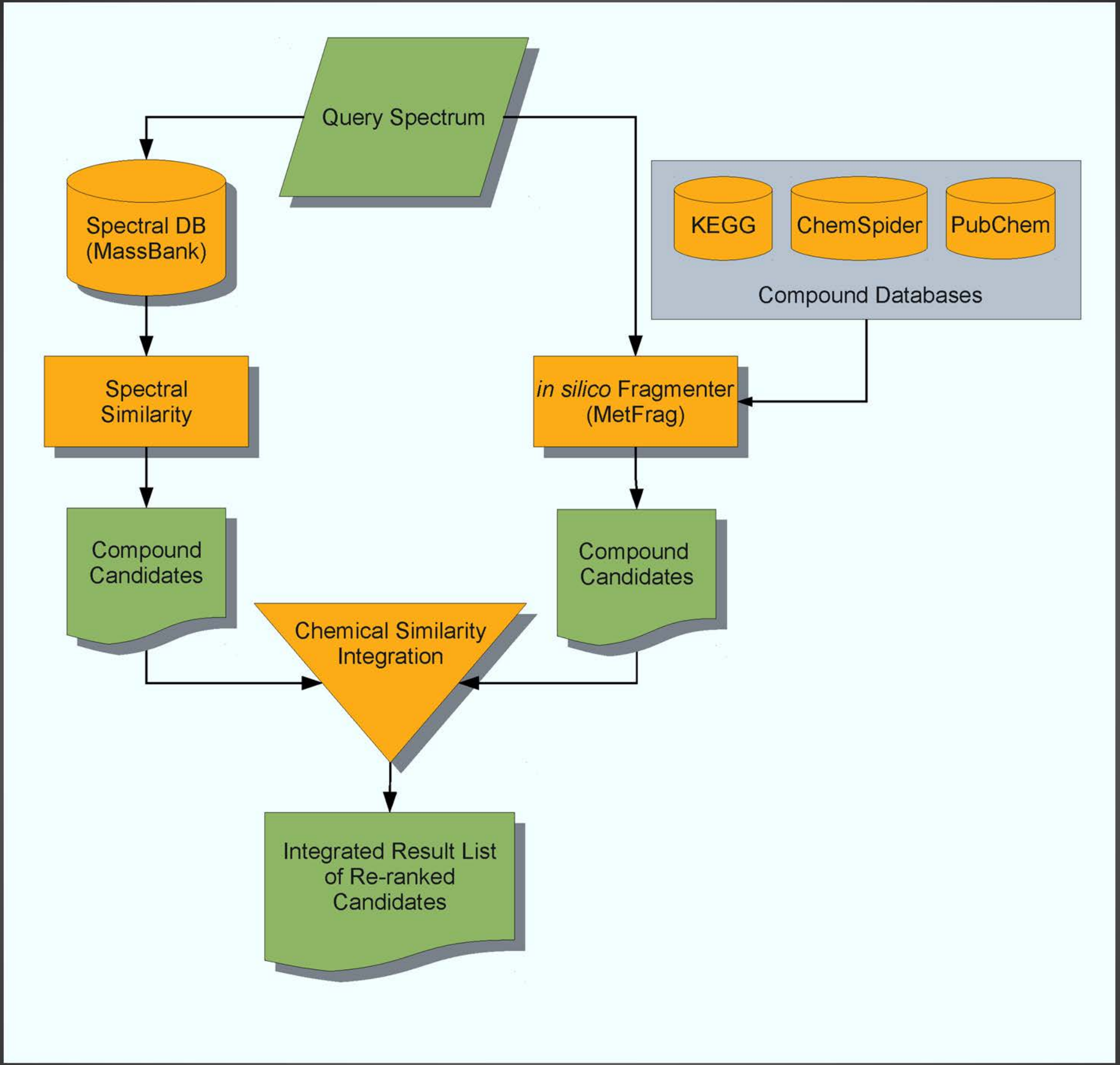
- Millions vs. thousands, overlap unknown
- Spectrum queries not possible for compound databases

Observation vs. Prediction

- Observation: MassBank
  - Search measured spectra with peak list
  - Find compounds with matching/similar spectrum
  - Few reliable results
- Prediction: MetFrag
  - Combinatorial Fragmenter
  - Generate fragments *in silico*, matches to peak list
  - Uses compound databases,
    many possible predicted results

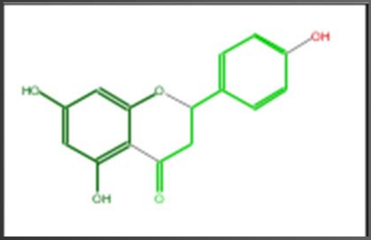
MetFusion = Observation, Prediction & Similarity

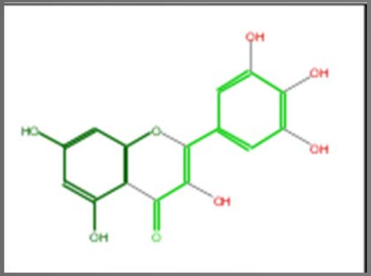
- Combine results via chemical similarity
  - Structural fingerprints
  - Use each information (score)
  - Avoid strict limits/thresholds
- Aim: improve identification
  - Assume that correct compound is present in compound database (larger coverage)
  - Enhance MetFrag results with spectral data
  - DOI 10.1002/jms.3123

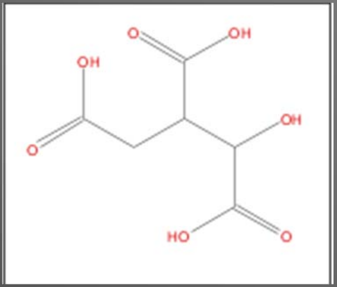

Query Spectrum
Spectral DB (MassBank)
KEGG
ChemSpider
PubChem
Compound Databases
Spectral Similarity
in silico Fragmenter (MetFrag)
Compound Candidates
Compound Candidates
Chemical Similarity Integration
Integrated Result List of Re-ranked Candidates

What is similar?

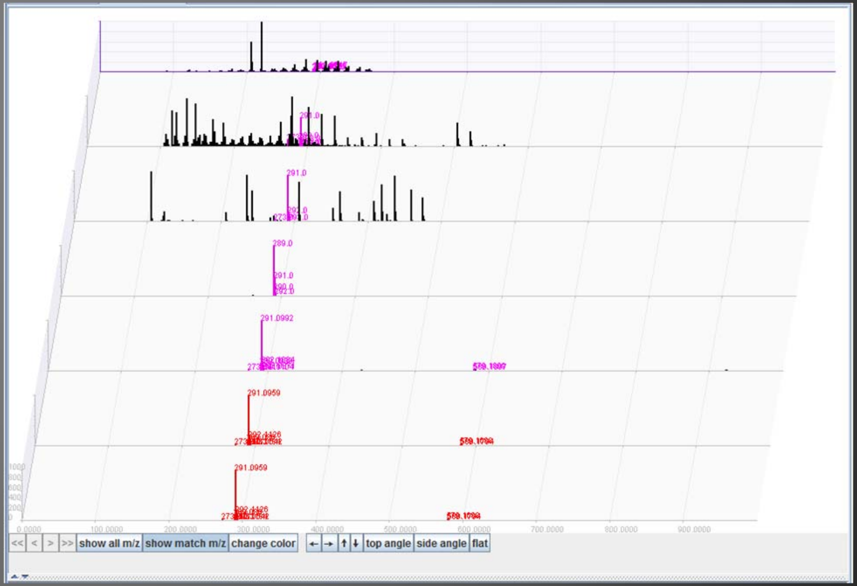
Chemical Similarity







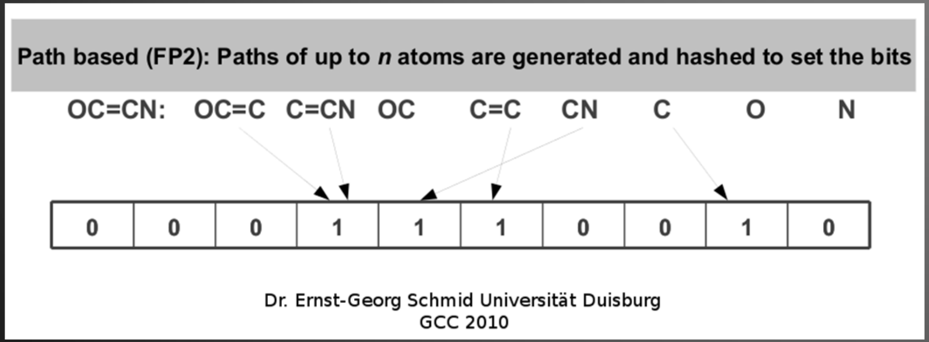
Spectral Similarity



Similarity Measures

Chemical Similarity

- Tanimoto coefficient matches properties
- $Tan = \frac{C}{A+B-C}$
- [0,1]



Path based (FP2): Paths of up to *n* atoms are generated and hashed to set the bits

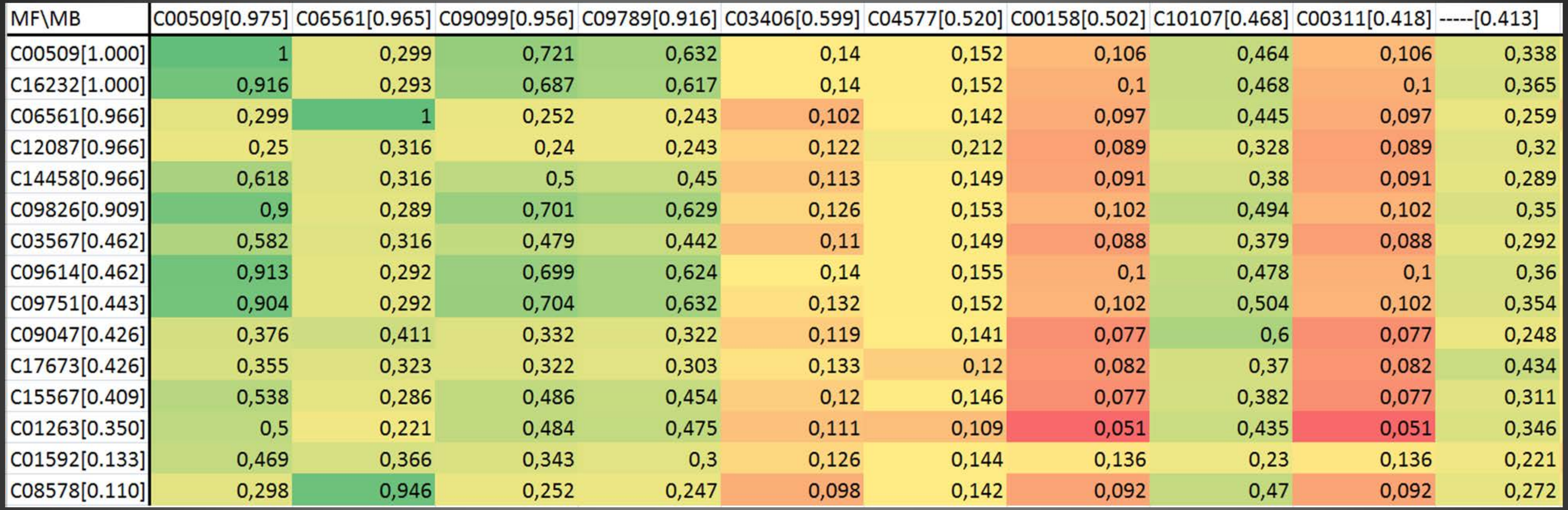
0	0	0	1	1	1	0	0	1	0

Dr. Ernst-Georg Schmid Universität Duisburg
GCC 2010

Spectral Similarity

- Modified Cosine distance
- Matches peak masses & intensities
- $W_i = int_i^m * m/z_i^n$
- [0,1]

Similarity Matrix



MF\MB	C00509[0.975]	C06561[0.965]	C09099[0.956]	C09789[0.916]	C03406[0.599]	C04577[0.520]	C00158[0.502]	C10107[0.468]	C00311[0.418]	-----[0.413]
C00509[1.000]	1	0,299	0,721	0,632	0,14	0,152	0,106	0,464	0,106	0,338
C16232[1.000]	0,916	0,293	0,687	0,617	0,14	0,152	0,1	0,468	0,1	0,365
C06561[0.966]	0,299	1	0,252	0,243	0,102	0,142	0,097	0,445	0,097	0,259
C12087[0.966]	0,25	0,316	0,24	0,243	0,122	0,212	0,089	0,328	0,089	0,32
C14458[0.966]	0,618	0,316	0,5	0,45	0,113	0,149	0,091	0,38	0,091	0,289
C09826[0.909]	0,9	0,289	0,701	0,629	0,126	0,153	0,102	0,494	0,102	0,35
C03567[0.462]	0,582	0,316	0,479	0,442	0,11	0,149	0,088	0,379	0,088	0,292
C09614[0.462]	0,913	0,292	0,699	0,624	0,14	0,155	0,1	0,478	0,1	0,36
C09751[0.443]	0,904	0,292	0,704	0,632	0,132	0,152	0,102	0,504	0,102	0,354
C09047[0.426]	0,376	0,411	0,332	0,322	0,119	0,141	0,077	0,6	0,077	0,248
C17673[0.426]	0,355	0,323	0,322	0,303	0,133	0,12	0,082	0,37	0,082	0,434
C15567[0.409]	0,538	0,286	0,486	0,454	0,12	0,146	0,077	0,382	0,077	0,311
C01263[0.350]	0,5	0,221	0,484	0,475	0,111	0,109	0,051	0,435	0,051	0,346
C01592[0.133]	0,469	0,366	0,343	0,3	0,126	0,144	0,136	0,23	0,136	0,221
C08578[0.110]	0,298	0,946	0,252	0,247	0,098	0,142	0,092	0,47	0,092	0,272

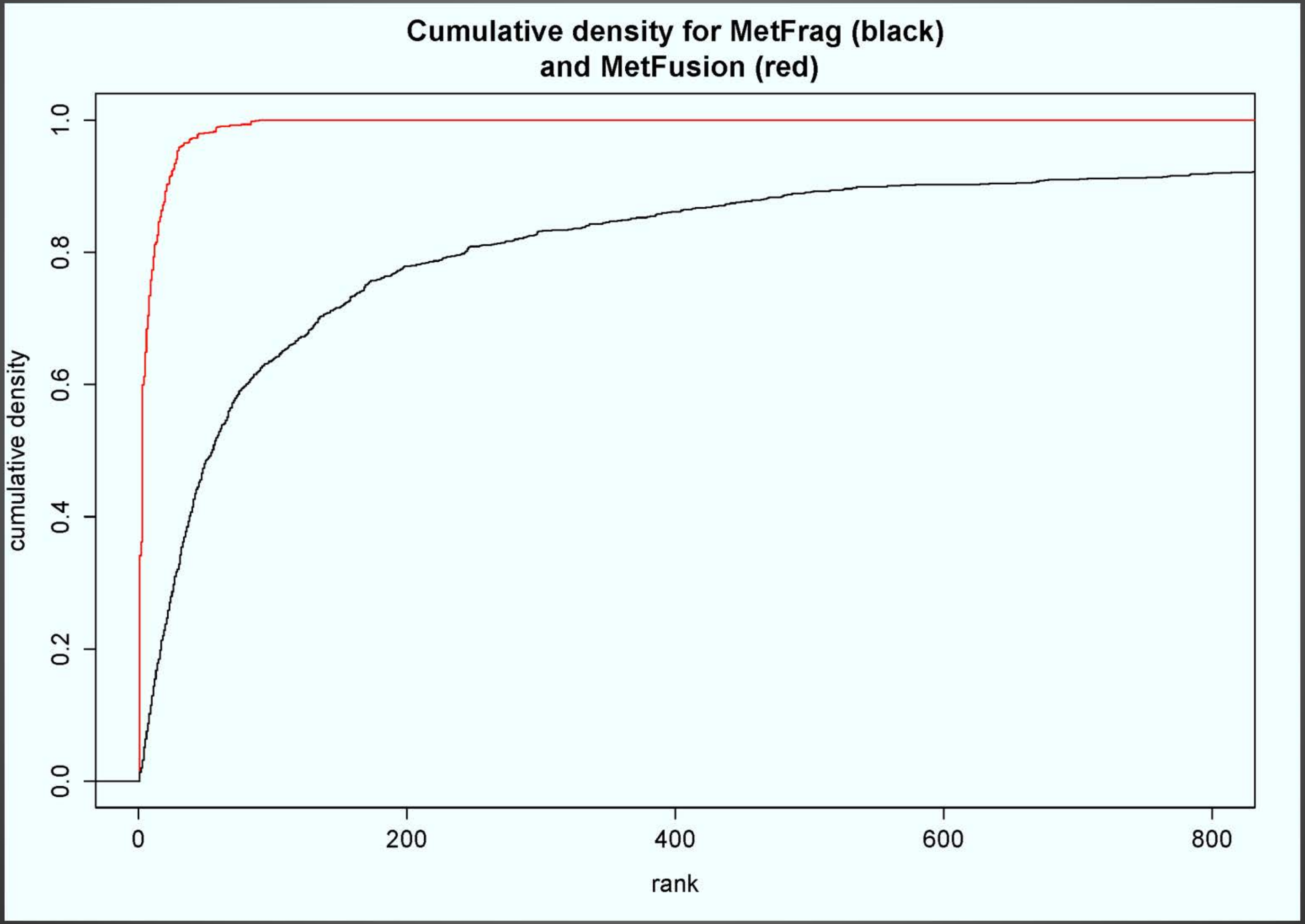
$$S_i = \alpha MF_i + (1-\alpha)\sum_{j=1}^{N} sig(MB_j * Tan_{i,j})$$

Results

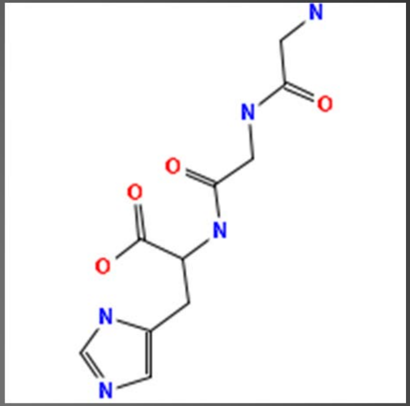
- Test data set: 1099 spectra
  - – Secondary metabolites, drugs, toxins, …
  - – 344 unique compounds
  - – Spectra from Hill et al., RIKEN & IPB
- Median rank of correct compound improved
  - – MetFrag: 28
  - – MetFusion: 7

➢ Works when informative spectra are present, but also when there is loss of information

MetFrag vs. MetFusion



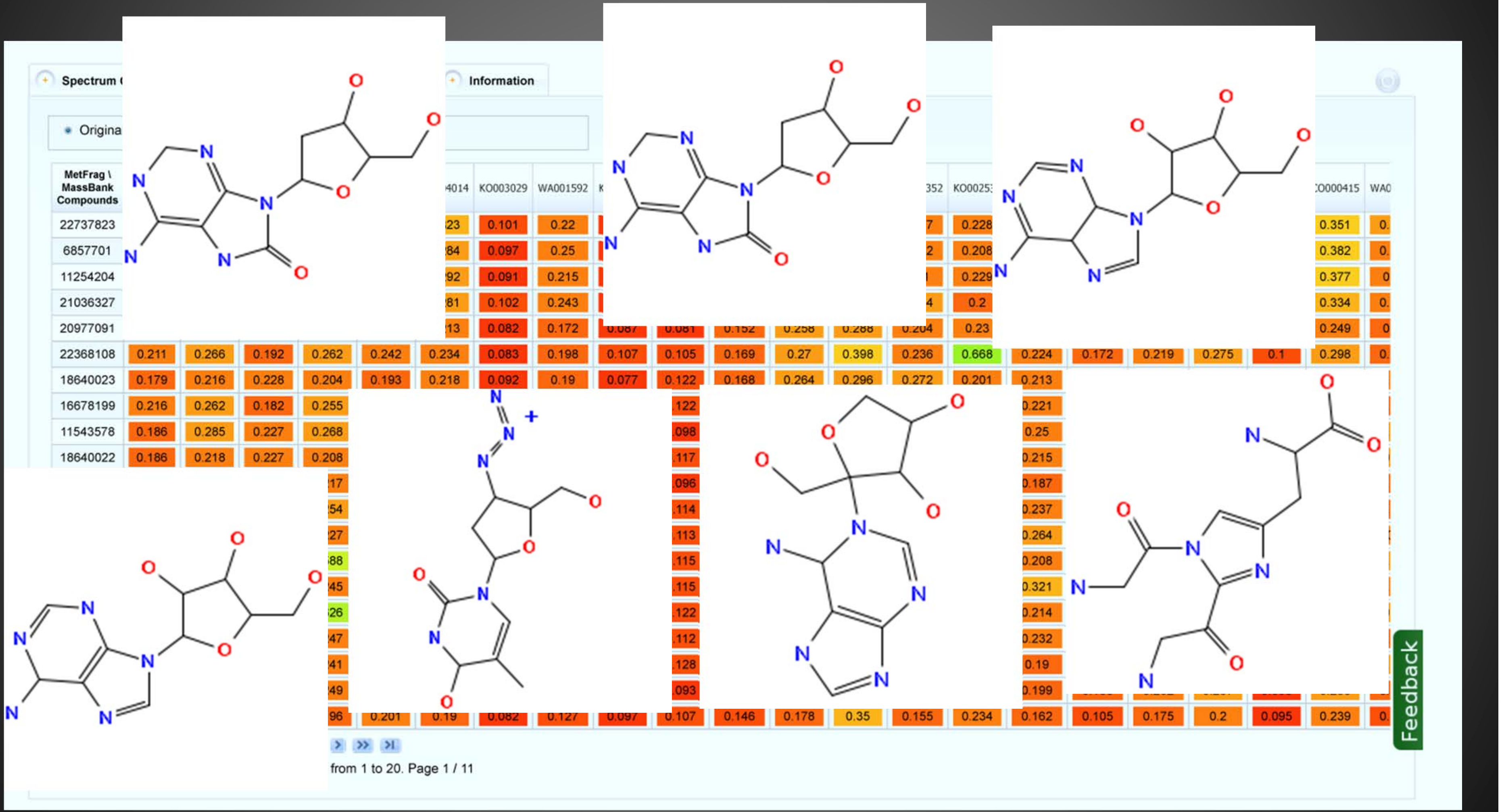
Example



- Tripeptide Gly-Gly-His
- 269.2572 Da
- QqQ spectrum from NIST
  - Nominal masses, modified MetFrag parameters
  - mzabs 0.1 da, mzppm 30ppm

- No tri- or polypeptides present in MassBank
  - But amino acids present

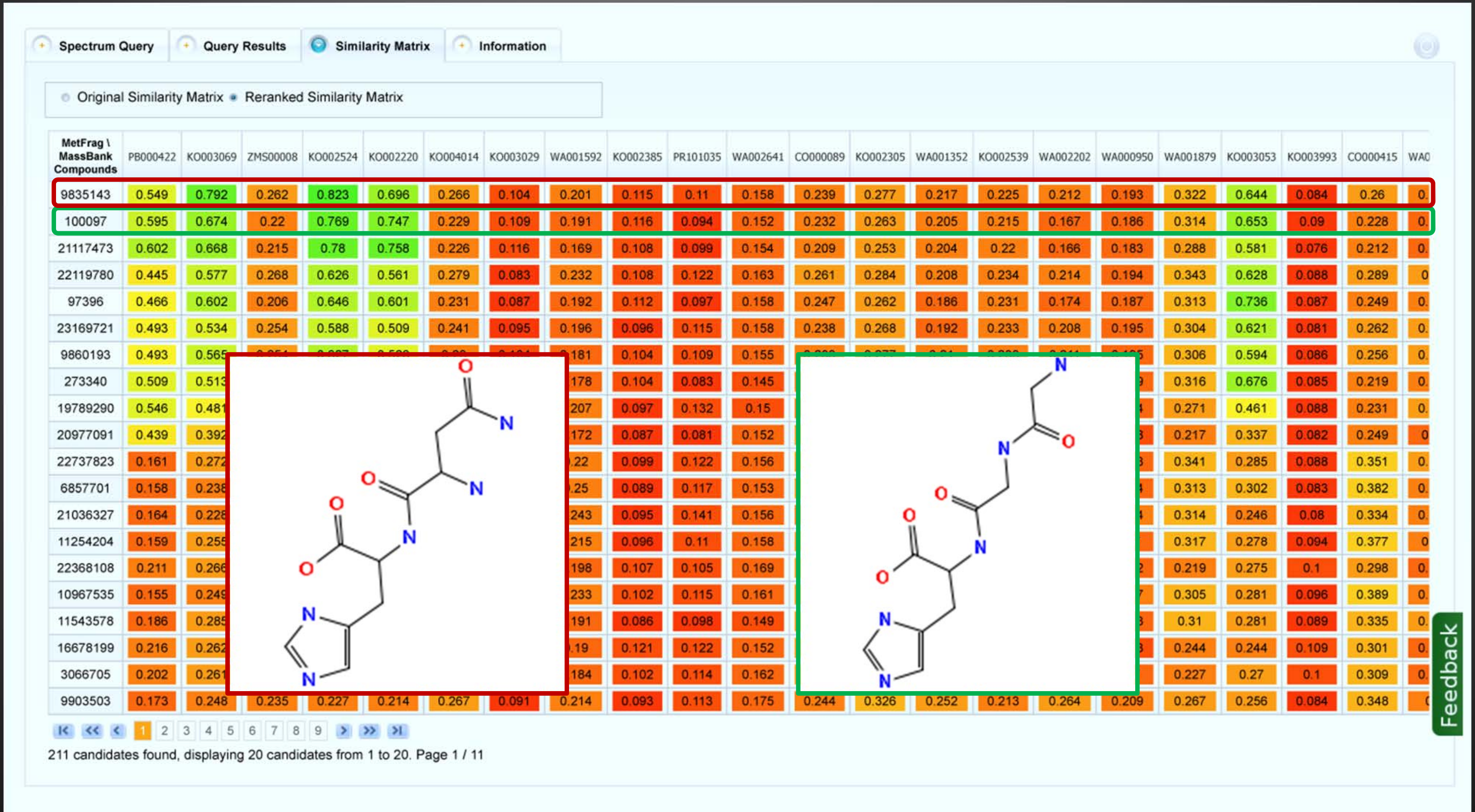
Original Similarity Matrix



MetFrag \ MassBank Compounds								KO003029	WA001592						KO0025							WA0
22737823								0.101	0.22						0.228						0.351	
6857701								0.097	0.25						0.208						0.382	
11254204								0.091	0.215						0.229						0.377	
21036327								0.102	0.243						0.2						0.334	
20977091								0.082	0.172						0.23						0.249	
22368108	0.211	0.266	0.192	0.262	0.242	0.234		0.083	0.198	0.107	0.105	0.169	0.27	0.398	0.236	0.668	0.224	0.172	0.219	0.275	0.1	0.298
18640023	0.179	0.216	0.228	0.204	0.193	0.218		0.092	0.19	0.077	0.122	0.168	0.264	0.296	0.272	0.201	0.213					
16678199	0.216	0.262	0.182	0.255																		
11543578	0.186	0.285	0.227	0.268																		
18640022	0.186	0.218	0.227	0.208																		
						0.201	0.19	0.082	0.127	0.097	0.107	0.146	0.178	0.35	0.155	0.234	0.162	0.105	0.175	0.2	0.095	0.239

from 1 to 20. Page 1 / 11

Re-ranked Similarity Matrix



○ Original Similarity Matrix ● Reranked Similarity Matrix

MetFrag \ MassBank Compounds	PB000422	KO003069	ZMS00008	KO002524	KO002220	KO004014	KO003029	WA001592	KO002385	PR101035	WA002641	CO000089	KO002305	WA001352	KO002539	WA002202	WA000950	WA001879	KO003053	KO003993	CO000415	WA0
9835143	0.549	0.792	0.262	0.823	0.696	0.266	0.104	0.201	0.115	0.11	0.158	0.239	0.277	0.217	0.225	0.212	0.193	0.322	0.644	0.084	0.26	0.
100097	0.595	0.674	0.22	0.769	0.747	0.229	0.109	0.191	0.116	0.094	0.152	0.232	0.263	0.205	0.215	0.167	0.186	0.314	0.653	0.09	0.228	0.
21117473	0.602	0.668	0.215	0.78	0.758	0.226	0.116	0.169	0.108	0.099	0.154	0.209	0.253	0.204	0.22	0.166	0.183	0.288	0.581	0.076	0.212	0.
22119780	0.445	0.577	0.268	0.626	0.561	0.279	0.083	0.232	0.108	0.122	0.163	0.261	0.284	0.208	0.234	0.214	0.194	0.343	0.628	0.088	0.289	0
97396	0.466	0.602	0.206	0.646	0.601	0.231	0.087	0.192	0.112	0.097	0.158	0.247	0.262	0.186	0.231	0.174	0.187	0.313	0.736	0.087	0.249	0.
23169721	0.493	0.534	0.254	0.588	0.509	0.241	0.095	0.196	0.096	0.115	0.158	0.238	0.268	0.192	0.233	0.208	0.195	0.304	0.621	0.081	0.262	0.
9860193	0.493	0.565	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.181	0.104	0.109	0.155	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.306	0.594	0.086	0.256	0.
273340	0.509	0.513	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.178	0.104	0.083	0.145	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.316	0.676	0.085	0.219	0.
19789290	0.546	0.481	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.207	0.097	0.132	0.15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.271	0.461	0.088	0.231	0.
20977091	0.439	0.392	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.172	0.087	0.081	0.152	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.217	0.337	0.082	0.249	0
22737823	0.161	0.272	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.22	0.099	0.122	0.156	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.341	0.285	0.088	0.351	0.
6857701	0.158	0.238	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.25	0.089	0.117	0.153	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.313	0.302	0.083	0.382	0.
21036327	0.164	0.228	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.243	0.095	0.141	0.156	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.314	0.246	0.08	0.334	0.
11254204	0.159	0.255	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.215	0.096	0.11	0.158	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.317	0.278	0.094	0.377	0
22368108	0.211	0.266	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.198	0.107	0.105	0.169	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.219	0.275	0.1	0.298	0.
10967535	0.155	0.249	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.233	0.102	0.115	0.161	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.305	0.281	0.096	0.389	0.
11543578	0.186	0.285	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.191	0.086	0.098	0.149	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.31	0.281	0.089	0.335	0.
16678199	0.216	0.262	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.19	0.121	0.122	0.152	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.244	0.244	0.109	0.301	0.
3066705	0.202	0.261	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.184	0.102	0.114	0.162	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.227	0.27	0.1	0.309	0.
9903503	0.173	0.248	0.235	0.227	0.214	0.267	0.091	0.214	0.093	0.113	0.175	0.244	0.326	0.252	0.213	0.264	0.209	0.267	0.256	0.084	0.348	[illegible]

1 2 3 4 5 6 7 8 9

211 candidates found, displaying 20 candidates from 1 to 20. Page 1 / 11

Recent additions

- MassBank alternatives
  - Metlin, 54.000 MS2 spectra
  - GMD, 8.800 GC-MS spectra with RI
- HMDB access pending

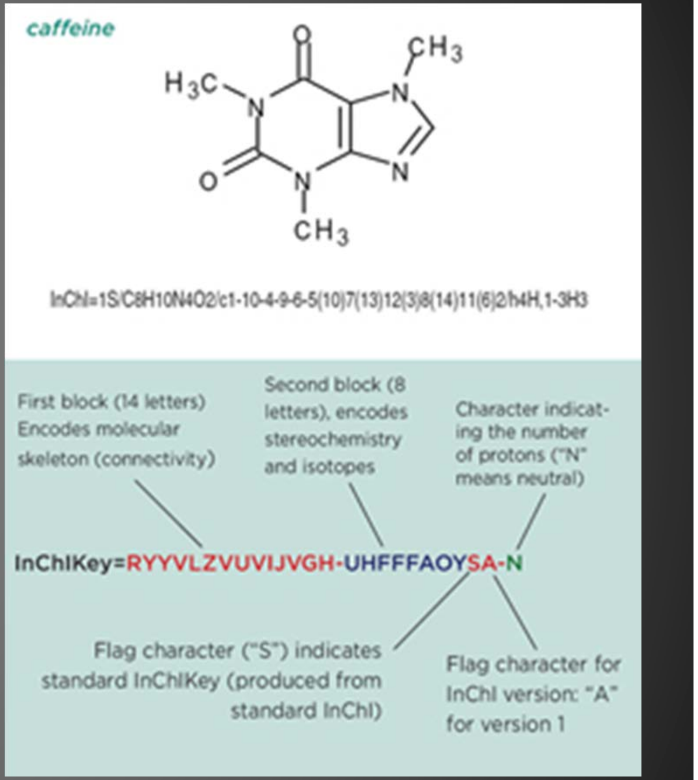
InChIKey-based filtering

- MetFrag candidate list often > 1000
- Lots of stereoisomers per candidate
- Results in clusters, valuable information scarce

- Use connectivity information from first part of InChIKey to retain only one representative per candidate
- Smaller list of candidates

InChIKey

- First part connectivity
- Second part stereochemistry



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

First block (14 letters) Encodes molecular skeleton (connectivity)

Second block (8 letters), encodes stereochemistry and isotopes

Character indicating the number of protons ("N" means neutral)

InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N

Flag character ("S") indicates standard InChIKey (produced from standard InChI)

Flag character for InChI version: "A" for version 1

- Image taken from http://www.iupac.org/publications/ci/2009/3105/iw6_inchi.htm

Summary

- Combine reference data & prediction
- Improves rank of correct compound
- Access multiple tools within one webpage
  - SDF & XLS export

- Available as web app
  http://msbi.ipb-halle.de/MetFusion/

Acknowledgements

- Dr. Steffen Neumann
- Dr. Emma Schymanski, EAWAG
- Members of MassBank consortium
- Sebastian Wolf, MetFrag developer
- Group members Carsten & Christoph

Thank you for your kind attention