Project topics for the course CS-E4880 Machine Learning in Bioinformatics

March 28, 2019

1 Comparison of the predictive performance between fingerprint-based drug kernels and graph-based drug kernels in the task of drug-protein binding affinity modelling

Instructor: Anna Cichonska (anna.cichonska@aalto.fi)

Background: Drug-like chemical compounds execute their actions mainly by modulating cellular targets, such as proteins. Despite the availability of modern high-throughput screening assays, experimental determination of interactions between chemical compounds and protein targets is still time consuming and expensive. Therefore, in the recent years, a lot of effort has been placed on the development of computational methods that could provide fast, large-scale and systematic pre-screening of chemical probes. In particular, a lot of work has been devoted to compound-based interaction prediction methods, including quantitative structure-activity relationship (QSAR) models, which aim to relate structural properties of chemical molecules to their bioactivity profiles. Another class of computational methods, so called protein-based methods, focus on finding the relationship between bioactivity profiles and protein descriptors. Systems-based approaches, also known as proteochemometric models, unify the above frameworks by exploiting the properties of both drug compounds and protein targets under the assumption that similar drugs are likely to interact with similar proteins. A proper representation and use of similarities, equivalent to a kernel choice, is therefore a first critical prerequisite for the achievement of high-quality drug-protein interaction (DPI) predictions.

A variety of different kernels have been introduced for calculating similarities between drugs and proteins. Among drug kernels, fingerprint-based Tanimoto kernels are the standard

choice for modelling purposes. Fingerprint encodes a molecular structure into a binary vector where each bit represents the presence (1) or absence (0) of a specific substructure in the molecule. Tanimoto kernel is computed based on the number of common substructures of the two drug molecules represented by their fingerprints. Graph kernels, on the other hand, measure similarities between graphs. They can be roughly categorised into three main groups, namely, graph kernels based on walks and paths, graph kernels based on limited-size subgraphs, and graph kernels based on subtree patterns. Graph kernels are applicable to measuring similarities between drug compounds, since a chemical molecule can be naturally represented as unlabeled or labeled graph, where a node corresponds to an atom, and an edge indicates a bond between two atoms.

Goal: The goal of the project is to compare the performance of several drug kernels calculated based on fingerprint and graph representations of chemical molecules in the task of prediction of drug-protein binding affinities.

Materials and Methods: The data set consists of 100 drug compounds and 100 protein targets, which is a subset of the data from the experimental study by Metz *et al.* (2011). DPIs are represented as real values reflecting binding affinities. The student will calculate several fingerprint-based drug kernels and graph-based drug kernels implemented in `ChemmineR` and `Rchemcpp` R packages. For proteins, Smith-Waterman amino acid sequence alignment will be adopted. The student will implement Kernel Ridge Regression (KRR) that uses algebraic properties of the Kronecker product to avoid the explicit computation of the pairwise kernel (KronRLS). KronRLS will work with drug kernel and protein kernel as input instead of pairwise kernel. The student will implement nested cross validation to tune the regularisation parameters λ of KronRLS and assess the predictive performance of the model with each combination of drug and protein kernels.

Prerequisite: Programming skills (MATLAB, R, Python), basic knowledge of machine learning. Some knowledge of chemoinformatics is beneficial.

References

[1] Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics* 2014; 15(5): 734–47.

[2] Cichonska A, Rousu J, Aittokallio T. Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opinion on Drug Discovery* 2015; 10(12): 1333–45.

[3] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008; 24(13): i232–40.

[4] Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T. Toward

more realistic drug-target interaction predictions. *Briefings in Bioinformatics* 2014; 16(2): 325–337.

[5] Cichonska A, Ravikumar B, Parri E, Timonen S, Pahikkala T, Airola A, Wennerberg K, Rousu J, Aittokallio T. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLOS Computational Biology* 2017; 13(8): e1005678.

[6] Metz JT, Johnson EF, Soni NB *et al.* Navigating the kinome. *Nature Chemical Biology* 2011; 7(4): 200–2.

2 Multiple kernel learning for drug-protein binding affinity prediction

Instructor: Anna Cichonska (`anna.cichonska@aalto.fi`)

Background: Drug-like chemical compounds execute their actions mainly by modulating cellular targets, such as proteins. Despite the availability of modern high-throughput screening assays, experimental determination of interactions between chemical compounds and protein targets is still time consuming and expensive. Therefore, in the recent years, a lot of effort has been placed on the development of computational methods that could provide fast, large-scale and systematic pre-screening of chemical probes. In particular, a lot of work has been devoted to compound-based interaction prediction methods, including quantitative structure-activity relationship (QSAR) models, which aim to relate structural properties of chemical molecules to their bioactivity profiles. Another class of computational methods, so called protein-based methods, focus on finding the relationship between bioactivity profiles and protein descriptors. Systems-based approaches, also known as proteochemometric models, unify the above frameworks by exploiting the properties of both drug compounds and protein targets under the assumption that similar drugs are likely to interact with similar proteins. A proper representation and use of similarities, equivalent to a kernel choice, is therefore a first critical prerequisite for the achievement of high-quality drug-protein interaction (DPI) predictions.

Classical kernel-based methods rely on a single kernel. However, such approaches are unlikely to be optimal when a growing variety of biological and molecular data sources become available simultaneously. Multiple kernel learning (MKL) methods, which search for an optimal combination of several kernels, enabling the use of different information sources simultaneously and learning their importance for the prediction task, are therefore receiving increasing attention.

Goal: The goal of the project is to compute several protein kernels as well as drug kernels, and then use them in MKL regression framework to predict drug-protein binding affinities.

Materials and Methods: The data set consists of 50 drug compounds and 50 protein targets, which is a subset of the data from the experimental study by Metz *et al.* (2011). DPIs are represented as real values reflecting binding affinities. The student will calculate Tanimoto kernels for drug compounds based on several fingerprints implemented in `rcdk` R package. For proteins, Smith-Waterman amino acid sequence alignment as well as Generic String kernel will be adopted. The student can also choose to compute other molecular descriptors. Then, pairwise kernels that directly relate drug-protein pairs will be constructed by taking a Kronecker product of each pair of drug kernel and protein kernel. The student will use pairwise kernels with two-stage MKL algorithm ALIGNF. In the first stage, kernel mixture weights are determined based on maximising the centred alignment, i.e., matrix similarity measure, between the final combined kernel and so-called *ideal* response kernel derived from the label values. In the second stage, the combined kernel is used with Kernel Ridge Regression (KRR) as a prediction model. The student will be provided a Python script for calculating kernel mixture weights (first stage; a modification in one equation will be needed in this script) but should implement KRR (second stage). UNIMKL algorithm will form a baseline model, where all kernel mixture weights are equal to $1/P$, P being the number of input kernels. The student will implement nested cross validation to tune the regularisation parameters λ of KRR and assess the predictive performance of the model.

Prerequisite: Programming skills (MATLAB, R, Python), basic knowledge of machine learning. Some knowledge of chemoinformatics is beneficial.

References

[1] Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics* 2014; 15(5): 734–47.

[2] Cichonska A, Rousu J, Aittokallio T. Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opinion on Drug Discovery* 2015; 10(12): 1333–45.

[3] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008; 24(13): i232–40.

[4] Giguere S, Marchand M, Laviolette F, Drouin A, Corbeil J. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinformatics* 2013; 14(1): 82.

[5] Cortes C, Mohri M, Rostamizadeh A. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research* 2012; 13(Mar): 795-828.

[6] Metz JT, Johnson EF, Soni NB *et al.* Navigating the kinome. *Nature Chemical Biology* 2011; 7(4): 200–2.