Metabolite Identification through Machine Learning

Juho Rousu, Associate Professor
Helsinki Institute for Information Technology HIIT
Aalto University, Finland

University of Helsinki, December 2015
Metabolites

- Small molecules inside biological cells, 1000s different types in each living cell
- Functions: energy transport, signaling, building blocks of cells, inhibition/catalysis (drugs)
- Numerous applications in biomedicine, pharmaceuticals, biotechnology
- Identification of metabolites is a major bottleneck
Metabolite identification from Tandem Mass Spectrometric (MS/MS) data

- MS/MS fragments the ionized metabolite
- Resulting daughter ion spectrum has peaks corresponding to molecular fragments
- Our task: predict the metabolite from the daughter ion spectrum
Metabolite identification through machine learning

Two step approach:

1. From a set of MS/MS spectra \( (x = \text{structured input}) \), learn a model to predict molecular fingerprints \( (y = \text{multilabel output}) \).

2. With predicted fingerprints retrieve candidate molecules from a large molecular database.
Building blocks

- Data: Set of (MS/MS spectra, molecule) pairs for training
- Kernel representations of the inputs
- Molecular fingerprints as the outputs
- Learning algorithm to predict inputs from the outputs
Kernel methods: key characteristics

- **Embedding:** Data items $z$ are embedded into a feature space via a non-linear feature map $\varphi(z)$; potentially very high dimensional

- **Linear models:** are built for the patterns in the feature space (typical form: $w^T \varphi(z)$); efficient to find the optimal model, convex optimization

- **Kernel trick:** Algorithms work with kernels, inner products of feature vectors $K(x, z) = \langle \varphi(z), \varphi(z) \rangle$ rather than the original features $\varphi(z)$; side-step the efficiency problems of high-dimensionality

- **Regularized learning:** To avoid overfitting, large feature weights are penalized, separation by large margin is favoured
Kernel = similarity metric

• A kernel function is an inner product (in a Hilbert space)
• If $\varphi(x)$ is a feature vector describing object $x$, the following is called the linear kernel
  $$K(x, z) = \varphi(x)^T\varphi(z) = \sum_j \varphi_j(x)\varphi_j(z)$$
• Geometric interpretation of the linear kernel: cosine angle between two normalized feature vectors
• Non-linear kernels enable learning complex feature spaces with out extra computational cost:
  – Polynomial kernel: $K_{\text{poly}}(x,z) = (K(x,z) + c)^d$
  – Gaussian kernel: $K_{\text{Gaussian}}(x,z) = \exp(||\varphi(x) - \varphi(z)||^2/2\sigma^2)$
  – ...
Kernels for MS/MS Spectra

- MS/MS spectra consists 2D information (mass/charge, intensity)
- Simple approach (ok, but not perfect)
  - divide m/z range into bins, each bin will give a feature \( \phi_j \)
  - take peak intensity in spectrum \( x \) in a bin the feature value \( \phi_j(x) \)
  - Use linear kernel of the feature vectors
- Problems:
  - Noise arising from too wide bins
  - Mass error causes alignment errors if bins too narrow
Probability product kernel

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

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

\[ K(X, X') = \int_{\mathbb{R}^2} p_X(x)p_{X'}(x)dx. \quad (\text{Jebara et al., 2004}) \]
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
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 (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)

\[ K = W_1 \text{PPK} + W_2 \text{NB} + W_3 \text{NI} + \ldots + W_{12} \text{CPC} \]
Kernel Alignment based MKL (ALIGNF)

- Kernel Alignment = Normalized Frobenius inner product of centered kernel matrices
- Weight of input kernel = alignment to the target
  \[ K'_c = \text{fingerprint kernel} \]

\[
K_c(x,x') = (\Phi(x) - \frac{1}{m}\mathbb{E}[\Phi])^T(\Phi(x') - \frac{1}{m}\mathbb{E}[\Phi])
\]

\[
K_c = \left[ I - \frac{11^T}{m} \right] K \left[ I - \frac{11^T}{m} \right].
\]

\[
\hat{\rho}(K,K') = \frac{\langle K_c, K'_c \rangle_F}{\|K_c\|_F \|K'_c\|_F}.
\]
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
  - OpenBabel fingerprints
  - PubChem fingerprints
Multiple kernel learning

Kernels

Multiple Kernel Learning (MKL)
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
Scoring schemes: fingerprint weights

- Uniform scoring: \( w_j = 1 \)
- Accuracy scoring: \( w_j \sim \text{cross-validation accuracy} \)
- Maximum likelihood scoring: \( w_j \sim \log P(y_j, \hat{y}_j) \)
- Platt scoring: \( w_j \sim \exp(a f(x) + b)^{-1} \), where \( f(x) \) is the SVM margin

<table>
<thead>
<tr>
<th>candidate y</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted ŷ</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>weight w</td>
<td>0.1</td>
<td>1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.8</td>
<td>0.9</td>
<td>0.2</td>
<td>0.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>
For each unknown metabolite, we score and rank candidate identifications.

- 80% of spectra have correct hit at top 10.
- ALIGNF (Cortes et al. 2012) is the best MKL method here.

Performance of fingerprint scoring methods

- Uniform weights and accuracy weights behave similarly
- Maximum likelihood scores in the middle
- Platt scoring is the best
  - Heuristic modification improves Platt a bit
Metabolite identification performance

- Plotting the fraction of the MS/MS spectra that have correct Pubchem molecule in top k
- 33% in top 1, 66% in top 10 – out of millions of molecules in Pubchem
- Better than the competition by a large margin (pun intended!)
Fingerprint prediction performance w.r.t training data size

- Fingerprint prediction accuracy (blue) and F1 score (green)
- Varying training set size (100% ~3000 molecules)
- More data = better fingerprint predictions

Related work

Traditional methods rely on reference MS/MS database.

A multi-label prediction task rely on molecular fingerprint prediction (Heinonen et al., 2012).
Metabolite identification performance w.r.t training set size / fingerprints

- Proportion of correctly identified molecules (top 1 rank) as function of
  - number of fingerprints (100% ~ 2800) (green)
  - training set size (100% ~ 2800) (blue)

- More training data = better identifications
Summary and future work

• Metabolite identification is an important problem in molecular biology
• Significant progress in recent years in automatic identification of metabolites
• Machine learning a key technology behind recent progress
• Future work includes
  – One-step metabolite identification through structured prediction
  – Identification of novel metabolites; calls for a combinatorial search in molecular spaces
  – Joint identification of metabolites
Acknowledgements

KEPACO research group
• Dr. Celine Brouard
• MSc Huibin Shen
• Mr Eric Bach

Friedrich Schiller Universität Jena
• Prof. Sebastian Böcker
• MSc Kai Dührkop

Funding
Academy of Finland (MIDAS grant)

http://research.ics.aalto.fi/kepaco/
References


