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

Small Molecule Identification in Untargeted Metabolomics

- Liquid chromatography (LC) coupled with tandem mass spectrometry (MS^2) widely utilized in untargeted metabolomics studies
- Challenge: Annotation of LC-MS peaks with potential molecular structures
- Most automated machine learning based approaches utilize MS information only **[CITATION]**
- LC retention time (RT) is valueable additional information for the annotation **[CITATION]**, e.g.

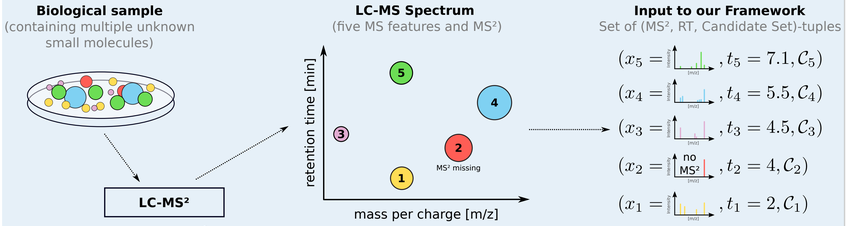


Fig. 1: LC-MS^2 analysis pipeline and resulting data used as input for our framework.

Retention Time (RT) Utilization

- 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]**

LC-MS Experiment Data and its Formal Representation

- Assume data arises from LC-MS experiment (after peak-picking and alignment)
- Available information: MS^1, RT and (etwaige only for some peaks) MS^2
- Molecular candidate lists are assumed to be given as well
- MS^2scores, e.g. MetFrag **[CITATION]**or CSI:FingerID **[CITATION]**, computed
- Data from LC-MS considered as set of N MS features:

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

- x_i: MS^2 spectrum (or MS^1, if no fragmentation available)
- t_i: Measured RT
- $\mathcal{C}_i$: Potentail molecular annotations for feature i, e.g. exact mass search

Probabilistic Framework to integrate MS and Retention Orders

- Definition of a probabilsitic graphical model superimposed on the LC-MS data
- Let $G = (E, V)$ be a complete graph
- Nodes $i \in V$ represent the MS features, Edes $(i, j) \in E$ the feature pairs
- Association of each node with discrete random variable $z_i \in \mathcal{Z}_i = \{1, \ldots, n_i\}$ ($n_i = |\mathcal{C}_i|$ number of candidates)
- Molecule annotation for complete data $\mathbf{z} = \{z_i \,|\, i \in V\} \in \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_N = \mathbf{Z}$
- Intuitively: Random variable denotes which candidate is assigned to each feature.
- Pairwise Markov Random Field as probabilsitic model **?**:

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

- Ranking molecular candidates via max-marginals:

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

- Intuitively, maxmimum probabilsity a candidate assignment with $z_i = r$ can achive
- Rank all candidates $r \in \{1, \ldots, n_i\}$ according to there max-marginals

Markov Random Field | **Inference of Max-marginal Probabilties** for all candidate molecules | **Ranked Candidates for each MS feature** using their max-marginal probabilities

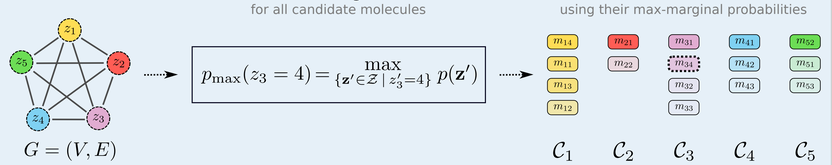
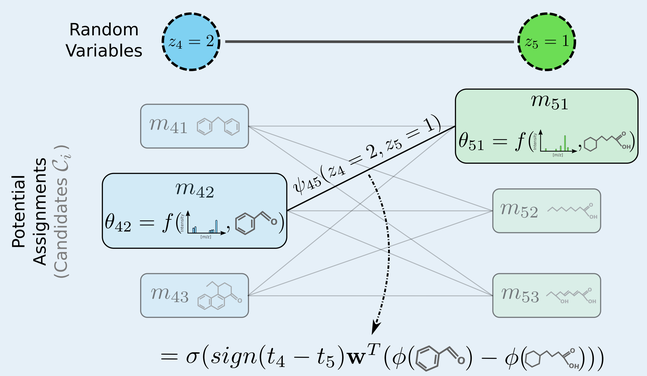


Fig. 2: From the MRF probability distribution to the candidate ranking: MS feature $i = 3$ and candidate 4 (m_{34}).

Node and Edge Potentials

- Node potential function $\psi_i : \mathcal{Z}_i \to \mathbb{R}_{>0}$: goodness of the match between measured spectrum x_i and candidates of feature i
- Edge potential function $\psi_{ij} : \mathcal{Z}_i \times \mathcal{Z}_j \to \mathbb{R}_{>0}$: consistency between the observed retention order of feature i and j with the predicted retention order of the candidates z_i and z_j

Node and Edge Potential Calculation



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

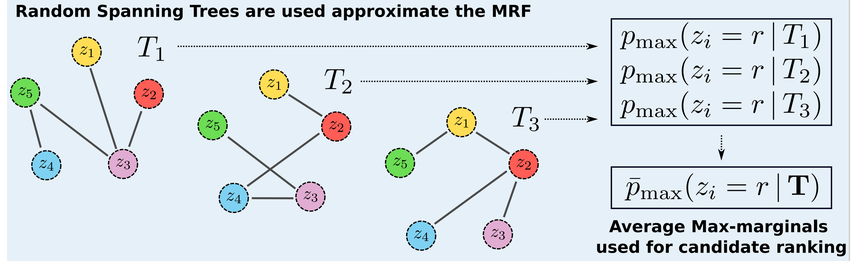


Fig. 3: TODO

5. Experiments and Results

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

References

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.