How many fingers does a compound have?
The various ways to define molecular similarity.
Eugen Lounkine, Novartis Institutes for Biomedical Research
ACS meeting Boston, August 18 2015
Overview

- Projecting bioactivity onto chemical fingerprints
  - Activity-weighted molecular clustering

- Biological molecular fingerprints
  - HTSFP and Bioturbo

- Literature molecular fingerprints
  - What are my compounds “about”?

- Heterogeneous similarity networks
Activity-aware molecular similarity

Translating Bayesian weights to molecular fingerprints

- Molecular fingerprint bits are weighted using Naive Bayesian model
- Quantitative comparison of activity-characteristic molecular cores

Activity-aware clustering

Compounds cluster around automatically identified bioactive cores

- Weighted clusters often spread across many conventional clusters
- Large clusters that both methods agree on

HTS Fingerprints

And Bioturbo similarity searching

HTS-FPs describe compounds in terms of their biological activity

<table>
<thead>
<tr>
<th>PROBE</th>
<th>ECFP4</th>
<th>OVERLAP</th>
<th>HTSfp</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epothilone</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>GRNN1</td>
<td></td>
<td></td>
<td>0.03</td>
</tr>
<tr>
<td>Dizocilpine</td>
<td></td>
<td></td>
<td>0.03</td>
</tr>
</tbody>
</table>


Literature fingerprint

*What are my hits about? – Exploratory annotation of phenotypic hits.*

**Compounds**

- Text mining (Abstracts) and curated sources. E.g., ChEMBL provides reference for each compound activity

**Pubmed Articles**

- What articles mention compound?

**MeSH terms**

- Filter for informative terms

**Hits**

- Naive Bayes model, TFIDF

**Reference compound**

- Similarity search

**Collection of terms**

- Similarity search, TFIDF

**What are the themes around my hits?**

- Compounds mentioned in same biological / clinical context

**Do we have tools compounds for phenotype of interest?**

- Collection of terms
Fingerprints without compounds

A signature is a fingerprint template endowed with meaning

Within one project

<table>
<thead>
<tr>
<th>Raw data</th>
<th>Raw data</th>
<th>Raw data</th>
</tr>
</thead>
<tbody>
<tr>
<td>QC&amp;Norm</td>
<td>QC&amp;Norm</td>
<td>QC&amp;Norm</td>
</tr>
</tbody>
</table>

Specific question 1

Specific question 2

Fingerprint

Meaning of particular fingerprint

Signature 1 (e.g., toxicity)

Signature 2 (e.g., AC)

Hits are defined by comparing a fingerprint with a signature

Data mining across projects

Ontology

Encode commonalities between readouts

Group by readout metadata

Fingerprint

Encode meaning of signature

Signature DB

Phenotype DB

Expansion

Fingerprint

Similarity metric

Fingerprint

Scoring

Fingerprint

Similarity or model

Signature

Quantitative?

Meaning

Signature

Encode meaning of signature

Meaning

Signature DB

Phenotype DB
Graph structure integrates distinct relationships

Turbo-similarity across molecular representations

- Sparse fingerprints benefit from turbo-expansion
  - Unlike in chemical fingerprints, HTS fingerprints have unknown bit positions

- Biological similarity enables MOA elucidation for mixtures / compounds with unknown structure
  - Natural product extracts $\rightarrow$ HTSFP $\rightarrow$ chemical analogs $\rightarrow$ targets

- Distinct relationships can be put on equal footing in graph structure
Heterogeneous similarity network

Seeds

Selection of painkillers, connected only by literature relationships
Heterogeneous similarity network

First expansion

All compounds have neighbors, but they come from distinct similarity measures.

**Literature**: more painkillers and Quercetin / Doxorubicin

**Chemical**: Diclofenac analogs

**HTSFP**: Ibuprofen similars
Heterogeneous similarity network

Connecting neighbors

The neighbors themselves are connected among each other, sometimes with more than one method.
Heterogeneous similarity network

**Adding interesting neighbors of neighbors (connectors)**

- **Morphine analogs**
- **NSAIDs**

**Oncology pain management**

- Literature
- Chemical
- Biological

Distinct clusters emerging
Heterogeneous similarity network

Identify pairs that are connected by more than one method

These voting schemes are intuitive in graphs, and harder to formalize in conventional approaches.

Alternatively, one can use degree, number of distinct edge types, etc.
Heterogeneous similarity network

Use flow algorithm to distribute scores

Standard graph neighborhood scoring algorithm. Intuitive to carry out and visualize, easily scalable.
Glitazones

![Graph showing Glitazones network]

- Rosiglitazone
- Pioglitazone
- Troglitazone

Literature
Chemical
Biological
Antidepressants
Statins and Warfarin

First expansion

Connected
With increasing knowledge, compounds can be encoded with fingerprints reflecting biology and phenotypes associated with them.

Similarity can be used to elucidate **compound themes**, in addition to classical virtual screening.

Sparse fingerprints bring with them challenges that can be addressed using **turbo-similarity** approaches.

**Graph representations** provide a unique opportunity to combine distinct similarity domains in an interoperable way.
Acknowledgements

**NIBR / iSLD**
- Anne Mai Wassermann*
- Stephan Reiling
- Miguel Camargo
- Ansgar Schuffenhauer
- Xian Zhang
- Meir Glick*
- John W Davies

**NIBR / DMP**
- Yuan Wang
- Ben Cornett
- Florian Nigsch
- Jeremy Jenkins