PIMKL: Pathway Induced Multiple Kernel Learning

IBM Research

ETH zürich

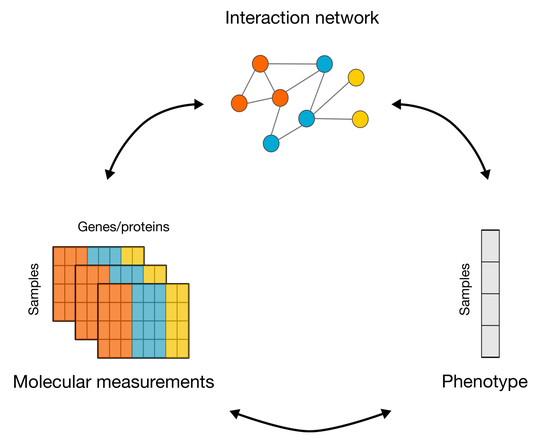
Matteo Manica[1,2,*], Joris Cadow[1,*], Roland Mathis[1,*], María Rodríguez Martínez[1]

[1] IBM Research Zurich, [2] ETH Zurich, [*] Shared first authors

Motivation

What

- Accurate **phenotype classification**
- Biological **interpretability**
- **Robustness** to noise
- Handling **curse of dimensionality**

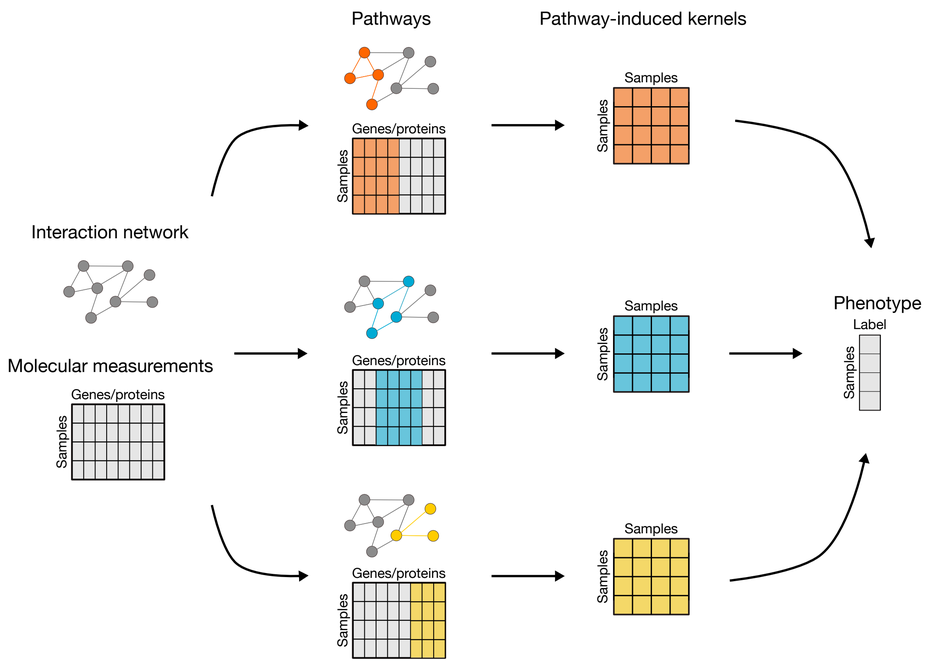


How

- Exploit **prior knowledge** from biological networks
- Apply **multiple kernel learning** for feature encoding
- Use **pathway annotations** to enable interpretability

PIMKL[1]

Concept



Formulation

Given:
$X \in \mathbb{R}^{n,m}$ molecular measurements
$y \in \mathbb{R}^{n}$ phenotype of interest
$\mathscr{P}$ set of pathways

Build *pathway-induced* kernels:

$$G_p = (V_p, E_p) \qquad L_p = D_p - A_p$$

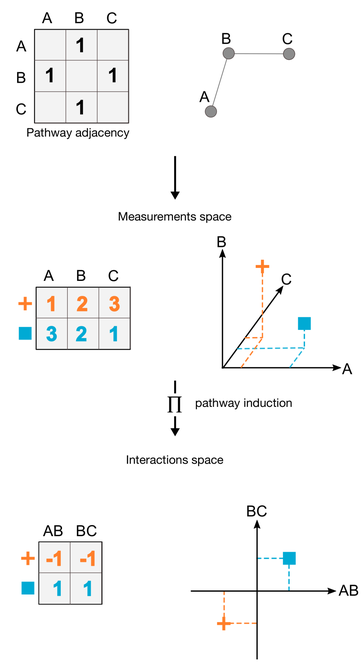
$$X_p \in \mathbb{R}^{n,m_p} \qquad K_p = X_p L_p X_p^t$$

Consider the mixture of kernels:

$$K = \sum_{p \in \mathscr{P}} w_p K_p$$

Optimize the weights for phenotype classification (custom implementation of EasyMKL[2,3])

Pathway induction



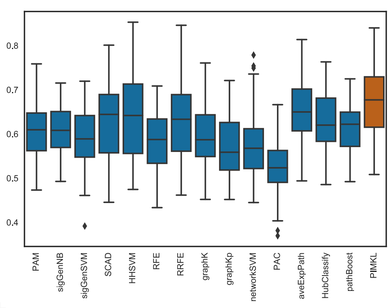
Results

Data

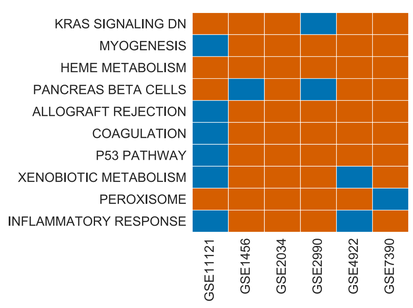
- Affymetrix Human Genome U133A Array from 6 breast cancer microarray cohorts[4]
- A collection of 50 hallmark gene sets from Molecular Signatures Database (MSigDB) version 5.2[5]
- Molecular networks from KEGG[6,7,8] and Pathway Commons[9]
- Illumina Human v3 microarray data and Affymetrix SNP 6.0 copy number data from METABRIC[10]

Benchmark

Comparison on a benchmark[4] against 14 methods exploiting prior knowledge

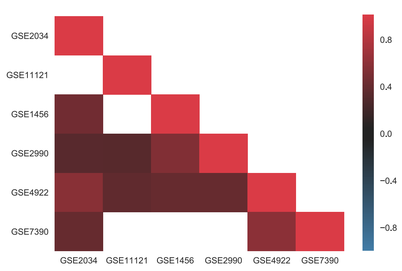


Weights learned in different cohorts highlight relevant pathways for breast cancer

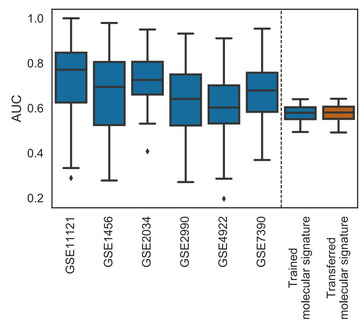


Transfer learning

Pathway signatures are significantly correlated across cohorts

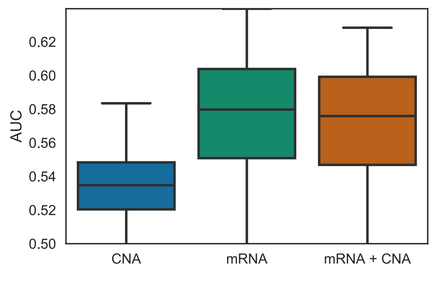


Signature learned on other cohorts successfully transferred to METABRIC



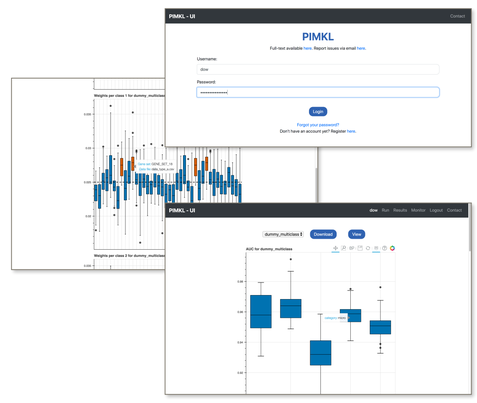
Multi-omics

In METABRIC case noisy kernels from CNA are integrated without significant performance loss



Availability

PIMKL is available as an open access web service on IBM Cloud



Find out more on the project website

https://ibm.biz/ibmpimkl

1) Manica M., et al., PIMKL: Pathway Induced Multiple Kernel Learning, npj Systems Biology and Applications, 2019
2) Aiolli F., et al., A kernel method for the optimization of the margin distribution, Lecture Notes in Computer Science, 2008
3) Aiolli F., et al., EasyMKL: A scalable multiple kernel learning algorithm, Neurocomputing, 2015
4) Cun Y., et al., Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge, BMC bioinformatics, 2012
5) Liberzon A., et al., The Molecular Signatures Database Hallmark Gene Set Collection, Cell Systems, 2015
6) Kanehisa M., et al., KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Research, 2000
7) Kanehisa M., et al., KEGG as a reference resource for gene and protein annotation, Nucleic Acids Research, 2016
8) Kanehisa M., et al., KEGG: new perspectives on genomes, pathways, diseases and drugs, , Nucleic Acids Research, 2017
9) Cerami E. G., et al., Pathway Commons, a web resource for biological pathway data, Nucleic Acids Research, 2011
10) Curtis C., et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature, 2012

This work is part of the PrECISE project. PrECISE combines hypothesis-driven strategies with data-driven analysis in a novel mathematical and computational methodology for the integration of genomic, epigenetic, transcriptomic, proteomic, and clinical data with the goal of risk-stratifying patients and suggesting personalized therapeutic interventions. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858.
Project website: www.precise-project.eu