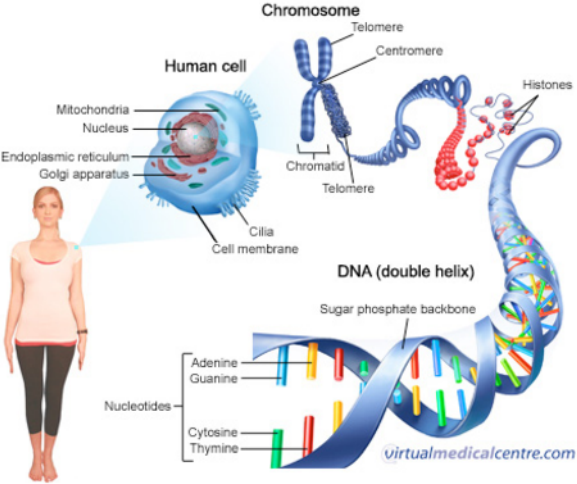
Machine Learning for Personalized Medicine

Jean-Philippe Vert

MascotNum 2014, ETH Zurich, April 24, 2014

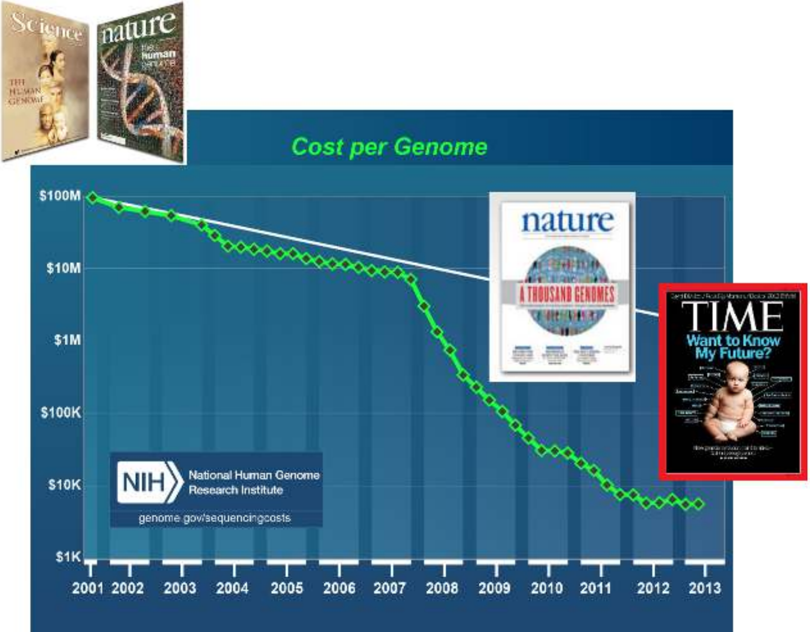
Complexity of life



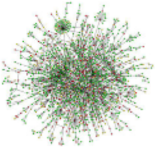
1 body = 10^{14} cells

1 cell = 6×10^9 ACGT coding for $20,000$ genes

Sequencing revolution



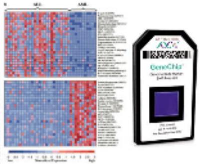
A flood of *omics* data



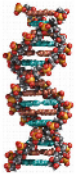
Interactome

PubMed

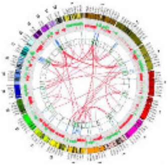
Publications



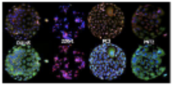
Transcriptome



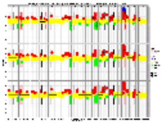
Genome



Mutations
Structural variations

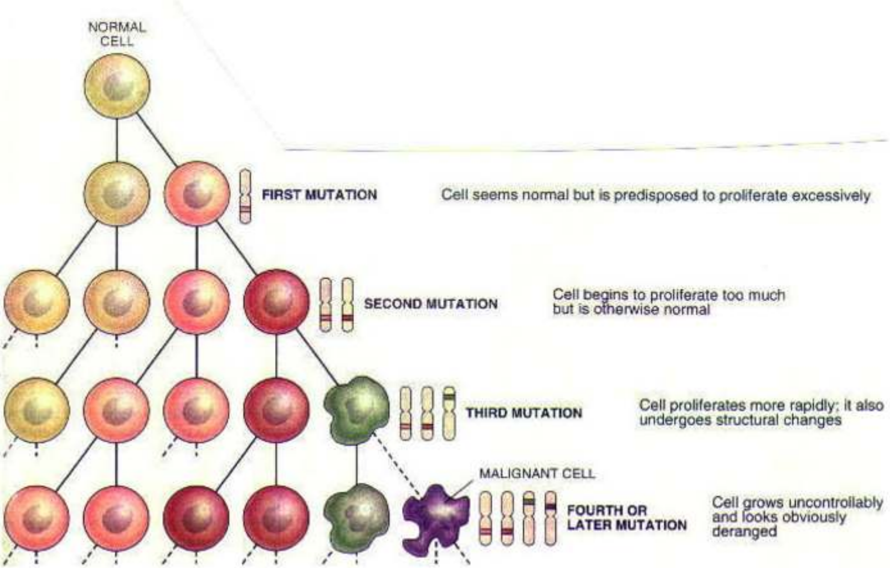


Phenome



Epigenome

Cancer



NORMAL CELL

FIRST MUTATION

Cell seems normal but is predisposed to proliferate excessively

SECOND MUTATION

Cell begins to proliferate too much but is otherwise normal

THIRD MUTATION

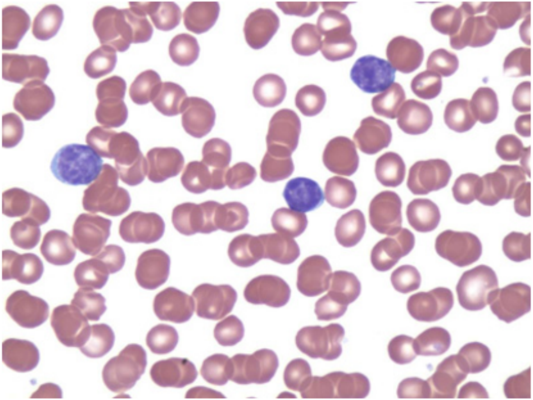
Cell proliferates more rapidly; it also undergoes structural changes

MALIGNANT CELL

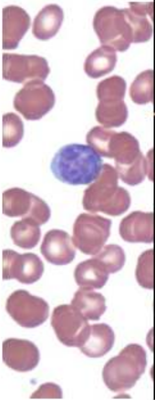
FOURTH OR LATER MUTATION

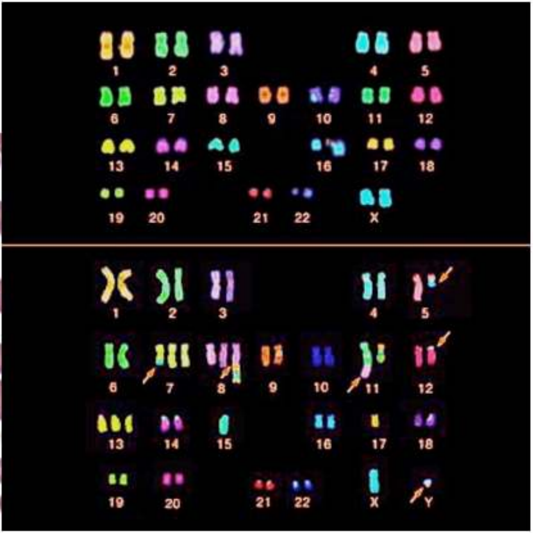
Cell grows uncontrollably and looks obviously deranged

A cancer cell

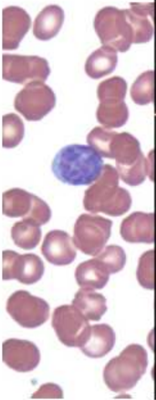


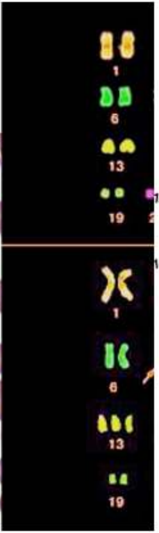
A cancer cell

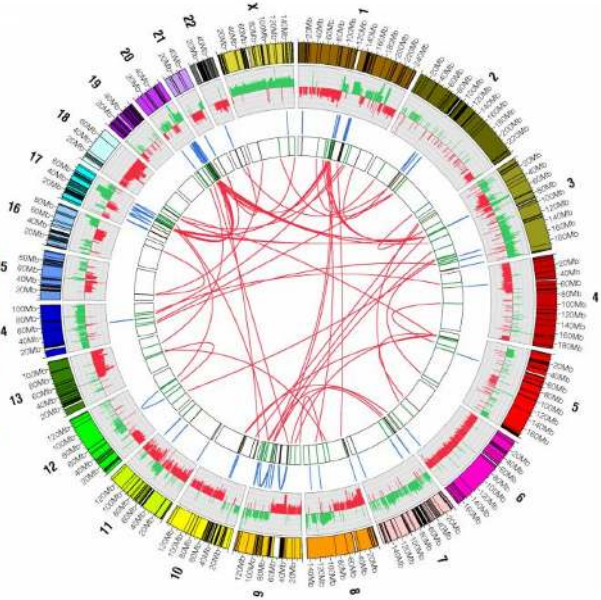




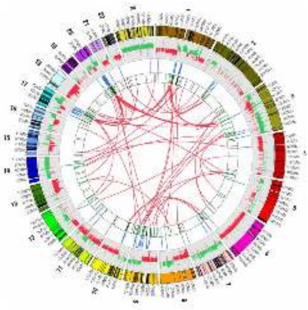
A cancer cell







Opportunities



- What is your risk of developing a cancer? (*prevention*)
- After diagnosis and treatment, what is the risk of relapse? (*prognosis*)
- What specific treatment will cure your cancer? (*personalized medicine*)

Outline

1. Learning molecular classifiers with network information

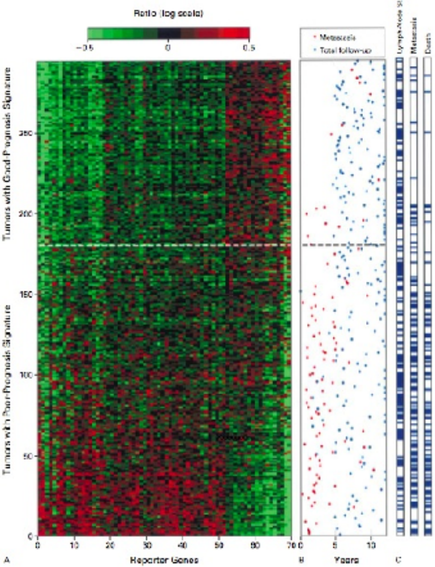
2. Kernel bilinear regression for toxicogenomics

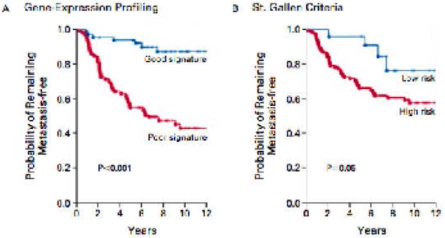
Outline

1. Learning molecular classifiers with network information

2. Kernel bilinear regression for toxicogenomics

Breast cancer prognosis





No. at Risk							
Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	20	17	9

No. at Risk							
Low risk	22	22	21	17	9	5	2
High risk	129	107	88	69	48	34	19

Learning with regularization

Given a training set $(x_i, y_i)_{i=1,\ldots,n}$ where $x_i \in \mathbb{R}^p$ (typically, $n = 200, p = 20,000$), we estimate a linear predictor

$$f_\beta(x) = \beta^\top x$$

by solving

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \Omega(\beta)$$

where:

- $R(\beta)$ is a convex empirical risk, typically

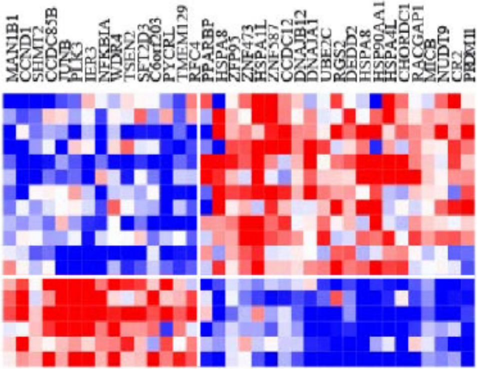
  $$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\beta^\top x_i, y_i)$$

  for some loss function ℓ (squared error, logistic loss, hinge loss...)
- $\Omega(\beta)$ is a regularization term, typically $\| \beta \|_2$ (ridge regression, SVM...) or $\| \beta \|_1$ (lasso...)

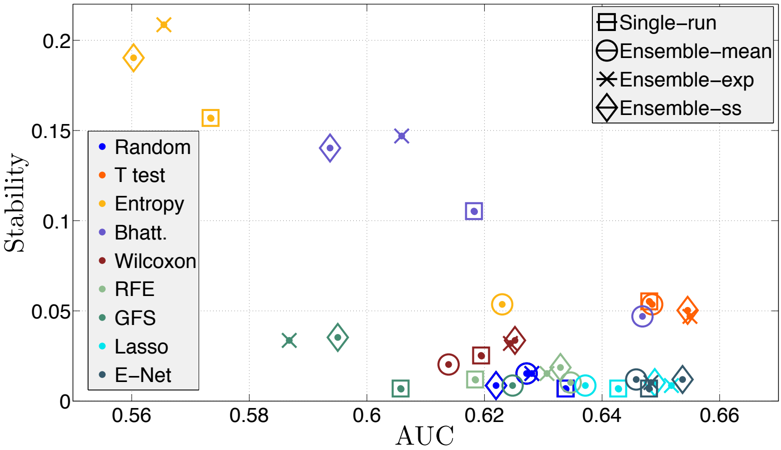
Gene selection, molecular signature

The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology

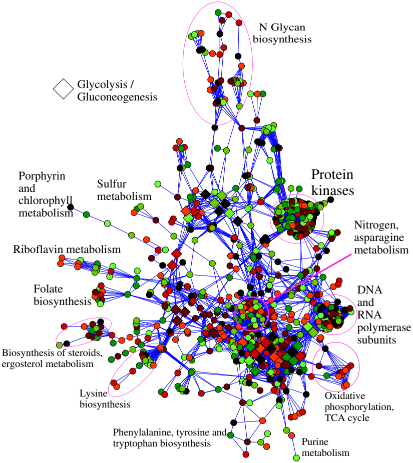


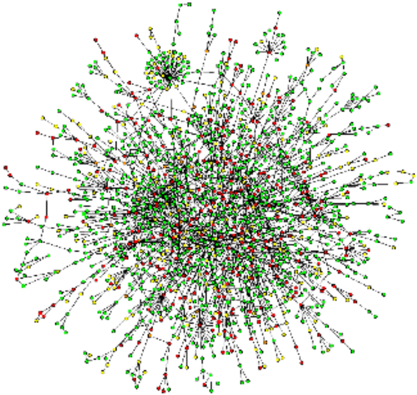
Lack of stability of signatures



Haury et al. (2011)

Gene networks

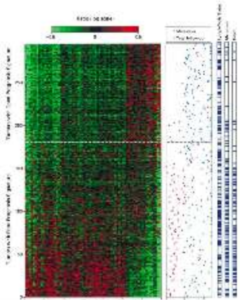

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative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

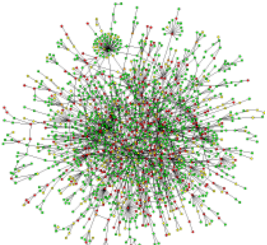


Gene networks and expression data

Motivation

- Basic biological functions usually involve the coordinated action of several proteins:
  - Formation of protein complexes
  - Activation of metabolic, signalling or regulatory pathways
- Many pathways and protein-protein interactions are already known
- Hypothesis: the weights of the classifier should be “coherent” with respect to this prior knowledge





Graph based penalty

$$f_\beta(x) = \beta^\top x \qquad\qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

$$\Omega(\beta) = \sum_{i\sim j} (\beta_i - \beta_j)^2 ,$$

$$\min_{\beta\in\mathbb{R}^p} R(f_\beta) + \lambda \sum_{i\sim j} (\beta_i - \beta_j)^2 .$$

Graph based penalty

$$f_\beta(x) = \beta^\top x \qquad \min_\beta R(f_\beta) + \lambda\Omega(\beta)$$

Prior hypothesis

Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

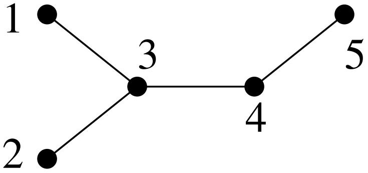
$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 ,$$

$$\min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2 .$$

Graph Laplacian

Definition

The Laplacian of the graph is the matrix $L = D - A$.



$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Spectral penalty as a kernel

Theorem

The function $f(x) = \beta^\top x$ where β is solution of

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\beta^\top x_i, y_i\right) + \lambda \sum_{i \sim j} \left(\beta_i - \beta_j\right)^2$$

is equal to $g(x) = \gamma^\top \Phi(x)$ where γ is solution of

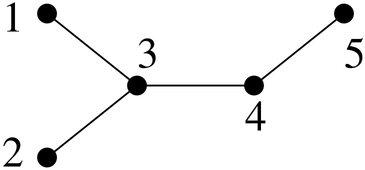
$$\min_{\gamma \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\gamma^\top \Phi(x_i), y_i\right) + \lambda \gamma^\top \gamma \,,$$

and where

$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

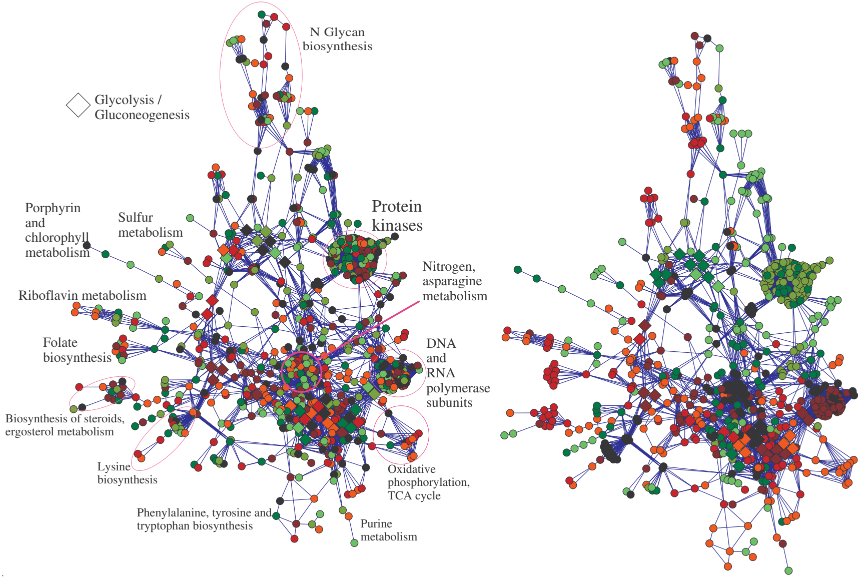
for $K_G = L^*$, the pseudo-inverse of the graph Laplacian.

Example

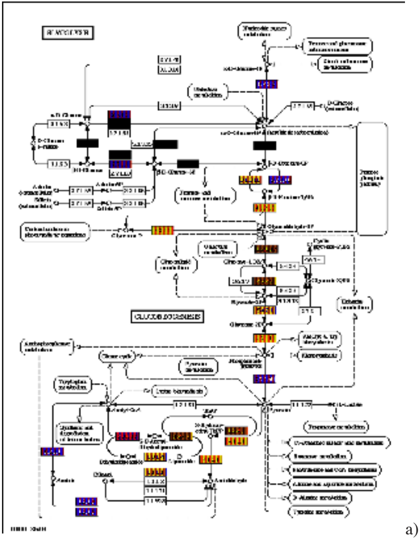


$$L^* = \begin{pmatrix} 0.88 & -0.12 & 0.08 & -0.32 & -0.52 \\ -0.12 & 0.88 & 0.08 & -0.32 & -0.52 \\ 0.08 & 0.08 & 0.28 & -0.12 & -0.32 \\ -0.32 & -0.32 & -0.12 & 0.48 & 0.28 \\ -0.52 & -0.52 & -0.32 & 0.28 & 1.08 \end{pmatrix}$$

Classifiers


N Glycan biosynthesis
◇ Glycolysis / Gluconeogenesis
Protein kinases
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Nitrogen, asparagine metabolism
Riboflavin 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

Classifier



Other penalties with kernels

$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

with:

- $K_G = (c + L)^-1$ leads to

$$\Omega(\beta) = c \sum_{i=1}^{p} \beta_i^2 + \sum_{i \sim j} \left(\beta_i - \beta_j\right)^2 .$$

- The diffusion kernel:

$$K_G = \exp_M(-2tL) .$$

penalizes high frequencies of β in the Fourier domain.

Other penalties without kernels

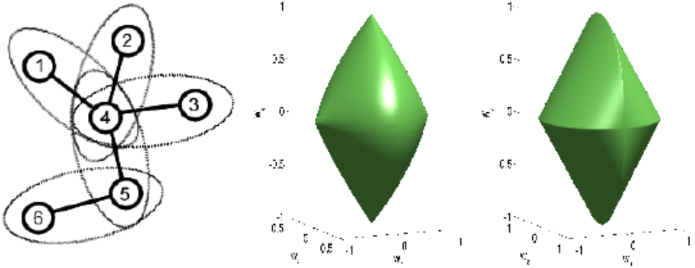
- Gene selection + Piecewise constant on the graph

$$\Omega(\beta) = \sum_{i \sim j} \left| \beta_i - \beta_j \right| + \sum_{i=1}^{p} | \beta_i |$$

- Gene selection + smooth on the graph

$$\Omega(\beta) = \sum_{i \sim j} \left(\beta_i - \beta_j \right)^2 + \sum_{i=1}^{p} | \beta_i |$$

Graph lasso



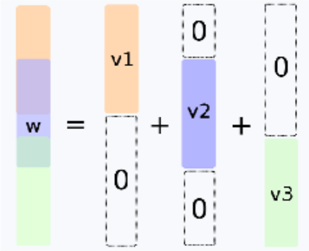
Two solutions

$$\Omega_{intersection}(\beta) = \sum_{i\sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

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

Generalization: Group lasso with overlapping groups

$$\Omega^{\mathcal{G}}_{\text{latent}}(w) \triangleq \begin{cases} \min_{v} \sum_{g\in\mathcal{G}} \|v_g\|_2 \\ w = \sum_{g\in\mathcal{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Equivalent to group lasso when there is no overlap

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{latent}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda\Omega^{\mathcal{G}}_{\text{latent}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

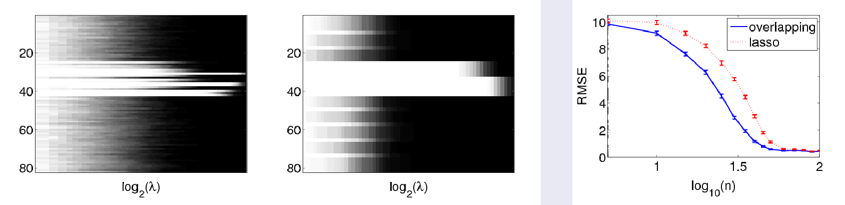
the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g\in\mathcal{G}}$ such that

$$\left\{g \in \mathcal{G} | \hat{v}_g \neq 0\right\} = \left\{g \in \mathcal{G} | \bar{v}_g \neq 0\right\}.$$

Experiments

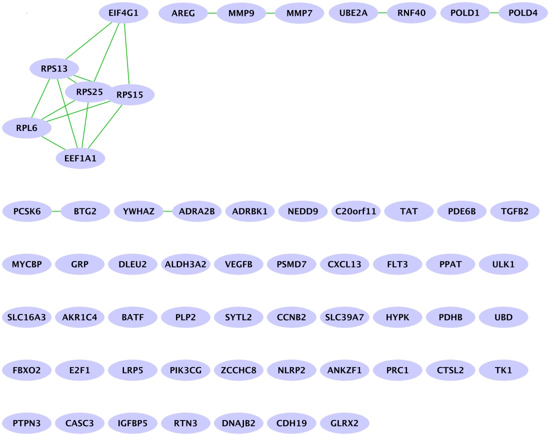
Synthetic data: overlapping groups

- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1,\ldots,10\},\{9,\ldots,18\},\ldots,\{73,\ldots,82\}$.
- Support: union of $4th$ and $5th$ groups.
- Learn from 100 training points.

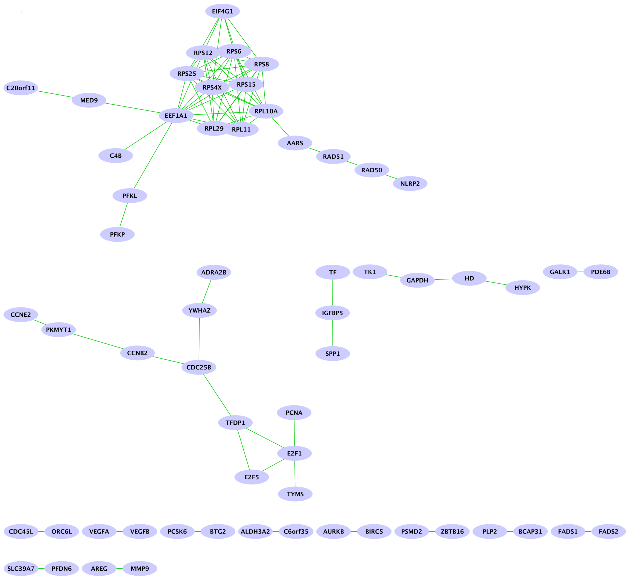


Frequency of selection of each variable with the lasso (left) and $\Omega^{\mathcal{G}}_{\text{latent}}(.)$ (middle), comparison of the RMSE of both methods (right).

Lasso signature (accuracy 0.61)



Graph Lasso signature (accuracy 0.64)

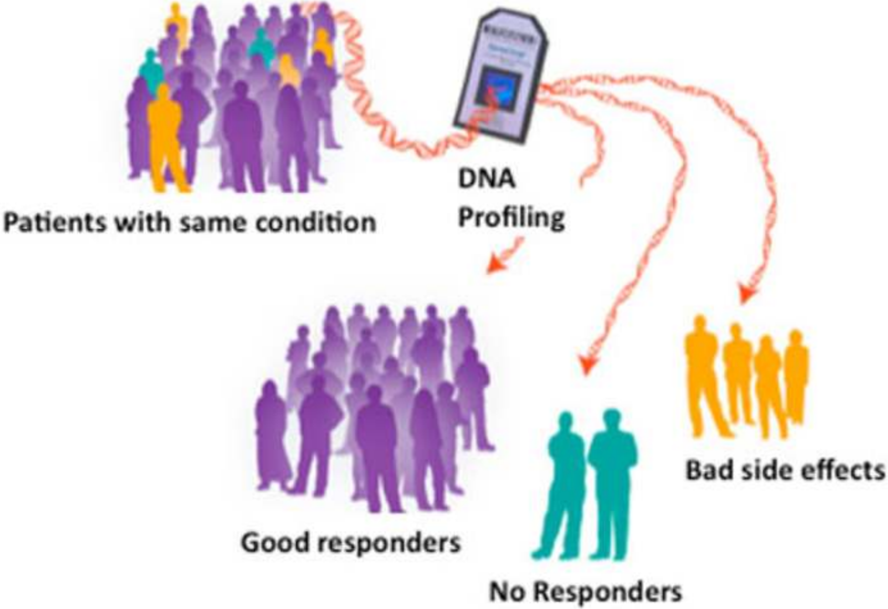


Outline

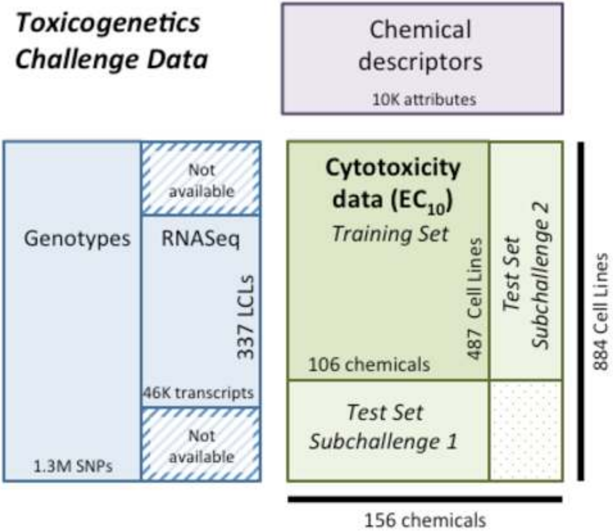
1. Learning molecular classifiers with network information

2. Kernel bilinear regression for toxicogenomics

Pharmacogenomics / Toxicogenomics



DREAM8 Toxicogenetics challenge



Genotypes from the 1000 genome project
RNASeq from the Geuvadis project

Bilinear regression

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

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

- Estimation by kernel ridge regression:

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

Solving in $O(max(n, p)^3)$

Theorem 1. *Let $Z \in \mathbb{R}^{n\times p}$ be the response matrix, and $K_X \in \mathbb{R}^{n\times n}$ and $K_Y \in \mathbb{R}^{p\times p}$ be the kernel Gram matrices of the n cell lines and p chemicals, with respective eigenvalue decompositions $K_X = U_X D_X U_X^\top$ and $K_Y = U_Y D_Y U_Y^\top$. Let $\gamma = U_X^\top \mathbf{1}_n$ and $S \in \mathbb{R}^{n\times p}$ be defined by $S_{ij} = 1/\left(\lambda + D_X^i D_Y^j\right)$, where D_X^i (resp. D_Y^i) denotes the i-th diagonal term of D_X (resp. D_Y). Then the solution (f^*, b^*) of (2) is given by*

$$b^* = U_Y Diag\left(S^\top \gamma^{\circ 2}\right)^{-1}\left(S^\top \circ \left(U_Y^\top Z^\top U_X\right)\right)\gamma \tag{3}$$

and

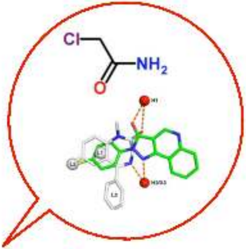
$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad f^*(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{p} \alpha^*_{i,j} K_X(x_i, x) K_Y(y_i, y)\,, \tag{4}$$

where

$$\alpha^* = U_X\left(S \circ \left(U_X^\top\left(Z - \mathbf{1}_n b^{*\top}\right)U_Y\right)\right)U_Y^\top\,. \tag{5}$$

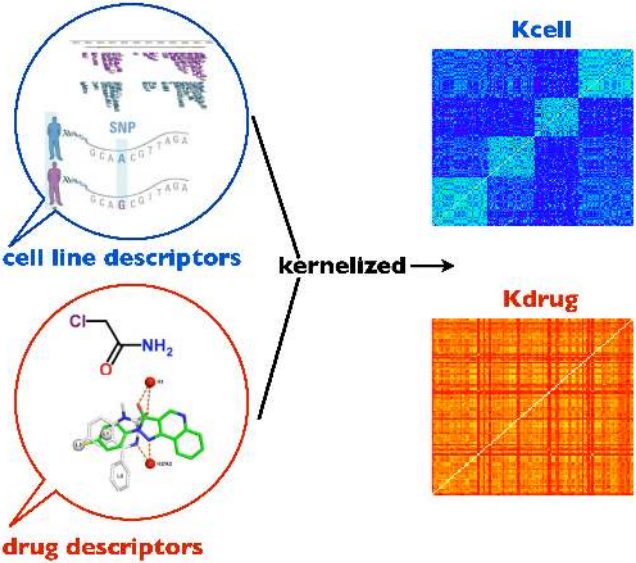
Kernel Trick

cell line descriptors



drug descriptors

Kernel Trick



cell line descriptors

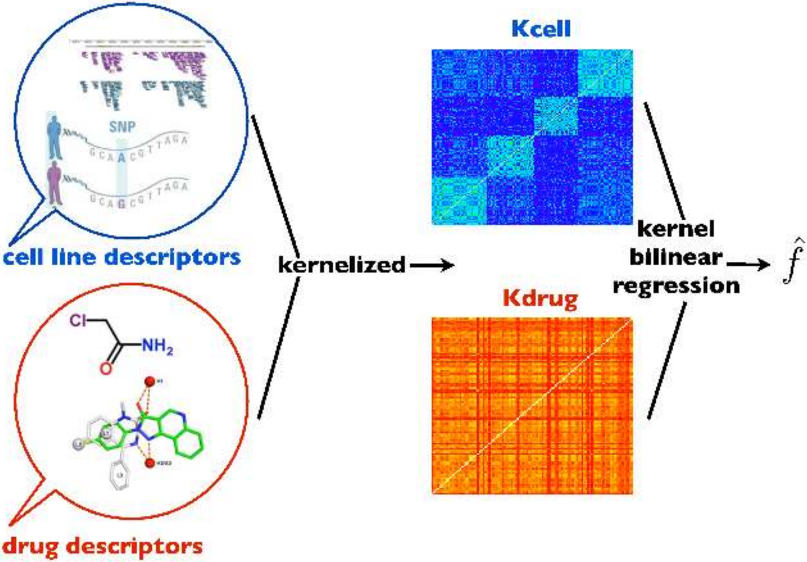
drug descriptors

kernelized →

Kcell

Kdrug

Kernel Trick



cell line descriptors

drug descriptors

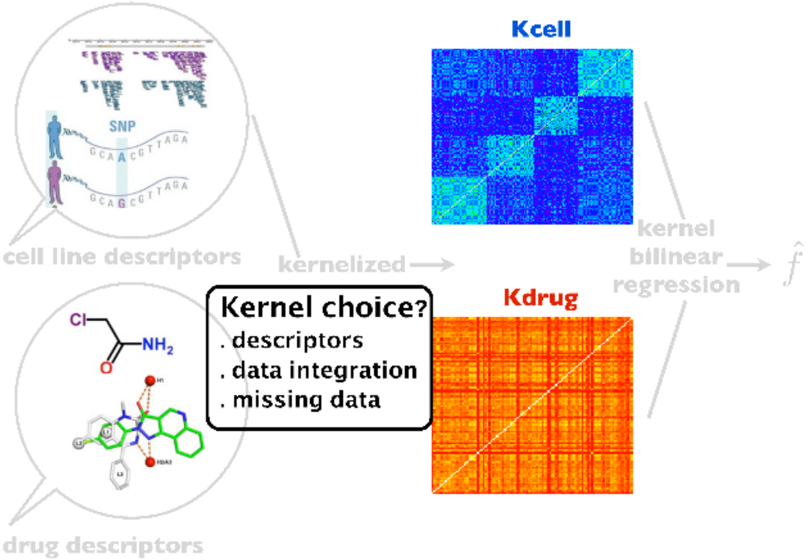
kernelized →

Kcell

Kdrug

kernel bilinear regression → $\hat{f}$

Kernel Trick



cell line descriptors

kernelized →

Kcell

kernel bilinear regression → $\hat{f}$

Kdrug

drug descriptors

Kernel choice?

. descriptors

. data integration

. missing data

Kernel choice

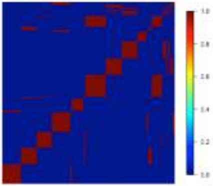
1. $\mathbf{K}_{\text{cell}}$:
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. $\mathbf{K}_{\text{drug}}$:
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Kernel choice

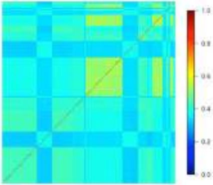
1. **K_{cell}** :
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. **K_{drug}** :
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Cell line data integration

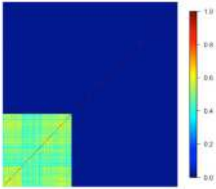
Covariates
. linear kernel



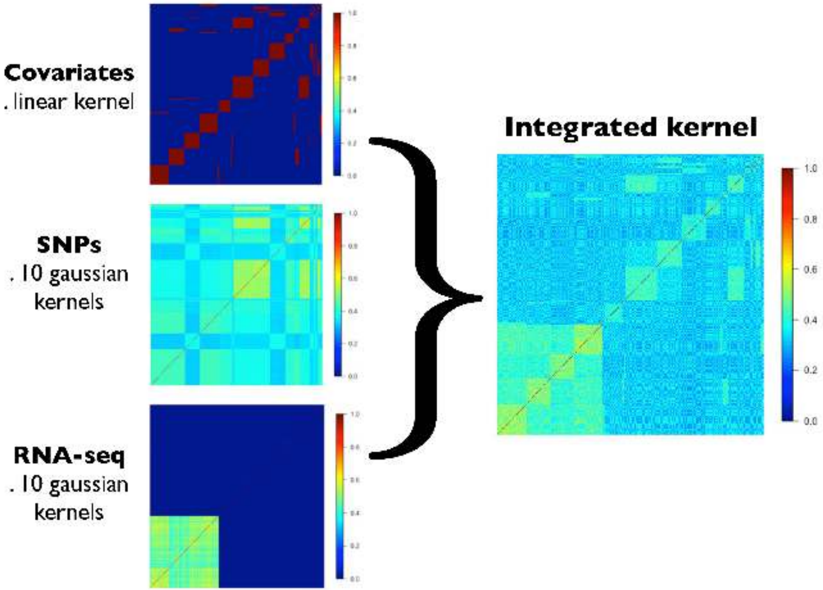
SNPs
. 10 gaussian kernels



RNA-seq
. 10 gaussian kernels



Cell line data integration



Covariates
. linear kernel

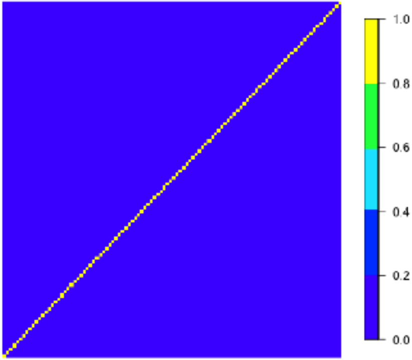
SNPs
. 10 gaussian kernels

RNA-seq
. 10 gaussian kernels

Integrated kernel

Multi-task drug kernels

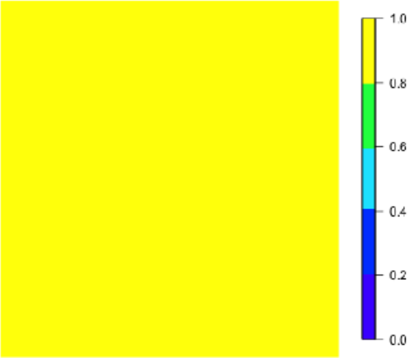
1. **Dirac**
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated



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

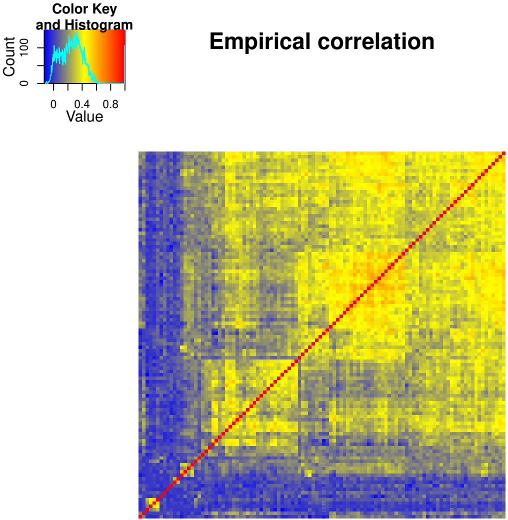
Multi-task drug kernels

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

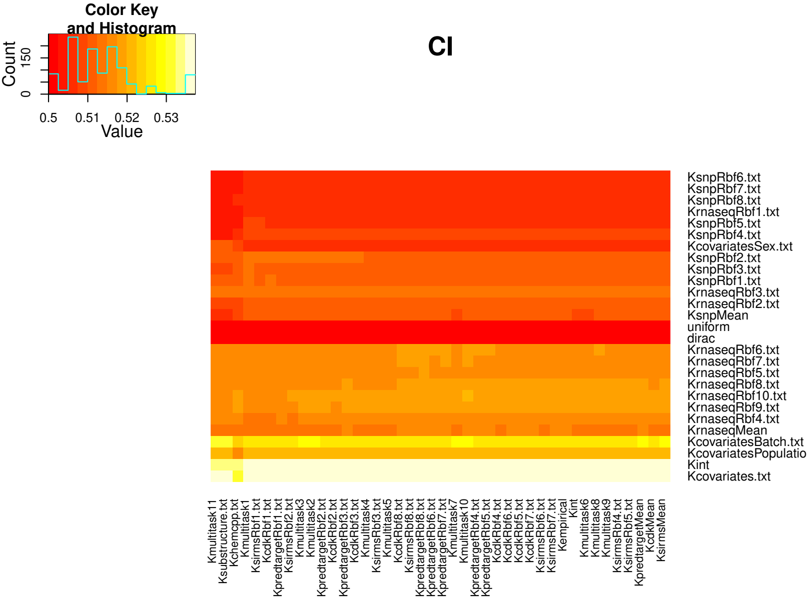
Multi-task drug kernels

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

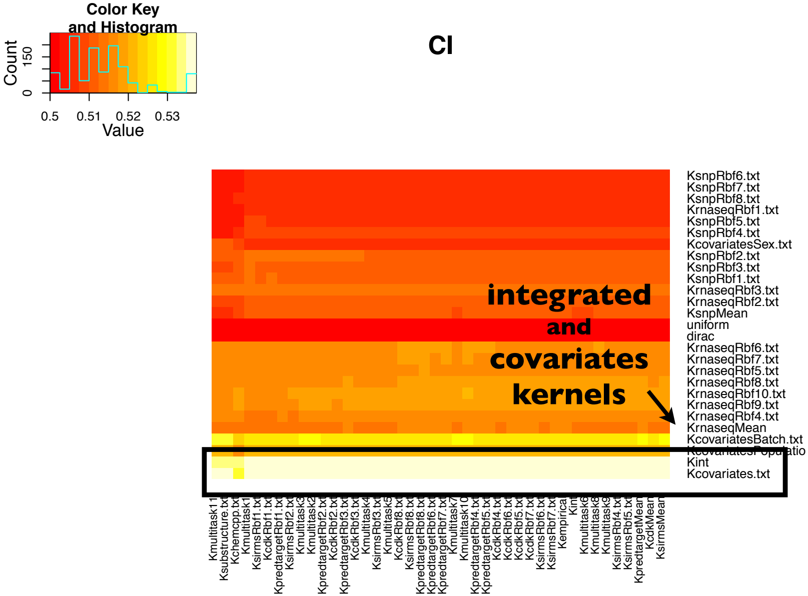
Integrated kernel:

- Combine all information on drugs

29x48 kernel combinations: CV results

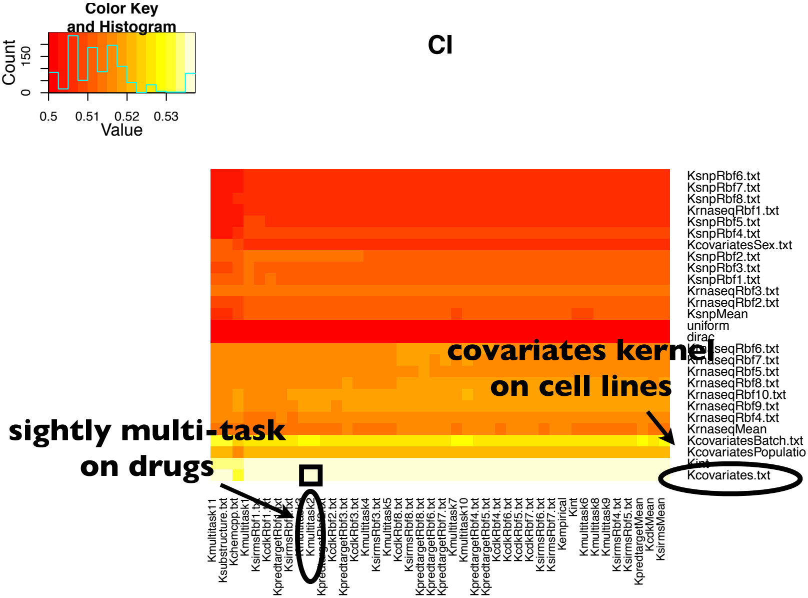


29x48 kernel combinations: CV results



integrated and covariates kernels

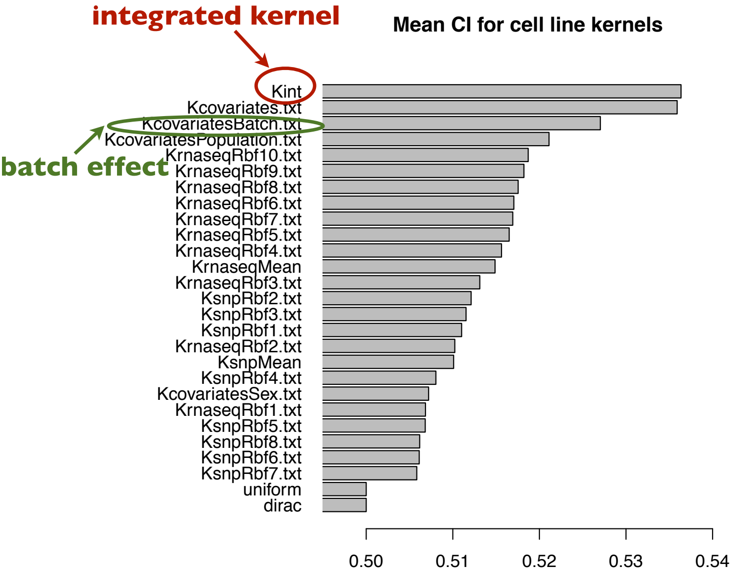
29x48 kernel combinations: CV results



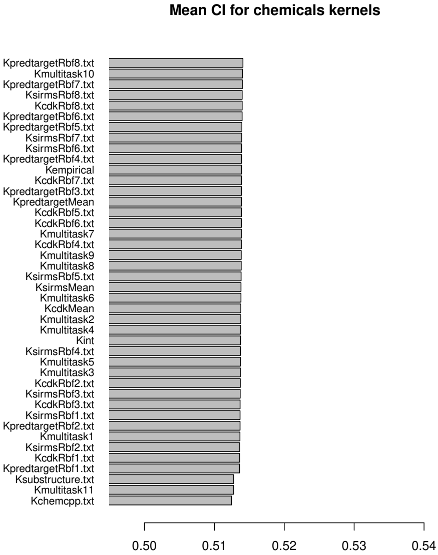
covariates kernel on cell lines

sightly multi-task on drugs

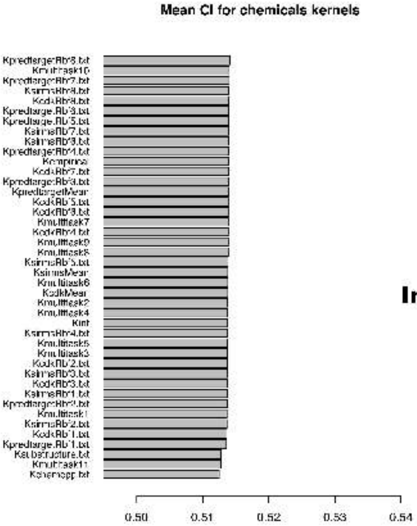
Kernel on cell lines: CV results



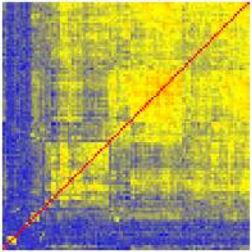
Kernel on drugs: CV results



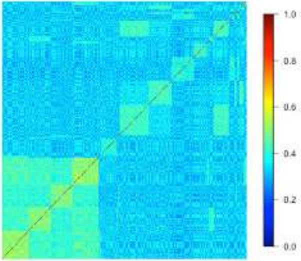
Final Submission (ranked 2nd)



Empirical kernel on drugs



Integrated kernel on cell lines



Thanks

institutCurie

European Research Council

erc

Inserm

Institut national de la santé et de la recherche médicale

MINES ParisTech