Aalto University
School of Science
and Technology

Machine Learning for Metabolite Identification

Huibin Shen

Department of Computer Science
School of Science, Aalto University

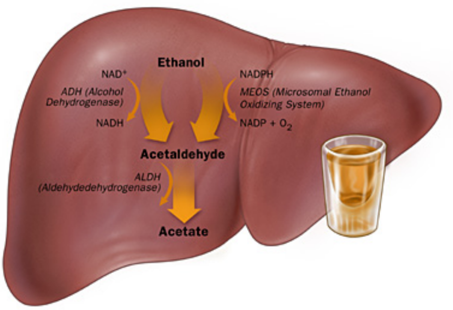
March 30, 2017

Content

- Metabolite identification
- Machine learning
- Machine learning for metabolite identification
- Conclusion

What is metabolite?

- Metabolism is the set of life-sustaining chemical transformations within the cells of of living organisms.


Ethanol
NAD+
ADH (Alcohol Dehydrogenase)
NADH
NADPH
MEOS (Microsomal Ethanol Oxidizing System)
NADP + O2
Acetaldehyde
ALDH (Aldehydedehydrogenase)
Acetate

- Metabolites are the intermediates and products of metabolism.

Why to identify the metabolites?

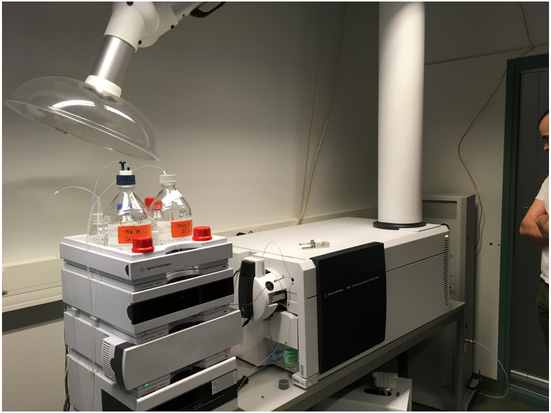
Because they relate to many things:

- health & nutrition
- pharmaceuticals
- biotechnology
- regulatory affairs (drug trafficking, anti-doping).

It is a pre-requisite step for many subsequent analyses.

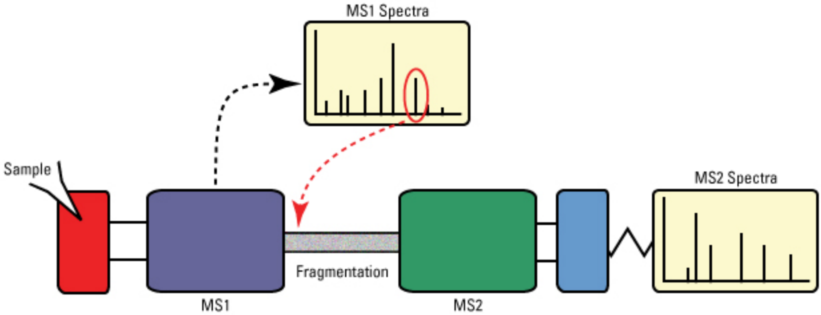
How to identify the metabolites?

The main technology is the tandem mass spectrometry.



How to identify the metabolites?

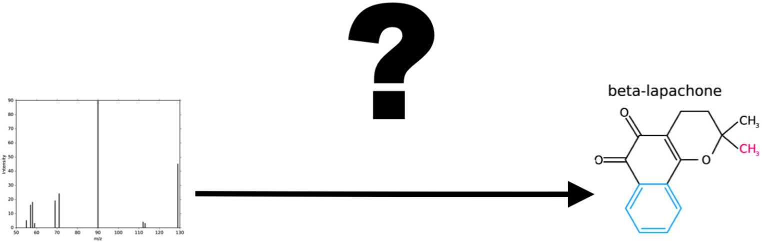
The main technology is the tandem mass spectrometry.



The peaks in the end characterize the structure of metabolite.

Metabolite identification

Given a tandem MS/MS spectrum, what is the molecular structure?

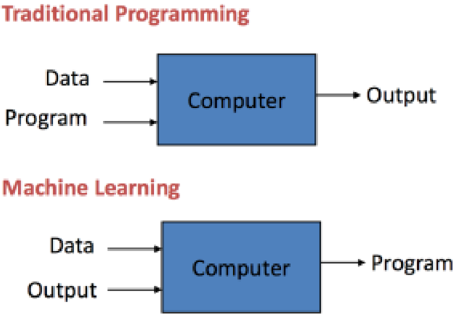


This is the core problem this dissertation trying to improve.

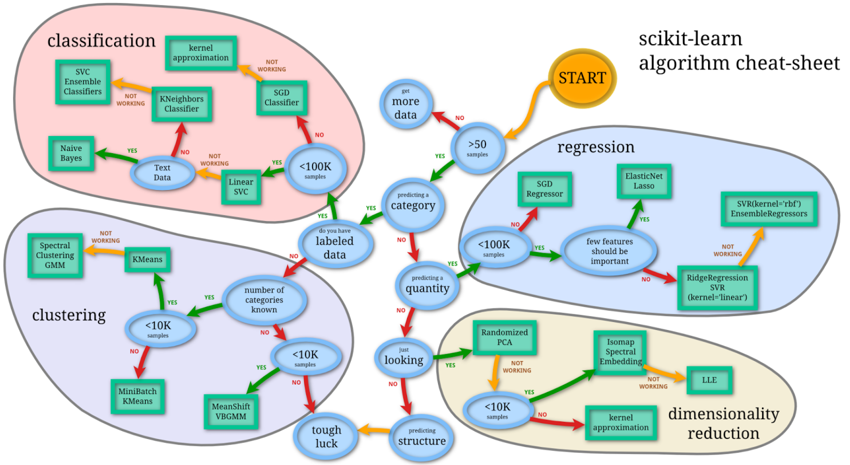
Content

- Metabolite identification
- Machine learning
- Machine learning for metabolite identification
- Conclusion

Machine learning

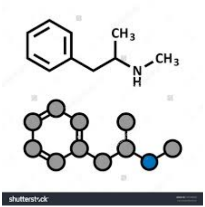


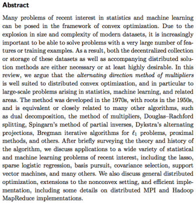
Machine learning



Machine learning

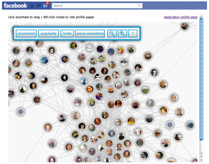
- We use supervised learning, where labels are available.
- We focus on a special family of the supervised learning, where non-tabular format of data can be handled.





Abstract

Many problems of recent interest in statistics and machine learning can be posed in the framework of convex optimization. Due to the explosion in size and complexity of modern datasets, it is increasingly important to be able to solve problems with a very large number of features or training examples. As a result, both the decentralized collection or storage of these datasets as well as accompanying distributed solution methods are either necessary or at least highly desirable. In this review, we argue that the *alternating direction method of multipliers* is well suited to distributed convex optimization, and in particular to large-scale problems arising in statistics, machine learning, and related areas. The method was developed in the 1970s, with roots in the 1950s, and is equivalent or closely related to many other algorithms, such as dual decomposition, the method of multipliers, Douglas–Rachford splitting, Spingarn's method of partial inverses, Dykstra's alternating projections, Bregman iterative algorithms for ℓ_1 problems, proximal methods, and others. After briefly surveying the theory and history of the algorithm, we discuss applications to a wide variety of statistical and machine learning problems of recent interest, including the lasso, sparse logistic regression, basis pursuit, covariance selection, support vector machines, and many others. We also discuss general distributed optimization, extensions to the nonconvex setting, and efficient implementation, including some details on distributed MPI and Hadoop MapReduce implementations.



- They are known as kernel methods.

Kernel methods

- Data as $D = \{(x_i, y_i)\}_{i=1}^{n}$, $x \in \mathcal{X}$ and $y \in \{+1, -1\}$.
- A kernel function $k(x, x')$ measures similarity.
- For example, support vector machine (SVM) solves:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{k(x_i, x_j)} \\ s.t. \quad & 0 \leq \alpha_i \leq C, i = 1, \ldots, n, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

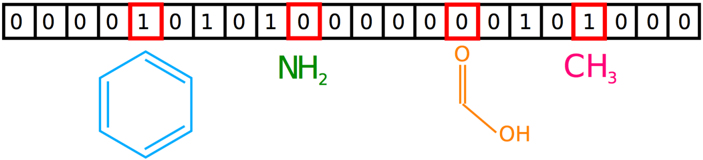
- We only need to plug in the similarities!

Content

- Metabolite identification
- Machine learning
- Machine learning for metabolite identification
- Conclusion

Molecular fingerprint

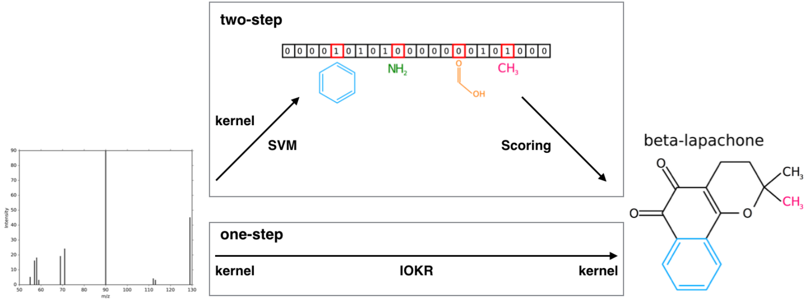
Molecular fingerprints are a representation of the molecules, encoding structures or other properties.



0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0

The proposed methods

Two step approach CSI-FingerID [Heinonen et al., 2012, Shen et al., 2013, Shen et al., 2014, Dührkop et al., 2015].



two-step

0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0

kernel

SVM

Scoring

one-step

kernel

IOKR

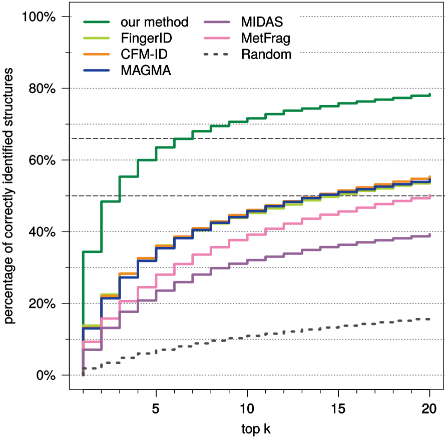
kernel

beta-lapachone

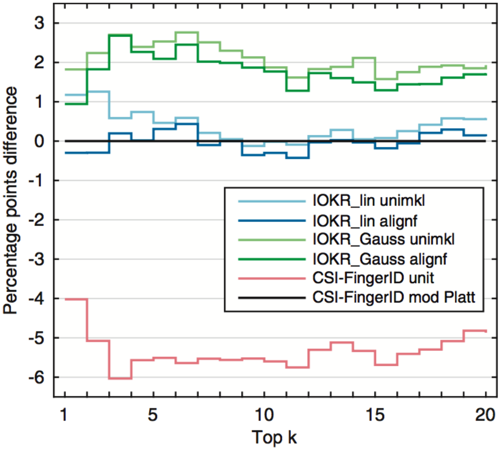
One-step approach IOKR [Brouard et al., 2016].

Results of the two-step approach CSI-FingerID

Cross validation on 5923 mass spectra and compounds pairs.



Results of the one-step apparoch IOKR

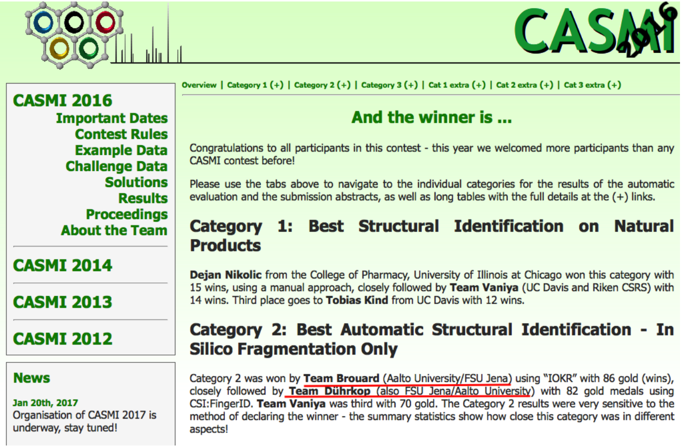


Results of the one-step apparoch IOKR

Table: Running time evaluation

	Training time	Test time
CSI:FingerID	82 h 28 min 23 s	1 h 11 min 31 s
IOKR linear	**42 s**	**1 min 15 s**
IOKR polynomial	**38 s**	21 min 58 s
IOKR Gaussian	**41 s**	33 min 15 s

Results on the CASMI 2016 challenge



And the winner is ...

Congratulations to all participants in this contest - this year we welcomed more participants than any CASMI contest before!

Please use the tabs above to navigate to the individual categories for the results of the automatic evaluation and the submission abstracts, as well as long tables with the full details at the (+) links.

Category 1: Best Structural Identification on Natural Products

Dejan Nikolic from the College of Pharmacy, University of Illinois at Chicago won this category with 15 wins, using a manual approach, closely followed by **Team Vaniya** (UC Davis and Riken CSRS) with 14 wins. Third place goes to **Tobias Kind** from UC Davis with 12 wins.

Category 2: Best Automatic Structural Identification - In Silico Fragmentation Only

Category 2 was won by **Team Brouard** (Aalto University/FSU Jena) using "IOKR" with 86 gold (wins), closely followed by **Team Dührkop** (also FSU Jena/Aalto University) with 82 gold medals using CSI:FingerID. **Team Vaniya** was third with 70 gold. The Category 2 results were very sensitive to the method of declaring the winner - the summary statistics show how close this category was in different aspects!

News

Jan 20th, 2017
Organisation of CASMI 2017 is underway, stay tuned!

Conclusion

- Metabolite identification is a major bottleneck in computational metabolomics.
- Kernel based machine learning for metabolite identification is the new state of the art.
- There are many possibilities to improve in the future.

Reference

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016).
Fast metabolite identification with input output kernel regression.
Bioinformatics, 32(12):i28–i36.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015).
Searching molecular structure databases with tandem mass spectra using csi: Fingerid.
Proceedings of the National Academy of Sciences, 112(41):12580–12585.

Heinonen, M., Shen, H., Zamboni, N., and Rousu, J. (2012).
Metabolite identification and molecular fingerprint prediction through machine learning.
Bioinformatics, 28(18):2333–2341.

Shen, H., Dührkop, K., Böcker, S., and Rousu, J. (2014).
Metabolite identification through multiple kernel learning on fragmentation trees.
Bioinformatics, 30(12):i157–i164.

Shen, H., Zamboni, N., Heinonen, M., and Rousu, J. (2013).
Metabolite identification through machine learning—tackling casmi challenge using fingerid.
Metabolites, 3(2):484–505.