May Institute on computation and statistics for mass spectrometry and proteomics

May 7, 12:30pm – May 9, 5:00pm 2025. **Interpretation of proteomic experiments in the context of biomolecular networks with INDRA**

Venue:

Northeastern University main campus, West Village H, second floor rooms: 210A/B, 212.
Please see the annotated map.

Lead instructor:

Benjamin Gyori, b.gyori@northeastern.edu

Teaching assistant

Anthony Wu, Klas Karis

Description

Interpreting proteomics data using networks representing cellular mechanisms is a powerful approach to gaining actionable insights. In this course, we first introduce software for annotating data sets using identifiers, a crucial step to prepare data for analysis. Next, we learn about widely used pathway databases, and perform enrichment against networks from such databases using the iQuery tool. We then learn about the INDRA system which automatically assembles mechanisms into networks from both pathway databases and biomedical literature, and learn about the INDRA Database, the INDRA Network Search tool, and the INDRA Biomedical Discovery Engine which each facilitate gaining insights from data and generating hypotheses based on large-scale networks. Most of the course will be conducted as hands-on exercises using sample data sets, making use of web-based UIs as well as Python libraries. Software tools used in this course are available under permissive open-source licenses. Note that some of the knowledge sources such as pathway databases may require licenses for use outside this course in a commercial setting.

References

Some of the tools used in this course are described in the following publications:

- “Automated assembly of molecular mechanisms at scale from text mining and curated databases” https://doi.org/10.15252/msb.202211325
- “NDEx IQuery: a multi-method network gene set analysis leveraging the Network Data Exchange” https://doi.org/10.1093/bioinformatics/btad118
- “Unifying the Identification of Biomedical Entities with the Bioregistry” https://doi.org/10.1038/s41597-022-01807-3
- “Gilda: biomedical entity text normalization with machine-learned disambiguation as a service” https://doi.org/10.1093/bioadv/vbac034
- “Assembling a corpus of phosphoproteomic annotations using ProtMapper to normalize site information from databases and text mining” https://doi.org/10.1101/822668

Relevant websites and software repositories include:

- INDRA DB at https://db.indra.bio
- INDRA Network Search at https://network.indra.bio
- INDRA Biomedical Discovery Engine at https://discovery.indra.bio
- Gilda grounding service at https://grounding.indra.bio
- iQuery network-based enrichment service at https://www.ndexbio.org/iquery/
- Protmapper at https://github.com/gyorilab/protmapper

Acknowledgements

The development of these tools is supported by the DARPA grant HR00112220036 and CZI grant 2023-329850.

Target audience

Target audience are computational scientists (including bioinformaticians, computational biologists, data scientists), as well as experimental scientists interested in learning about network-based omics data analysis. Note that several of the tools used in this course require Python programming. Attendees should complete the "Modern software development practice with Python" course before this one.

Tentative schedule

Wednesday, May 7, 2024
12:30 p.m. Registration
1:30 p.m. **Introductory lecture and Q&A**
3:00 p.m. Refreshments
3:30 p.m. **Hands on: Introduction to data annotation using identifiers**
5:00 p.m. Adjourn

Thursday, May 8, 2024
9:00 a.m. **Hands on: interacting with human-curated networks in pathway databases**
10:30 a.m. Refreshments
11:00 a.m. **Hands on: network-based enrichment analysis of proteomics experiments**
12:30 p.m. Lunch on your own
1:30 p.m. **Hands on: exploring literature-derived networks for data interpretation with INDRA**
3:00 p.m. Refreshments
3:30 p.m. **Hands on: causal path finding for data interpretation using INDRA Network Search**
5:00 p.m. Adjourn

Friday, May 9, 2024
9:00 a.m. **Hands on: Uncovering regulatory patterns from experimental observations using the INDRA Biomedical Discovery Engine**
10:30 a.m. Refreshments

11:00 a.m. **Hands on: Creating context-specific mechanistic networks from experimental data and prior knowledge**
12:30 p.m. Lunch on your own
1:30 p.m. **Bring your own data: hands-on guidance on applying these tools to your own data sets**
3:00 p.m. Refreshments
3:30 p.m. **Bring your own data continued, final recap, and Q&A**
5:00 p.m. Wrap-up

Pre-course setup

Data files used for this course can be downloaded from Google Drive here.

Interactive notebooks are available on GitHub at: https://github.com/gyorilab/indra_may_institute. The code can be cloned from GitHub if you have git set up locally, or it can be downloaded as a zip file.

This course uses a combination of web based UIs which don't require special setup, and Python notebooks. Setting up Python:

- On Mac, we recommend using Homebrew: https://docs.python-guide.org/starting/install3/osx/
- On Windows, use the installer at https://www.python.org/downloads/release/python-31210/

Once Python is set up, the dependencies for running the notebooks can be installed using the requirements.txt in the code repository as `pip install -r requirements.txt`.

Slides are available at: https://www.dropbox.com/scl/fi/dwslymjo9eyi1eu1kge4x/2025-05-07_may_institute.pptx?rlkey=kl3buyxutlm4k58gye86gaeve&dl=0

Link to participant lunch order pickup

If you would like us to pick up your prepaid lunch order from Tatte or Anna's Taqueria, fill this form: https://forms.gle/5E8apFjvjW3qZhgQA

Course evaluation

Please help us improve the program in the future by filling in this form: https://forms.gle/s6bbgAUGhdDy6GjG9