Probabilistic framework for integration of mass spectrum and retention time information in small molecule identification

Eric Bach [1,✉], **Simon Rogers** [2], **John Williamson** [2], **and Juho Rousu** [1]

[1]Helsinki institute for Information Technology (HIIT), Department of Computer Science, Aalto University, Espoo, Finland
[2]School of Computing Science, University of Glasgow, Glasgow, UK

Source code available

Check out our paper

1. Small Molecule Identification in Untargeted Metabolomics

- Challenge in untargeted metabolomics studies: **Identification of the small molecules** present in a biological sample
- **LC-MS²** widely used analysis platform: Liquid chromatography (LC) coupled with tandem mass spectrometry (MS²) (Fig. 1)
- Most machine learning approaches for small molecule identification only utilize MS² information [2, 1]
- LC retention times (RT) can improve the small molecule annotation [4, 5]
- **Challenges utilizing RT information:** (1) LC-system specific RT measurements and (2) public RT databases are limited in size and coverage

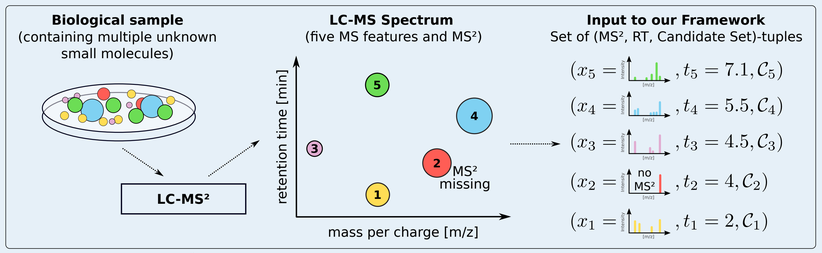

Fig. 1: LC-MS² analysis pipeline and resulting data used as input for our framework.

2. Retention Time (RT) Utilization for Small Molecule Identification

- Different approaches proposed in the literature
- Multiple approaches to utilize RT for molecule annotation exist
- (utilization of RT information, scalable, cross laboratories (LC-systems), RT reference free)

1)	Compare measured RTs with in-house reference RTs	✓, ✗, ✗, ✗
2)	Compare measured RTs with projected reference RTs	✓, ✗, ●, ✗
3)	Compare measured RTs with predicted RTs	✓, ✓, ●, ●
4)	Compare measured RTs with predicted RTs proxies, e.g. LogP	✓, ✓, ✓, ✗
5)	Compare measured retention orders with predicted ones	●, ✓, ✓, ✓

- Fully supported: ✓, Partially supported: ●, Not supported: ✗
- RT comparison to prune candidate lists or (re)ranking **[CITATION]**

3. LC-MS² Experiment Data: Input and Output of our Framework

- **Input:** Preprocessed LC-MS² data, i.e. after peak-picking and alignment (Fig.1):

$$\mathcal{D} = \{(x_i, t_i, \mathcal{C}_i)\}_{i=1}^{N}$$

x_i : MS Information; MS² or MS¹ (precursor m/z), if no fragmentation available
t_i : Measured RT
$\mathcal{C}_i$: Molecular candidate sets, e.g. molecular structures found in PubChem by exact mass search
N : Number of MS features

- **Precomputed MS scoring assumed:** MS¹ deviation of candidate and precursor mass or MS² scores, e.g. by CSI:FingerID [2], MetFrag [4] or IOKR [1]
- **Output:** Ranking of the molecular candidates in $m_{ir} \in \mathcal{C}_i$ for each MS feature i
- Ranking based on MS and RT information

4. Probabilistic Framework to integrate MS and RT Information

- **Graphical model** G superimposed on the LC-MS² data (Fig. 2)
- Let $G = (V, E)$ be complete graph with a **node** $i \in V$ for each MS feature, and an **edge** $(i, j) \in E$ for each feature pair
- Discrete random variable $z_i \in \mathcal{Z}_i = \{1, \ldots, n_i\}$ associated with each node ($n_i = |\mathcal{C}_i|$)
- Candidate annotation for the complete data $\mathbf{z} = \{z_i \,|\, i \in V\} \in \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_N = \mathcal{Z}$
- Intuitively: Random variable z_i denotes the candidate $m_{ir} \in \mathcal{C}_i$ assigned to feature i.
- Pairwise **Markov Random Field** as probabilistic model [3]:

$$p(\mathbf{z}) = \frac{1}{Z} \prod_{i \in V} \psi_i(z_i) \prod_{(i,j) \in E} \psi_{ij}(z_i, z_j)$$

- Potential functions: $\psi_i(z_i)$ MS score and $\psi_{ij}(z_i, z_i)$ match of observed and **predicted retention order**
- Molecular **candidates ranked** based on max-marginals [3] (Fig. 2):

$$p_{\max}(z_i = r) = \max_{\{\mathbf{z}' \in \mathcal{Z} \,|\, z'_i = r\}} p(\mathbf{z}')$$

- Intuitively: Maximum marginal probability of a candidate assignment with $z_i = r$.

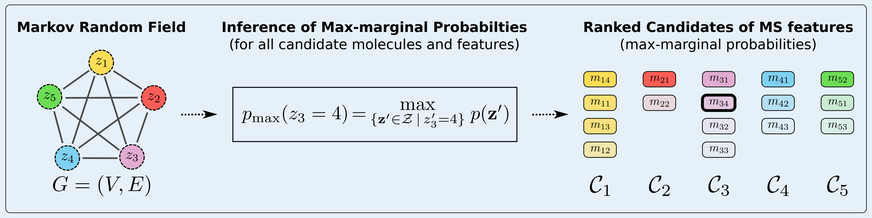

Fig. 2: MRF probability distribution and candidate ranking, e.g. MS feature $i = 3$ and candidate 4 (m_{34}).

5. Encoding MS and Retention Order Information: ψ_i and ψ_{ij}

- Node potential $\psi_i : \mathcal{Z}_i \to \mathbb{R}_{>0}$: $\psi_i(z_i = r) = f(x_i, m_{ir})$
- f returns the MS matching score $\in (0, 1]$ of spectrum x_i and candidate m_{ir}
- Edge potential $\psi_{ij} : \mathcal{Z}_i \times \mathcal{Z}_j \to \mathbb{R}_{>0}$, with σ being the sigmoid function:

$$\psi_{ij}(z_i = r, z_j = s) = \sigma(\underbrace{\text{sign}(t_i - t_j)}_{\text{observed retention order}} \cdot \underbrace{\langle \mathbf{w}, \phi(m_{ir}) - \phi(m_{js}) \rangle}_{\text{predicted retention order}})$$

- Intuitively: Matching observed and predicted retention orders receive high scores.
- Utilization of a Ranking Support Vector Machine $\mathbf{w}$ to predict retention orders [**?**]
- Candidate molecules m_{ir} representation using non-linear features ϕ

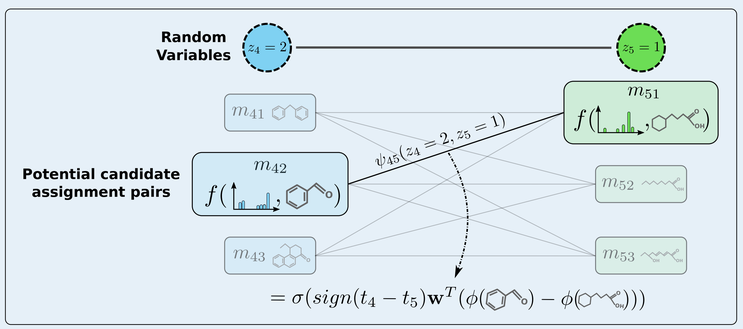

Fig. 3: Example: Node and edge score for all candidate pairs of feature $i = 4$ and $j = 5$.

Spanning Tree Approximation

- Marginal inference intractable in practice due to exponentail sized candidate assignment space $\mathcal{Z}$
- Exact inference is feasible if G is tree-like **[CITATION]**
- Resort to infer the max-marginals a set of trees $\mathbf{T} = \{T_t\}_{t=1}^{L}$ sampled from G
- Each tree $T_t = (V, E_t)$ is connected graph with all nodes of G but reduces edges set $E_t \subseteq E$
- Avergaged marginals used for ranking

$$\bar{p}_{\max}(z_i = r \,|\, \mathbf{T}) = \frac{1}{L} \sum_{t=1}^{L} p_{\max}(z_i = r \,|\, T_t)$$

Random Spanning Trees are used approximate the MRF

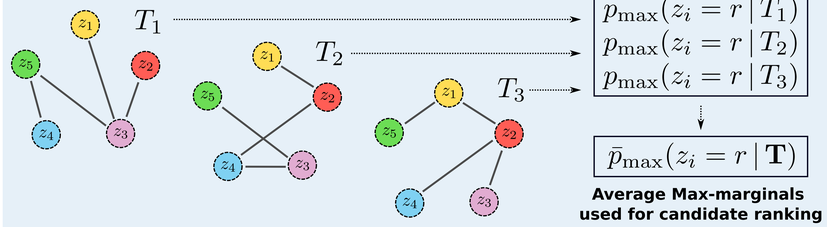

Fig. 4: TODO

5. Experiments and Results

- Dataset description
- Show table 4 from the paper
- Show figure 3 from the paper

References

[1] E. Bach, S. Szedmak, C. Brouard, S. Böcker, and J. Rousu. Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics*, 34(17):i875–i883, 2018.

[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*, 32(12):i28–i36, 2016.

[3] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, and S. Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 2019. Doi 10.1038/s41592-019-0344-8.

[4] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2005.

[5] 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, Jan 2016.

[6] J. Stanstrup, S. Neumann, and U. Vrhovsek. Predret: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry*, 87(18):9421–9428, 2015. PMID: 26289378.

Acknowledgements: This work has been supported by the Academy of Finland grant 310107 (MACOME); and the Aalto Science-IT infrastructure. SR and JHW were supported by EPSRC (EP/R018634/1). This work started during a visit by JR to SR, funded through the Scottish Informatics and Computing Science Alliance (SICSA) distinguished visiting fellow scheme.