Data Science and Cheminformatics Tools to Support Exposomics and Metabolomics

Dinesh Kumar Barupal
Assistant Professor
Department of Environmental Medicine and Public Health
Icahn School of Medicine at Mt Sinai
New York, USA
Overview

- Opportunities in non-targeted analyses (NTA)
- Chemical to publication mapping
- Prioritizing chemicals for hazard assessments
Opportunities in non-targeted analyses
NTA for the disease prevention

Discussion point: How to prioritize NTA assays for identifying risk factors or discovering new metabolic reactions?

Low signal prevalence is important

Discussion point: NTA studies should **avoid thresholding** signal prevalence so we don’t miss rare signals with high PAFs.
Discussion point: Raw LC/GC MS raw from NTA studies should be indexed in enterprise databases to support basic queries as well as advanced signal processing.

Annotation capacity building needs an integrated approach

**Sample processing**
- compound enrichment
- derivatization

**Mass spectral library**
- coverage and diversity
- in-silico prediction
- spectra quality

**Annotation workflows**
- scalable
- automated
- filtering criteria
- multiple-evidences
- context-dependent

**Discussion point**: How to rank experimental and in-silico evidences for a peak annotation?
Poor coverage of NTA data in pathway DBs

Discussion points:
1) A background database does not exist for NTA.
2) Assuming a statistical independence of chemicals is false.

Chemical similarity graph for NTA data

Tanimoto = \frac{AB}{(A + B - AB)}

Substructure decomposition for calculations of chemical similarity

Discussion point: How to interpret large-scale network visualization for NTA data?
ChemRICH uses the MeSH ontology

Discussion points:
1) Prioritization of MeSH chemical ontology terms of biomonitoring
2) How to include unidentified metabolites into the set analysis?

- Node color indicate the proportion of node had a positive (red) or negative (blue) association with a phenotype.
- The Kolmogorov–Smirnov was used compute set level p-values (y-axis)

1) A large number of signals (50-95%) remains **unknown**
2) **Slow** signal processing for a large batch of samples
3) **Errors** in peak grouping and deconvolution
4) Correction of retention time **drifts** for large sample sizes
5) Presence of **missing** values
6) **Low** frequency signals are often ignored
7) Presence of **artifacts** and background signals
8) **Issues** with data normalization
9) Challenging biological **interpretation**
10) Ethical issues in data sharing for **sensitive** analytes such as illicit drugs

**Discussion point**: How and when to address these issues in the NTA data processing?
Chemical to Literature Mapping
### Abstract

**Plasma Biomarkers of Inflammation, the Kynurenine Pathway, and Risks of All-Cause, Cancer, and Cardiovascular Disease Mortality: The Hordaland Health Study.**

Zer H. Ueland PM, Uanh A, Fussen SJ, Vollset SF, Nørgård O, Midtun Ø, Theodorsen T, Meyer K, Tall G.

**Abstract**

We aimed to evaluate 10 biomarkers related to inflammation and the kynurenine pathway, including neopterin, kynurenine tryptophan ratio, C-reactive protein, tryptophan, and 6 kynurenines, as potential predictors of all-cause and cause-specific mortality in a general population sample. The study cohort was participants involved in a community-based Norwegian study, the Hordaland Health Study (HUSK). We used Cox proportional hazards models to assess associations of the biomarkers with all-cause mortality and competing-risk models for cause-specific mortality. Of the 7,015 participants, 1,466 deaths were recorded after a median follow-up time of 14 years (1998-2012). Plasma levels of inflammatory markers (neopterin, kynurenine-tryptophan ratio, and C-reactive protein), anthranilic acid, and 3-hydroxykynurenine were positively associated with all-cause mortality, and tryptophan and anthranilic acid were inversely associated. Multivariate-adjusted hazard ratios for the highest (versus lowest) quartiles of the biomarkers were 1.19-1.60 for positive associations and 0.73-0.87 for negative associations. All of the inflammatory markers and most kynurenines, except kynurenic acid and 3-hydroxyanthranilic acid, were associated with cardiovascular disease (CVD) mortality. In this general population, plasma biomarkers of inflammation and kynurenines were associated with risk of all-cause, cancer, and CVD mortality. Associations were stronger for CVD mortality than for mortality due to cancer or other causes.

### Table 1. Descriptive statistics for study participants

**Table**

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Median (Range)</th>
<th>% Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDE-26835</td>
<td>3.22 (0.7-25.85)</td>
<td>100</td>
</tr>
<tr>
<td>BDE-47</td>
<td>46.57 (10.8-90.84)</td>
<td>97</td>
</tr>
<tr>
<td>BDE-99</td>
<td>9.19 (0.5-60.63)</td>
<td>89</td>
</tr>
<tr>
<td>BDE-100</td>
<td>9.96 (0.5-93.94)</td>
<td>97</td>
</tr>
<tr>
<td>BDE-153</td>
<td>59.6 (20.31-180.91)</td>
<td>100</td>
</tr>
<tr>
<td>BDE-209</td>
<td>18.39 (20.31-204.22)</td>
<td>97</td>
</tr>
</tbody>
</table>

### Supplementary data

- 4-methylcatechol sulfate: Xenobiotics
- 4-methylglucuronic acid: Xenobiotics
- 4-vinylphenol sulfate: Xenobiotics
- benzene: Xenobiotics
- catechol sulfate: Xenobiotics
- goyalac acid: Xenobiotics
- hippurate: Xenobiotics
- methyl-4-hydroxybenzoate sulfate: Xenobiotics
- o-cresol sulfate: Xenobiotics
- p-cresol sulfate: Xenobiotics
- propyl-4-hydroxybenzoate sulfate: Xenobiotics
- 1,2,3-benzenetriol sulfate (2): Xenobiotics
- 2,3-Methylendioxy(6-tert-butyl-p-cresol): Xenobiotics
- 2-aminophenol sulfate: Xenobiotics
- 2-methoxyresorcinol sulfate: Xenobiotics
- 3-acetylphenol sulfate: Xenobiotics
- 3-hydroxypryidine sulfate: Xenobiotics
- 4-hydroxychlorothroid: Xenobiotics
- 4-methylbenzenesulfonate: Xenobiotics
- 6-hydroxyindole sulfate: Xenobiotics
- benzoylacetamide*: Xenobiotics
- bromine: Xenobiotics

### Figure

![Figure](PMID:16169030)

**In-paragraph**

“A lagging in serum folate concentrations was moderately associated with increased risk of UCC (OR: 1.18; 95% CI: 0.98–1.43)” - PMC6899898

### Discussion point

How far we can go in developing a chemical to publication mapping resource?
Discussion points: 1) How publication count for a chemical can improve peak annotation in NTA? 2) How to cover compounds that are not reported in an abstract text?

Number of identified compounds in a blood metabolomics dataset by metabolon.

**Discussion point:** We should ensure that existing mass spectral libraries have EI/ESI spectra for these compounds.
Prioritizing chemicals for hazard assessments
Most exposures are chemicals

Mechanisms are in place to identify, monitor and regulate exposure to a specific chemical.

**Evidence based hazard assessments**

<table>
<thead>
<tr>
<th>Evidence in Experimental Animals</th>
<th>Group 1 (120 agents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient</td>
<td>1 strong evidence in exposed humans</td>
</tr>
<tr>
<td>Limited</td>
<td>2A belongs to a mechanistic class where other members are classified in Groups 1 or 2A</td>
</tr>
<tr>
<td>Inadequate</td>
<td>2B with supporting evidence from mechanistic and other relevant data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence in Humans</th>
<th>Group 2A</th>
<th>Group 2B</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient</td>
<td>1 strong evidence in exposed humans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited</td>
<td>2A belongs to a mechanistic class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate</td>
<td>2B with strong evidence from mechanistic and other relevant data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **ESLC**: Evidence suggesting lack of carcinogenicity

Text mining for prioritizing chemicals

- Individual pesticides are represented as nodes on the chemical similarity maps. The node size is proportional to the number of publications overall on a pesticide and cancer: larger nodes represent more publications.

- The node border width represents the number of publications on epidemiology, cancer and the pesticide: a thicker border represents more papers. The node color, ranging from yellow to red, also represents the number of publications on epidemiology, cancer and the pesticide: red represents the highest count of publications.

- The node shape indicates whether results for a particular pesticide were available in the ToxRefDB database (circle = absent; square = present).

- The node border color represents the KEGG pesticide classification: green = Organochlorine, navy blue = Phenoxy, light blue = Organophosphorus, white = Others.

Discussion points:
1) Chemically similar agents can be evaluated together as they might have similar toxicological profile.
2) We can develop a similar approach for the California Biomonitoring program chemical list?

IARC Monographs on the Evaluation of Carcinogenic Risks to Humans

Meeting 112: Some Organophosphate Insecticides and Herbicides: Diclofenac, Glyphosate, Malathion, Parathion, and Tetraclorvinphos
(3-10 March 2015)

Meeting 113: Some Organochlorine Insecticides and Some Chlorophenoxy Herbicides (2-9 June 2015)

Conclusions

- Non-targeted analysis has a great potential for detecting high-priority chemicals for exposome research in biospecimens.

- However, a proper combination of analytical chemistry and data science needs to be planned ahead.

- Indexing raw data into enterprise databases and avoiding a signal prevalence threshold are needed for exposomics.

- Computational text mining can improve the prioritization process by linking chemicals to publications.

- Interpretational bias remains a major challenges in mining NTA.
Thanks to current and former collaborators at:

Icahn School of Medicine at Mount Sinai

Institute for Exposomic Research

International Agency for Research on Cancer

World Health Organization

University of California

Special thanks to NIH for funding these initiatives

NIH Common Fund Metabolomics Program

HHEAR Human Health Exposure Analysis Resource