Metabolite identification and molecular fingerprint prediction

via machine learning

Markus Heinonen[1,3], Huibin Shen[1], Nicola Zamboni[4], Juho Rousu[2,3]

(1) Department of Computer Science, University of Helsinki
(2) Department of Information and Computer Science, Aalto University
(3) Helsinki Institute for Information Technology
(4) Institute of Molecular Systems Biology, ETH Zurich

September 9, 2012

Aalto University

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Outline

1. Motivation
   - Metabolite identification
   - Mass spectrometry
2. Kernel framework
   - Mass kernels
   - Poisson-Binomial model
3. Experiments
   - SVM performance
   - Metabolite matching

Summary

- We present a "FingerID"[1] machine learning framework for metabolite identification using tandem mass spectral data
  1. We introduce novel kernels for mass spectra for prediction of intermediate binary metabolite properties
  2. We introduce a statistical model to search metabolites with matching properties

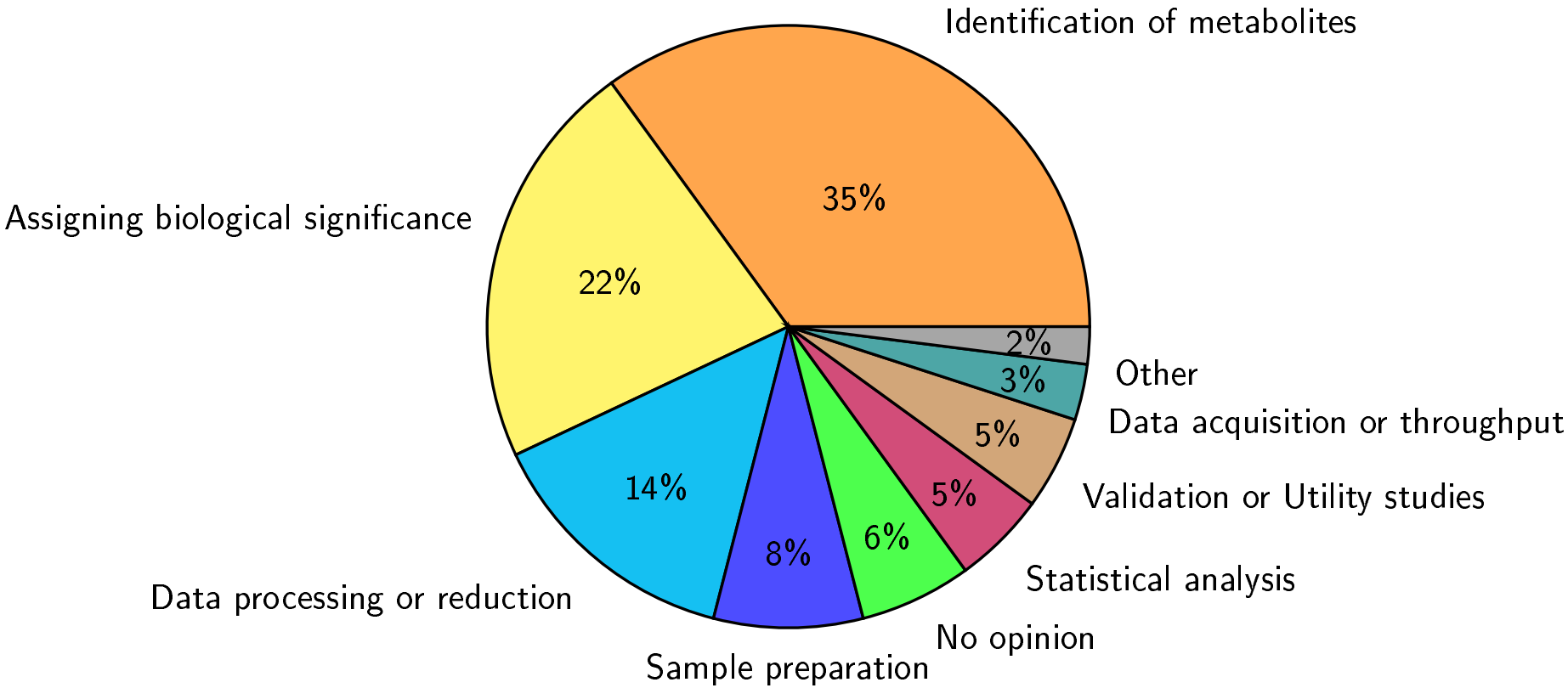
[1] sourceforge.net/p/fingerid

Contents

Metabolomics bottlenecks

- At the American Society for Mass Spectrometry (ASMS) conference 2009, a survey among the 600 participants asked [http://metabolomicssurvey.com]:

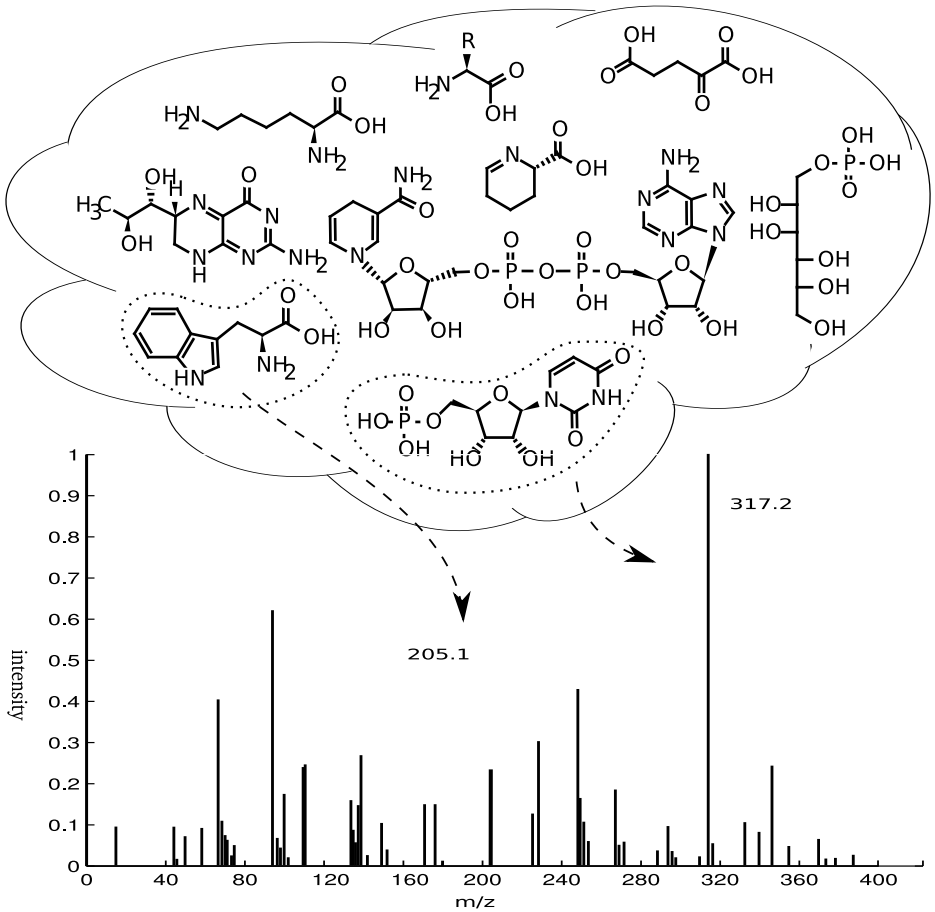
  "From your perspective, what is the biggest bottleneck in metabolomics today?"



Metabolite identification

- Determination of the metabolic contents of the cell
- Requirement for further metabolomic analysis
- Mass spectrometry
  - Offers a "wide" view on the cell contents
  - Reveals only mass-to-charges (m/z), not structures
  - Average measurement error ε: true mass in range $[m - \varepsilon, m + \varepsilon]$

[Kind & Fiehn 2006: *Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm*]

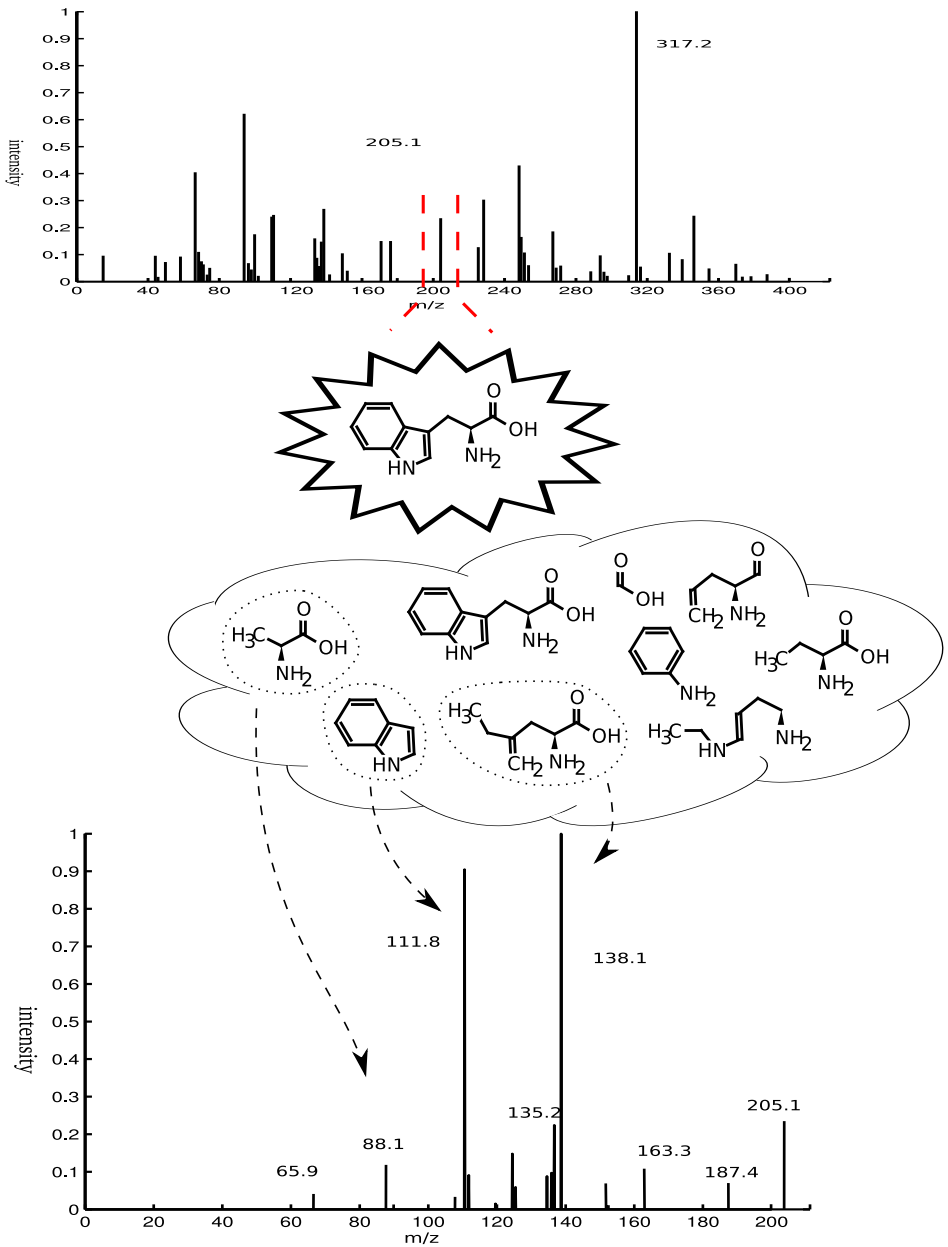


Tandem mass spectrometry (MS/MS)

- Filter a single unknown compound by mass
  - Fragment the compound by high-energy collision into sub-structures called *fragments*
  - Measure the m/z of the *fragments*
- Each molecule produces a 'unique' set of fragments, and hence peaks
- The collision energy can be varied to produce more or less fragmented products
- $\Rightarrow$ structural information

Data:

- The mass of the unknown metabolite (*precursor mass*)
- A list of (m/z,int) pairs of the fragments of the unknown metabolite



Current metabolite identification methods

Reference databases: Given an MS/MS spectrum of an unknown metabolite, search matching spectra from reference databases [Wiley, NIST, MassBank]

- Fails if the spectrum is not in the database, or if the measurement conditions/energies differ too much

Simulation: Simulate the fragmentation of candidate metabolites and match the observed spectrum against the simulated *in silico* spectra

- MetFrag software: exhaustively cleave the bonds to produce possible fragments

Machine learning: Use the MS/MS peaks as a characterizing pattern to predict the structure of the metabolite

- No need for databases or simulation of the fragmentation process

Contents

Machine learning problem

- Given a MS/MS spectrum measurement $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \in \mathcal{X}$ as a collection of peaks $\mathbf{x} = (mass, intensity)^T$ with average mass error ε, predict the measured unknown metabolite (a labeled graph) $M \in \mathcal{M}$
  - $\Rightarrow$ A structured prediction problem from sets to graphs

$$f : \mathcal{X} \to \mathcal{M}$$

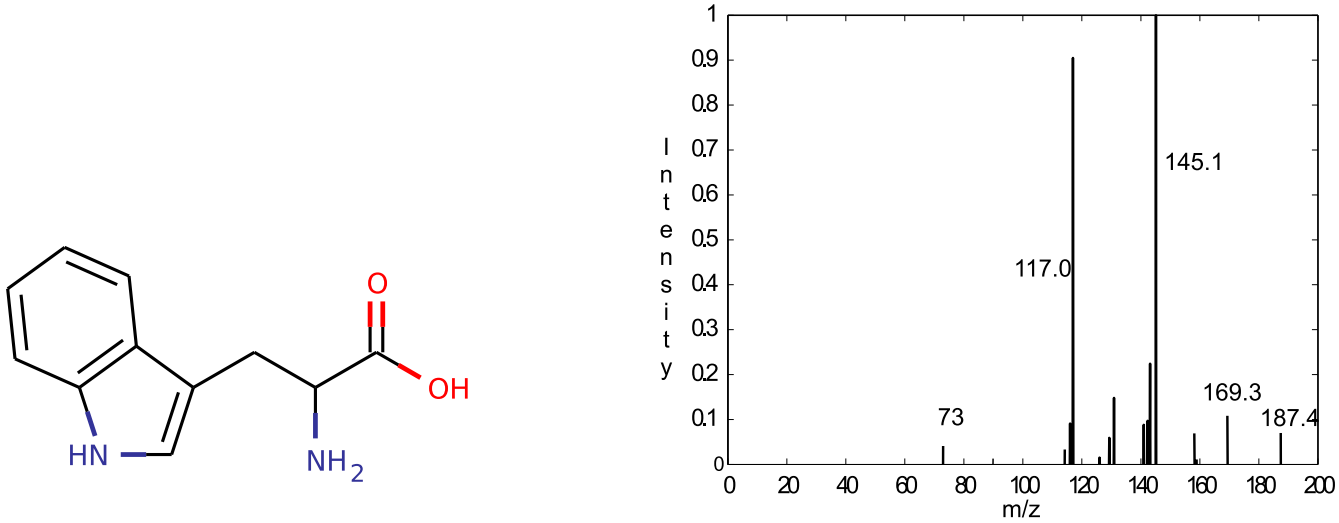
- We opt for a two-phase scheme instead

1. An intermediate prediction target: a vector of m binary and independent structural properties ("fingerprints") $\mathbf{y} = (y_i)_{i=1}^m$, which characterizes the unknown metabolite structure
   - $\Rightarrow$ A set of standard binary prediction problems (we use SVM's)

$$f_i : \mathcal{X} \to \{0, 1\}^m \qquad i = 1, \ldots, m$$

2. Reconstruct M from fingerprints: We introduce a statistical model to find matching metabolite candidate's based on the predicted property vector $\hat{y}$

Overview of the framework




```
true: 11000101...
pred: 11100101...
```

Fingerprints

- We use 528 structural fingerprints as a prediction targets
- Generated from OpenBabel's FP3, FP4 and MACCS fingerprint sets
- The fingerprints should be predictable from MS/MS data, and be informative regarding the metabolite structure

SMILES	Interpretation
('[N,n]~[C,c](~[O,o])~[N,n]',0)	NC(O)N
('[N,n]~[C,c](~[C,c])~[N,n]',0)	NC(C)N
('[O,o]~[S,s](~[O,o])~[O,o]',0)	OS(O)O
('[C,c]-[O,o]',0)	C-O
('[C,c]-[N,n]',0)	C-N
[+]	cation
[CX3H1](=O)[\#6]	aldehyde
[\#6][CX3](=O)[\#6]	ketone
[\#6][CX3](=[SX1])[\#6]	Thioketone
[SX2H][c]	Arylthiol
...	...

Mass spectral kernels

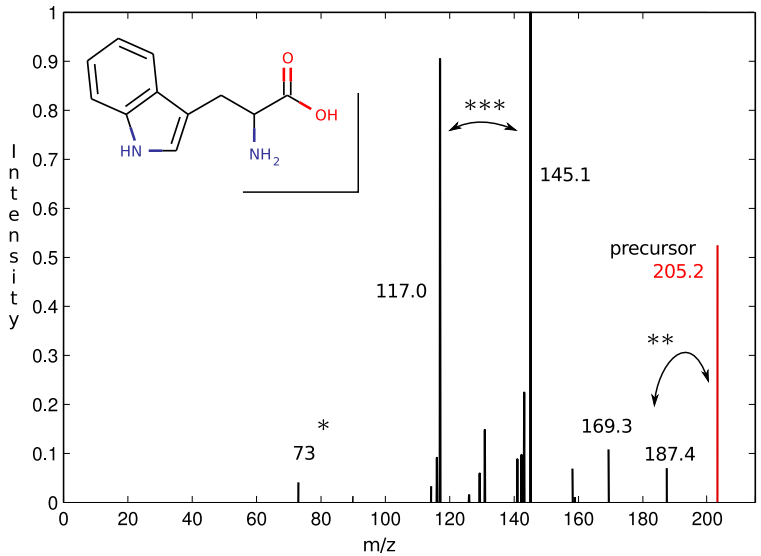
- We introduce kernels for mass spectral data $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$
- We extract three classes of features from MS/MS spectra into sparse vectors with 'bins' of fixed width of 1

$$\phi_{peaks}(\chi)_i = \sum_{(mass,int)\in\chi} \delta_{i\pm 0.5}(mass) \cdot int \qquad i = 1, 2, 3, \ldots$$

$$\phi_{nloss}(\chi)_i = \sum_{(mass,int)\in\chi} \delta_{i\pm 0.5}(prec(\chi) - mass) \cdot int$$

$$\phi_{diff}(\chi)_i = \sum_{\substack{(mass,int)\in\chi \\ (mass',int')\in\chi}} \delta_{i\pm 0.5}(|mass - mass'|) \cdot int \cdot int'$$

where δ is an indicator function



$$\phi_{peaks}(\chi)_{73} = 0.04^{*}$$

$$\phi_{nloss}(\chi)_{18} = 0.11^{**}$$

$$\phi_{diff}(\chi)_{28} = 1.0 * 0.90 = 0.90^{***}$$

Integral mass kernel

- The *integral mass kernels* are

$$K_{peaks}(\chi, \chi') = \langle \phi_{peaks}(\chi, \chi') \rangle$$
$$K_{nloss}(\chi, \chi') = \langle \phi_{nloss}(\chi, \chi') \rangle$$
$$K_{diff}(\chi, \chi') = \langle \phi_{diff}(\chi, \chi') \rangle$$

  A summed kernel

$$K_{full} = K_{peaks} + K_{nloss} + K_{diff}$$

  correspond to a concatenation of the feature sets

$$[\phi_{peaks}; \phi_{nloss}; \phi_{diff}].$$

- An explicit feature mapping $\phi : \mathcal{X} \to \mathbb{R}^D$
- An alignment problem: does a peak 70.493m/z belong to bin 70 or 71 with mass error $\varepsilon = 0.5$?

Spectral density model

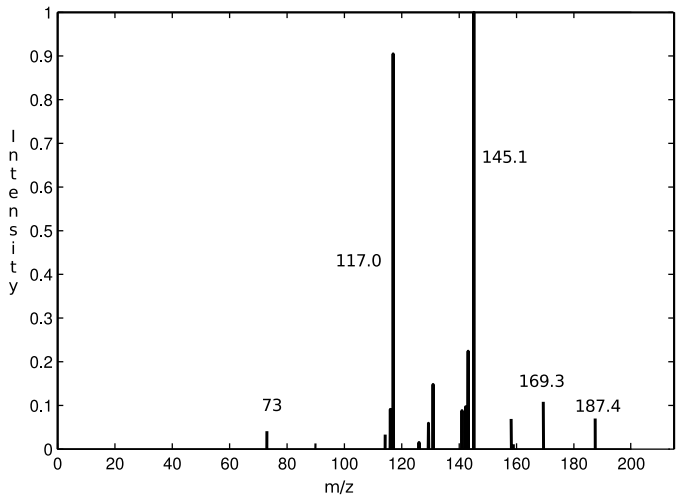
- We incorporate the mass measurement error directly into the features
- We model each peak as a 2-dimensional gaussian

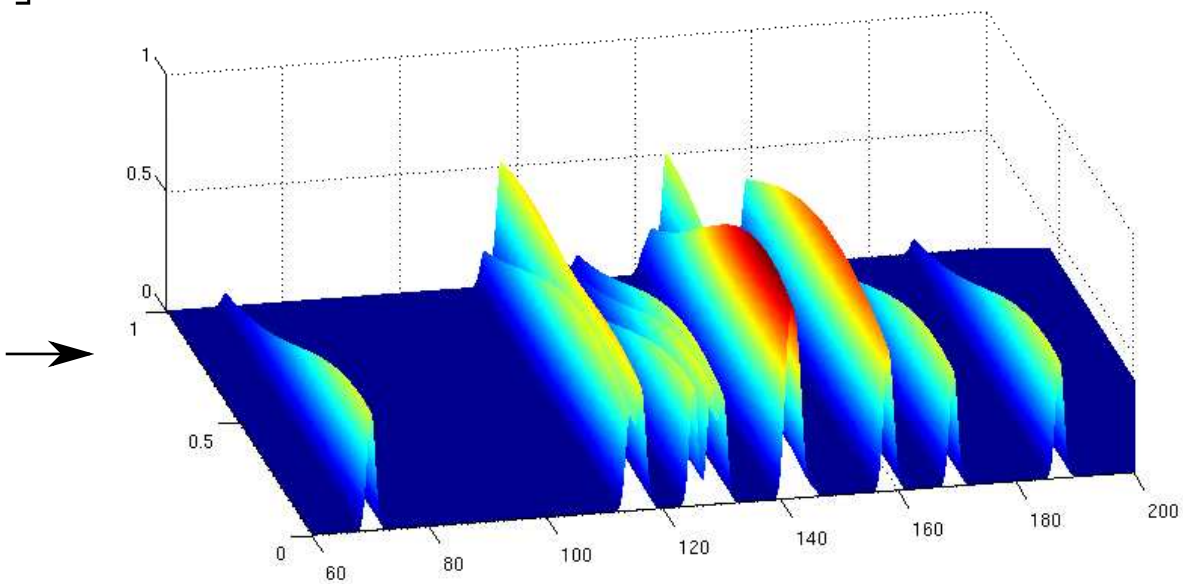
$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}, \Sigma).$$

The spectrum becomes a gaussian mixture model

$$p(\chi) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{N}(\mathbf{x}_i, \Sigma)$$

The $\Sigma = \begin{bmatrix} \sigma_{mass} & 0 \\ 0 & \sigma_{int} \end{bmatrix}$ models the error





High resolution probability product kernel

- Kernels between sets or distributions [Jebara & Kondor 2004]
- Represent a spectrum $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ of peaks with a probability distribution $p(\chi)$
- The kernel $K(\chi, \chi') \equiv K(p, p')$ is then a similarity between probability distributions as the integral of the product distribution:

$$K(p, p') = \int_{\mathbb{R}^2} p(\mathbf{x}) p'(\mathbf{x}) d\mathbf{x}$$

- Interpretation as expectation of one distribution under the other (*expectation likelihood kernel*):

$$\int_{\mathbb{R}^2} p(\mathbf{x}) p'(\mathbf{x}) d\mathbf{x} = \mathbb{E}_p[p'(\mathbf{x})] = \mathbb{E}_{p'}[p(\mathbf{x})]$$

- Feature map: $\varphi : \chi \to p(\chi)$, the kernel $K(p, p') = \langle p, p' \rangle$ in ℓ_2 space
- Closed form solution for gaussian mixtures (fast)
- We use the probability product kernel over the three features

Fingerprints into metabolites

- We predict the fingerprint vector $\hat{\mathbf{y}}$ of the unknown metabolite using SVM's and the mass spectral kernels
- Next, we find candidate metabolites with matching fingerprints from molecular databases (PubChem)
- The fingerprint predictions contain almost always errors and thus the candidate metabolite with exactly matching fingerprints is rarely correct
  - We list candidates according to how confident we are in specific predictions
  - The cross-validation prediction accuracies $(p_i)_{i=1}^m$ of a fingerprint i being correctly predicted are used to determine which fingerprints we allow to mismatch

Poisson-Binomial model

- Poisson-Binomial model for a particular fingerprint vector $\mathbf{y}$ being true given the prediction $\hat{\mathbf{y}}$ and the prediction accuracies $\mathbf{p} = (p_i)_{i=1}^m$:

$$P(\mathbf{y}|\mathbf{p}, \hat{\mathbf{y}}) = \prod_{i=1}^{m} p_i^{[y_i = \hat{y}_i]} (1 - p_i)^{[y_i \neq \hat{y}_i]}$$

  - Maximum value at $\mathbf{y} = \hat{\mathbf{y}}$
  - A high p_i indicates that a candidate with non-matching i'th fingerprint is unlikely to be true
  - A low p_i indicates that a candidate with non-matching i'th fingerprint might be true

- Each candidate metabolite gets a score based on its fingerprint vector:

$$score(M) = P(\mathbf{y}(M)|\mathbf{p}, \hat{\mathbf{y}})$$

- We rank metabolites by score (success = true metabolite in top10)

Contents

Experiments

- Three datasets from MassBank
  - 'QqQ' ($n = 514, m = 286$): A low-accuracy Quadrupole dataset with repeated measurements at collision energies 10eV, 20eV, ..., 50eV
  - 'Ltq' ($n = 293, m = 128$): A high-accuracy LTQ Orbitrap dataset
  - 'Lipids' ($n = 403, m = 20$): A high-accuracy LTQ Orbitrap dataset of non-common phosphatidylethanolamines
- Standard SVM's, 5-fold crossvalidation, C parameter from $\{10^0, \ldots, 10^4\}$
- Candidate metabolites are queried from
  - KEGG (a small database of over 14,000 metabolites)
  - PubChem (a large general-purpose repository of over 30 million molecules)

1. We evaluate the accuracy of fingerprint prediction using different kernels
2. We evaluate the ranks of true metabolites using fingerprint predictions

Fingerprint prediction accuracy

		QqQ							Ltq	Lipids
		Single spectra (CE eV)					Multiple spectra			
	Kernel	10	20	30	40	50	$\sum_e K_e$	merge		
Integral	K_p, linear	87.8	88.2	88.8	89.3	89.5	89.5	89.2	85.5	98.4
	quadr.	87.9	88.3	88.8	89.4	89.6	89.9	89.8	84.4	98.1
	K_{nl}	88.4	88.8	88.8	88.7	89.2	89.4	89.0	86.3	98.8
		88.4	88.9	88.8	88.9	89.2	89.6	89.3	86.1	98.7
	K_{df}	87.8	88.0	87.7	87.8	88.2	88.0	87.9	82.6	97.1
		87.8	88.0	87.8	87.9	88.3	87.9	87.9	82.9	96.9
	K_{p+nl}	88.5	89.5	89.9	90.1	90.3	**90.7**	90.3	88.3	**99.5**
		88.4	89.4	90.0	90.0	90.3	90.5	90.6	88.1	99.3
	K_{p+df}	88.2	88.6	89.0	89.4	89.6	89.4	89.2	85.6	98.7
		88.1	88.7	89.2	89.6	89.8	89.3	89.7	84.8	98.4
	$K_{p+nl+df}$	88.5	89.5	90.1	90.1	90.3	90.5	90.3	88.3	99.5
		88.6	89.8	90.3	90.3	**90.5**	90.3	90.7	87.6	99.3
High resolution	K_p^{φ}	88.0	88.6	89.1	89.1	89.4	89.3	89.4	86.7	98.6
		88.2	89.1	89.5	89.7	89.9	89.3	90.0	85.5	97.3
	K_{nl}^{φ}	88.8	89.5	89.3	89.2	89.2	89.8	89.6	88.8	99.1
		89.0	89.8	89.7	89.5	89.6	90.0	90.0	88.1	98.0
	K_{df}^{φ}	88.5	88.9	88.6	88.4	88.4	89.2	89.3	83.7	97.8
		88.6	89.0	88.9	88.6	88.6	89.2	89.5	83.9	97.1
	K_{p+nl}^{φ}	89.0	89.9	90.1	90.1	90.2	90.5	90.5	**91.1**	99.3
		89.2	**90.1**	90.3	90.3	90.4	90.1	90.8	89.6	97.9
	K_{p+df}^{φ}	88.8	89.4	89.5	89.5	89.5	90.0	90.0	86.5	98.8
		88.9	89.5	89.7	89.8	89.8	89.8	90.4	84.9	97.5
	$K_{p+nl+df}^{\varphi}$	89.1	90.0	90.3	90.2	90.2	90.6	90.7	90.5	99.3
		89.2	**90.1**	**90.4**	**90.5**	90.4	90.2	**91.1**	88.6	98.0
	random	87.3	87.2	87.2	87.2	87.7		87.3	78.7	88.3

Table : The classification accuracies (in %) of the three datasets with various kernels. Abbreviations: p is peaks, nl is neutral loss, and df is difference kernel.

Fingerprint prediction accuracy cont.

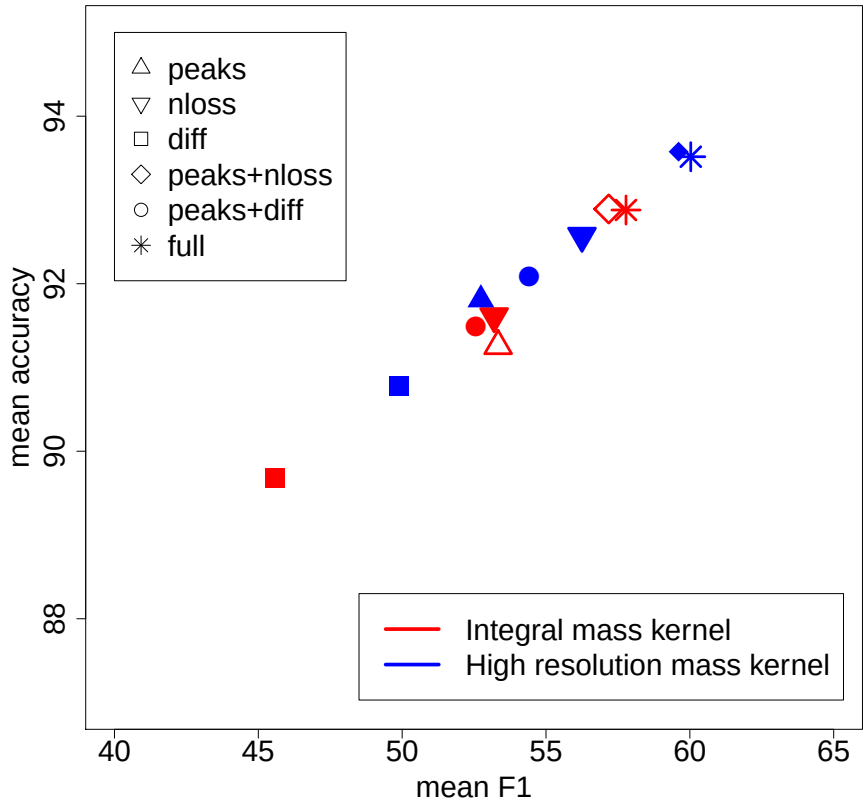


Figure : Scatter plot of the aggregate average accuracy/F_1 across the three datasets with different kernel features. The open markers represent higher accuracy/F_1 ratio in a linear kernel.

Individual fingerprint prediction accuracies

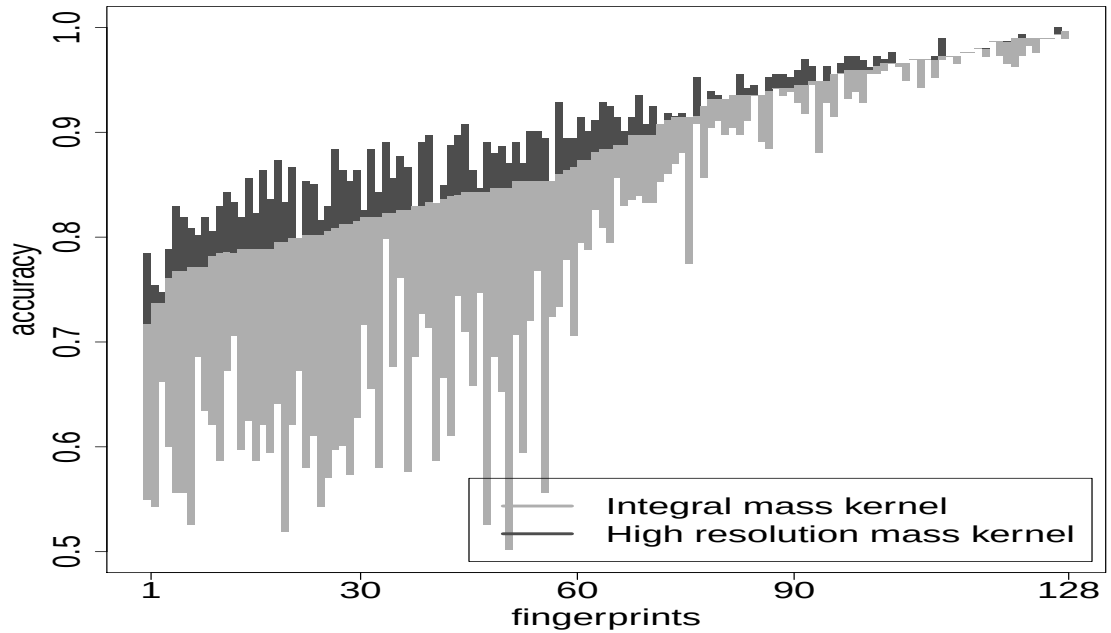


Figure : SVM prediction accuracies of individual fingerprints of the LTQ dataset with high resolution and integral mass kernels. The bottom of the bars is the baseline classifier.

Ranks

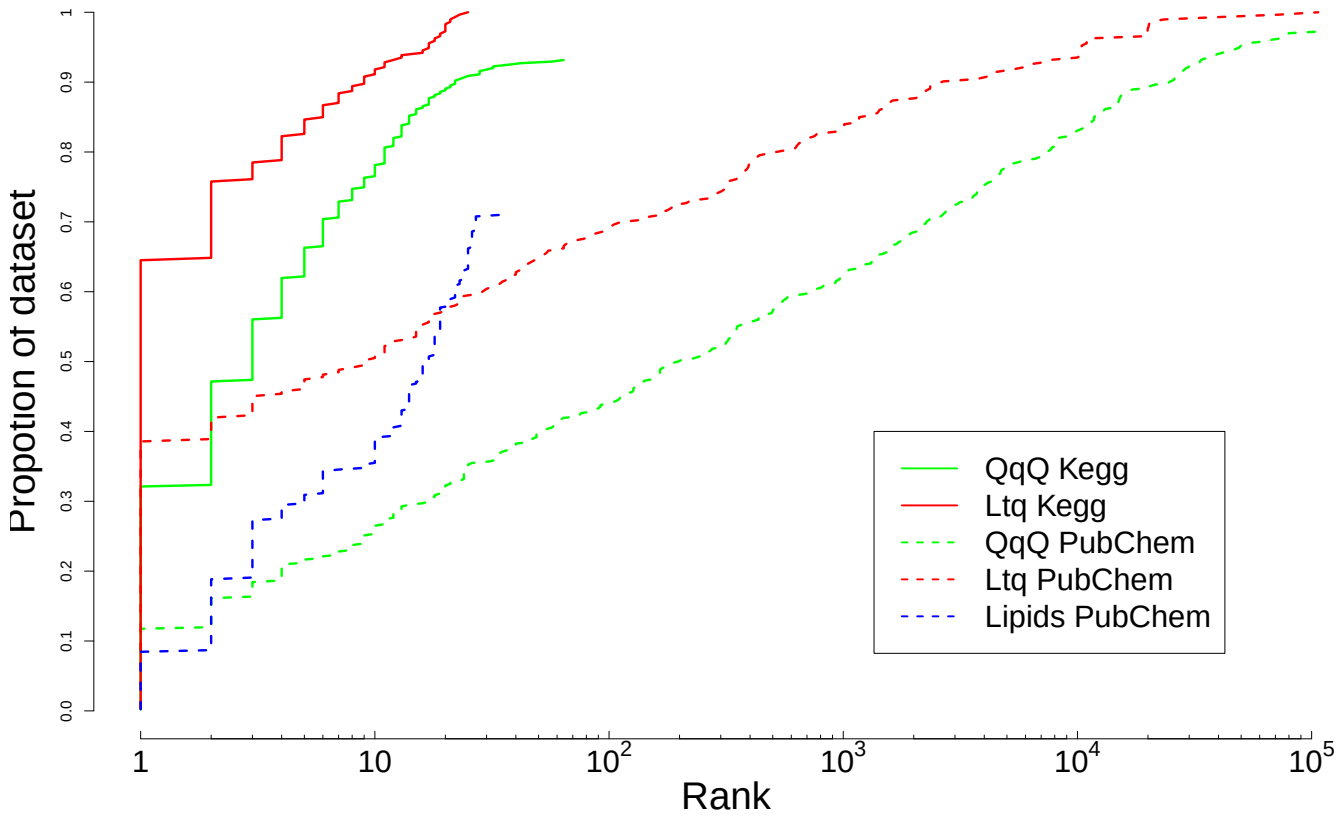


Figure : The ranks of the true metabolite according to the high resolution kernel and the Poisson-Binomial matching model with three datasets and two molecular repositories.

Comparison to MetFrag

- MetFrag is a state-of-the-art computational metabolite identification package[2]
- MetFrag simulates the fragmentation process and tries to match the simulated spectra against the observed
- MetFrag also extracts candidate metabolites from KEGG or PubChem

Molecular database	Spectral dataset	FingerID match	FingerID Avg. rank	FingerID $rank \leq 10$	MetFrag match	MetFrag Avg. rank	MetFrag $rank \leq 10$
Kegg	QqQ	17	3.2	16/17	16	5.1	9/16
	Ltq	20	3.8	18/20	12	5.6	11/12
PubChem	QqQ	11	905	8/11	2	68	0/2
	Ltq	20	58	9/20	1	20	0/1

Table : Comparison of metabolite identification against MetFrag on a subset of 20 spectra from both 'QqQ' and 'Ltq', respectively.

[2] Wolf, Schmidt, Muller-Heinemann & Neumann 2010; msbi.ipb-halle.de/MetFrag/

Conclusions

- Software FingerID: sourceforge.net/p/fingerid
- A machine learning framework for metabolite identification
- Probability product kernels provide a flexible model for mass spectra
- Future work: explore structured prediction, feature selection (L1)

Thank you

Thank you!