

# Structured Output Prediction of Anti-Cancer Drug Activity

Hongyu Su, Markus Heinonen and Juho Rousu  
Department of Computer Science  
University of Helsinki, Finland

## Abstract

We present a structured output prediction approach for classifying potential anti-cancer drugs. Our QSAR model takes as input a description of a molecule and predicts the activity against a set of cancer cell lines in one shot. Statistical dependencies between the cell lines are encoded by a Markov network that has cell lines as nodes and edges represent similarity according to an auxiliary dataset. Molecules are represented via kernels based on molecular graphs. Margin-based learning is applied to separate correct multilabels from incorrect ones. The performance of the multilabel classification method is shown in our experiments with NCI-Cancer data containing the cancer inhibition potential of drug-like molecules against 59 cancer cell lines. In the experiments, our method outperforms the state-of-the-art SVM method.

## 1 Introduction

Machine learning has become increasingly important in drug discovery where viable molecular structures are searched or designed for therapeutic efficacy. In particular, Quantitative Structure-Activity Relationship (QSAR) models, relating the molecular structures to bioactivity (therapeutical effect, side-effects, toxicity, etc.) are routinely built using state-of-the-art machine learning methods. In particular, the costly pre-clinical *in vitro* and *in vivo* testing of drug candidates can be focused to the most promising molecules, if accurate *in silico* models are available [16].

Molecular classification—the task of predicting the presence or absence of the bioactivity of interest—has been tackled with a variety of methods, including inductive logic programming [9] and artificial neural networks [1]. During the last decade kernel methods [11, 16, 4] have emerged as a computationally effective way to handle the non-linear properties of chemicals. In numerous studies, SVM-based methods have obtained promising results [3, 16, 20]. However, classification methods focusing on a single target variable are probably not optimally suited to drug screening applications where large number of target cell lines are to be handled.

In this paper we propose, to our knowledge, the first multilabel learning approach for molecular classification. Our method belongs to the structured output prediction family [15, 17, 12, 13], where graphical models and kernels

have been successfully married in recent years. In our approach, the drug targets (cancer cell lines) are organized in a Markov network, drug molecules are represented by kernels and discriminative max-margin training is used to learn the parameters. Alternatively, our method can be interpreted as a form of multitask learning [5] where the Markov network couples the tasks (cell lines) and joint features are learned for pairs of similar tasks.

## 2 Methods

### 2.1 Structured output learning with MMCRF

The model used in this paper is an instantiation of the structured output prediction framework MMCRF [13] for associative Markov networks and can also be seen as a sibling method to HM<sup>3</sup>[12], which is designed for hierarchies. We give a brief outline here, the interested reader may check the details from the above references.

The MMCRF learning algorithm takes as input a matrix  $K = (k(x_i, x_j))_{i,j=1}^m$  of kernel values  $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$  between the training patterns, where  $\phi(x)$  denotes a feature description of an input pattern (in our case a potential drug molecule), and a label matrix  $Y = (\mathbf{y}_i)_{i=1}^m$  containing the multilabels  $\mathbf{y}_i = (y_1, \dots, y_k)$  of the training patterns. The components  $y_j \in \{-1, +1\}$  of the multilabel are called microlabels and in our case correspond to different cancer cell lines. In addition, the algorithm assumes an associative network  $G = (V, E)$  to be given, where node  $j \in V$  corresponds to the  $j$ 'th component of the multilabel and the edges  $e = (j, j') \in E$  correspond to a microlabel dependency structure.

The model learned by MMCRF takes the form of a conditional random field with exponential edge-potentials,

$$P(\mathbf{y}|x) \propto \prod_{e \in E} \exp(\mathbf{w}_e^T \varphi_e(x, \mathbf{y}_e)) = \exp(\mathbf{w}^T \varphi(x, \mathbf{y})),$$

where  $\mathbf{y}_e = (y_j, y_{j'})$  denotes the pair of microlabels of the edge  $e = (j, j')$ . A joint feature map  $\varphi_e(x, \mathbf{y}) = \phi(x) \otimes \psi_e(\mathbf{y}_e)$  for an edge is composed via tensor product of input  $\phi(x)$  and output feature map  $\psi(\mathbf{y})$ , thus including all pairs of input and output features. The output feature map is composed of indicator functions  $\psi_e^u(\mathbf{y}) = \mathbb{I}[\mathbf{y}_e = u]$  where  $u$  ranges over the four possible labelings of an edge given binary node labels. The corresponding weights are denoted by  $\mathbf{w}_e$ . The benefit of the tensor product representation is that context (edge-labeling) sensitive weights can be learned for input features and no prior alignment of input and output features needs to be assumed.

The parameters are learned by maximizing the minimum loss-scaled margin between the correct training examples  $(x_i, \mathbf{y}_i)$  and incorrect pseudo-examples  $(x_i, \mathbf{y}), \mathbf{y} \neq \mathbf{y}_i$ , while controlling the norm of the weight vector. The primal soft-margin optimization problem takes the form

$$\begin{aligned} \underset{\mathbf{w}, \xi \geq 0}{\text{minimize}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \mathbf{w}^T \varphi(x_i, \mathbf{y}_i) - \mathbf{w}^T \varphi(x_i, \mathbf{y}) \geq \ell(\mathbf{y}_i, \mathbf{y}) - \xi_i, \\ & \text{for all } i \text{ and } \mathbf{y}, \end{aligned} \tag{1}$$

where  $\xi_i$  denote the slacks allotted to each example. The effect of loss-scaling is to push high-loss pseudo-examples further away from the correct example than the low-loss pseudo-examples, which, intuitively, decreases the risk of incurring high-loss. We use *Hamming loss*

$$\ell_{\Delta}(\mathbf{y}, \mathbf{u}) = \sum_j \llbracket y_j \neq u_j \rrbracket$$

that is gradually increasing in the number of incorrect microlabels so that we can make a difference between 'nearly correct' and 'clearly incorrect' multilabel predictions.

The MMCRF algorithm [13] optimizes the model (1) in the so called marginal dual form, that has several benefits: the use of kernels to represent high-dimensional inputs, and polynomial-size of the optimization problem with respect to the size of the output structure. Efficient optimization is achieved via the conditional gradient algorithm [2] with feasible ascent directions found by loopy belief propagation over the Markov network  $G$ .

## 2.2 Kernels for drug-like molecules

A major challenge for any statistical learning model is to define a measure of similarity. In chemical community, widely researched quantitative structure-activity relationship (QSAR) theory asserts that compounds having similar physico-chemical and geometric properties should have related bioactivity [7]. Various descriptors have been used to represent molecules with fixed-length feature vectors, such as atom counts, topological and shape indices, quantum-chemical and geometric properties [19]. Kernels computed from the structured representation of molecules extend the scope of the traditional approaches by allowing complex derived features to be used (walks, subgraphs, properties) while avoiding excessive computational cost [11].

In this paper, we experiment with a set of graph kernels designed for classification of drug-like molecules, including walk kernel [6], weighted decomposition kernel [10] and Tanimoto kernel [11]. All of them rely on representing the molecule as a labeled graph with atoms as nodes and bonds between the atoms as the edges.

**Walk kernel** [8, 6] computes the sum of matching walks (a sequence of labeled nodes so that there exists an edge for each pair of adjacent nodes) in a pair of graphs. The contribution of each matching walk is downscaled exponentially according to its length. We consider finite-length walk kernel where only walks of length  $p$  are counted. The finite walk kernel can be efficiently computed using dynamic programming.

**Weighted decomposition kernel** [4] is an extension of the substructure kernel by weighting identical parts in a pair of graphs based on contextual information [4]. The kernel looks at matching subgraphs (*contextor*) in the neighborhood of *selector* atoms.

**Tanimoto kernel** [11] is a kernel computed from two molecule fingerprints by checking the fraction of features that occur in both fingerprints of all features. *Hash fingerprints* enumerates all linear fragments of a given length, while

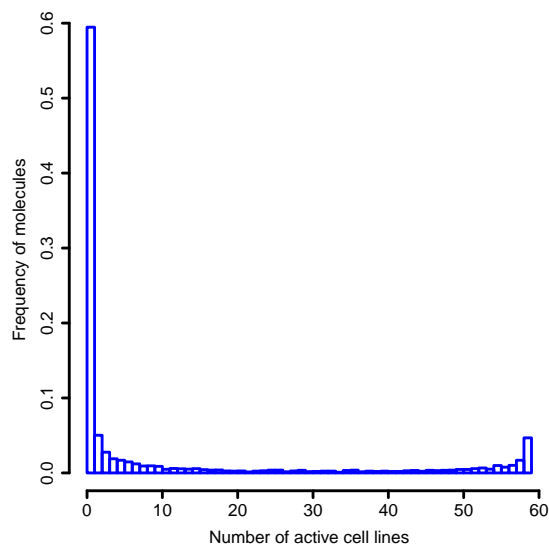


Figure 1: Skewness of the multilabel distribution.

*substructure keys* correspond to molecular substructures in a predefined set designed by domain experts. Based on good performance in preliminary studies, in this paper we concentrate on hash fingerprints.

### 2.3 Markov network generation for cancer cell lines

In order to use MMCRF to classify drug molecules we need to build a Markov network for the cell lines used as the output, with nodes corresponding to cell lines and edges to potential statistical dependencies. To build the network we used auxiliary data (e.g. mRNA and protein expression, mutational status, chromosomal aberrations, DNA copy number variations, etc) available on the cancer cell lines from NCI database<sup>1</sup>. The basic approach is to construct from this data a correlation matrix between the pairs of cell lines and extract the Markov network from the matrix by favoring high-valued pairs. The following methods of network extraction were considered:

- Maximum weight spanning tree. Take the minimum number of edges that make a connected network whilst maximizing the edge weights.
- Correlation thresholding. Take all edges that exceed fixed threshold. This approach typically generates a general non-tree graph.

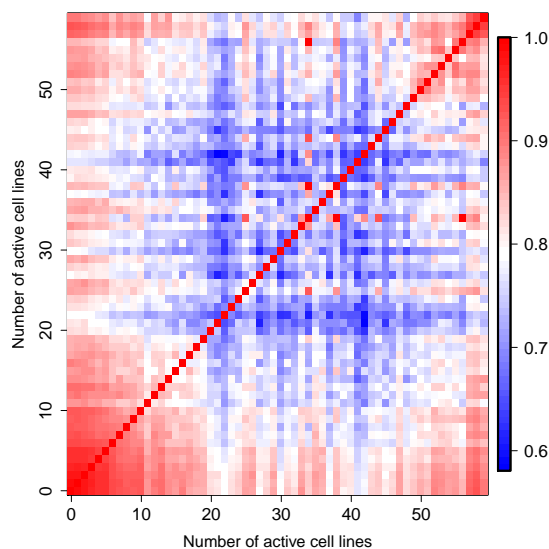


Figure 2: Heatmap of the kernel space for the molecules sorted by the multilabel distribution.

## 3 Experiments

### 3.1 NCI-Cancer dataset

In this paper we use the NCI-Cancer dataset obtained through PubChem Bioassay<sup>2</sup> [18] data repository. The dataset initiated by National Cancer Institute and National Institutes of Health (NCI/NIH) contains bioactivity information of large number of molecules against several human cancer cell lines in 9 different tissue types, including leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney. For each molecule tested against a certain cell line, the dataset provide a bioactivity outcome that we use as the classes (active, inactive).

### 3.2 Data preprocessing

Currently, there are 43884 molecules in the PubChem Bioassay database together with anti-cancer activities in 73 cell lines. 59 cell lines have screen experimental results for most molecules and 4554 molecules have no missing data in these cell lines, therefore these cell lines and molecules are selected and employed in our experiments.

However, molecular activity data are highly biased over the cell lines. Figure 1 shows the molecular activity distribution over all 59 cell lines. Most of the molecules are inactive in all cell lines, while a relatively large proportion of molecules are active against almost all cell lines, which can be taken as toxics. These molecules are less likely to be potential drug candidates than the ones in

<sup>1</sup><http://discover.nci.nih.gov/cellminer/home.do>

<sup>2</sup><http://pubchem.ncbi.nlm.nih.gov>

the middle part of the histogram.

Figure 2 shows a heatmap of normalized Tanimoto kernel, where molecules have been sorted by the number of cell lines they are active in. The heatmap shows that the molecules in the two extremes of the multilabel distribution form groups of high similarity whereas the molecules in the middle are much more dissimilar both to each other and to the extreme groups. The result seems to indicate that the majority of molecules in the dataset are either very specific or very general in the targets they are active against. Other kernels mentioned in section 2.2 produce a similar heatmap indicating that the phenomenon is not kernel-specific.

Because of the above-mentioned skewness, we prepared different versions of the dataset:

**Full.** This dataset contains all 4554 molecules in the NCI-Cancer dataset that have their activity class (active vs. inactive) recorded against all 59 cancer cell lines.

**No-Zero-Active.** From this dataset, we removed all molecules that are not active towards any of the cell lines (corresponding to the leftmost peak in Figure 1). The remaining 2305 molecules are all active against at least one cell line.

**Middle-Active.** Here, we followed the preprocessing suggested in [14], and selected molecules that are active in more than 10 cell lines and inactive in more than 10 cell lines. As a result, 544 molecules remained and were employed in our experiments.

### 3.3 Experiment setup

We conducted experiments to compare the effect of various kernels, as well as the performances of support vector machine (SVM) and MMCRF. We used the SVM implementation of the LibSVM software package written in C++<sup>3</sup>. We tested SVM with different margin  $C$  parameters, relative hard margin ( $C = 100$ ) emerging as the value used in subsequent experiments. The same value was used for MMCRF classifier as well.

Because of the skewness of the multilabel distribution (c.f. 1) we used the following *stratified 5-fold cross-validation* scheme in all experiments reported: we group the molecules in equivalence classes based on the number of cell lines they are active against. Then each group is randomly split among the five folds. This ensures that also the smaller groups have representation in all folds.

### 3.4 Kernel setup

For the three kernel methods, walk kernel (WK) was constructed using parameters  $\lambda = 0.1$  and  $p = 6$  as recommended in [6]. The Weighted decomposition kernel (WDK) used context radius 3 as in [4], and a single attribute (atom type) was sufficient to give the best performance. We also used hash fragments as molecular fingerprints generated by OpenBabel<sup>4</sup> (using default value  $n = 6$

<sup>3</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

<sup>4</sup><http://openbabel.org>

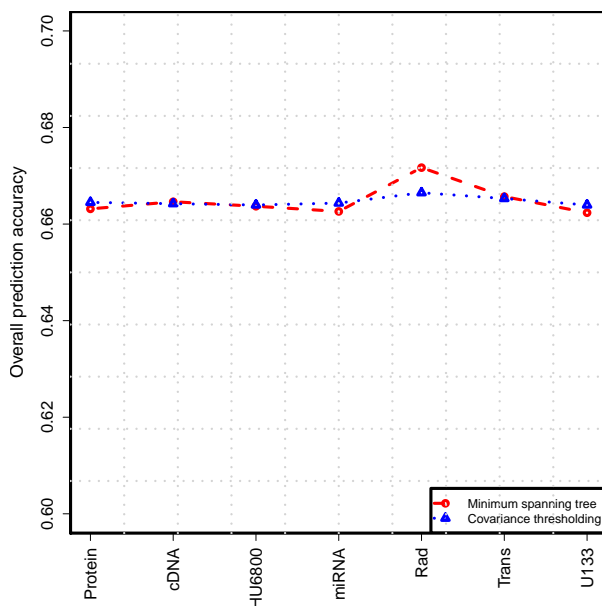


Figure 3: Effects of Markov network construction methods and type of auxiliary data (from left to right: reverse-phase lysate arrays, cDNA arrays, Affymetric HU6800 arrays, miRNA arrays, RNA radiation arrays, transporter arrays, and Affymetrix U133 arrays).

for linear structure length), which is a chemical toolbox available in public domain. All kernels were normalized.

## 4 Results

### 4.1 Effect of Markov network generation methods

We report overall prediction accuracies on the Middle-Active dataset from various Markov networks shown in Figure 3. X-axis corresponds to different microarray experiments. The accuracies from different Markov networks differ slightly. The best accuracy was achieved by using maximum weighted spanning tree approach on RNA radiation arrays dataset, shown in Figure 4, which describes profiles of radiation response in cell lines. This meets our expectations since cancer cells mostly mutated from normal cells and normal cells with radiation treatments can possibly explain the mutations.

### 4.2 Effect of molecule kernels

In Table 1, we report overall accuracies and microlabel F1 scores using SVM with different kernels on the Middle-Active dataset. The results were from a five-fold cross validation procedure. Here, the three kernel methods achieve almost the same accuracies in SVM classifier, while Tanimoto kernel is slightly better than others in microlabel F1 score. Thus we deemed Tanimoto kernel to be the best

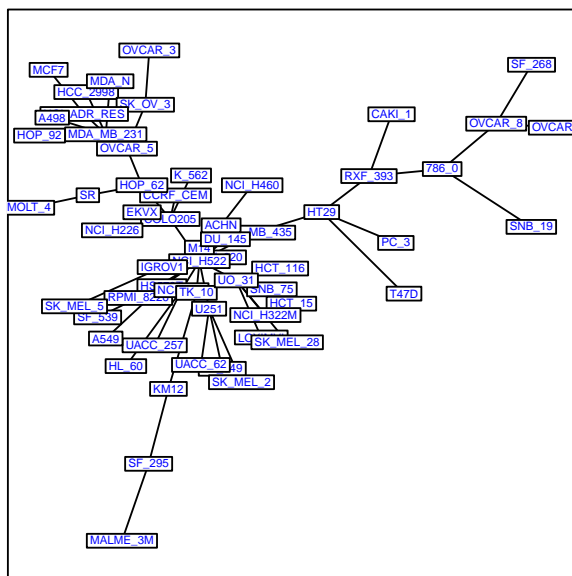


Figure 4: Markov network constructed from maximum weighted spanning tree method on RNA radiation array data. The labels correspond to different cancer cell lines.

Table 1: Accuracies and microlabel F1 scores of MMCRF and SVM with different kernels.

| Classifier | Kernel   | Accuracy | F1 score |
|------------|----------|----------|----------|
| SVM        | WK       | 64.6%    | 49.0%    |
|            | WDK      | 63.9%    | 51.6%    |
|            | Tanimoto | 64.1%    | 52.7%    |
| MMCRF      | Tanimoto | 67.6%    | 56.2%    |

kernel in this experiment and chose it for the subsequent experiments.

### 4.3 Effect of dataset versions

Figure 5 gives overall accuracy and microlabel F1 score of MMCRF versus SVM for each cell line on the three versions of the data. Points above the diagonal line correspond to improvements in accuracies or F1 scores by MMCRF classifier. MMCRF improves the F1 score over SVM on each version of the data in statistically significant manner, as judged by the two-tailed sign test. Accuracy is improved in two versions, No-Zero-Actives and the Middle-Active molecules, again in statistically significant manner. Among the Middle-Active dataset, the difference in accuracy (bottom, left of Figure 5) is sometimes drastic, around 10 percentage units in favor of MMCRF for a significant fraction of the cell lines.

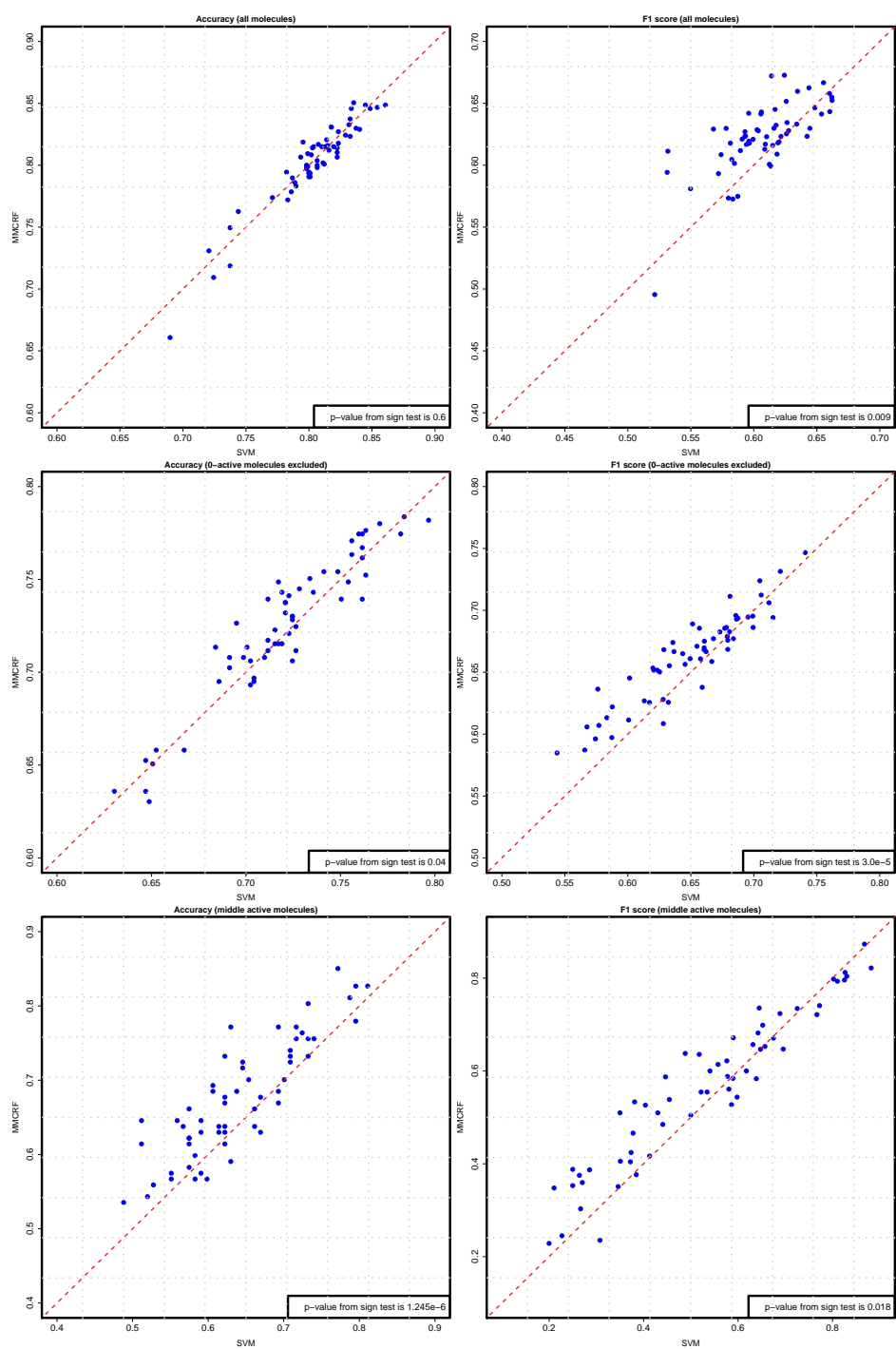


Figure 5: Accuracy (left) and F1 score (right) of MMCRF vs. SVM on Full data (top), No-Zero-Active (middle) and Middle-Active molecules (bottom).

Table 2: Agreement of MMCRF and SVM on the positive (left) and negative (right) classes.

|                 | Positive class   |                  | Negative class   |                 |
|-----------------|------------------|------------------|------------------|-----------------|
|                 | SVM Correct      | SVM Incorrect    | SVM Correct      | SVM Incorrect   |
| MMCRF Correct   | $48.6 \pm 4.1\%$ | $7.1 \pm 2.6\%$  | $88.0 \pm 4.9\%$ | $2.2 \pm 1.2\%$ |
| MMCRF Incorrect | $3.4 \pm 1.3\%$  | $40.9 \pm 3.4\%$ | $3.8 \pm 1.7\%$  | $6.1 \pm 3.0\%$ |

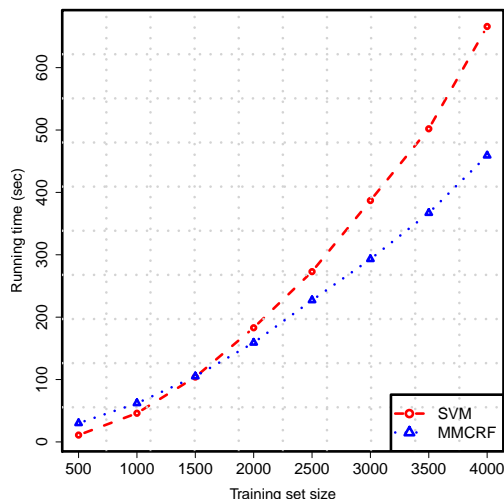


Figure 6: Training time for SVM and MMCRF classifiers on training sets of different sizes.

#### 4.4 Agreement of MMCRF and SVM predictions

For a closer look at the predictions of MMCRF and SVM, Table 2 depicts the agreement of the two models among positive and negative classes. Both models were trained on the Full dataset. Overall, the two models agree on the label most of the time (close to 90% of positive predictions and close to 95% of the negative predictions). MMCRF is markedly more accurate than SVM on the positive class while SVM is slightly more accurate among the negative class. Qualitatively similar results are obtained when the zero-active molecules are removed from the data (data not shown).

#### 4.5 Computation time

Besides predictive accuracy, training time of classifiers is important when a large number of drug targets need to be processed. The potential benefit of multilabel classification is the fact that only single model needs to be trained instead of a bag of binary classifiers.

We compared the running time needed to construct MMCRF classifier (implemented in native MATLAB) against libSVM classifier (C++). We conducted the experiment on a 2.0GHz computer with 8GB memory. Figure 6 shows that MMCRF scales better when training set increases.

## 5 Conclusions

We presented a multilabel classification approach to drug activity classification using the Max-Margin Conditional Random Field algorithm. In the experiments against a large set of cancer lines the method significantly outperformed SVM in training time and accuracy. In particular, drastic improvements could be seen in the setup where molecules with extreme activity (active against no or a very small fraction, or a very large fraction of the cell lines) were excluded from the data. The remaining middle ground of selectively active molecules is in our view more important from drug screening applications point of view, than the two extremes.

The MMCRF software and preprocessed versions of the data are available from <http://cs.helsinki.fi/group/sysfys/software>.

## Acknowledgements

This work was financially supported by Academy of Finland grant 118653 (AL-GODAN) and in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-2007-216886. This publication only reflects the authors' views.

## References

- [1] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, and M.R. Tine. Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. *J. Chem. Inf. Model.*, 46:2030–2042, 2006.
- [2] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
- [3] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.*, 43:1882–1889, 2003.
- [4] A. Ceroni, F. Costa, and P. Frasconi. Classification of small molecules by two- and three-dimensional decomposition kernels. *Bioinformatics*, 23:2038–2045, 2007.
- [5] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *KDD'04*, pages 109–117. ACM, 2004.
- [6] T. Gärtner. A survey of kernels for structured data. *SIGKDD Explor. Newsl.*, 5(1):49–58, 2003.
- [7] M. Karelson. *Molecular Descriptors in QSAR/QSPR*. Wiley-Interscience, 2000.
- [8] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the 20<sup>th</sup> International Conference on Machine Learning (ICML)*, Washington, DC, United States, 2003.

- [9] R. King, S. Muggleton, A. Srinivasan, and M. Sternberg. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *PNAS*, 93:438–442, 1996.
- [10] S. Menchetti, F. Costa, and P. Frasconi. Weighted decomposition kernels. In *International Conference on Machine Learning*, pages 585–592. ACM Press, 2005.
- [11] L. Ralaivola, S. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Networks*, 18:1093–1110, 2005.
- [12] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-Based Learning of Hierarchical Multilabel Classification Models. *JMLR*, 7:1601–1626, 2006.
- [13] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Efficient algorithms for max-margin structured classification. *Predicting Structured Data*, pages 105–129, 2007.
- [14] P. Shivakumar and M. Krauthammer. Structural similarity assessment for drug sensitivity prediction in cancer. *Bioinformatics*, 10:S17, 2009.
- [15] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Neural Information Processing Systems 2003*, 2003.
- [16] M. Trotter, M. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comp. and Chem.*, 26:1–20, 2001.
- [17] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML'04*, pages 823–830, 2004.
- [18] Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B.A. Shoemaker, T.O. Suzek, J. Wang, J. Xiao, J. Zhang, and S.H. Bryant. An overview of the pubchem bioassay resource. *Nucleic Acids Research*, 38:D255–D266, 2009.
- [19] Y. Xue, Z. Li, C. Yap, et al. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.*, 44:1630–1638, 2004.
- [20] V. Zernov, K. Balakin, A. Ivaschenko, N. Savchuk, and I. Pletnev. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.*, 43:2048–2056, 2003.