CFM-ID: A Web Server for Annotation, Spectrum Prediction and Metabolite Identification from MS/MS

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

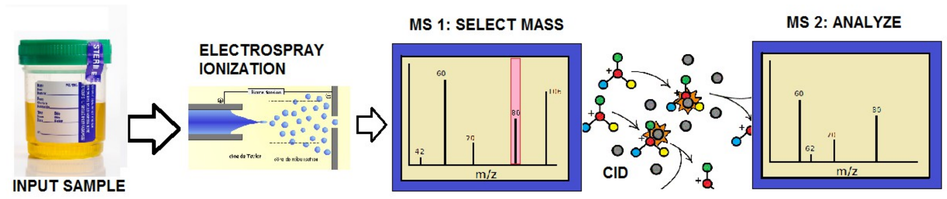
Felicity Allen, Allison Pon, Michael Wilson, Russ Greiner, David Wishart
Department of Computing Science, University of Alberta

Contact: felicity.allen@ualberta.ca

Summary

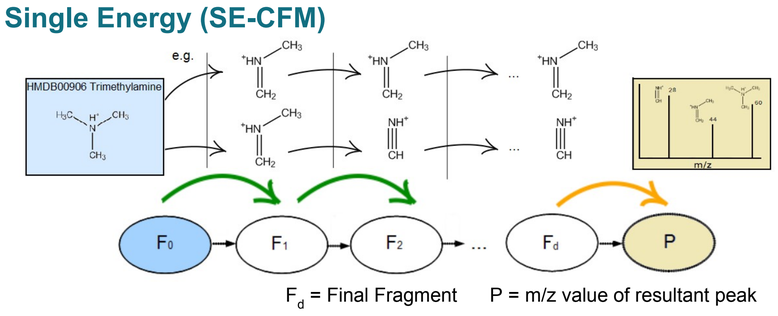
- **Goal:** Automated identification of metabolites from tandem mass spectra (MS/MS).
- **Existing Methods:**
  - Search against reference databases of measured spectra [3,4,5] – but limited coverage!
  - Enumerate all ways molecules could break[6,10], and/or make a heuristic selection of likely breaks[6,11,12] to predict spectra – usually predict far more peaks than actually occur.
- **Our approach:**
  - Design Competitive Fragmentation Modeling (CFM)[2], a model for Electrospray (ESI) MS/MS fragmentation. Derive parameters for CFM from MS/MS data.
- **CFM-ID:** A web server that uses CFM to provide three utilities associated with interpretation of MS/MS spectra:
  - Spectrum Prediction, Peak Assignment and Compound Identification.
- **Experimental Results:**
  - Spectrum Prediction: Better Jaccard scores vs full enumeration of possible peaks.
  - Compound Identification: Better ranking results vs existing methods MetFrag [6] and FingerID[7] querying KEGG[8] and PubChem[9] for possible candidates.

Competitive Fragmentation Modeling (CFM) [2]



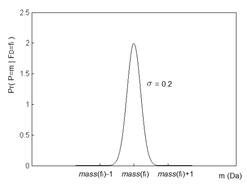
- Model ESI-MS/MS (above) fragmentation as a stochastic, homogeneous, Markov process of state transitions between charged fragments (below).

Single Energy (SE-CFM)



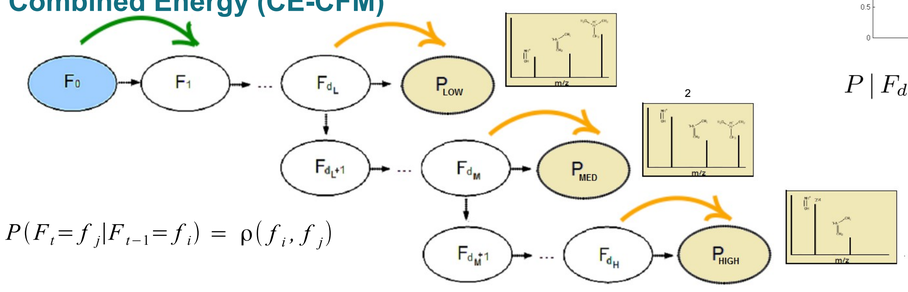
F_d = Final Fragment P = m/z value of resultant peak

- Observation model links F_d to P via Gaussian distribution.



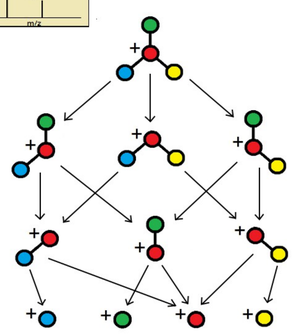
$$P \mid F_d \sim \mathcal{N}(mass(F_d), \sigma)$$

Combined Energy (CE-CFM)



$$P(F_t = f_j \mid F_{t-1} = f_i) = \rho(f_i, f_j)$$

- The initial molecule (F_0) and the output peak (P) are observed.
- All intermediate fragments (F_1..F_d) are latent.
- Possible transitions: Enumerate a graph of all possible fragmentations for each molecule (right), similar to [6,10].

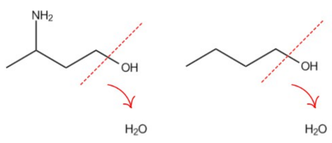


- Softmax transition function is competitive:

→ a particular break is likely to occur only if no other breaks are substantially more likely.

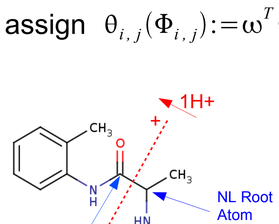
$$\rho(f_i, f_j) = \begin{cases} \frac{\exp \theta_{i,j}}{1 + \sum_k \exp \theta_{i,k}} & : f_i \neq f_j \text{ and } f_i \rightarrow f_j \text{ is possible} \\ \frac{1}{1 + \sum_k \exp \theta_{i,k}} & : f_i = f_j \\ 0 & : f_i \rightarrow f_j \text{ is not possible} \end{cases}$$

e.g. the NH_2 group in the left molecule reduces the likelihood of the H_20 loss.



- Given $\Phi_{i,j}$ = chemical features associated with break (f_i,f_j), assign $\theta_{i,j}(\Phi_{i,j}) := \omega^T \Phi_{i,j}$.

e.g. $\Phi_{i,j}$
- Break Pair:
  - C-C? true, C-N? false...etc
- Root paths (length 2 and 3):
  - C-N on ion side? true ...
- Gasteiger Charges of root atoms
- Hydrogen Movement
  - 1H+ moves from NL to Ion
- Ring Break Features
  - Size of ring, broken bond distance, aromatic ...

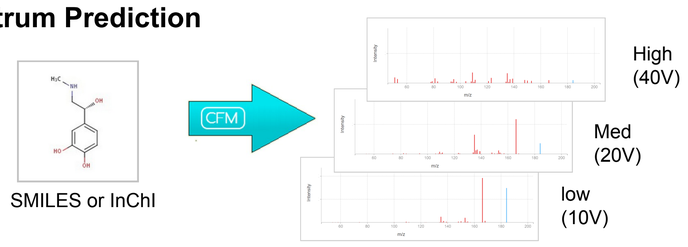


- Set model parameters ω using a maximum likelihood approach applied to a training set using the Expectation Maximization (EM) algorithm.

CFM-ID Web Server [1]

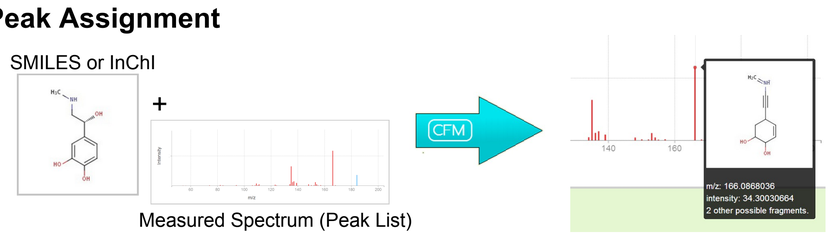
- Supports three sub-tasks for automated metabolite identification from MS/MS data:

Spectrum Prediction



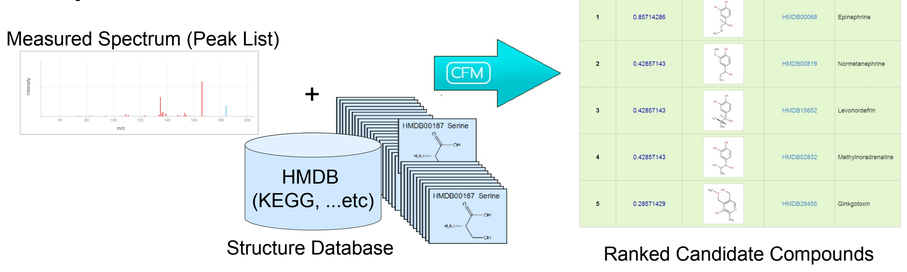
- Runs trained CFM model forward to predict spectra for low, medium and high collision energies.

Peak Assignment



- Assigns fragments within mass tolerance of each peak.
- Orders fragments according to CFM likelihoods.

Compound Identification



- Predicts spectra for all candidate compounds.
- Ranks compounds by Jaccard Score between measured and predicted spectra.

- Over 300,000 precomputed spectra for compounds in HMDB[5] and KEGG[8]!!

Available free at http://cfmid.wishartlab.com

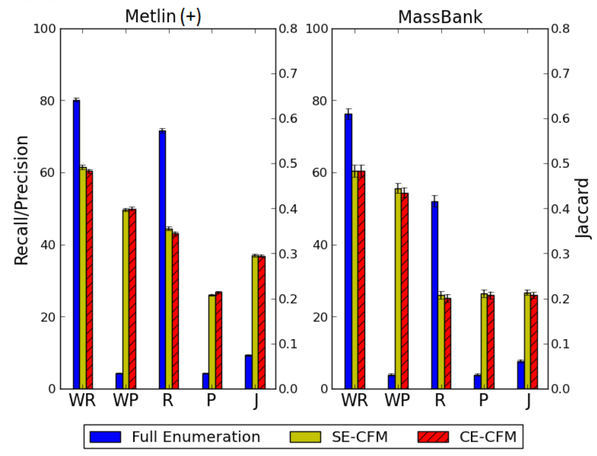
Experimental Validation

Data Sets:

Data Set	# Mols	Mode	Device	Energies
Metlin (+)	1491	+	Agilent 6510 Q-TOF	10V, 20V, 40V
MassBank	192	+	Agilent 6520 Q-TOF	10V, 20V, 40V
HMDB	500	+	Quattro QqQ	10V, 25V, 40V
Metlin (-)	976	-	Agilent 6510 Q-TOF	10V, 20V, 40V

- Metlin[3] tests used a 10-fold cross validation framework. MassBank[4] and HMDB[5] tests used a model trained on the Metlin data.

Spectrum Prediction:
- Compare vs full enumeration of all possible fragments (right).
- Low energy (10V) spectra better predicted.
- Positive though imperfect correlation between measured and predicted intensity values - Pearson correlations of 0.7 (10V), 0.6 (20V) and 0.45 (40V).



(R)ecall: Percent of measured peaks predicted.
(P)recision: Percent of predicted peaks measured.
(WR) and (WP): Recall and Precision weighted by peak intensity.
(J)accard: Predicted (A) vs measured (B) peaks | A ∩ B | / | A U B|

Compound Identification:
- Query KEGG[8] and PubChem[9] for candidates within tolerance of the known mass of the target.
- Compare against other methods (below).

	Data Set	Querying KEGG (±0.5 Da, # cand. ≈ 22) R = 1	R ≤ 5	MF = 1	Querying PubChem (±5 ppm, # cand. ≈ 1025) R = 1	R ≤ 10	MF = 1
CFM-ID	Metlin (+)	76.5	96.2	94.8	10.9	40.7	89.3
	MassBank	72.8	97.5	97.5	7.3	46.9	93.2
	HMDB	23.1	58.1	39.0	4.1	24.9	88.4
	Metlin (-)	72.1	96.5	95.2	13.4	51.4	93.8
MetFrag	Metlin (+)	51.9	89.9	72.2	5.7	30.5	82.6
	MassBank	48.1	88.9	71.6	4.7	20.8	85.4
	HMDB	13.3	43.6	28.3	2.6	13.4	88.0
	Metlin (-)	44.7	80.7	62.3	7.5	28.8	81.8
FingerID	Metlin (+)	8.7	36.1	17.0	1.3	9.3	67.7
	MassBank	14.8	37.0	19.8	0.5	5.7	71.9

Values are % of data set (restricted to those with correct structure in queried candidate list).
R : Ranking of the correct molecule in the candidate list
MF : Ranking of the correct molecular formula
cand. ≈ N : The median number of molecules in the candidate list

- **MetFrag**[6]: Combinatorial method with heuristics to rank likely breaks.
- **FingerID**[7]: uses SVMs to predict each bit of a molecular fingerprint, then searches using that fingerprint.

Support: This work was supported by Alberta Innovates Technology Futures, Natural Sciences and Engineering Research Council of Canada, Canadian Institute of Health Research and The Metabolomics Innovation Centre supported by Genome Alberta and Genome Canada, and made possible by the Compute Canada Westgrid Facility.

References

[1] F. Allen, *et al.* "CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra", Nucleic Acids Research, Web Server Edition 2014.
[2] F. Allen, R. Greiner, D. Wishart, "Competitive Fragmentation Modeling of ESI-MS/MS spectra for putative metabolite identification", Metabolomics, 10(3): 2014.
[3] C. Smith, *et al.* "METLIN: a metabolite mass spectral database". Therapeutic drug monitoring, 27(6):747-51, December 2005.
[4] H. Horai, *et al.* "MassBank: a public repository for sharing mass spectral data for life sciences". J. of Mass Spectrometry 45(7):703-14, 2010.
[5] D. Wishart, *et al.* "HMDB: A knowledge base for the human metabalome", Nucleic Acids Resarch, 37:D603-610, 2009.
[6] S. Wolf, *et al.* "In silico fragmentation for computer assisted identication of metabolite mass spectra". BMC bioinformatics, 11:148, January 2010.
[7] M. Heinonen, *et al.* "Metabolite identication and molecular fingerprint prediction through machine learning". Bioinformatics, 28(18):2333-41, September 2012.
[8] M. Kanehisa, *et al.* "From genomics to chemical genomics: new developments in KEGG". Nucleic Acids Research 34:D354-7, 2006.
[9] E Bolton, *et al.* "PubChem: Integrated Platform of Small Molecules and Biological Activities". Chapeter 12 in Annual Reports in Computational Chemistry, 4, 2008.
[10] M. Heinonen *et al.* "FiD : a software for ab initio structural identication of product ions from tandem mass spectrometric data". p3043-52, 2008.
[11] ACD Labs. ACD/MS Fragmenter. http://www.acdlabs.com/products
[12] Thermo Scientic. Mass Frontier Software.