Metabolite Identification via Machine Learning

Huibin Shen

Department of Information and Computer Science
Aalto University

February 7, 2013

Outline

General picture

What is the metabolites identification?

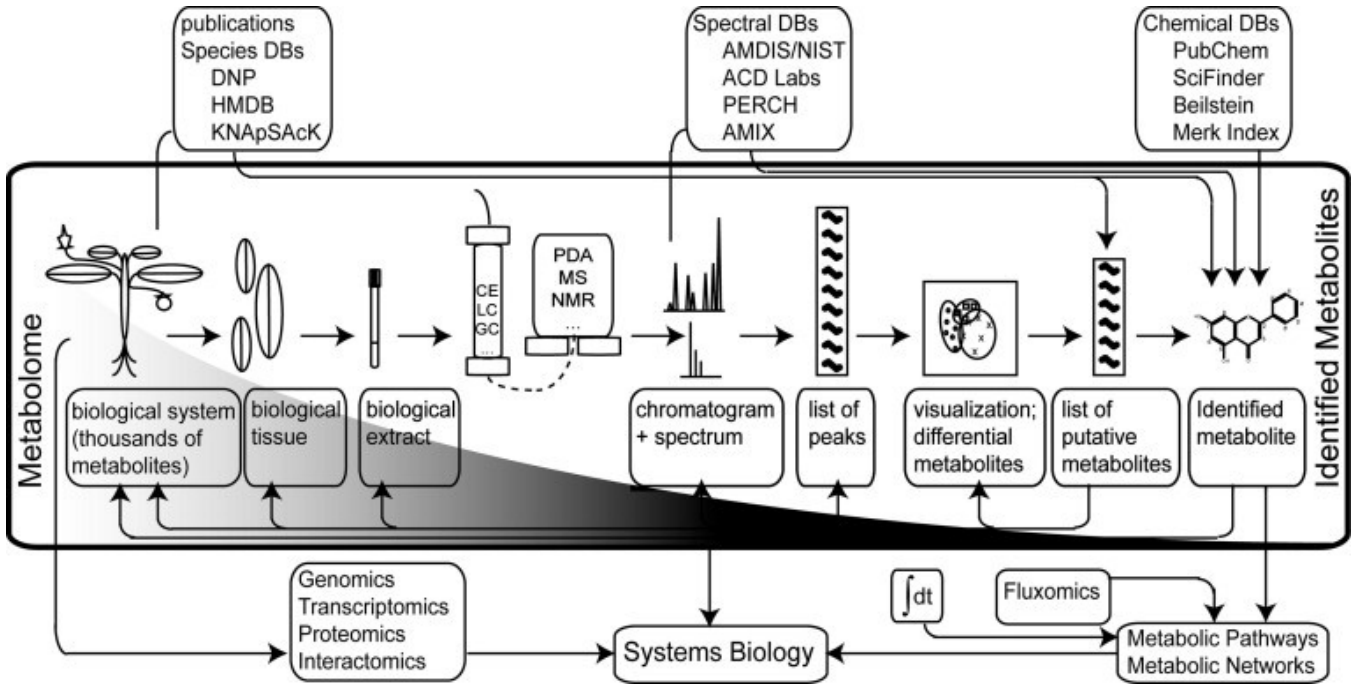

Figure 1: Metabolomics pipeline towards a systems biology approach: from the whole metabolome to identified metabolites [M. Sofia, 2007].

Standard computational method

Matching reference spectral database.

Problems:

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.
- Seldom public.

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.
- Seldom public.
- Limited number.

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.
- Seldom public.
- Limited number.
- Diversity of mass spectrometer.

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.
- Seldom public.
- Limited number.
- Diversity of mass spectrometer.
- Similarity definition.

Standard computational method

Matching reference spectral database.

Problems:

- Quality of data.
- Seldom public.
- Limited number.
- Diversity of mass spectrometer.
- Similarity definition.

Molecular fingerprint

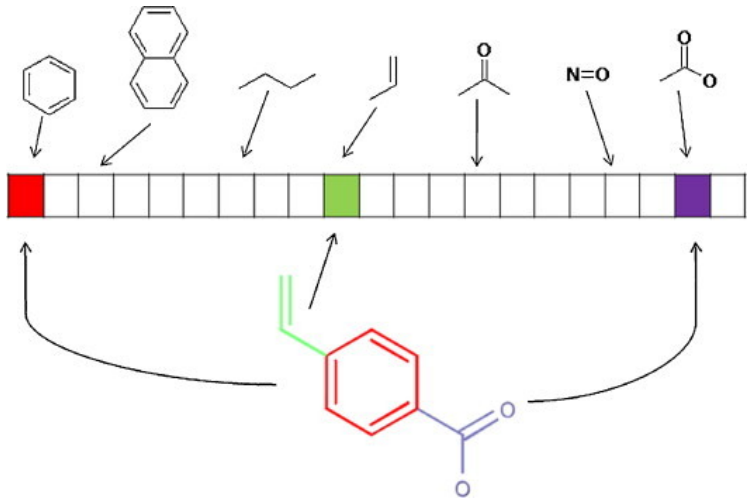

Figure 2: Representation of a molecular substructure fingerprint with a substructure fingerprint dictionary of given substructure patterns. This molecule is represented in a series of binary bits that represent the presence or absence of particular substructures in the molecules [D.S. Cao, 2012].

Machine learning method

We propose a new framework to identify metabolites through machine learning:

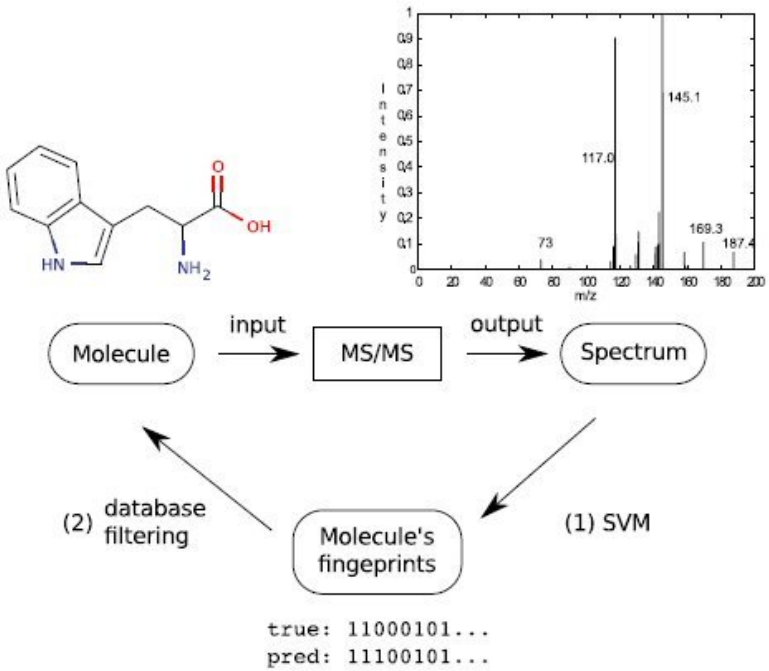

Figure 3: The overview of the two-step metabolite identification framework.

Support Vector Machine (SVM)

SVM, a supervised machine learning method for classification and regression.

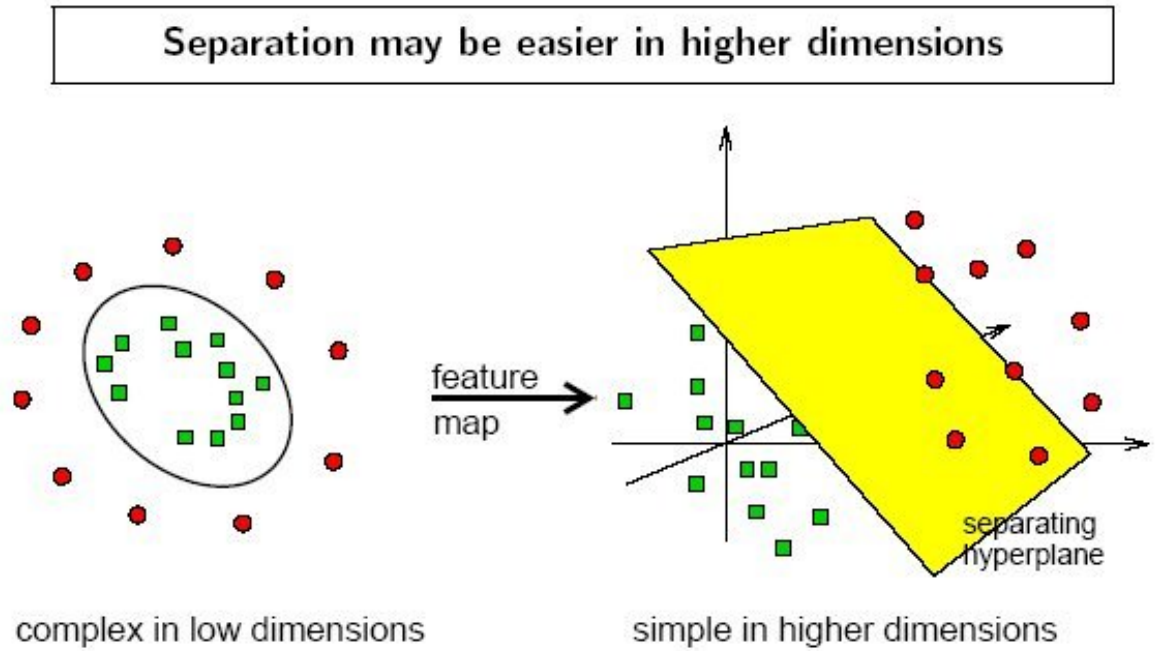

Figure 4: Three dimensional case for SVM[1].

[1] Figure from http://www.dtreg.com/svm.htm

kernels for mass spectrum

- Feature mapping $\approx$ kernel function.

kernels for mass spectrum

- Feature mapping $\approx$ kernel function.
- Three basic features and their combination.

kernels for mass spectrum

- Feature mapping $\approx$ kernel function.
- Three basic features and their combination.
- Two families of kernels: integral mass kernel and probability product kernel.

Integral mass kernels

$$k(x, x') = \langle \mathbf{x}, \mathbf{x}' \rangle$$

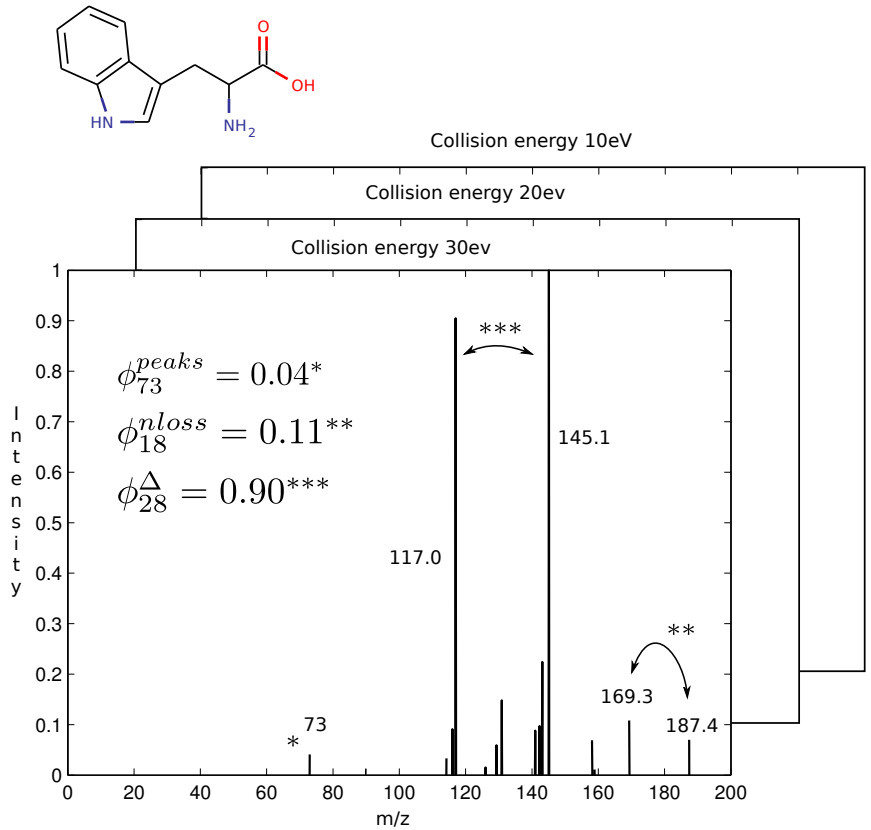

Figure 5: Three basic features and integral mass kernel.

Probability product kernel

$$k(x, x') = k^{prob}(p(x), p'(x')) = \int_{\mathcal{X}} p(\mathbf{x})p'(\mathbf{x}')dx$$

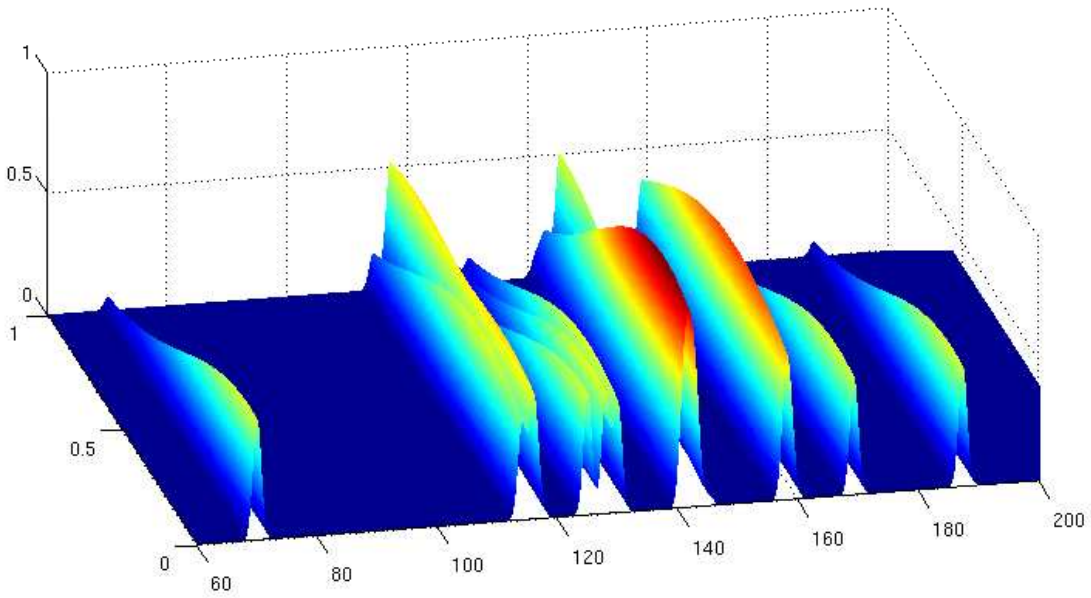

Figure 6: Probability product kernel.

Scoring

Given the cross validation accuracy $\mathbf{p} = (p_i)_{i=1}^m \in \mathbb{R}^m$ over m fingerprints $\mathbf{y} = (y_i)_{i=1}^m$. The similarity score between two fingerprints $\mathbf{y}$ and $\mathbf{y}^*$ is:

$$p(\mathbf{y}|\mathbf{p}, \mathbf{y}^*) = \prod_{i=1}^{m} p_i^{1-|y_i - y_i^*|}(1 - p_i)^{|y_i - y_i^*|}.$$

Experiments data

A summary of the datasets is listed in this table

Data	Device	Size	Mode	Mass error	Std	Fingerprints
QqQ	*misc*	514	Pos			286
	- API3000	410	Pos	0.128	0.164	
	- QuattroPremier XE	82	Pos	-0.092	0.073	
	- TSQ 7000	17	Pos	-0.124	0.036	
	- TSQ Quantum AM	3	Pos			
	- Q-Trap	2	Pos			
Ltq	LTQ Orbitrap XL	293	Pos	0.0	0.049	128
Lipids	LTQ Orbitrap	403	Neg	-0.135	0.090	20

Table 1: The dataset statistics. Only a subset of fingerprints are exhibited in each dataset's molecules.

Fingerprint prediction

We show the predication accuracies for *ltq* dataset.

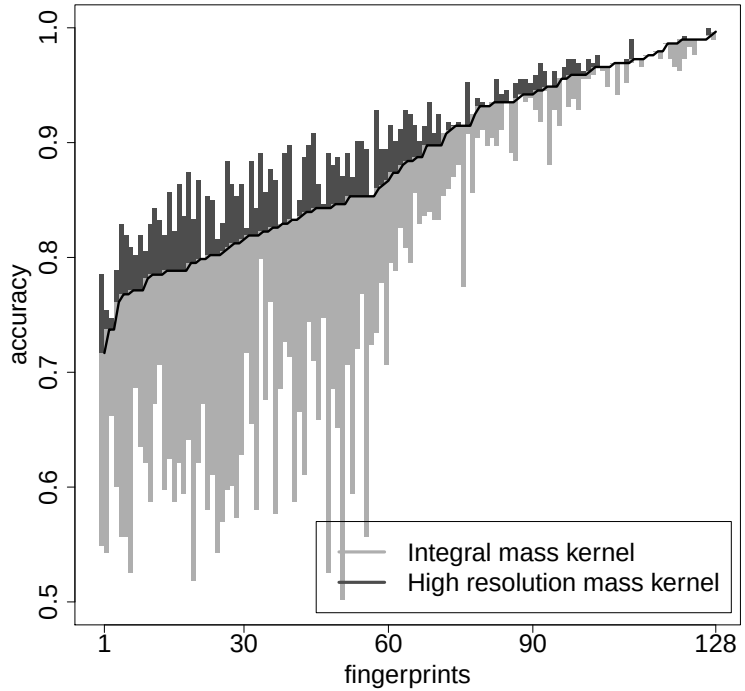

Figure 7: Light grey is improvement by integral kernel from default classifier. Dark grey is improvement by product probability kernel from integral kernel.

Feature selection

We show the effect of different features.

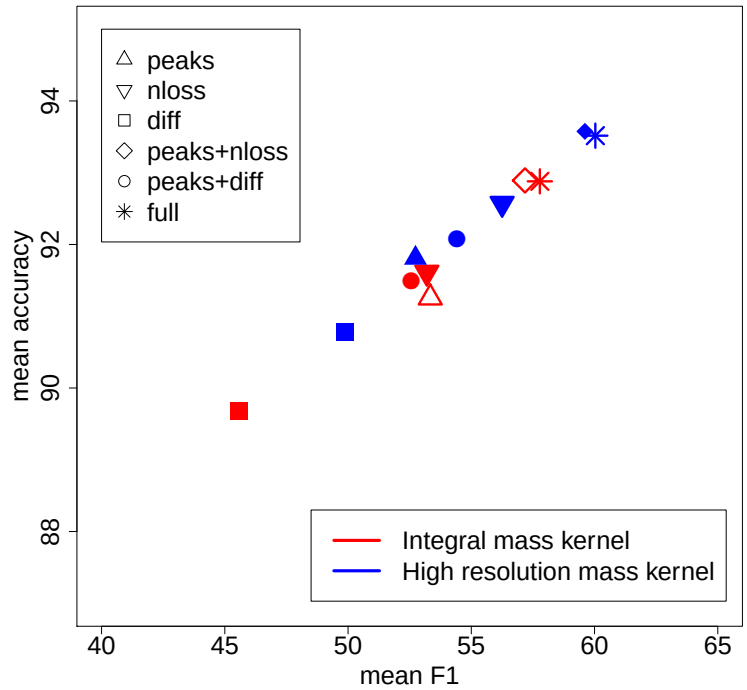

Figure 8: Scatter plot of the aggregate average accuracy/F_1 across three datasets. The non-filled marks represent higher accuracy/F_1 ratio in quadratic kernel.

Experiments data (for CASMI challenge)

MS2 spectra are used to train the model and MS1 spectra are used for comparing the result of isotopic patterns matching.

MS type	Instument type	Size	No. of Mol	Fingerprints
MS2	APCI-ITFT-CID	295	65	179
	APCI-ITFT-HCD	882	86	181
	LC-ESI-ITFT-CID	447	244	281
	LC-ESI-ITFT-HCD	2655	225	281
	LC-ESI-QTOF-CID	1027	523	290
MS1	LC-ESI-ITFT	41	41	
	LC-ESI-QTOF	62	62	

Table 2: The dataset statistics. Only a subset of fingerprints are exhibited in each dataset's molecules.

Learning curve

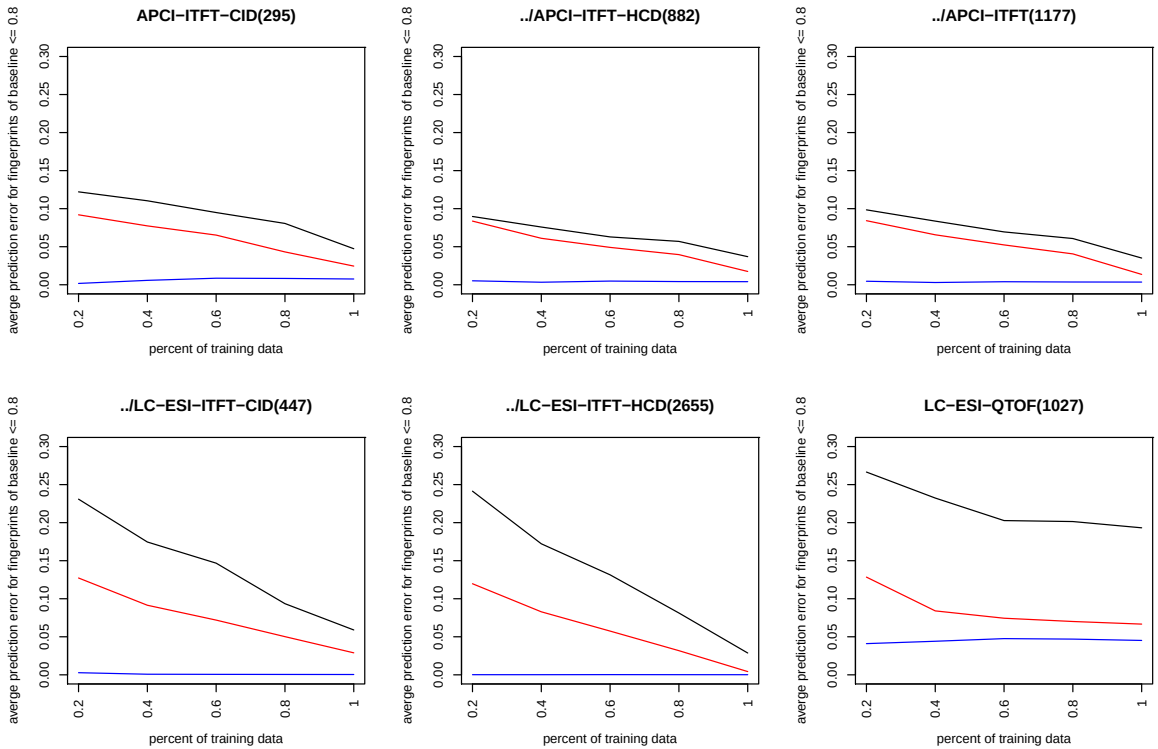

Figure 9: Blue line is training error; red line is cross validation prediction error; black line is the relative rank of the correct molecule. Matching database is Kegg.

Combine isotopic patterns matching (LC-ESI-ITFT)

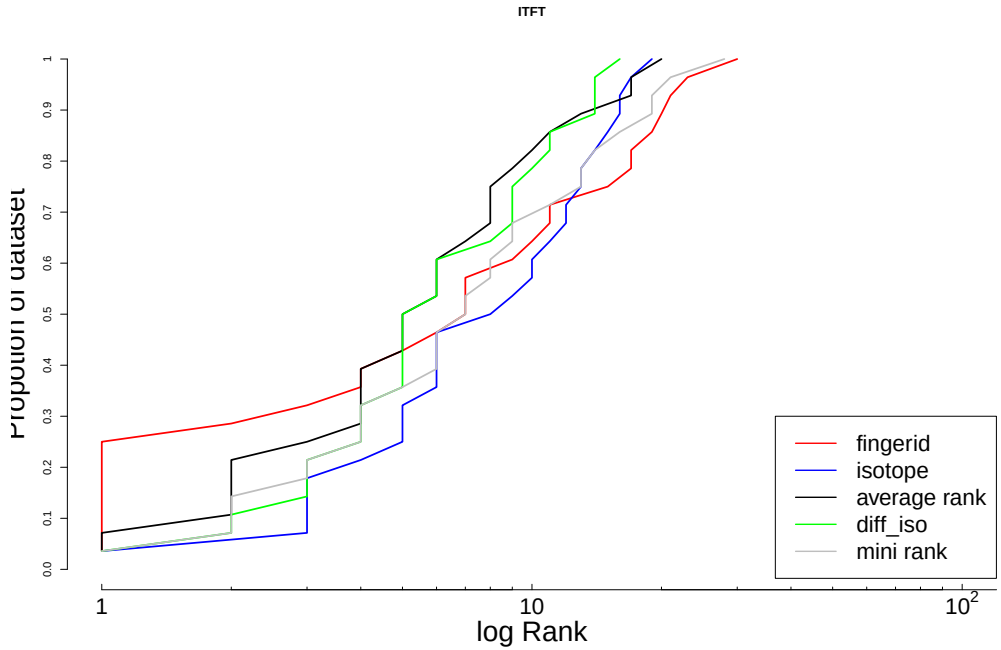

Figure 10: The fingerid red line rank the candidates only by fingerid scores while the isotope blue line rank by isotopic patterns matching scores. The average rank black line take the average of fingerid and isotopic matching while the diff_iso green line rank by isotopic matching scores first and then for the ones having the tie, rank them by fingerid scores. The mini rank grey line rank by taking the minimum rank of the fingerid and isotopic matching.

Combine isotopic patterns matching (LC-ESI-QTOF)

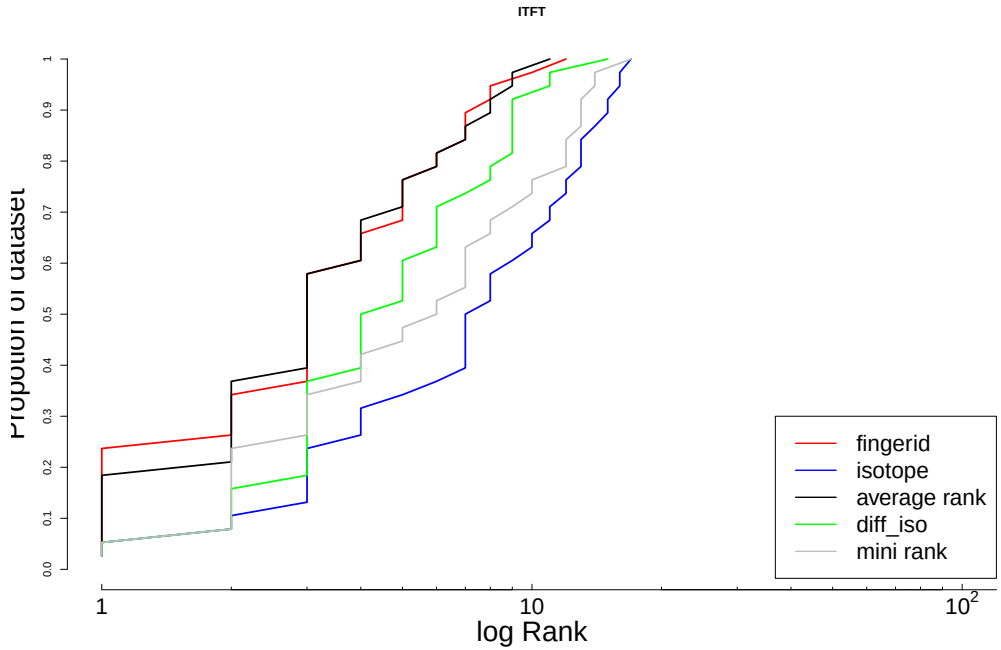

Figure 11: The fingerid red line rank the candidates only by fingerid scores while the isotope blue line rank by isotopic patterns matching scores. The average rank black line take the average of fingerid and isotopic matching while the diff_iso green line rank by isotopic matching scores first and then for the ones having the tie, rank them by fingerid scores. The mini rank grey line rank by taking the minimum rank of the fingerid and isotopic matching.

Result for CASMI challenge

Challenge	1	2	3	4	5	6	10	11	12	13	14	15	16	17
Category1	4	1		3	4	4	1				1	5		
Category2	5	1				4	11							

Table 3: The result for CASMI-2012 challenge. Category 1 is chemical formula identification and category 2 is molecule structure identification. For Challenge 11, we deducted the wrong exact mass of the molecule. For the others, we don't have that molecule in our database (most molecules in Kegg).

Conclusion

- Predicting the fingerprints with high accuracy. Product probability kernels and combined features are better in most cases.
- Isotopic patterns matching does not help a lot.
- Choosing the right molecular database can be a critical problem in our framework.