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.

Aalto University School of Science
HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY
ACADEMY OF FINLAND

1. Introduction

- Challenge in untargeted metabolomics studies: **Identification of the metabolites** present in a biological sample.
- Widely used analysis method: Liquid chromatography (LC) combined with tandem mass spectrometry (MS/MS)
- LC-MS/MS analysis produces (MS/MS, retention time)-tuples (Fig. 1).
- State-of-the-art machine learning metabolite identification methods use *only* MS/MS information to rank molecular candidate structures [2]
- Retention time (RT) is *valuable* orthogonal information [6, 7], e.g. distinction of diastereoisomers.
- **Challenges utilizing RTs**: Measurements are *LC-system specific*; Public datasets relatively *small* and originate from *heterogeneous systems*.

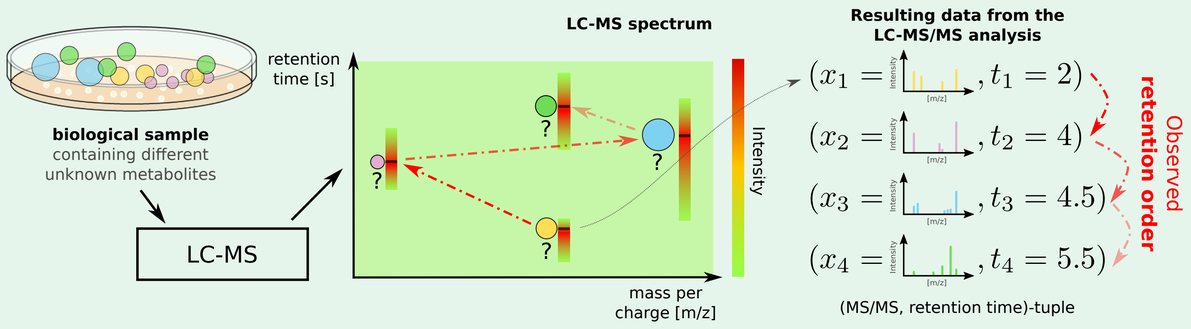

Fig. 1: LC-MS/MS analysis pipeline and resulting data. Observed retention order in red.

2. Proposed method: Utilizing observed retention orders

- We propose to use Ranking Support Vector Machine (RankSVM) [4] to **predict the pairwise retention order** of molecular candidate structures.
  - Retention orders are largely preserved across LC-systems [7].
  - RankSVM can be trained on *multiple* retention time datasets arising from *heterogeneous* LC-systems.
- We introduce a dynamic programming methodology for **integrating predicted candidate retention orders and MS/MS based scores** to *jointly* identify a set of metabolites arising, e.g., in a metabolomics experiment (Fig. 1).

3. Ranking Support Vector Machine (RankSVM)

- **Preference learning** using RankSVM [4] for retention order prediction.
- **Notation**: Molecule m_i from molecular space $\mathcal{M}$, $t_i \in \mathbb{R}_+$ its retention time, s_i LC-system it has been measured with. Set of training LC-systems S. Set of RTs measured with LC-system s is denoted with $\mathcal{T}(s)$.
- Molecule m_i is preferred over m_j when it elutes before m_j, i.e. $t_i < t_j$.
- Set of pairwise preferences of LC-system $s \in S$ 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:

$$\mathcal{P} = \bigcup_{s \in S} \mathcal{P}(s).$$

- **Kernel RankSVM**: Molecular structure encoded by kernel function $k_m : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, with feature-map $\phi : \mathcal{M} \to \mathcal{F}_m$ and feature space $\mathcal{F}_m$
- RankSVM preference prediction model:

$$f(m_i, m_j) = \text{sign}(\mathbf{w}^T(\phi(m_j) - \phi(m_i))) \in \{-1, 1\}$$

- Model parameters $\mathbf{w}$ are found by solving:

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

  with $C > 0$ being a regularization parameter.
- By solving the Problem (1): $\mathbf{w}$ is learned such that:

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

4. Integration of MS/MS scores & retention orders

- **Notation**: $n_{i,j}$ denotes molecular candidate j for spectrum i and $y_{i,j}$ its MS/MS based score.
- MS/MS scores predicted using Input Output Kernel Regression (IOKR) [2]
- Directed graph G with nodes representing the molecular candidates (Fig. 2)
- Edges connect the candidates $n_{i,j}$ and $n_{i+1,s}$ with weight:

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

  $D \geq 0$ weight on order penalty: $\max(\ldots) > 0$ if observed $\neq$ predicted order.
- Molecular candidates along the **shortest path** connecting the fist and the last layer are the **most consistent identifications**.

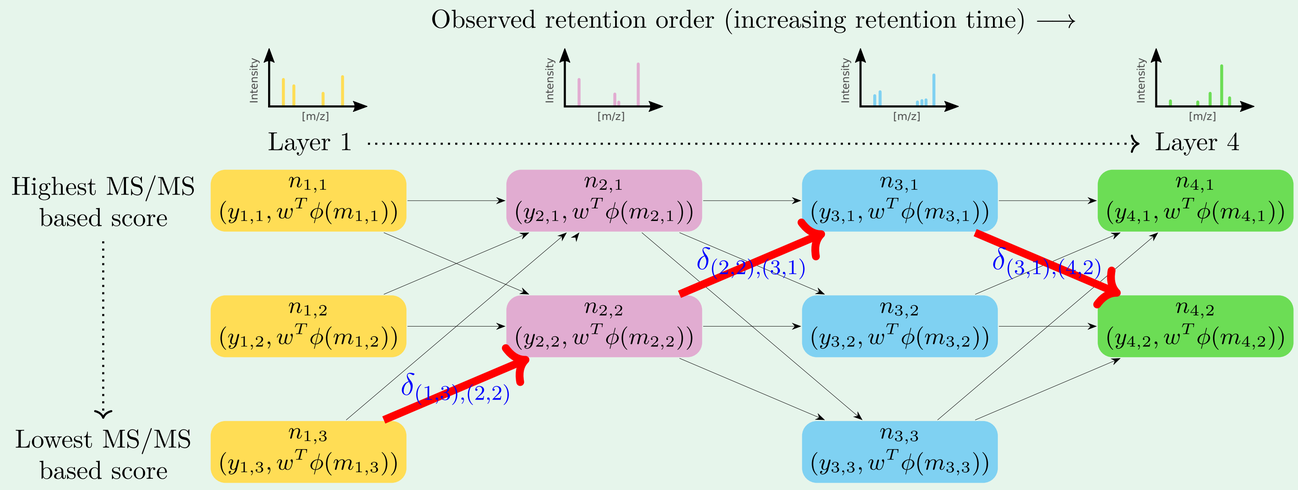

Fig. 2: G with layers corresponding to the molecular candidates per MS/MS. Shortest path in red.

5. Experiments

Retention order prediction:

- 1098 retention times from 5 different reversed phase LC-systems [7] ($\hat{S}$)
- Molecular representation: Counting fingerprints based on the MACCS dictionary [3] combined with MinMax-Kernel [5] k_m
- Competing method: Retention time predicting using Support Vector Regression (SVR) [1].
- Access prediction accuracy in target system $s \in \hat{S}$ by cross-validation (Fig. 3)
- Training sets for RankSVM and SVR for target LC-system $s \in \hat{S}$:

Single (only target) system	$\mathcal{P}(s)$	(RankSVM)	$\mathcal{T}(s)$	(SVR)
Multiple (all available) systems	$\bigcup_{s' \in \hat{S}} \mathcal{P}(s)$	(RankSVM)	$\bigcup_{s' \in \hat{S}} \mathcal{T}(s)$	(SVR)

Metabolite identification:

- 342 reversed phase LC RTs: for 120 MS/MS spectra available $\to$ (MS/MS, RT)-tuple, remaining 222 used for RankSVM training (s_{Impact})
- Identification performance for different D values accessed using repeated bootstrapping of 80 tuples (Fig. 4, black line: baseline with $D = 0$)
- Different RankSVM training sets:

(only) Target	$\mathcal{P}(s_{Impact})$	222 Mol.
(only) Others	$\bigcup_{s' \in \hat{S}} \mathcal{P}(s)$	1098 Mol.
Others & target	$\bigcup_{s' \in \hat{S}} \mathcal{P}(s) \cup \mathcal{P}(s_{Impact})$	1320 Mol.

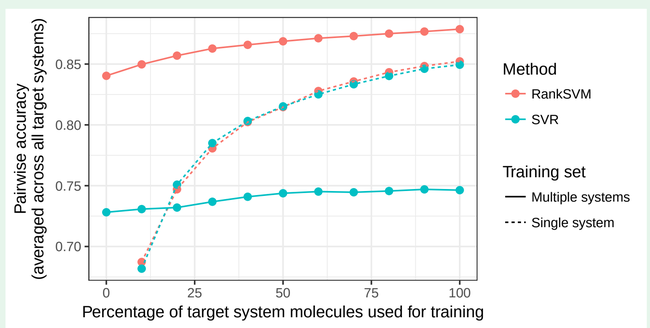

Fig. 3: Accuracy averaged over the 5 systems.

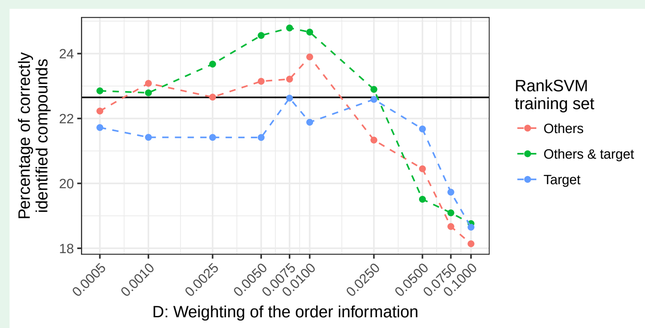

Fig. 4: Accuracy averaged over 1000 samples.

References

[1] F. Aicheler, J. Li, M. Hoene, R. Lehmann, G. Xu, and O. Kohlbacher. Retention time prediction improves identification in nontargeted lipidomics approaches. *Analytical chemistry*, 2015.

[2] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, and J. Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 2016.

[3] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 2002.

[4] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02. ACM, 2002.

[5] L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural networks*, 2005.

[6] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, 8(1):3, 2016.

[7] J. Stanstrup, S. Neumann, and U. Vrhovsek. Predret: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry*, 2015.

Acknowledgements: This work has been supported by Academy of Finland [268874/MIDAS, 310107/MACOME]; and the Aalto Science-IT infrastructure

✉: eric.bach@aalto.fi