Machine Learning for Toxicogenetics and Drug Response Prediction

Jean-Philippe Vert

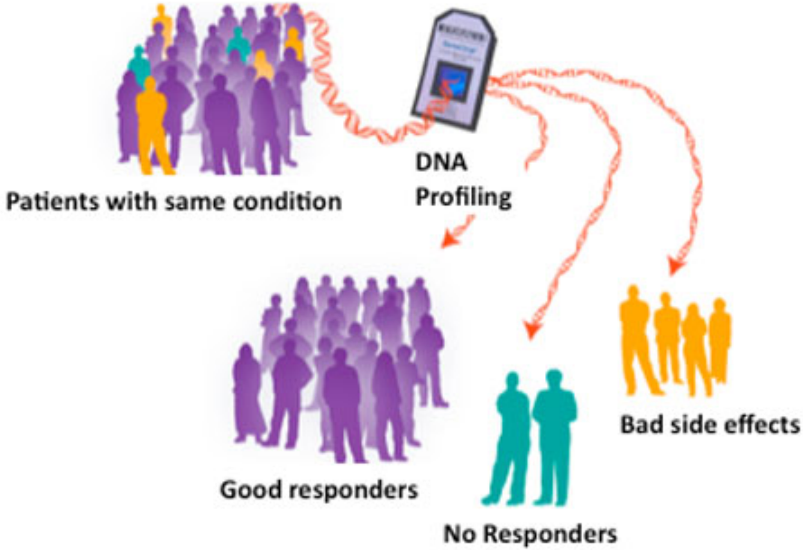
MINES ParisTech

institut Curie

Together, let's beat cancer.

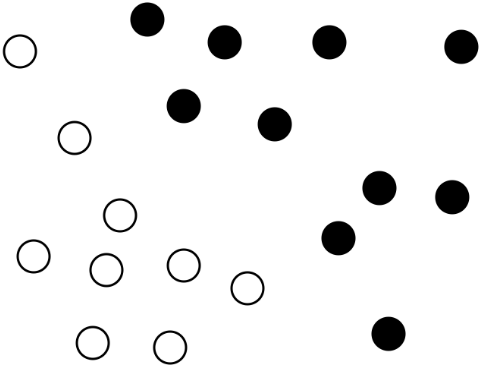
Festival of Genomics, San Mateo, Nov 6, 2015

Molecular stratification



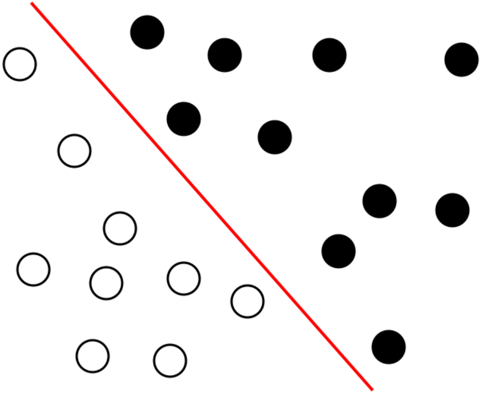
Diagnosis, prognosis, drug response prediction, ...

Machine learning formulation



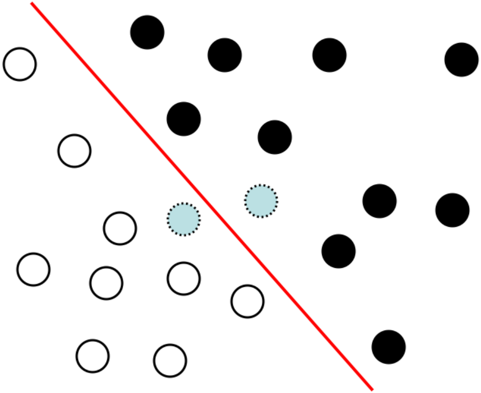
$n(= 19) >> p(= 2)$: easy

Machine learning formulation



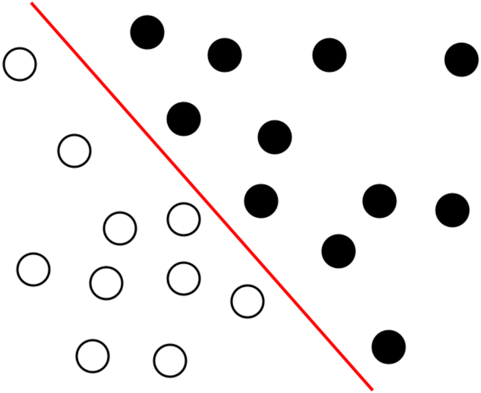
$n(=19) >> p(=2)$: easy

Machine learning formulation



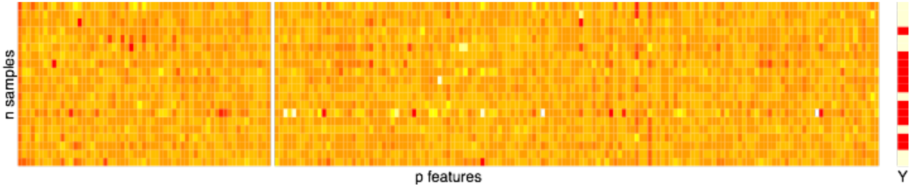
$n(=19) >> p(=2)$: easy

Machine learning formulation



$n (= 19) >> p (= 2)$: easy

Challenge: $n << p$



- $n = 10^2 \sim 10^4$ (patients)
- $p = 10^4 \sim 10^7$ (genes, mutations, copy number, ...)

Accuracy drops,

Outline

1. Learning molecular signatures with network information

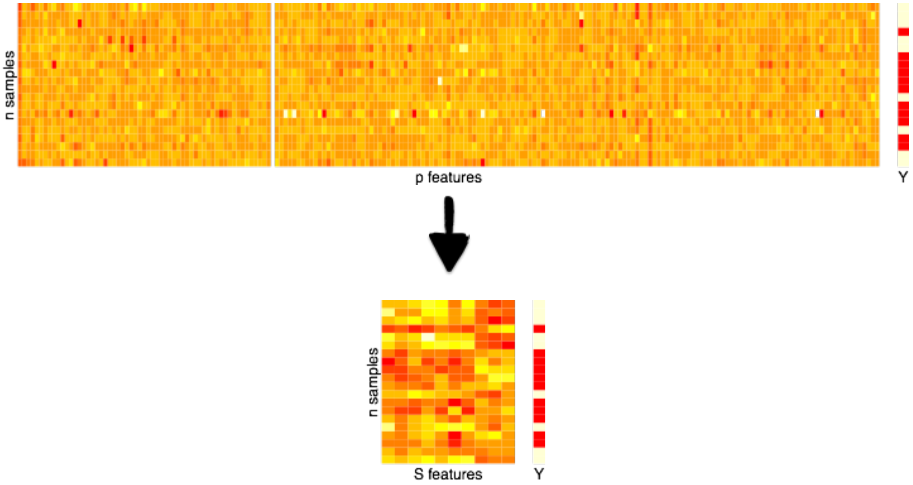
2. Multitask learning for toxicogenetics

Outline

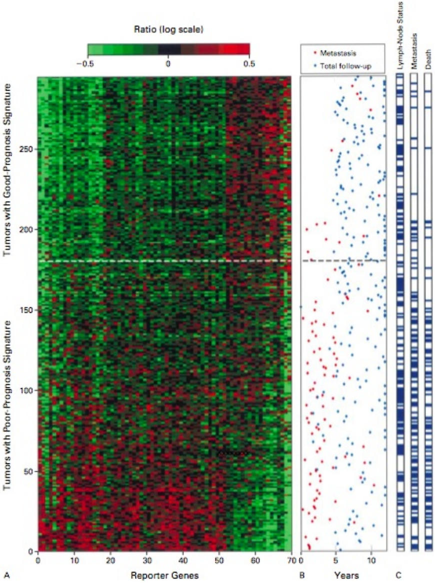
1. Learning molecular signatures with network information

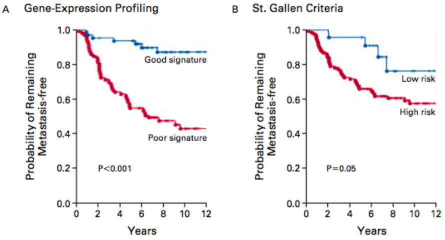
2. Multitask learning for toxicogenetics

Feature selection (a.k.a. *molecular signature*)



Example: Breast cancer prognostic signature





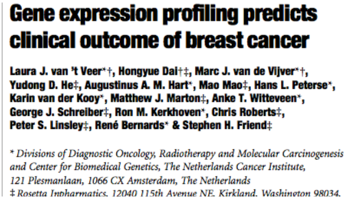
No. at Risk

Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

No. at Risk

Low risk	22	22	21	17	9	5	2
High risk	129	107	88	69	48	34	19

But...

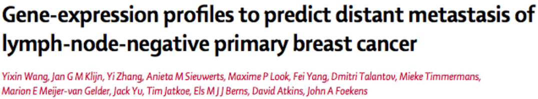


Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

** Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*

‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034,



Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

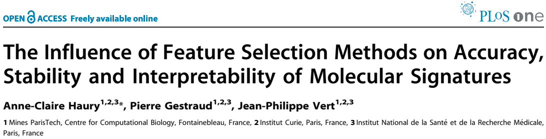
Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

70 genes (Nature, 2002)

76 genes (Lancet, 2005)

3 genes in common

3 genes is the best you can expect given *n* and *p*



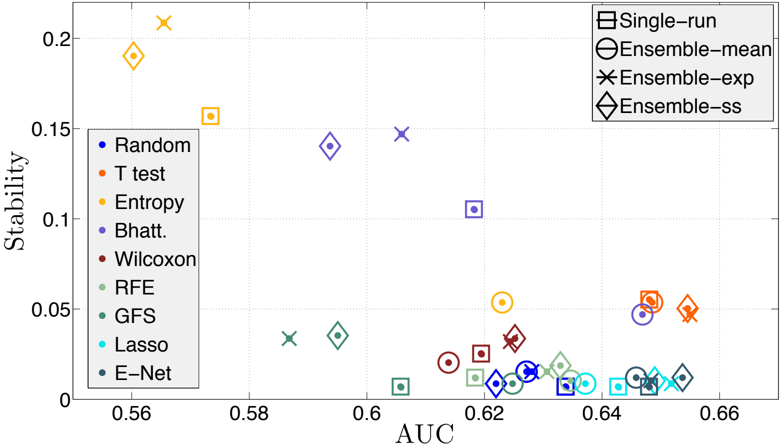
OPEN ACCESS Freely available online

PLoS one

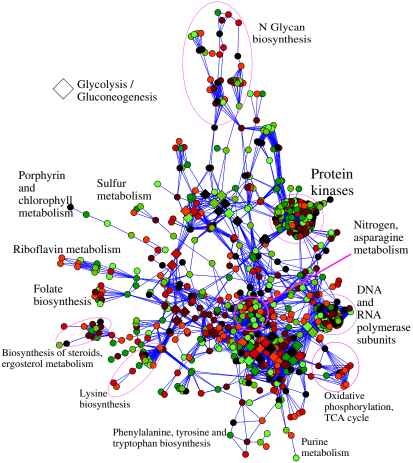
The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures

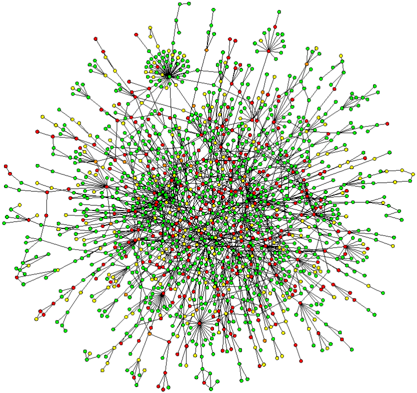
Anne-Claire Haury[1,2,3]*, Pierre Gestraud[1,2,3], Jean-Philippe Vert[1,2,3]

1 Mines ParisTech, Centre for Computational Biology, Fontainebleau, France, 2 Institut Curie, Paris, France, 3 Institut National de la Santé et de la Recherche Médicale, Paris, France



Gene networks as prior knowledge

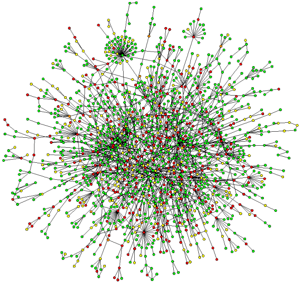

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Riboflavin metabolism
Nitrogen, asparagine metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism

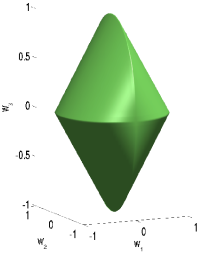


Can we force the signatures to be "coherent" with a known gene network?

Network-driven structured feature selection (Jacob et al., 2009)

1. Using the network, define a non-smooth and convex subset of "candidate" signatures compatible with it

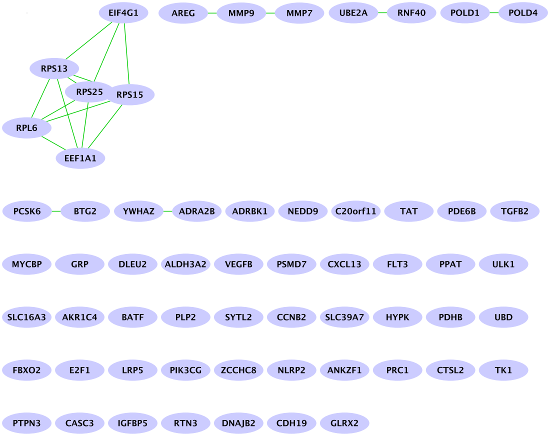




$$\Omega(\beta) = \sup_{\alpha\in\mathbb{R}^p:\forall i\sim j, \|\alpha_i^2+\alpha_j^2\|\leq 1} \alpha^\top \beta \,.$$

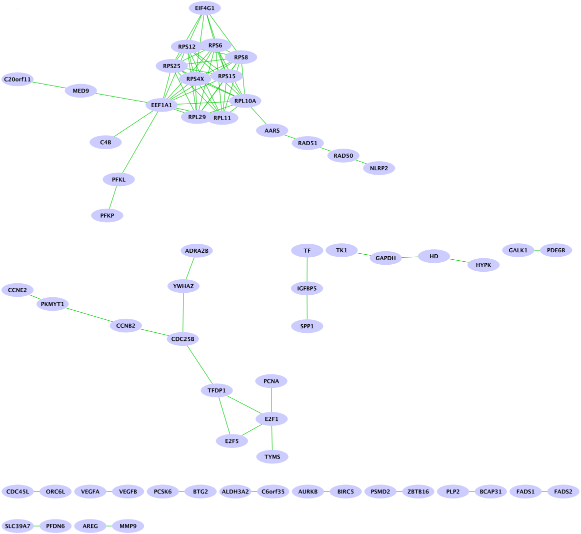
2. Among the candidates, find the best signature that explains the data (efficient optimization through convex programming)

Lasso signature (accuracy 0.61)



Breast cancer prognosis

Graph Lasso signature (accuracy 0.64)



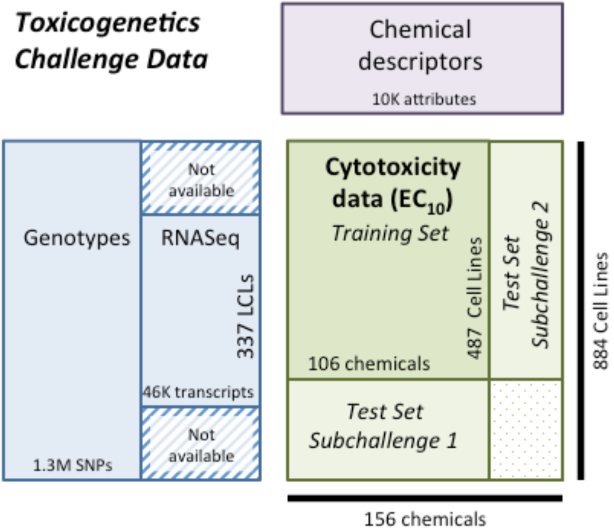
Breast cancer prognosis

Outline

1 Learning molecular signatures with network information

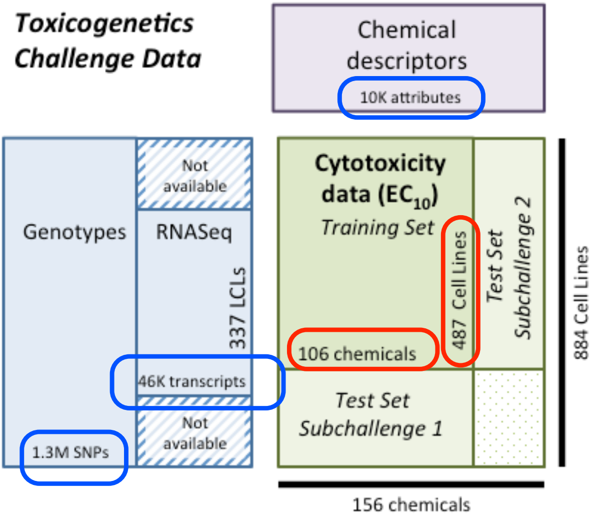
2 Multitask learning for toxicogenetics

Toxicogenetics



Genotypes from the 1000 genome project
RNASeq from the Geuvadis project

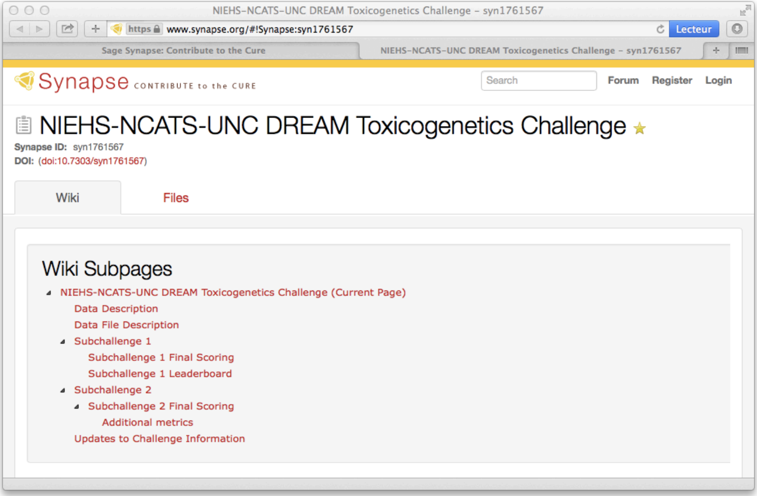
Again, $n << p$



n = 5E4

p = 1E10

Crowd-sourcing: the DREAM8 Toxicogenetics challenge



NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
https www.synapse.org/#!Synapse:syn1761567
Lecteur
Sage Synapse: Contribute to the Cure
NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge - syn1761567
Synapse CONTRIBUTE to the CURE
Search
Forum
Register
Login

NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge ★

Synapse ID: syn1761567

DOI: (doi:10.7303/syn1761567)

Wiki Subpages

- NIEHS-NCATS-UNC DREAM Toxicogenetics Challenge (Current Page)
  - Data Description
  - Data File Description
  - Subchallenge 1
    - Subchallenge 1 Final Scoring
    - Subchallenge 1 Leaderboard
  - Subchallenge 2
    - Subchallenge 2 Final Scoring
      - Additional metrics
  - Updates to Challenge Information

Bilinear regression

- Cell line X, chemical Y, toxicity Z.
- Bilinear regression model:

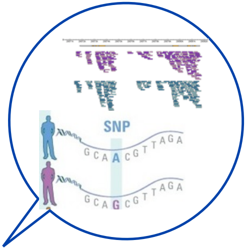
$$Z = f(X, Y) + b(Y) + \epsilon\,,$$

- Estimation by kernel ridge regression:

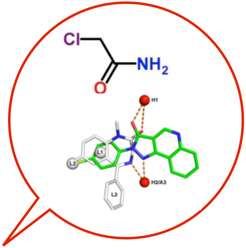
$$\min_{f \in \mathcal{H}, b \in \mathbb{R}^m} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(f(x_i, y_j) + b_j - z_{ij} \right)^2 + \lambda \|f\|^2\,,$$

- Solved in $O(max(n, p)^3)$

Kernel Trick

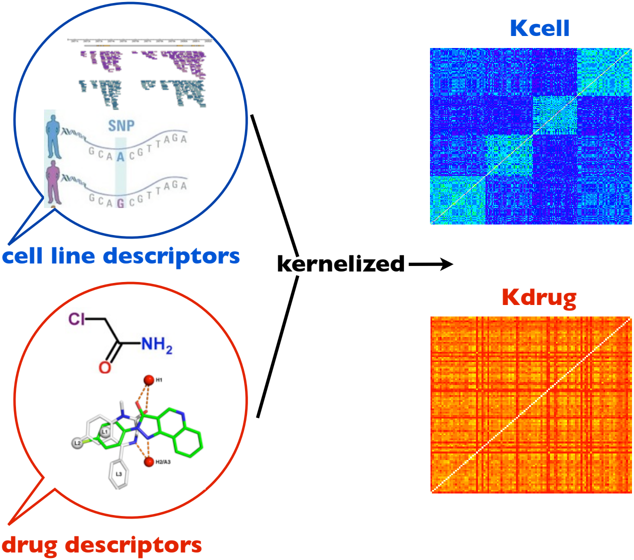


cell line descriptors



drug descriptors

Kernel Trick



cell line descriptors

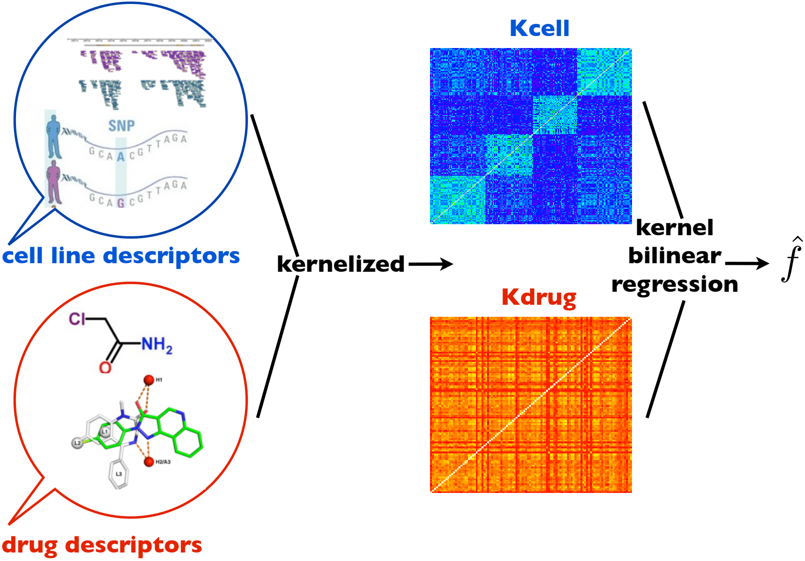
drug descriptors

kernelized →

Kcell

Kdrug

Kernel Trick



cell line descriptors

drug descriptors

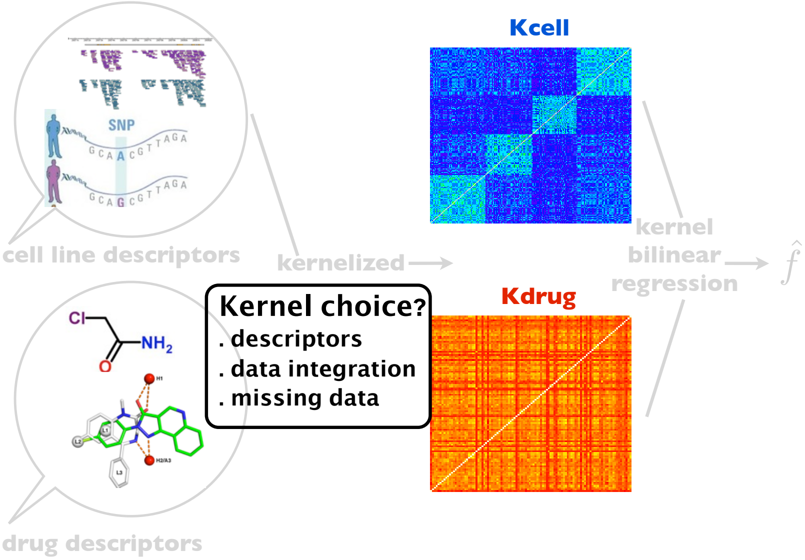
kernelized →

Kcell

Kdrug

kernel bilinear regression → $\hat{f}$

Kernel Trick



Kernel choice?

. descriptors

. data integration

. missing data

Kernel choice

1. $\mathbf{K}_{\text{cell}}$:

   $\Longrightarrow$ 29 cell line kernels tested

   $\Longrightarrow$ 1 kernel that *integrate all information*

   $\Longrightarrow$ deal with missing data

2. $\mathbf{K}_{\text{drug}}$:

   $\Longrightarrow$ 48 drug kernels tested

   $\Longrightarrow$ multi-task kernels

Kernel choice

1. $\mathbf{K}_{\text{cell}}$:

   $\Longrightarrow$ 29 cell line kernels tested

   $\Longrightarrow$ 1 kernel that *integrate all information*

   $\Longrightarrow$ deal with missing data

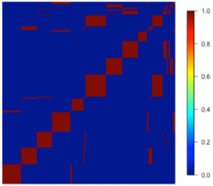
2. $\mathbf{K}_{\text{drug}}$:

   $\Longrightarrow$ 48 drug kernels tested

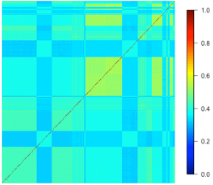
   $\Longrightarrow$ multi-task kernels

Cell line data integration

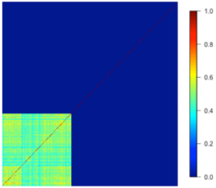
Covariates
. linear kernel



SNPs
. 10 gaussian
kernels

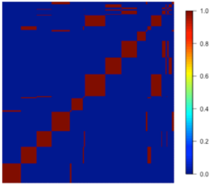


RNA-seq
. 10 gaussian
kernels

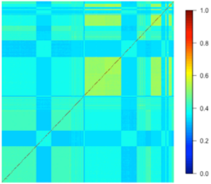


Cell line data integration

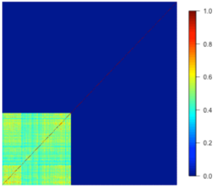
Covariates
. linear kernel



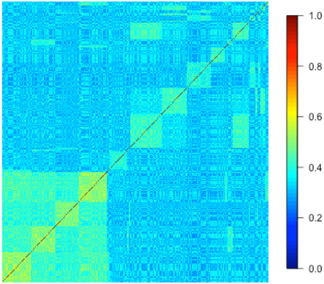
SNPs
. 10 gaussian
kernels



RNA-seq
. 10 gaussian
kernels

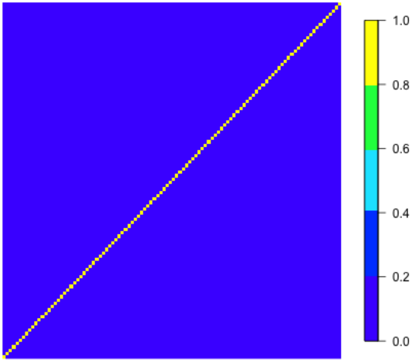


Integrated kernel



Multi-task drug kernels

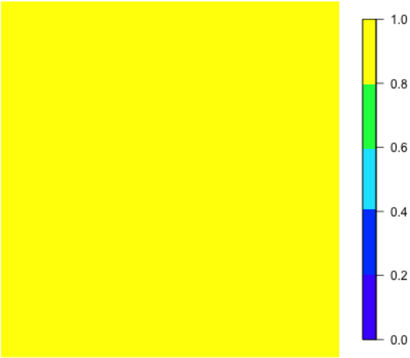
1. **Dirac**
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated



independent regression for each drug

Multi-task drug kernels

1. Dirac
2. **Multi-Task**
3. Feature-based
4. Empirical
5. Integrated



sharing information across drugs

Multi-task drug kernels

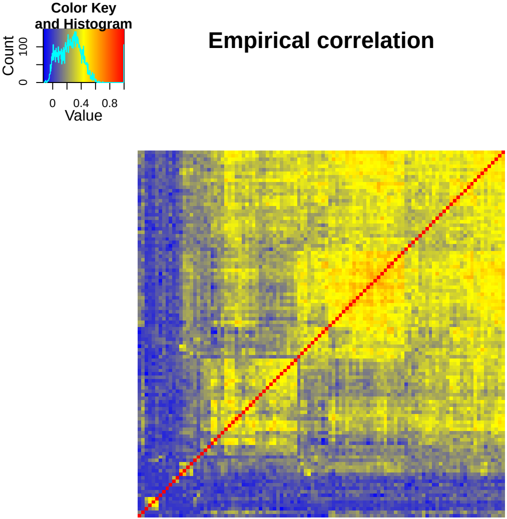
1. Dirac
2. Multi-Task
3. **Feature-based**
4. Empirical
5. Integrated

Linear kernel and 10 gaussian kernels based on features:

- CDK (160 descriptors) and SIRMS (9272 descriptors)
- Graph kernel for molecules (2D walk kernel)
- Fingerprint of 2D substructures (881 descriptors)
- Ability to bind human proteins (1554 descriptors)

Multi-task drug kernels

1. Dirac
2. Multi-Task
3. Feature-based
4. **Empirical**
5. Integrated



Multi-task drug kernels

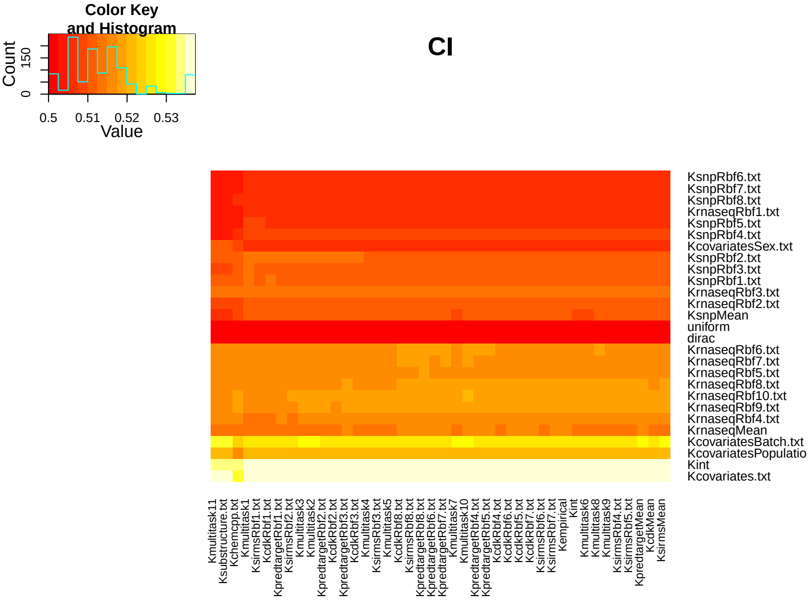
1. Dirac
2. Multi-Task
3. Feature-based
4. Empirical
5. **Integrated**

$$K_{int} = \sum_i K_i$$

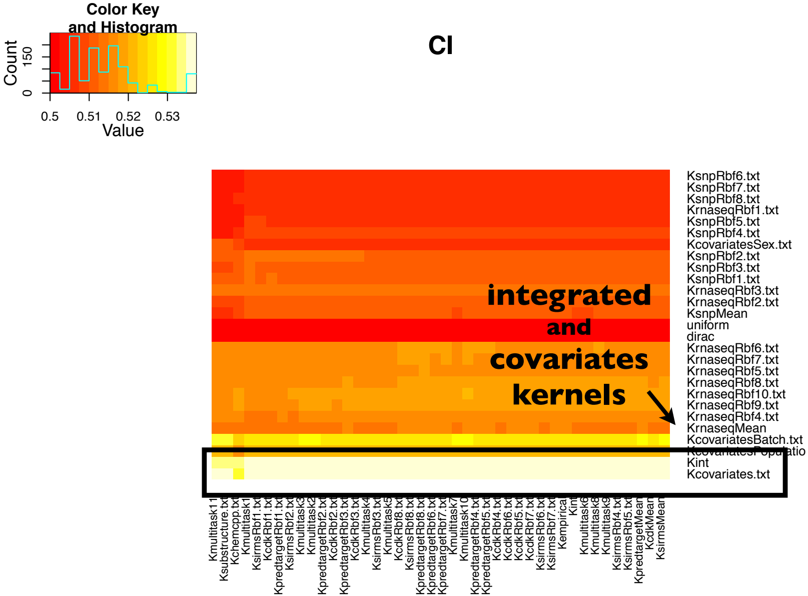
Integrated kernel:

- Combine all information on drugs

29x48 kernel combinations: CV results

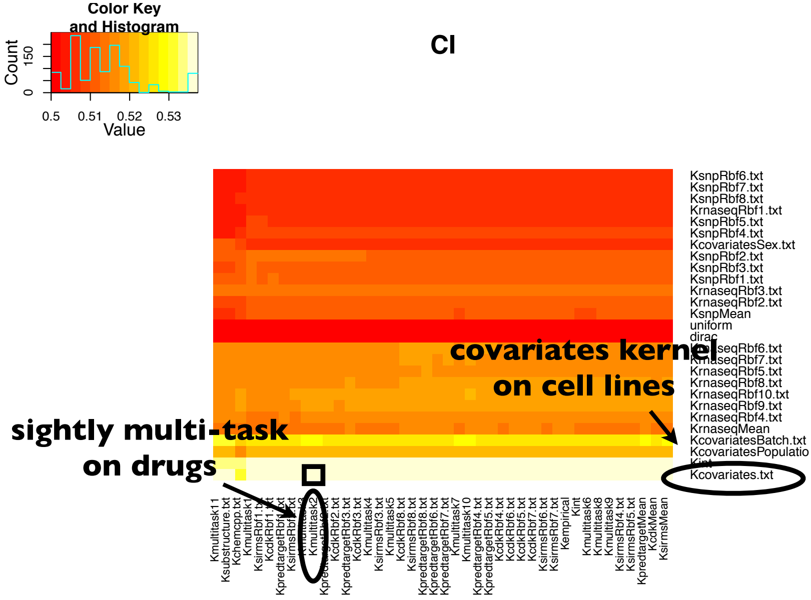


29x48 kernel combinations: CV results



integrated and covariates kernels

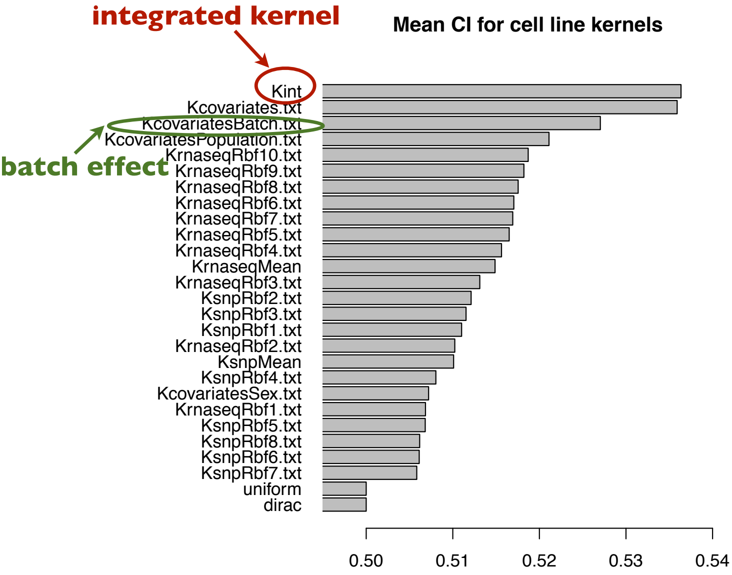
29x48 kernel combinations: CV results



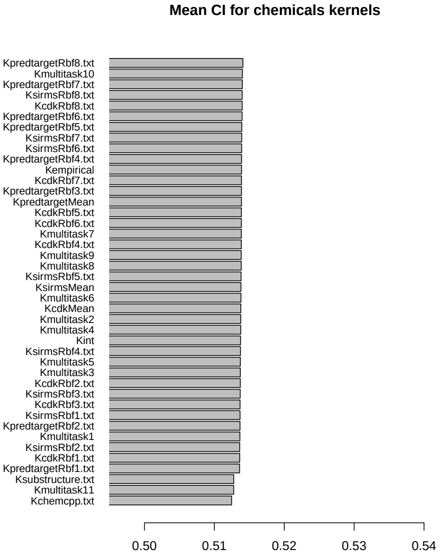
covariates kernel on cell lines

sightly multi-task on drugs

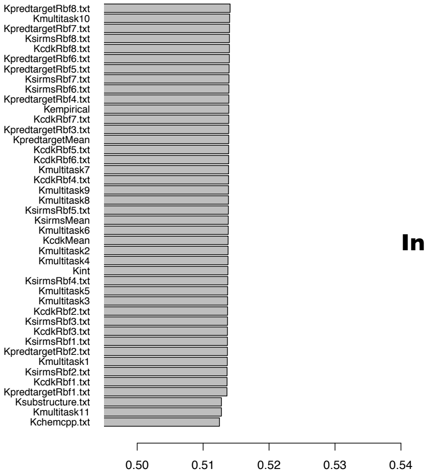
Kernel on cell lines: CV results



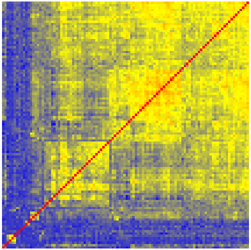
Kernel on drugs: CV results



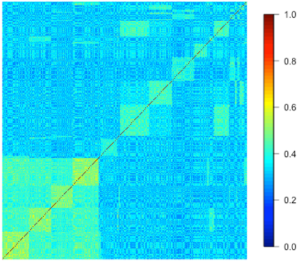
Final Submission (ranked 2nd)


Mean CI for chemicals kernels

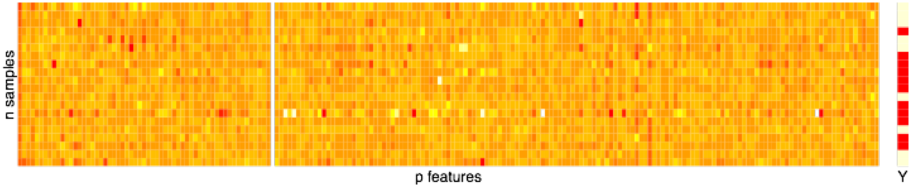
Empirical kernel on drugs



Integrated kernel on cell lines



Conclusion



- Small *n* large *p* $\Longrightarrow$ regularized models with prior knowledge
- Heterogeneous data integration $\Longrightarrow$ kernel methods
- Performance remains often disappointing!
- Progress arise by small steps

Thanks

cbio@mines-paristech.fr , u900@curie.fr, sandrine@stat.berkeley.edu

The Adolph C. and Mary Sprague
Miller Institute for Basic
Research in Science
University of California, Berkeley