IMPRS CHEMICAL COMMUNICATION IN ECOLOGICAL SYSTEMS

Project 7

Molecular networks rethought

Supervisors:
Prof. Dr. Sebastian Böcker (main supervisor)
Bioinformatics, Faculty of Mathematics and Computer Science, FSU Jena
Prof. Dr. Georg Pohnert (co supervisor),
Instrumental Analytics, Institute for Inorganic and Analytical Chemistry, Faculty of Chemistry and Earth Sciences, FSU Jena

Background:

Mass spectrometry (MS) is the analytical platforms of choice for high-throughput screening of small molecules. Molecular networks have found widespread application in untargeted metabolomics, natural products research and related areas. Molecular networking is basically a method of visualizing your data, based on the observation that similar tandem mass spectra (MS/MS) often correspond to compounds that are structurally similar. Constructing a molecular network allows us to propagate annotations through the network, and to annotate compounds for which no reference MS/MS data are available. Since its introduction in 2012, the computational method has received a few "updates", including Feature-Based Molecular Networks (FBMN, *Nat Methods* 2020) and Ion Identity Molecular Networks (IIMN, *Nat Commun* 2021), both of which are co-authored by the Böcker group. Yet, the fundamental idea of using the modified cosine to compare tandem mass spectra, has basically remained unchanged at the core of the method.

The Böcker group is one of the leading research groups developing computational methods for untargeted metabolomics. Numerous scientific approaches for this task were developed in our lab during the last decade, including CSI:FingerID for searching in molecular structure databases (*PNAS*, 2015), SIRIUS for molecular formula annotation and processing of full datasets (*Nat Methods*, 2019), CANOPUS for comprehensive compound class assignment (*Nat Biotechnol*, 2021), and COSMIC for assigning confidence in annotations (*Nat Biotechnol*, 2022). We have won numerous CASMI challenges on the topic, and our web services for small molecule annotation have processed more than half a billion queries. SIRIUS version 6 is to be released in Q1 of 2024.

Project description:

For mid-2024, it is planned that molecular networks will be made available through the SIRIUS software. Beyond the pure amenity that users can compute and visualize molecular networks as part of a SIRIUS analysis, this will allow us to quickly integrate and evaluate newly developed computational methods into the molecular networking setup.

Here, we will focus on the scientific side of molecular networks: As noted above, molecular networks currently use the modified cosine to compare mass spectra; more precisely, a heuristic (inexact) method is used for computing the modified cosine.

1. We will replace the inexact method for computing the modified cosine by an exact method, and study how this changes results.
2. We will complement the modified cosine by methods that estimate similarity using fragmentation trees computed as part of the SIRIUS analysis pipeline.

3. We will further complement the modified cosine by similarity measures based on molecular fingerprints, such as Tanimoto coefficients. Molecular fingerprints are computed as part of the CSI:FingerID analysis pipeline, and using those fingerprints has already proven beneficial for the Qemistry method (*Nat Chem Biol* 2021) that we also co-author.
4. We will complement the modified cosine by a False Discovery Rate (FDR) estimation for individual edges in the molecular network. This method will be derived from our method for FDR estimation in spectral library search (*Nat Commun*, 2017). This will allow users to overcome thresholding issues of molecular networks.
5. We will combine the different measures into a single visualization, and also try to integrate them into a single measure telling us whether a particular edge of the network is true or false: I.e., are the compounds behind the spectra structurally similar?

Since molecular networks are an interactive method for data exploration, it is extremely important to integrate feedback from experimental groups into the development of new computational methods. Any new computational method must indeed be of help in data exploration, and this can only be verified in biological studies. To this end, we will promptly apply developed methods to biological data, and we will do so in close collaboration with experimental research groups around the globe, including that of Prof. Georg Pohnert.

Candidate profile:

- M.Sc. in bioinformatics, cheminformatics, computer science, mathematics
- Expertise and interest in algorithmics and bioinformatics methods development
- Experience in software development (git, artifactory) is highly desirable
- Experience in biochemistry and machine learning is desirable
- Experience in Java, Python and ML frameworks is desirable
- Ability to interact with coworkers, collaboration partners and software users

Reading:

1. M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni. MSNovelist: de novo structure generation from mass spectra. *Nat Methods*, 2022.
2. M. A. Hoffmann, …, S. Böcker. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol*, 40(3):411–421, 2022.
3. K. Dührkop, …, S. Böcker. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*, 39(4):462–471, 2021.
4. R. Schmid, …, P. C. Dorrestein. Ion identity molecular networks for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun*, 12(1):3832, 2021.
5. Tripathi, …, P. C. Dorrestein. Chemically-informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat Chem Biol*, 17(2):146–151, 2021.
6. L.-F. Nothias, …, P. C. Dorrestein. Feature-based molecular networks in the GNPS analysis environment. *Nat Methods*, 17(9):905–908, 2020.
7. K. Dührkop, …, S. Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 16(4):299–302, 2019.
8. K. Scheubert, …, S. Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun*, 8:1494, 2017.
9. K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci USA*, 112(41):12580–12585, 2015.