Elementary, my dear Watson: Fingerprint search in molecular structure databases

Sebastian Böcker
Friedrich-Schiller-Universität Jena
We are talking about small molecules
Metabolites: Why care?

- metabolites closest to phenotype
- majority of drugs derived from natural products (that is, metabolites)
- vast majority of medical biomarker assays target metabolites
- vast majority of (plant, animal and human) diseases have a non-genetic cause

Stolen from a talk by David Wishart
A simple question

- Given the **tandem mass spectrum** of a compound, can we find it – in a molecular structure database?
**Why is this so complicated?**

- **SEQUEST**: Searching peptide sequence databases since 1994
- but metabolites are different

<table>
<thead>
<tr>
<th>proteins/peptides</th>
<th>metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td>molecules are...</td>
<td>structurally similar</td>
</tr>
<tr>
<td>genome information tells you...</td>
<td>everything but PTMs</td>
</tr>
<tr>
<td>molecules fragment...</td>
<td>at one fixed energy</td>
</tr>
<tr>
<td>fragmentation is...</td>
<td>“easily” predictable</td>
</tr>
</tbody>
</table>
The classic: rule-based prediction
Rule-based prediction

[Hill, ..., Grant, *Anal Chem* 2008]
MetFrag: combinatorial fragmentation
MetFrag (Neumann group)

use explained peaks to compute some score

[Wolf, ..., Neumann, *BMC Bioinf* 2010]
MetFrag web interface

In silico fragmentation for computer assisted identification of metabolite mass spectra
Competitive Fragmentation Modelling
Competitive fragmentation modelling

[Allen, Greiner, Wishart, Metabolomics 2014]
FingerID: predicting fingerprints
Molecular fingerprints

- when are two molecules “similar”?

- encode presence/absence of substructures in binary vector

- different types: MACCS, FP1 – FP4, PubChem, ...

- used for: virtual screening, estimating chemical similarity, ...
Can you **predict** the molecular fingerprint of an unknown compound directly from the tandem MS data?
FingerID (Rousu group)

[Heinonen, ..., Rousu, *Bioinformatics* 2012]
Precision, Recall, F-score

\[
\text{precision} = \frac{2}{2 + 2} = \frac{1}{2}
\]

\[
\text{recall} = \frac{2}{2 + 3} = \frac{2}{5}
\]

\[
\text{F-score} = \frac{2}{2/1 + 5/2} = \frac{4}{9}
\]
FingerID (Rousu group)

F-score from 0.49 to 0.67
Our method
1st step

Fragmentation trees

Sebastian Böcker
Chair for Bioinformatics, Friedrich-Schiller-University Jena
Our fragmentation trees

- tandem MS, **multiple MS not required**
- fully **automated** method
- best explains experimental data
- combinatorial optimization

MS\(n\) not required
Our fragmentation trees

- Böcker and Rasche, *Bioinf* 2008
- Dührkop and Böcker, unpublished
- White *et al.*, unpublished
Our fragmentation trees

- Böcker and Rasche, *Bioinf* 2008
- Dührkop and Böcker, unpublished
- White *et al.*, unpublished
Molecular formula prediction

Percentage of correct molecular formula predictions vs. rank.
Molecular formulas with isotope patterns

Percentage of correct molecular formula predictions vs. rank for different methods and isotope filtering thresholds.

- METLIN, 5% isotope filtering
- METLIN, 10% isotope filtering
- Agilent, 5% isotope filtering
- Agilent, 10% isotope filtering
**CASMI challenge 2013**

- we got **12 out of 14** molecular formulas correct
- **2nd place**, winner manually analyzed the challenges
- we were the only contestants that **did not search PubChem**, but instead considered **all possible molecular formulas**
2nd step
Fingerprints
Machine Learning

Sebastian Böcker
Chair for Bioinformatics, Friedrich-Schiller-University Jena

28
Support Vector Machines

- separate cats and dogs via features (weight, height, ...)
- map features so that linear separation is possible
Use fragmentation trees as input

FT structure kernels
- nodes binary
- nodes intensity
- loss binary
- loss count
- loss intensity
- root loss binary
- root loss intensity
- common path counting
- common paths of length 2
- common paths with peak scores
- common subtree counting

[Shen et al., ISMB 2014]

Sebastian Böcker  Chair for Bioinformatics, Friedrich-Schiller-University Jena
One example: Common Path Counting kernel

- tree kernels measure the **structural similarity of two fragmentation trees**
Multiple kernel learning

- combine predictions of all 12 kernels into one new kernel
- ALIGNF: Learn weights by comparing kernels to target kernel

\[
\hat{\beta}(K, K') = \frac{\langle K_c, K'_c \rangle_F}{\|K_c\|_F \|K'_c\|_F}.
\]
Was it all worth it?
Use fragmentation trees as input

F-score from 0.64 to 0.73

FT structure kernels
- loss binary
- loss intensity
- root loss binary
- root loss intensity
- common path counting
- common paths of length 2
- common paths with peak scores
- common subtree counting

[Shen et al., ISMB 2014]
3rd step
Searching PubChem
Searching a molecular structure database

• retrieve all compounds with correct molecular formula

• for each compound in the database, we know its structure and, hence, we know its correct molecular fingerprint

• compare predicted fingerprint to those of all candidates

• simplest score is unit costs

\[
\begin{array}{cccccccccccc}
\text{predicted fingerprint} & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
\text{candidate 1853} & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
\text{differences: 4} & \checkmark & \checkmark & \times & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \times & \checkmark & \checkmark & \times & \checkmark & \checkmark & \checkmark
\end{array}
\]

• rank candidates according to score

\[
\begin{array}{cccccccc}
\text{rank} & 1^{\text{st}} & 2^{\text{nd}} & 3^{\text{rd}} & 4^{\text{th}} & \ldots \\
\text{candidate} & 765 & 2271 & 1853 & 61 & \ldots \\
\text{differences} & 1 & 3 & 4 & 7 & \ldots
\end{array}
\]
Evaluation setup

• “Evaluation is the process of judging something or someone based on a set of standards.”

• retrieve all compounds with correct molecular formula

• for each compound in the evaluation dataset, we know its correct molecular structure

• at what position do we find the correct answer? (TOP-\(k\))

• we only evaluate plain structures (no stereochemistry etc)
Intermission: WWW search engines
Training and cross validation data

- GnPS database (UCSD, San Diego)
- Forensic database (Agilent Technologies)
- QTOF MS instruments
- $\approx 2800 + 2200 = 5000$ compounds
- tandem mass spectra (CID) at different frag. energies
- mass accuracy usually $10$ ppm or better
- used to train and evaluate: $10x$ cross validation
Where to search: molecular structure databases

- PubChem compounds that have a citation in PubMed
- plus HMDB, Knapsack, ChEBI, METLIN, contaminants
- total 400,000 compounds

- full PubChem: more than 50 million compounds
Conclusion

• searching in molecular structure dbs using tandem MS data has become an option

• for the complete PubChem dataset (40 million structures) our method currently reaches 35% hits (correct IDs)

Outlook

• better kernels, better scores, better search results

• significances: False Discovery Rates, q-values, p-values

• and much more to come...
Thank you

• GNPS db, Pieter Dorrestein
• Agilent Technologies

Funding

• Deutsche Forschungsgemeinschaft

Credits

Kai Dührkop
Huibin Shen
Marvin Meusel
Juho Rousu

Aalto University, Helsinki, Finland

Thank you for your attention!