Machine learning for molecular data

Juho Rousu
Computational Systems Biology & Bioinformatics group
Dept. of Computer Science, University of Helsinki

Algodan seminar, October 28, 2011

Current Research @ CSBB

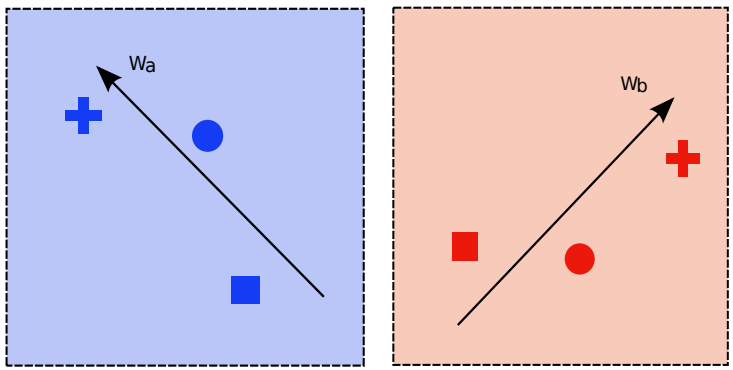
- **Machine learning for biomarker discovery (JR, collaboration with UCL/NIMR)**
- **Metabolite fingerprint prediction form MS/MS data (Markus Heinonen, Huibin Shen, JR, collaboration with ETH Zurich)**
- **Drug bioactivity prediction (Hongyu Su, Markus Heinonen, JR)**
- Kernels for molecular and reaction graphs (Markus Heinonen, JR, Niko Välimaki, Veli Mäkinen)
- Metabolic reconstruction and pathway analysis (GEOBIOINFO project, Esa Pitkänen, Yvonne Herrmann)

Biomarker discovery via sparse canonical correlations

- In biomarker discovery, one is concerned of finding a small set of features that are predictive of the condition of interest (here: tuberculosis)
- Supervised approaches (assume target classification known):
  - Feature selection with classification learning (vast literature)
  - ℓ_1-regularized learning (e.g. LASSO family)
- Here we consider an unsupervised scenario, where we have two paired datasets: proteomics and clinical profiles, but we lack the diagnostic labels at learning time.
- Sparse canonical correlation analysis (SCCA) is the method of choice

Biomarker discovery via sparse canonical correlations

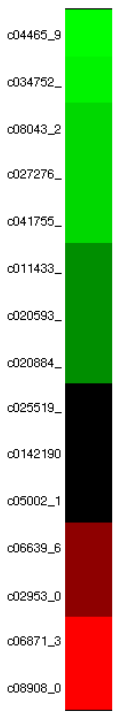
- The first view is represented by feature vector: $score_a(x) = w_a^T \phi_a(x)$
- The second view is represented by a kernel: $score_b(x) = \sum_i \beta_i K_b(x, x_i)$
- Learning aims to minimize the discrepancy between the two views
- The weights in the *first* view are penalized by ℓ_1-norm $||\mathbf{w}_a||_1 = \sum_j |w_j|$ to induce sparse weight vector (feature selection)

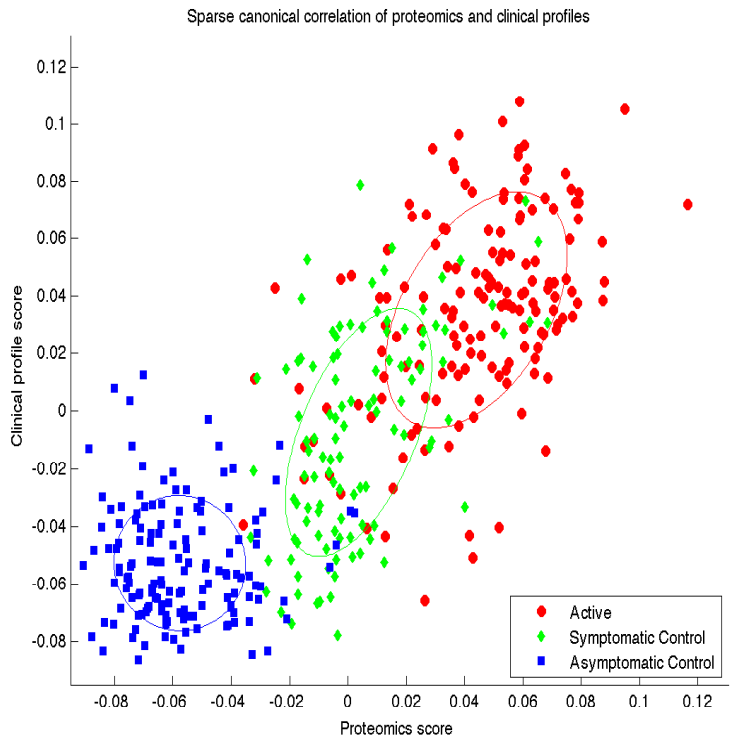


(Hardoon, Shawe-Taylor, 2011)

Biomarker discovery via sparse canonical correlations

- Heatmap of extracted proteomics features (right), corresponding to non-zero coefficients.
- Correlation of the projection direction proteomics and clinical views.
- Diagnostic labels have been inputed in postprocessing (not used in training)

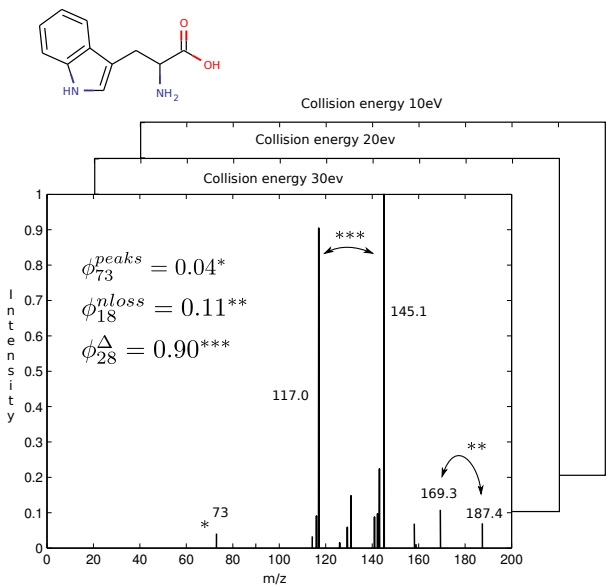


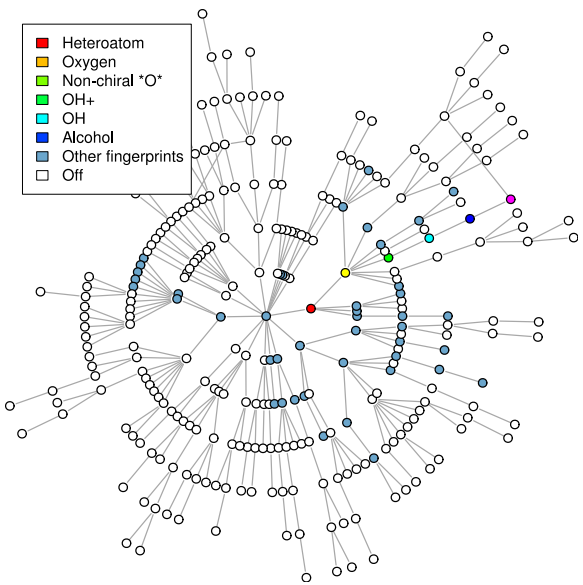


(Rousu, Agranoff, Shawe-Taylor, Fernandez-Reyes. Proc. MLSB-2011, Vienna, 2011)

Metabolite fingerprint prediction from MS/MS data

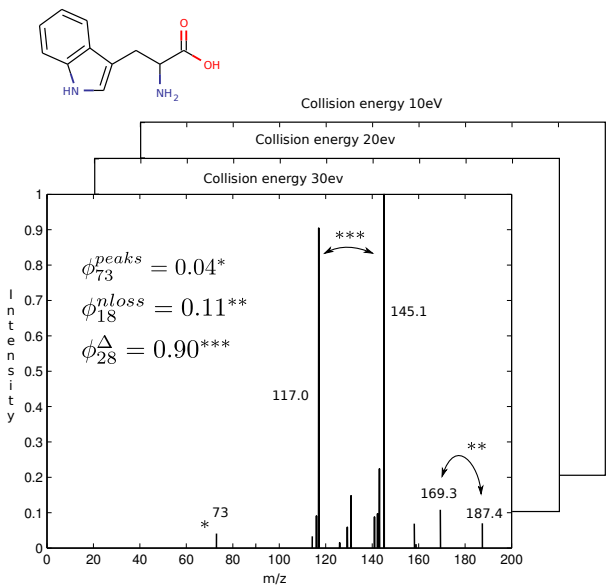
- Task: given a tandem MS spectrum of a small molecule, predict properties (the fingerprints) of the molecule
- Motivation: First step towards de novo metabolite identification, a major bottleneck in metabolomics
- Collaboration with ETH Zurich (N. Zamboni) and IPB Halle (S. Neumann), in talks with Agilent (large proprietary data)

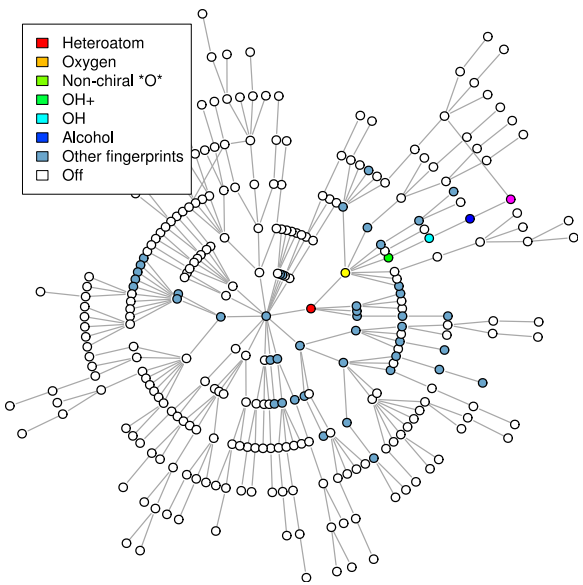




Metabolite fingerprint prediction from MS/MS data

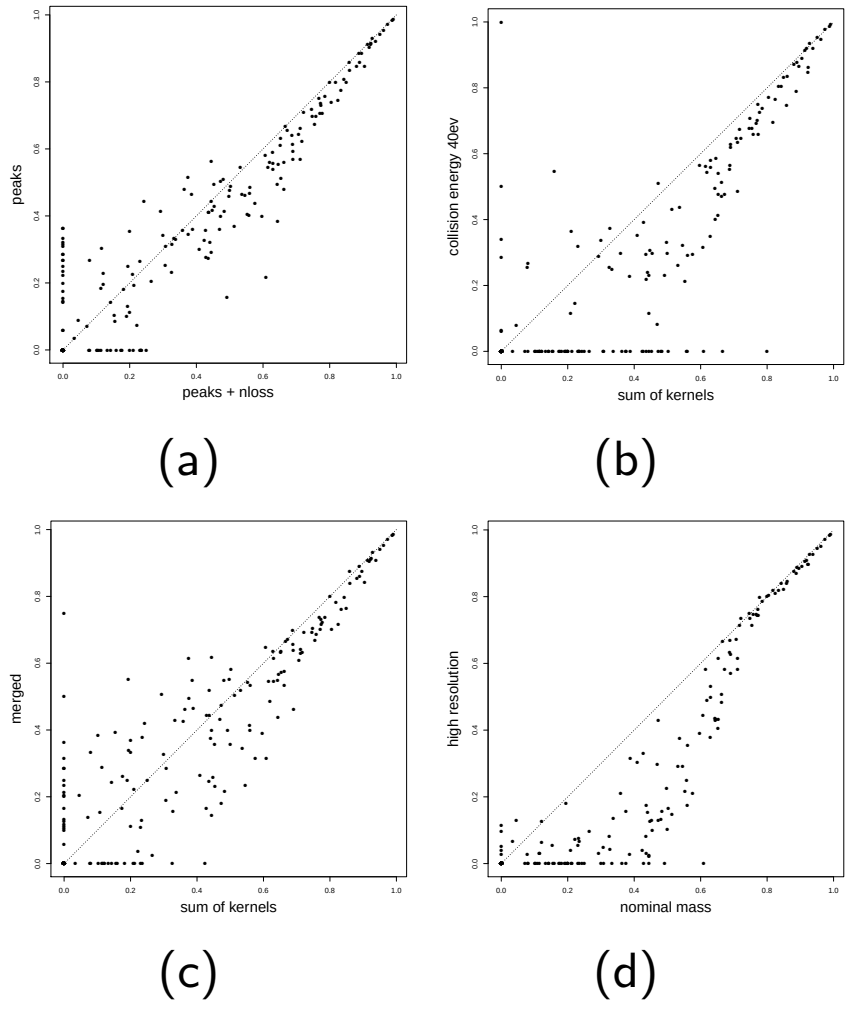
- Input: kernels for tandem MS/MS spectra, taking into account peak locations, intensities, neutral losses, different collision energies, different ways of combining the data
- Output: binary vector of fingerprint presence in the molecule
- Method: set of SVMs as the baseline, multi-task/multi-label classifiers as the final method





Metabolite fingerprint prediction from MS/MS data

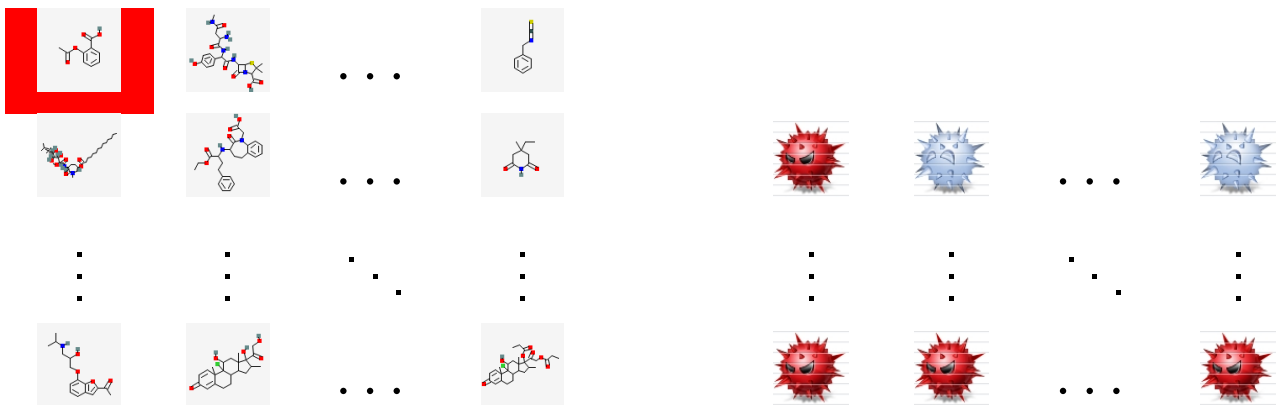
- Some initial results
- F1 score comparisons using different kernels in SVM.
  a. Neutral loss signal helps
  b. Combining several collision energies (CE) with kernel fusion helps
  c. Merging spectra of different CEs does not
  d. High resolution mass accuracy does not work well (yet!)



(Heinonen, Shen, Zamboni, Rousu, 2011, submitted)

Multi-Task Classification via Graph Labeling

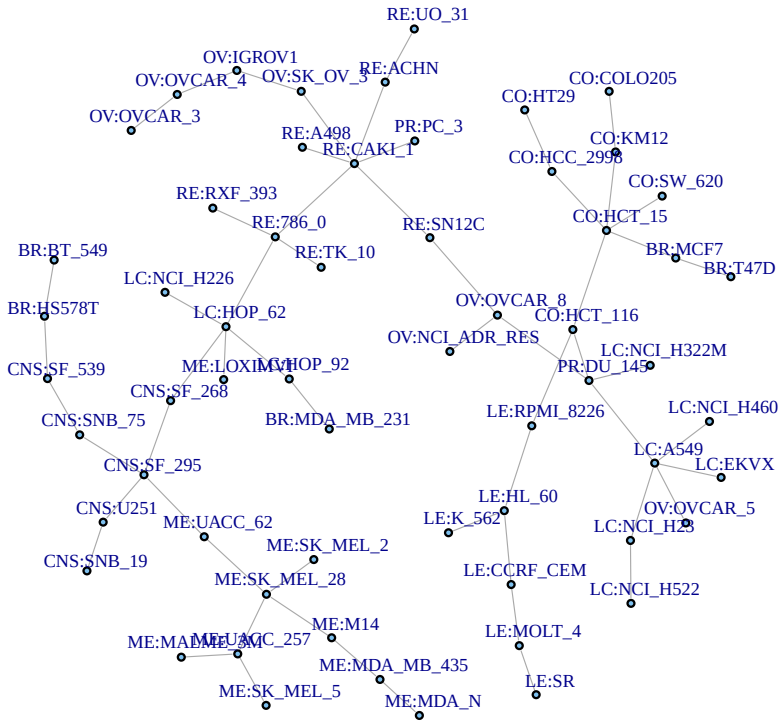
- Task: Given molecule, predict active/not active against a given target (a virus, cancer type,. . .)
- State of the art prior to 2010: SVM with graph kernels over the molecules, independently trained for each target
- Can we predict the activity better by learning against all available targets at the same time?
- Multi-task and Multi-label classification are machine learning methods developed for such scenarios



Multi-Task Classification via Graph Labeling

Our approach

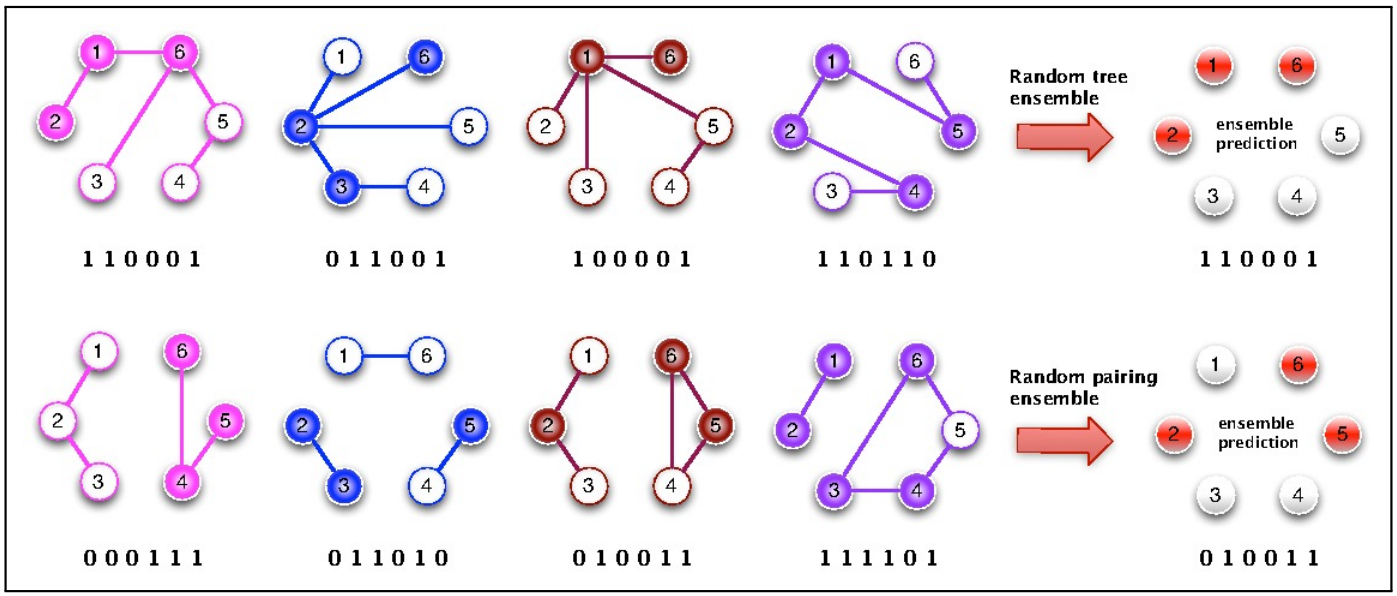
- We convert the multi-task learning setting to a graph labeling problem
- Output graph connecting the tasks is learned from an auxiliary dataset (different microarray datasets)
- Labeling of the graph is learned using the MMCRF method (Rousu et al. 2007)



(H. Su, M. Heinonen, J. Rousu. Pattern Recognition in Bioinformatics, 2010)

Multi-Task Classification via Graph Labeling

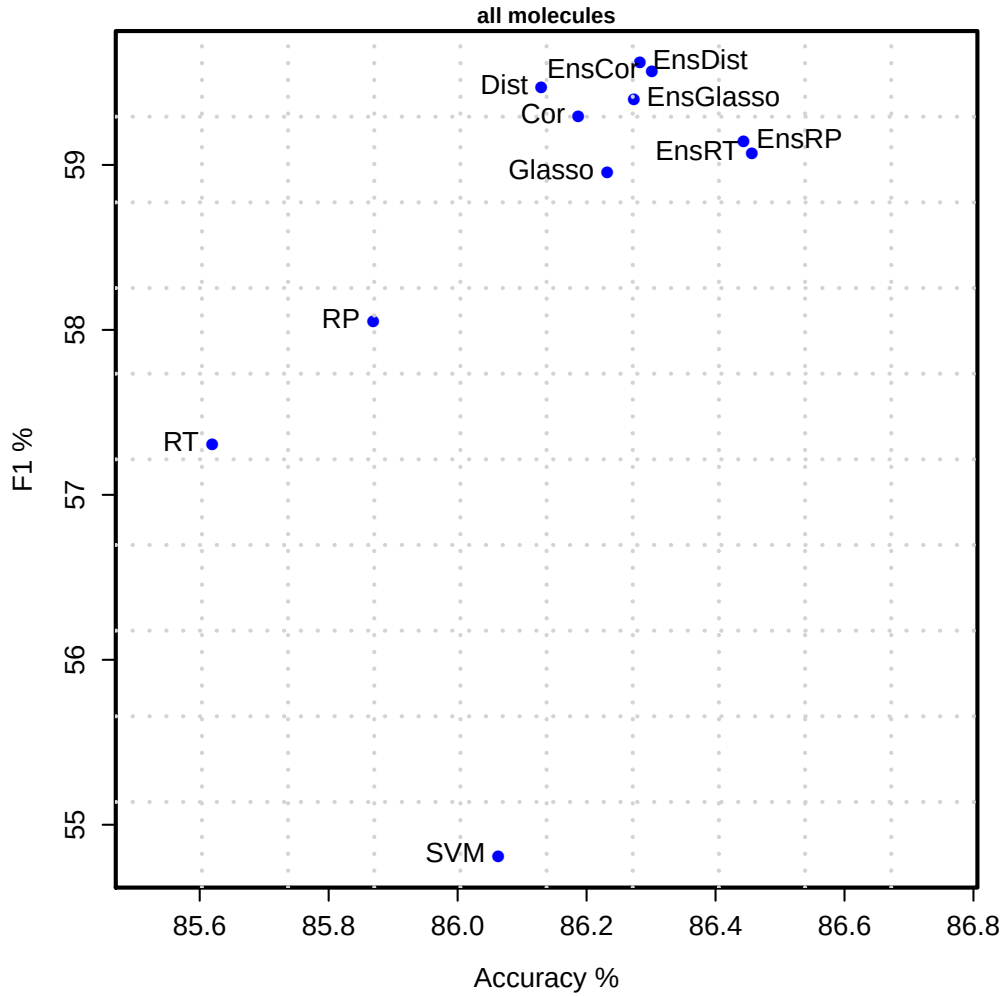
- There are several sources for learning the output graphs (13 datasets)
- Suggests an ensemble approach: train a set of graph labeling classifiers, with differnt graph structured and vote
- It turns out that random graphs can be used as well (no auxiliary data needed!)



(H. Su, J. Rousu. Pattern Recognition in Bioinformatics, November 2011, to appear)

Multi-Task Classification via Graph Labeling

- Scatter plot shows the F1 score (Y-axis) and accuracy (X-axis) for different methods
- SVM - support vector machine for each target individually
- MMCRF models with different output graphs
  - RP, RT - random graphs
  - Dist, Cor, Glasso - graph extraction from auxiliary data
  - Ens-*: Ensemble versions of the above



Plans for 2012

- Metabolite Fingerprint Prediction: Markus Heinonen, Huibin Shen, collaboration with ETH Zurich, IPB Halle
- Kernels for molecular data: Markus Heinonen
- Further development of graph based multi-task and ensemble learning: Hongyu Su
- Machine learning of protein functions and interactions: BIOLEDGE EU FP7 Project, collaboration with VTT, Cambridge, Malaga, and three SMEs (post-doctoral researcher to be hired)