Metabolite Identification through Machine Learning

Juho Rousu

Machine Learning Coffee, 23.1.2017
Metabolites

- **Small molecules** (< 1000 Da) inside biological cells, 1000s different types in each living cell,
- **Numerous functions**: energy transport, signaling, building blocks of cells, inhibition/catalysis (drugs)
- **Identification of metabolites is a major bottleneck**
  - **Needed in numerous applications**: in biomedicine, pharmaceuticals, biotechnology, regulatory affairs

Picture: Eric Bach
And the winner is ...

Congratulations to all participants in this contest - this year we welcomed more participants than any CASMI contest before!

Please use the tabs above to navigate to the individual categories for the results of the automatic evaluation and the submission abstracts, as well as long tables with the full details at the (+) links.

Category 1: Best Structural Identification on Natural Products

Dejan Nikolic from the College of Pharmacy, University of Illinois at Chicago won this category with 15 wins, using a manual approach, closely followed by Team Vaniya (UC Davis and Riken CSRS) with 14 wins. Third place goes to Tobias Kind from UC Davis with 12 wins.

Category 2: Best Automatic Structural Identification - In Silico Fragmentation Only

By Team Brouard (Aalto University/FSU Jena) using “TOKR” with 86 gold (wins), closely followed by Team Vaniya with 70 gold. The Category 2 results were very sensitive to the method of declaring the winner - the summary statistics show how close this category was in different aspects!
Machine learning for metabolite identification

Metabolite identification can be seen as a structured prediction problem:

- $\mathcal{X}$: set of MS/MS spectra
- $\mathcal{Y}$: set of molecules
- Training set: $(x_1, y_1), \ldots, (x_\ell, y_\ell) \in \mathcal{X} \times \mathcal{Y}$
- Learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that maps a MS$^2$ spectrum to a molecule
METHOD I: CSI:FINGERID

CSI:Fingerid: Two-step method for metabolite identification

1. From a set of MS/MS spectra (thousands) of known molecules, learn a model to predict molecular fingerprints ➔ Supervised machine learning

2. With predicted fingerprints retrieve candidate molecules from a large (millions of molecules) molecular database ➔ Ranking/information retrieval

Step 1: Learning molecular fingerprints

- Data: Set of (MS/MS spectra, molecule) pairs for training
- Kernel representations of the inputs
  - Similarity of spectra $\rightarrow$ Probability Product kernel (Jebara, 2004)
  - Similarity of fragmentation (fragmentation trees) $\rightarrow$ tree kernels
- Molecular fingerprints as the outputs
- Set of SVM classifiers to predict inputs from the outputs

Outputs: molecular fingerprints

- Describe molecular properties
  - different types atoms, bonds
  - substructures (e.g. aromatic rings)

- Counts or Binary indicators

- Standard sets used by computational chemistry community
  - PubChem fingerprints
  - Klekota-Roth fingerprints
  - Ca. 5000 properties in total

Inputs: Probability product kernel on spectra

- Models spectra as sets of 2D probability distributions (mass, intensity)
- Kernel: all-against-all matching of the distributions

\[ p_X = \frac{1}{\ell_X} \sum_{k=1}^{\ell_X} p_X(k), \quad p_{X'} = \frac{1}{\ell_{X'}} \sum_{k=1}^{\ell_{X'}} p_{X'}(k) \]

\[ K(x; x') = \int_{\mathbb{R}^2} p_X(x)p_{X'}(x)dx. \] (Jebara et al., 2004)
Inputs: Fragmentation trees

- Models of fragmentation of a molecule in MS/MS
  - Nodes ≈ peaks ≈ molecular formula of fragments
  - Edges ≈ losses ≈ putative uncharged fragments

- Trees can be predicted from spectra
  - Also gives us the predicted molecular formula of the unknown metabolite (but not the molecular structure!)
  - We take the predicted trees as input for our method

Inputs: Kernels for fragmentation trees

- Node kernels (picture): count nodes (peaks) with the same molecular formula (colors) in the two trees
- Edge kernels: count edges (losses) with the same molecular formula
- More complex ones, computed using dynamic programming:
  - Path kernels
  - Subtree kernels

Multiple kernel learning

• Idea: instead of trying to select the best kernel, learn combination weights for them

• Several methods have been proposed recent years
  – Centered alignment (ALIGNF, Cortes et al., 2012)
  – Quadratic combination (Li and Sun, 2010)
  – $L_p$-norm regularized combination

• Combined kernel is used in learning the final prediction model (here: SVM predicting fingerprints)

\[
K = W_1 \cdot \text{PPK} + W_2 \cdot \text{NB} + W_3 \cdot \text{NI} + \cdots + W_{12} \cdot \text{CPC}
\]
Overview of fingerprint prediction

Step 2: Scoring and ranking metabolites with fingerprints

- Goal: match the predicted fingerprints to fingerprints of known molecules in large database (e.g. Pubchem)
- Take uncertainty in the fingerprint predictions into account
  - Different scoring schemes for fingerprints
  - e.g. SVM confidence of prediction, error rate, …
Experimental results

Percentage of correctly identified 2D structures:


34.4% vs. 13.8%

2.5-fold increase
METHOD 2: INPUT-OUTPUT KERNEL REGRESSION

Input Output Kernel Regression

Decomposition of the regression problem in two tasks:

1. Estimation of the output feature map: $h : \mathcal{X} \rightarrow \mathcal{F}_y$
2. Computation of the pre-image: $g : \mathcal{F}_y \rightarrow \mathcal{Y} \approx \phi_y^{-1}$
1) Estimation of the output feature map

First step: approximation of the function $h : \mathcal{X} \to \mathcal{F}_y$. 

Kernel Ridge Regression generalized to the case of vector space images.

- **Optimization problem:**

  \[
  \text{arg min}_{h \in \mathcal{H}} \sum_{i=1}^{\ell} \| h(x_i) - \phi_y(y_i) \|_{\mathcal{F}_y}^2 + \lambda \| h \|_{\mathcal{H}}^2, \quad \lambda > 0.
  \]

- **Model:** 

  \[ h(x) = \sum_{i=1}^{\ell} \alpha_i k_x(x_i, x), \quad \alpha_i \in \mathcal{F}_y, \text{ where } k_x \text{ is an input kernel}. \]

- **Solution:** 

  \[ \hat{h}(x) = \Phi_{y_\ell} (\lambda I_{\ell \ell} + K_{x_\ell})^{-1} k_{x_\ell}^x, \]

  where $\Phi_{y_\ell} = (\phi_y(y_1), \ldots, \phi_y(y_\ell))$ and $k_{x_\ell}^x = (k_x(x_1, x), \ldots, k_x(x_\ell, x))^T$. 

2) Pre-image step

To predict the output molecule $f(x)$ associated to the spectra $x \in \mathcal{X}$, we must determine the pre-image of $h(x)$ by $\phi_y$.

- Search the space of molecules for one with image nearest to $h(x)$:

$$\hat{f}(x) = \arg \min_{y \in \mathcal{Y}^*} ||\hat{h}(x) - \phi_y(y)||^2_{\mathcal{F}_y}$$
2) Pre-image step

We replace \( \hat{h}(x) \) by the solution of the optimization problem:

\[
\hat{f}(x) = \arg \max_{y \in \mathcal{Y}^*} (k_{x_\ell}^x)^T (\lambda I_\ell + K_{x_\ell})^{-1} k_{y_\ell}^y.
\]

Remarks:

- We do not need to know explicitly the training output feature vectors to evaluate \( \hat{f}(x) \).
- \( \mathcal{Y}^* \): set of candidate molecules from molecular databases such as PubChem or KEGG.
- \( \mathcal{Y}^* \) can be filtered using the mass of the unknown molecule or its molecular formula if already known.
Comparison with CSI:FingerID: running times

- 4138 training compounds (GNPS) / 625 test compounds (MassBank)
- Fix the values of the parameters
- The computation of the fragmentation trees, input kernels and fingerprints was not taken into account

<table>
<thead>
<tr>
<th></th>
<th>Training time</th>
<th>Test time</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSI:FingerID</td>
<td>82 h 28 min 23 s</td>
<td>1 h 11 min 31 s</td>
</tr>
<tr>
<td>IOKR linear</td>
<td>42 s</td>
<td>1 min 15 s</td>
</tr>
<tr>
<td>IOKR polynomial</td>
<td>38 s</td>
<td>21 min 58 s</td>
</tr>
<tr>
<td>IOKR Gaussian</td>
<td>41 s</td>
<td>33 min 15 s</td>
</tr>
</tbody>
</table>

- IOKR is $\approx$7000 times faster to train than CSI:FingerID because CSI:FingerID needs to train 2765 SVMs (one for each molecular property).

- **IOKR linear**: avoid kernel computations in the pre-image step by computing explicitly the output feature vectors.
Comparison with CSI:FingerID

<table>
<thead>
<tr>
<th>Method</th>
<th>MKL</th>
<th>Top 1</th>
<th>Top 10</th>
<th>Top 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSI:FingerID unit</td>
<td>ALIGNF</td>
<td>24.82</td>
<td>60.47</td>
<td>68.2</td>
</tr>
<tr>
<td>CSI:FingerID mod Platt</td>
<td>ALIGNF</td>
<td>28.84</td>
<td>66.07</td>
<td>73.07</td>
</tr>
<tr>
<td>IOKR linear</td>
<td>UNIMKL</td>
<td>30.02</td>
<td>66.05</td>
<td>73.66</td>
</tr>
<tr>
<td>IOKR Gaussian</td>
<td>UNIMKL</td>
<td>30.66</td>
<td>67.94</td>
<td>75.00</td>
</tr>
</tbody>
</table>

![Comparison graph with CSI:FingerID](image-url)
Summary and future work

• Metabolite identification is an important problem in molecular biology
• Machine learning a key technology behind recent progress in untargeted metabolite identification
• Our current and future work includes
  – Improved metabolite identification accuracies through IOKR and related approaches
  – Identification of novel metabolites; calls for a combinatorial search in molecular spaces
  – Joint identification of metabolites
  – Applications in molecular biology, biomedicine and nutrition