Liquid-Chromatography Retention Order Prediction for Metabolite Identification

Eric Bach [1,✉], Sandor Szedmak [1], Céline Brouard [1], Sebastian Böcker [2] and Juho Rousu [1]

[1]Helsinki institute for Information Technology (HIIT), Department of Computer Science, Aalto University, Espoo, Finland
[2]Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany.

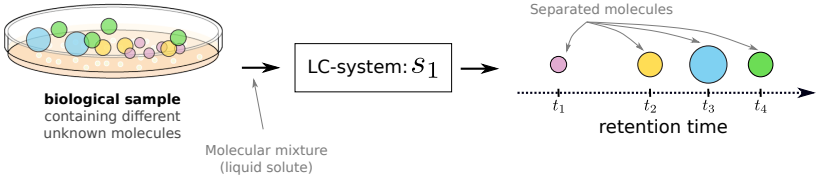
ECCB conference in Athens: September 11, 2018

Aalto University
School of Science

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

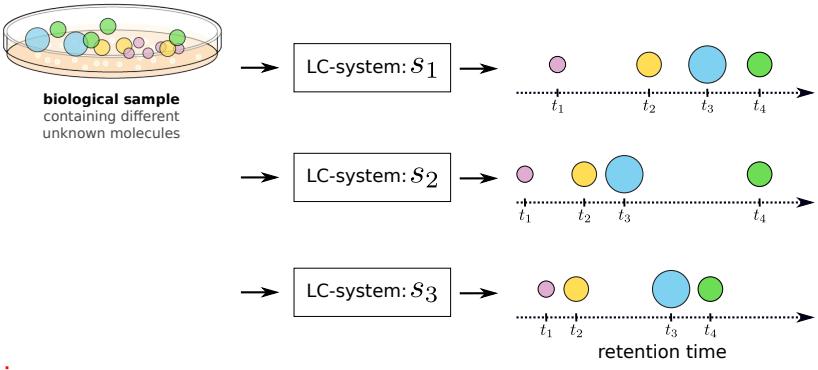
Liquid Chromatography (LC)

A method to reduce sample complexity when analyzing molecular mixtures.



Liquid Chromatography (LC)

A method to reduce sample complexity when analyzing molecular mixtures.



Observations

- Retention time of a molecule varies across chromatographic system.
- *Retention orders* are preserved.

Utilizing retention times

- Retention times are *valuable* information.
  Used to identify unknown molecular structures.
- Large number of methods to predict retention times exist. Regression
- Suffer from the different retention times across systems.

We propose to ...

- **predict the pairwise retention order** given molecular structures using preference learning.
  - Prediction model can be trained on *multiple* retention time datasets arising from *heterogeneous* LC-systems.
  - Retention orders are largely preserved across LC-systems [SNV15].

Retention order pairs for preference learning

Notation

- Molecule m_i from the molecular space $\mathcal{M}$
- $t_i \in \mathbb{R}_+$ its retention time
- $s_i \in \mathcal{S}$ chromatographic system it has been measured with

Pairwise molecule preference

- m_i is preferred over m_j when it *elutes before* m_j, i.e. $t_i < t_j$
- Set of pairwise preferences of given LC-system s is defined as:

$$\mathcal{P}(s) = \{(i,j) \,|\, s_i = s_j = s, t_i < t_j\}$$

- Set of pairwise preferences from *multiple* LC-systems: $\mathcal{P} = \bigcup_{s \in \mathcal{S}} \mathcal{P}(s)$

Preference learning: **Rank**ing **S**upport **V**ector **M**achine

We want to learn a pairwise retention order prediction function:

$$f(m_i, m_j) = \begin{cases} 1 & m_i \text{ elutes before } m_j \\ -1 & \text{otherwise} \end{cases}$$

RankSVM prediction model

$$f(m_i, m_j) = \text{sign}(\mathbf{w}^T(\phi(m_j) - \phi(m_i)))$$

- $\mathbf{w}$ are the RankSVM [Joa02; KLL14] model parameters
- $\phi : \mathcal{M} \rightarrow \mathcal{F}_m$ function to embed the *representation* of the molecular structure m_i into a feature space.
- In the following the molecular representations is denoted with m_i.

Molecular graphs, molecular fingerprints

Training the RankSVM for retention order prediction

Optimizing $\mathbf{w}$ considering the pairwise preferences $\mathcal{P}$ from different systems.

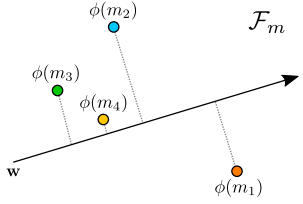
Optimization problem

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{(i,j)\in\mathcal{P}} \xi_{ij} \\ s.t. \quad & \mathbf{w}^T(\phi(m_j) - \phi(m_i)) \geq 1 - \xi_{ij}, \forall (i,j) \in \mathcal{P} \\ & \xi_{ij} \geq 0, \forall (i,j) \in \mathcal{P}, \end{aligned}$$

with $C > 0$ being the regularization parameter.

Learned model

$$\mathbf{w}^T\phi(m_i) < \mathbf{w}^T\phi(m_j), \text{ if } (i,j) \in \mathcal{P}$$



Evaluating retention order prediction

Dataset and molecule representation

- 1098 retention times of 946 unique molecular structures
- 5 different reversed phase LC-systems (denoted with $\hat{S}$)
- Molecules represented using counting MACCS fingerprints:

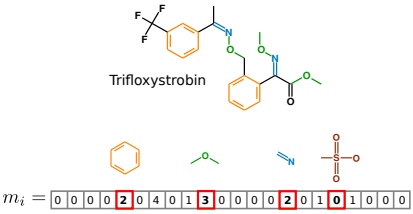


Figure: Counting fingerprint

Evaluation measure and protocol

- Pairwise prediction accuracy for a target system $s \in \hat{S}$:

$$Acc(s) \equiv \frac{|\{(i,j) \in \mathcal{P}(s) \,|\, \mathbf{w}^T \phi(m_i) < \mathbf{w}^T \phi(m_j)\}|}{\mathcal{P}(s)}$$

- Accuracy accessed using repeated 10-fold cross-validation.

Train model with preferences from different systems

Can pairwise predictor benefit from information of different systems?

Compare performance of different training sets

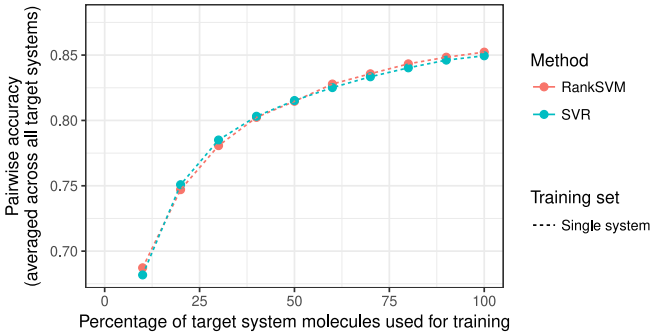
- Single system, target data only: $\mathcal{P}(s)$
- Multiple systems, *no* target data: $\mathcal{P} \setminus \mathcal{P}(s)$
- Multiple systems, all available data: $\mathcal{P}$
- Vary percentage of target system molecules used for training

Comparison method

- Support Vector Regression (SVR) trained on retention times directly [Aic+15].
- Multiple systems: Retention times are considered jointly.

Train model with preferences from target system

Application setting: Training retention times only available from single target system.

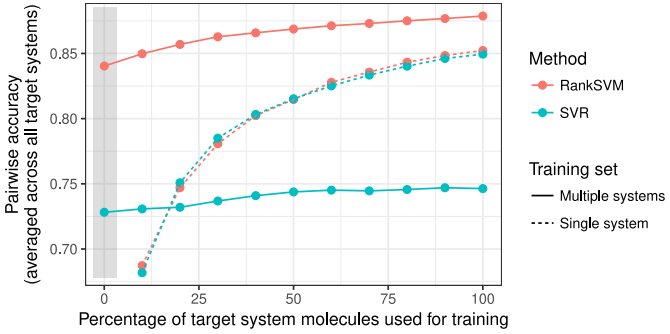


Observations

- Increasing amount of training data improves prediction.
- RankSVM and SVR perform equally.

Train model with preferences from different systems

Application Setting: Training retention times only available *from not target* system.

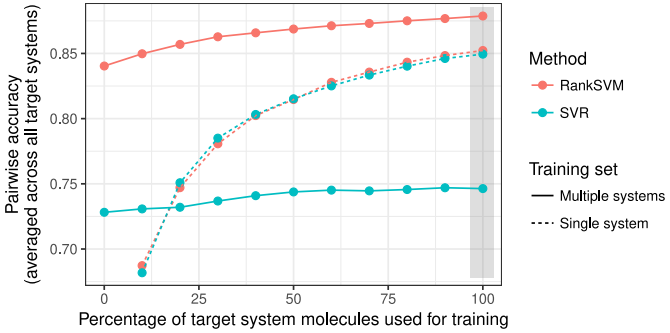


Observations

- Performance of single system *without* data from the target.
- RankSVM outperforms SVR by considering retention *orders*.

Train model with preferences from different systems

Application Setting: Training retention times from target *and* others systems available.

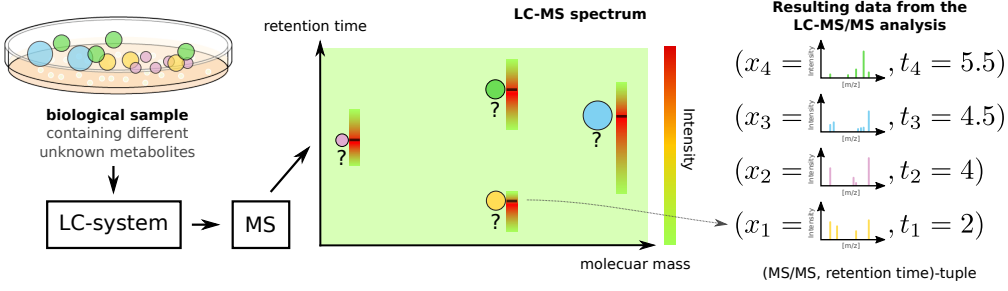


Observations

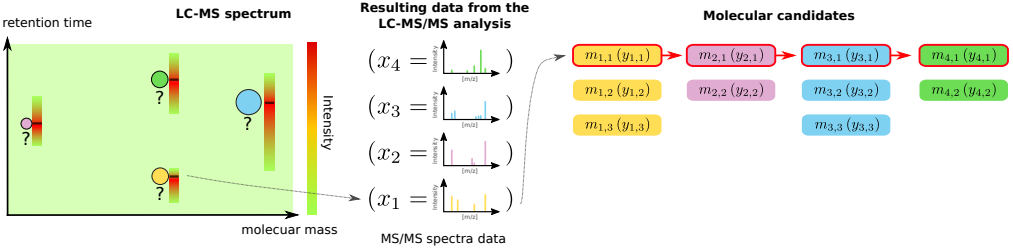
- Considering target *and* non-target systems' data outperforms single system.
- RankSVM again outperforms SVR.

Metabolite identification

- Small molecules (< 1000 Da) involved in biological processes
- Identification of metabolites present in a biological sample
- Widely used analysis workflow: Liquid chromatography (LC) combined with tandem mass spectrometry (MS/MS)
- Molecular structure is not measured, but inferred from the MS/MS spectrum.



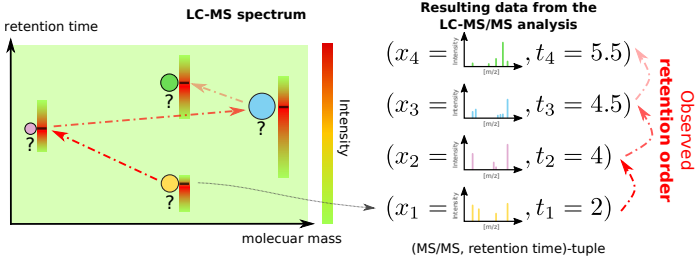
State-of-the-art MS/MS based metabolite identification



Identification workflow

1. For each MS/MS spectrum x_i define a set of *molecular candidate structures* $\{m_{i,1}, m_{i,2}, \ldots\}$. using the molecular mass
2. Assign a "MS/MS matching score" $y_{i,j}$ to each candidate. Input-Ouput-Kernel-Regression[Bro+16]
3. Highest scoring candidate $m_{i,j}$ is the identification for spectrum x_i.

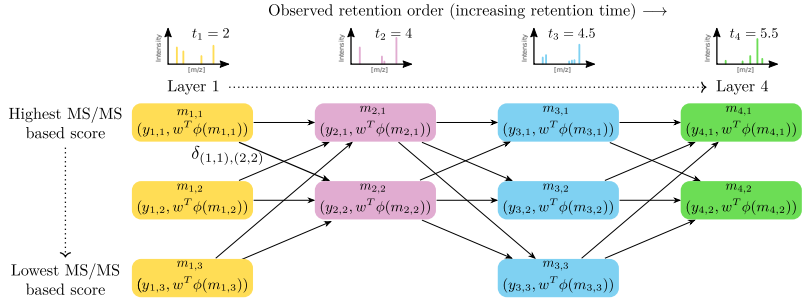
Predicted retention orders for metabolite identification



Identification workflow

1. execute 1. (query candidates) and 2. (predict matching scores)
2. Construct a layered graph:
   - Nodes: Molecular candidate structures $m_{i,j}$
   - Edges: Encode matching scores and predicted retention orders
3. Find the *overall* most consistent metabolite identification using the shorest path algorithm.

Predicted retention orders for metabolite identification



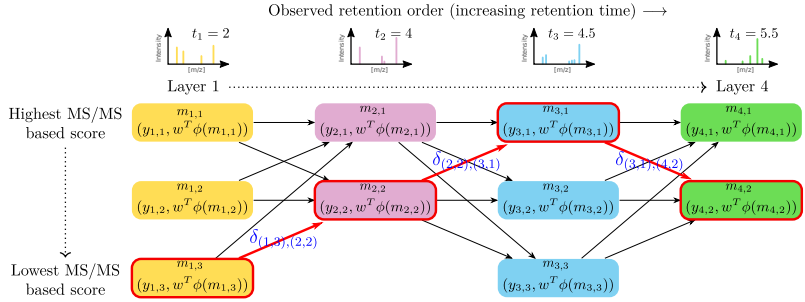
- Edges connecting candidates of consecutive layers with edge weight:

$$\delta_{(i,j),(i+1,s)} = -y_{i+1,s} + D \cdot \max(0, \underbrace{\mathbf{w}^T(\phi(m_{i,j}) - \phi(m_{i+1,s}))}_{\text{RankSVM order penalty}}),$$

  $D \geq 0$ weight on order penmalty: $\max(\ldots) > 0$ if observed $\neq$ predicted order.

- Candidates along the shortest path from first to last layer: *most consistent identification.*

Predicted retention orders for metabolite identification



- Edges connecting candidates of consecutive layers with edge weight:

$$\delta_{(i,j),(i+1,s)} = -y_{i+1,s} + D \cdot \max(0, \underbrace{\mathbf{w}^T(\phi(m_{i,j}) - \phi(m_{i+1,s}))}_{\text{RankSVM order penalty}}),$$

  $D \geq 0$ weight on order penmalty: $\max(\ldots) > 0$ if observed $\neq$ predicted order.

- Candidates along the shortest path from first to last layer: *most consistent identification*.

Experiments metabolite identification

Dataset

- 342 reversed phase LC-retention times
  - for 120 MS/MS spectra available → (MS/MS, RT)-tuple
  - remaning 222 RTs are used for RankSVM training (*target*)
- 5 datasets (*others*) of previous experiments also used for RankSVM training

Evaluation measure and protocol

- randomly sample 1000 times 80 (MS/MS, RT)-tuples
- Construction of the graph containing the candidates to run the shortest path algorithm.
- Percentage of correct identifications for different values of D
- Comparison to baseline performance when $D = 0$

Experiments metabolite identification

Baseline performance 22.7%: ($D = 0$, only MS/MS spectra used, black line)

Observations

- Improved identification accuracy for *Others* (23.9%) and *Others & target* (24.8%)
- RankSVM trained only on the *target* data cannot improve.

Summary

- Proposed a method for predicting liquid chromatographic orders using RankSVM.
- Prediction model can be trained on retention time data from different chromatographic systems.
- Proposed method to integrate predicted retention orders and MS/MS scores for metabolite identification in LC-MS setting
- Metabolite identification accuracy can be improved using predicted retention orders.

Acknowledgement

This work has been supported by Academy of Finland and the Aalto Science-IT infrastructure. Travel fellowship was granted from ECCB with support of the ISCB society.

Visit the poster at 18:30!

ID: P_Da080

Source code available

`https://version.aalto.fi/gitlab/bache1/retention_order_prediction`

Fabian Aicheler et al. "Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches". In: *Analytical chemistry* 87.15 (2015), pp. 7698–7704.

Céline Brouard et al. "Fast metabolite identification with Input Output Kernel Regression". In: *Bioinformatics* 32.12 (2016), pp. i28–i36.

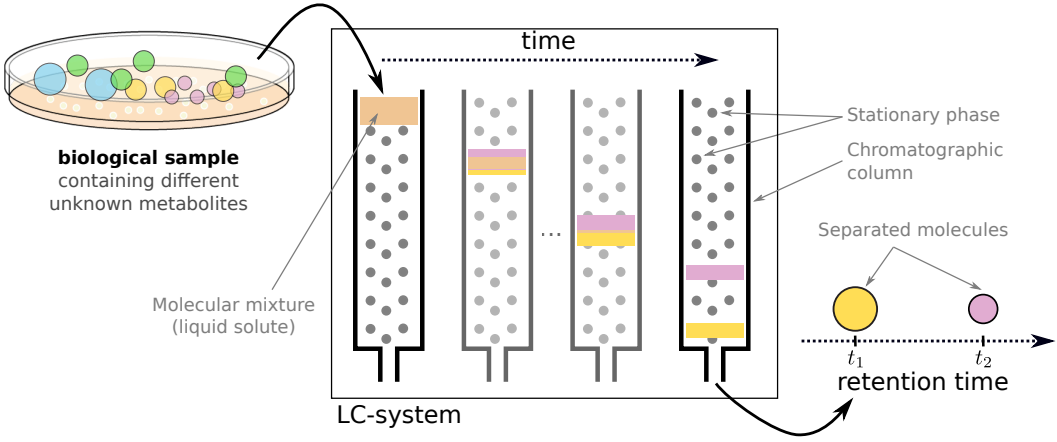
Thorsten Joachims. "Optimizing Search Engines Using Clickthrough Data". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, 2002, pp. 133–142.

Tzu-Ming Kuo, Ching-Pei Lee, and Chih-Jen Lin. "Large-scale kernel rankSVM". In: *Proceedings of the 2014 SIAM international conference on data mining*. SIAM. 2014, pp. 812–820.

Liva Ralaivola et al. "Graph kernels for chemical informatics". In: *Neural networks* 18.8 (2005), pp. 1093–1110.

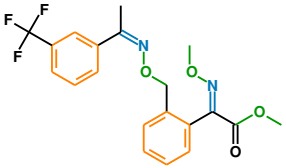
Jan Stanstrup, Steffen Neumann, and Urška Vrhovšek. "PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems". In: *Analytical Chemistry* 87.18 (2015). PMID: 26289378, pp. 9421–9428.

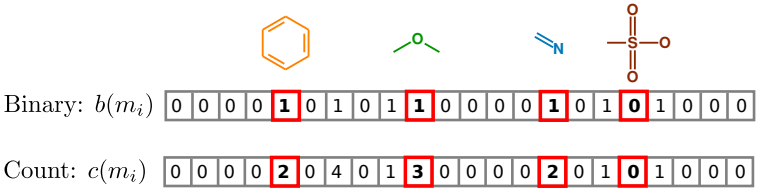
Liquid-Chromatography principle



Molecules represented using MACCS dictionary fingerprints

m_i = Trifloxystrobin





MACCS fingerprint key set

Kernels used for the feature embedding in RankSVM

- Binary: Tanimoto kernel [Ral+05]

$$k_m(m_i, m_j) = \frac{|b(m_i) \cap b(m_j)|}{|b(m_i) \cup b(m_j)|}$$

- Count: MinMax kernel [Ral+05]

$$k_m(m_i, m_j) = \frac{\sum_{s=1}^{N_{sub}} \min(c_s(m_i), c_s(m_j))}{\sum_{s=1}^{N_{sub}} \max(c_s(m_i), c_s(m_j))}$$

Compare binary and counting molecular fingerprints

- Pairwise prediction accuracy ($\pm 2\sigma$) for different target systems
- RankSVM models trained using single system $\mathcal{P}(s)$.

Target system s	Binary MACCS	Counting MACCS
Eawag_XBridgeC18	0.796($\pm$0.015)	**0.844**($\pm$**0.011**)
FEM_long	0.882($\pm$0.016)	**0.905**($\pm$**0.015**)
RIKEN	0.826($\pm$0.024)	**0.848**($\pm$**0.017**)
UFZ_Phenomenex	0.790($\pm$0.027)	**0.802**($\pm$**0.017**)
LIFE_old	0.842($\pm$0.050)	**0.862**($\pm$**0.035**)