Machine Learning for Personalized Medicine

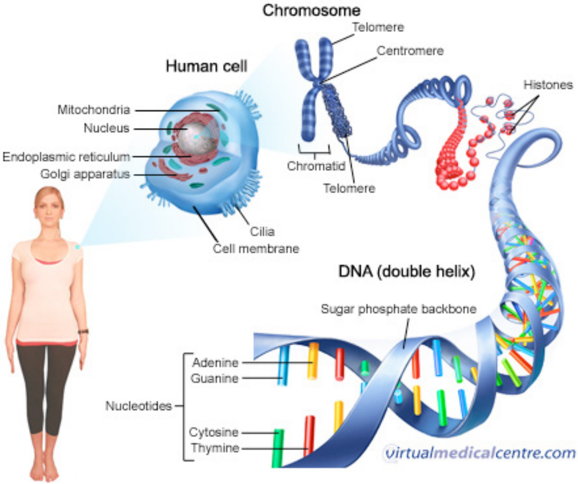
Jean-Philippe Vert

MINES ParisTech

institutCurie
Together, let's beat cancer.

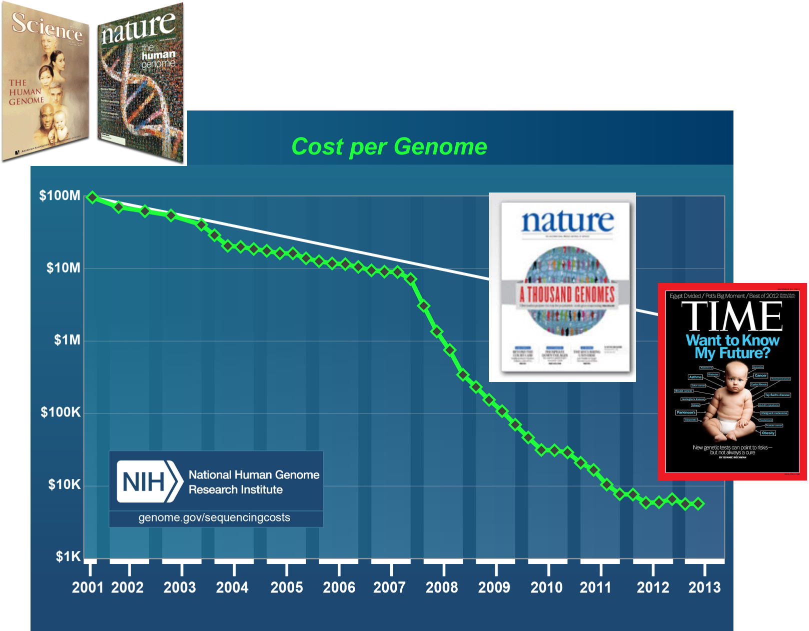
Genentech, July 24, 2014

What’s in your body

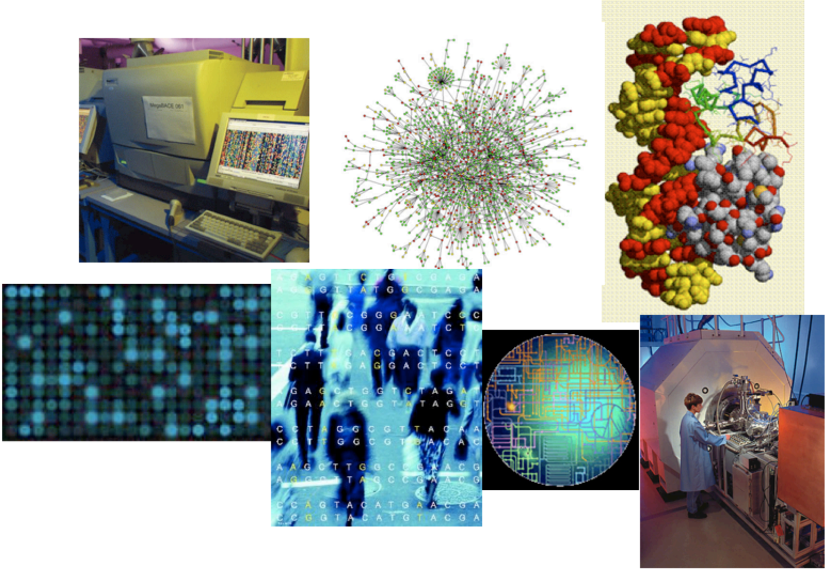


1 body = 10^{14} human cells (and 100x more non-human cells)
1 cell = 6×10^9 ACGT coding for $20,000$ genes

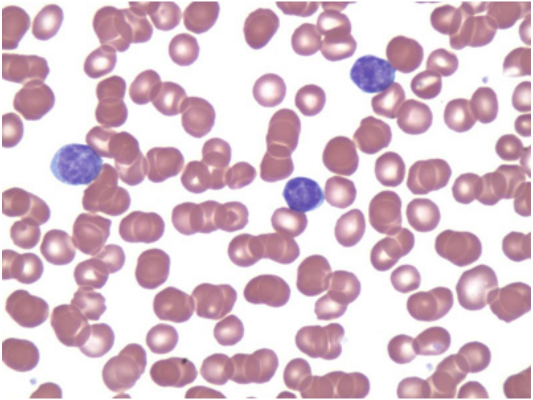
Sequencing revolution



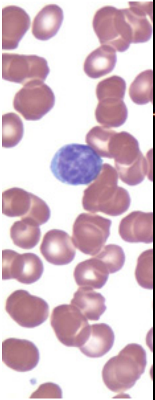
Many various data

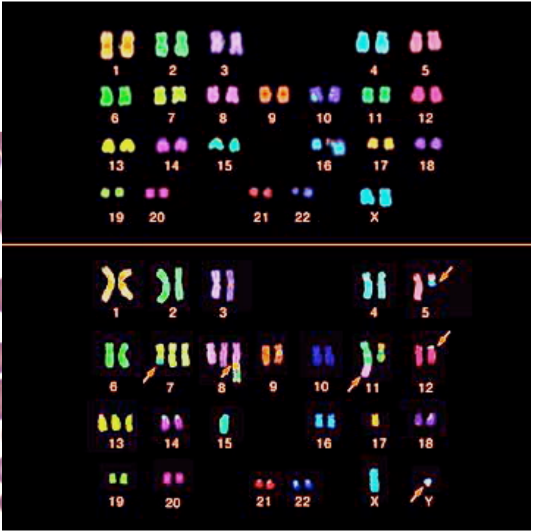


A cancer cell

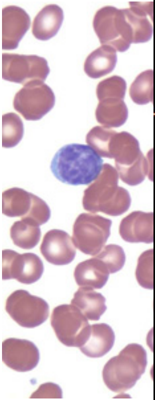


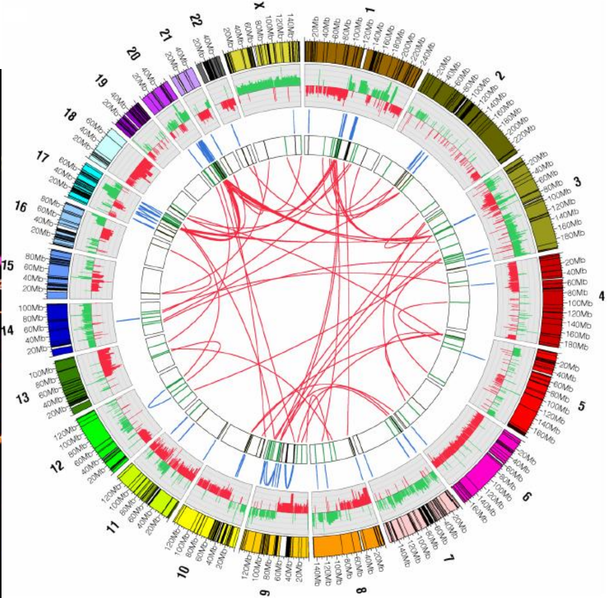
A cancer cell



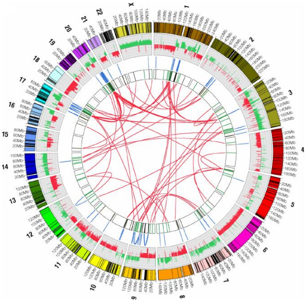


A cancer cell



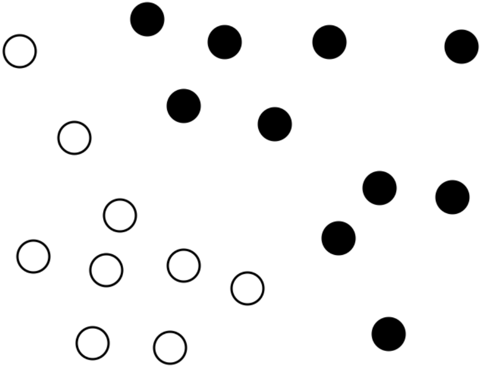


Opportunities

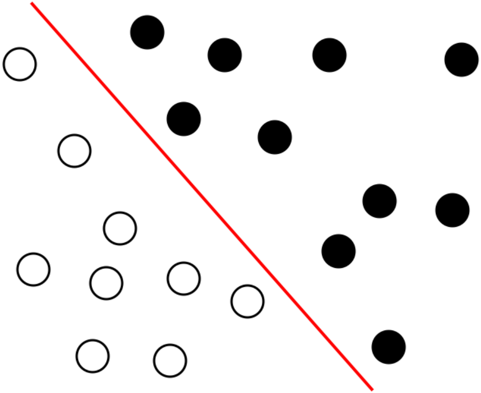


- What is your risk of developing a cancer? (*prevention*)
- After diagnosis and treatment, what is the risk of relapse? (*prognosis*)
- What specific treatment will cure your cancer? (*personalized medicine*)

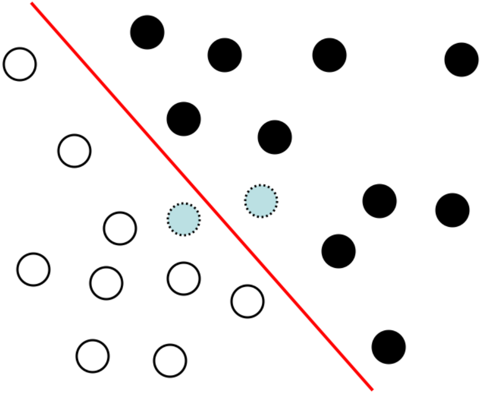
Machine learning formulation



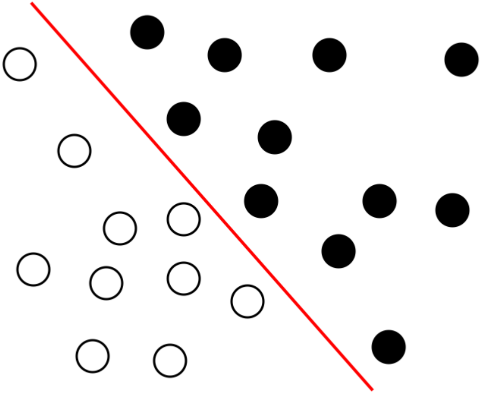
Machine learning formulation



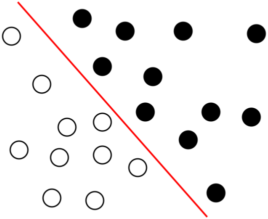
Machine learning formulation

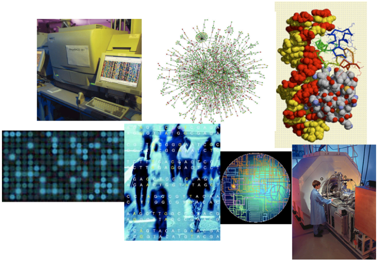


Machine learning formulation



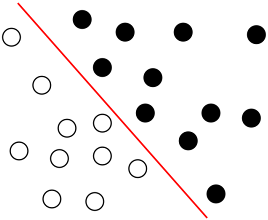
Challenges





- High dimension
- Few samples
- Structured data
- Heterogeneous data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models

Learning with regularization



Learn

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

by solving

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

- $R(f_\beta)$ empirical risk
- $\Omega(\beta)$ penalty

Outline

1. FlipFlop: fast isoform prediction from RNA-seq data

2. Learning molecular classifiers with network information

3. Kernel bilinear regression for toxicogenomics

Outline

1. FlipFlop: fast isoform prediction from RNA-seq data

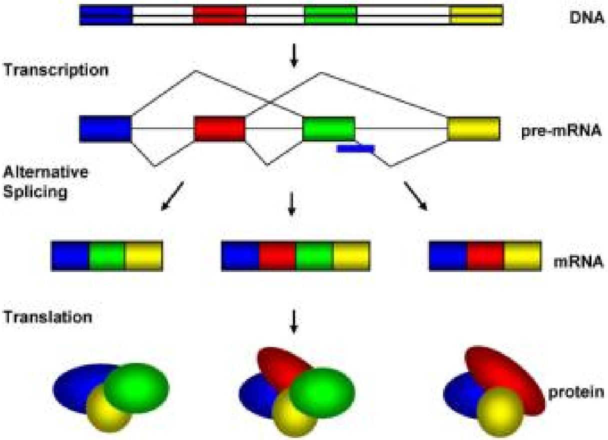
2. Learning molecular classifiers with network information

3. Kernel bilinear regression for toxicogenomics

Joint work with...

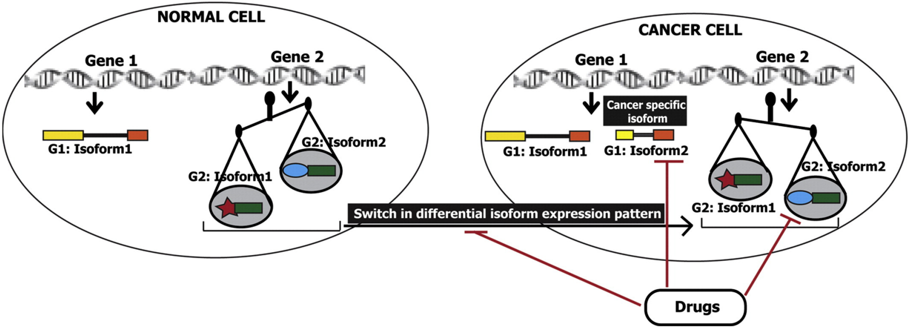
Elsa Bernard (Mines ParisTech / Institut Curie), Laurent Jacob (CNRS / LBBE), Julien Mairal (INRIA)

Alternative splicing: 1 gene = many proteins



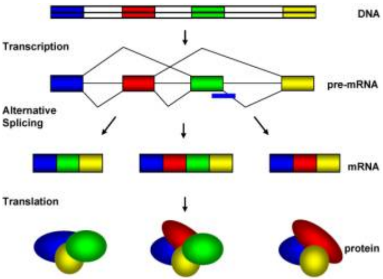
In human, 28k genes give 120k known transcripts (*Pal et al., 2012*)

Opportunities for drug developments...



(*Pal et al., 2012*)

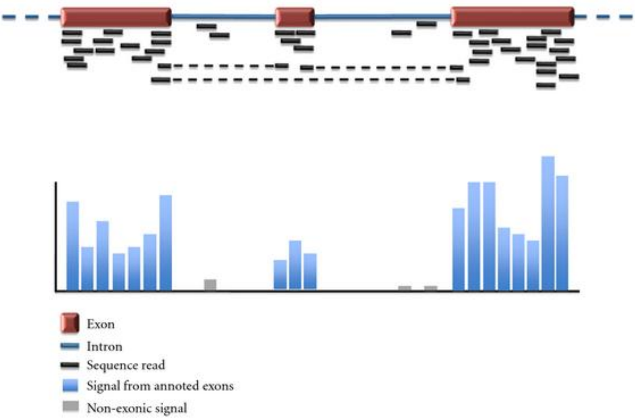
The isoform identification and quantification problem



Given a biological sample (e.g., cancer tissue), can we:

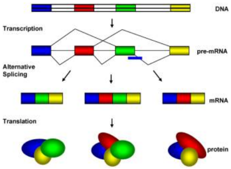
1. identify the isoform(s) of each gene present in the sample?
2. quantify their abundance?

RNA-seq data



(Costa et al., 2011)

Lasso-based estimation of isoforms

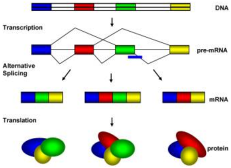


- Let a gene with e exons
- Suppose there are c candidate isoform (c large, up to 2^e)
- Let $\phi \in \mathbb{R}^c$ the unknown c-dimensional vector of abundance
- Let $L(\phi)$ quantify whether ϕ explains well the observed read counts (e.g., minus log-likelihood)
- Find a sparse vector of abundances by solving (e.g., IsoLasso, SLIDE, NSMAP...)

$$\min_{\phi \in \mathbb{R}_+^c} L(\phi) + \lambda \| \phi \|_1$$

- Computational problem: Lasso problem with 2^e variables

Lasso-based estimation of isoforms



- Let a gene with e exons
- Suppose there are c candidate isoform (c large, up to 2^e)
- Let $\phi \in \mathbb{R}^c$ the unknown c-dimensional vector of abundance
- Let $L(\phi)$ quantify whether ϕ explains well the observed read counts (e.g., minus log-likelihood)
- Find a sparse vector of abundances by solving (e.g., IsoLasso, SLIDE, NSMAP...)

$$\min_{\phi \in \mathbb{R}_+^c} L(\phi) + \lambda \| \phi \|_1$$

- Computational problem: Lasso problem with 2^e variables

Fast isoform deconvolution with the Lasso (FlipFlop)

Theorem (Bernard, Mairal, Jacob and V., 2014)

The isoform deconvolution problem

$$\min_{\phi \in \mathbb{R}_+^c} L(\phi) + \lambda \| \phi \|_1$$

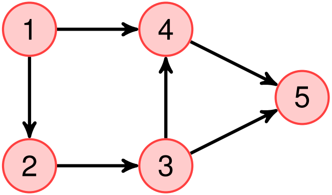
can be solved in polynomial time in the number of exon.

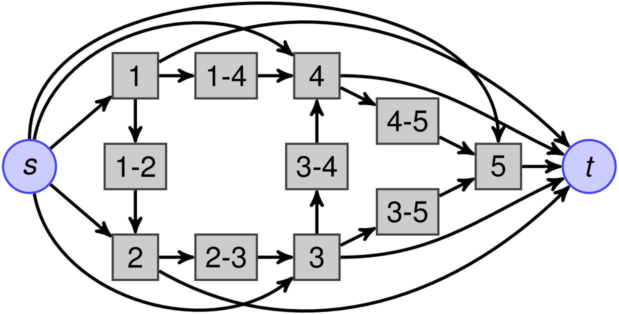
Key ideas

1. Reformulation as a convex cost flow problem (Mairal and Yu, 2012)
2. Recover isoforms by flow decomposition algorithm

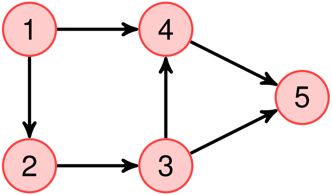
"Feature selection on an exponential number of features
in polynomial time"

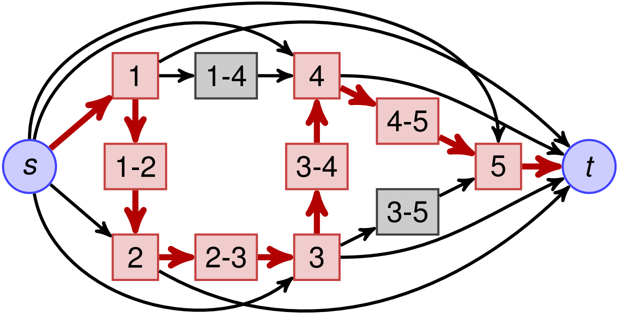
Isoforms are Paths in a Graph



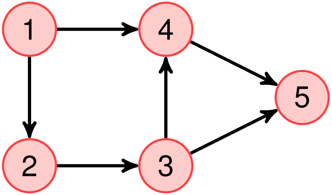


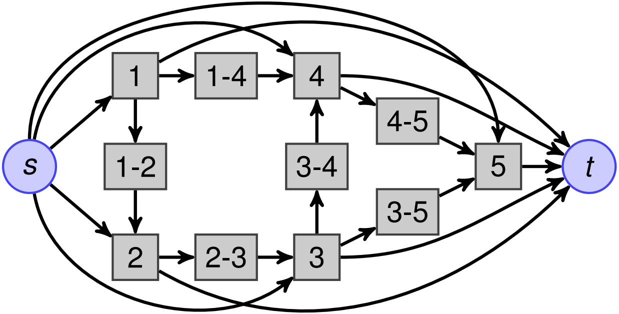
Isoforms are Paths in a Graph



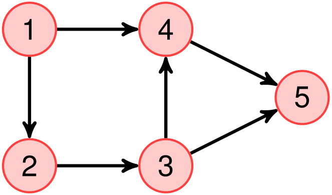


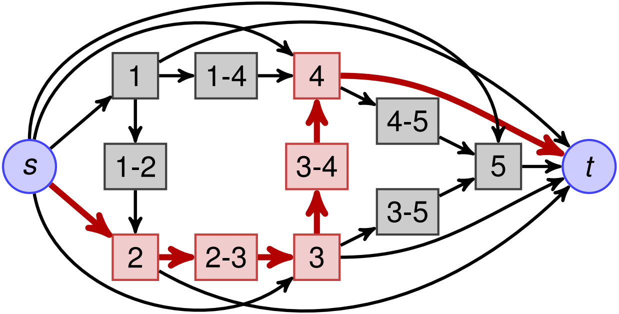
Isoforms are Paths in a Graph



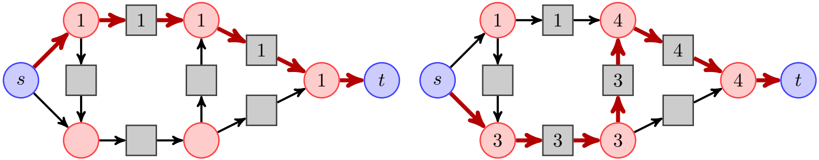


Isoforms are Paths in a Graph





Combinations of isoforms are flows



(a) Reads at every node corresponding to one isoform.

(b) Reads at every node after adding another isoform.

- $L(\phi)$ depends only on the values of the flow on the vertices
- $||\phi||_1 = f_t$

Therefore,

$$\min_{\phi \in \mathbb{R}_+^c} L(\phi) + \lambda \| \phi \|_1$$

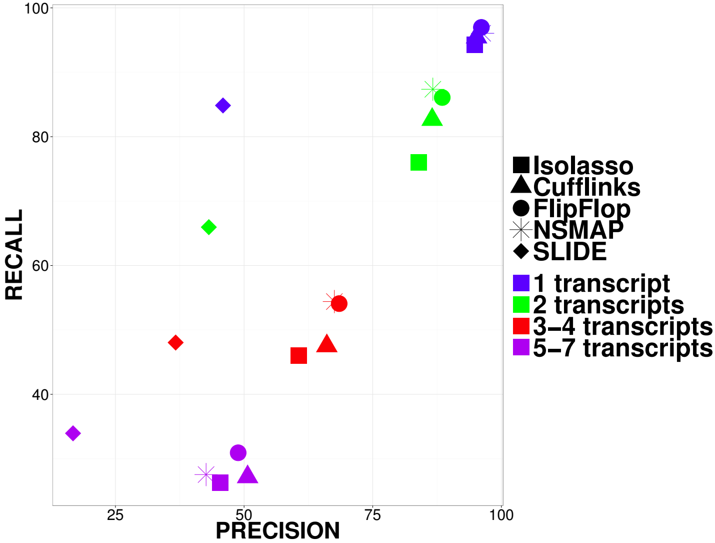
is equivalent to

$$\min_{\text{f flow}} R(f) + \lambda f_t$$

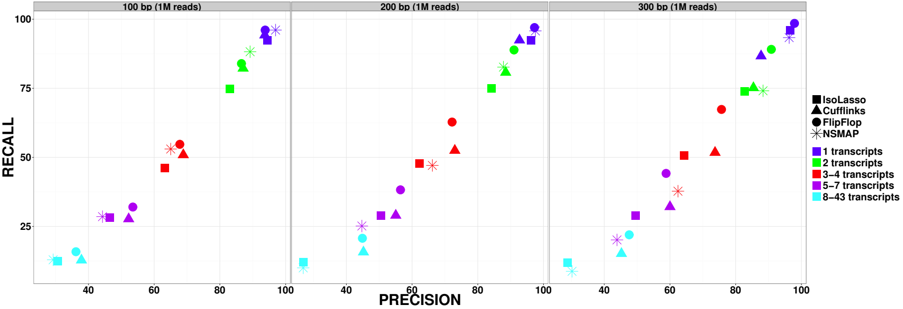
Human Simulation: Precision/Recall

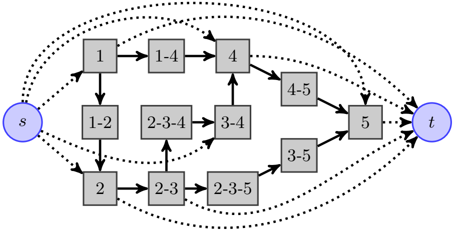
hg19, 1137 genes on chr1, 1million 75 bp single-end reads by transcript levels.

Simulator: `http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html`

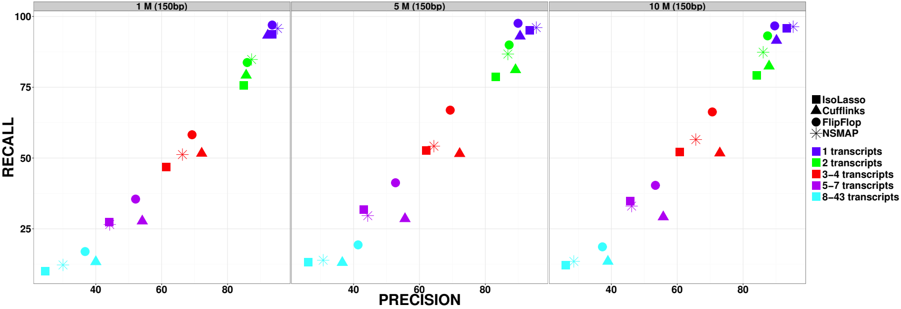


Performance increases with read length

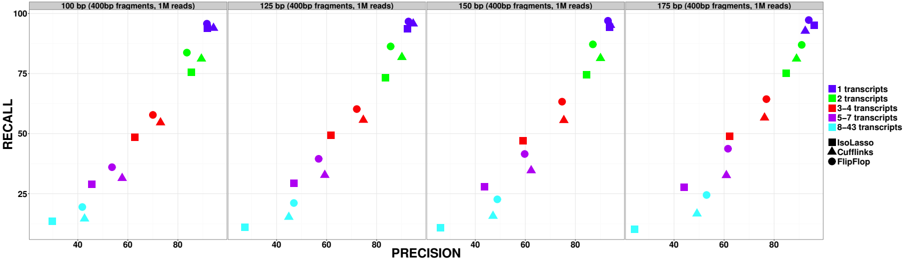




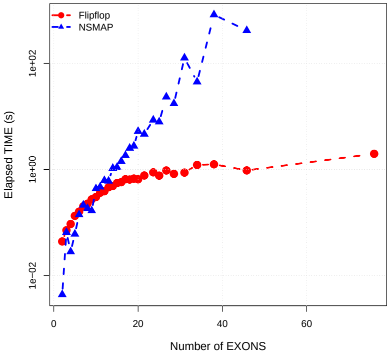
Performance increases with coverage

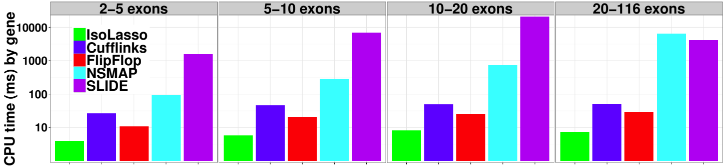


Extension to paired-end reads OK.



Speed trial


Flipflop
NSMAP
Elapsed TIME (s)
1e+02
1e+00
1e−02
0
20
40
60
Number of EXONS


2–5 exons
5–10 exons
10–20 exons
20–116 exons
CPU time (ms) by gene
10000
1000
100
10
IsoLasso
Cufflinks
FlipFlop
NSMAP
SLIDE

FlipFlop summary

- Fast method for exact Lasso-based isoform detection and quantification
- `http://cbio.mines-paristech.fr/flipflop`
- Available as an R package

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("flipflop")
```

- Reference: E. Bernard, L. Jacob, J. Mairal and J.-P. Vert. Efficient RNA isoform identification and quantification from RNA-seq data with network flows. *Bioinformatics*, 2014.
- Ongoing: extension to multiple samples and differential analysis

Outline

1. FlipFlop: fast isoform prediction from RNA-seq data

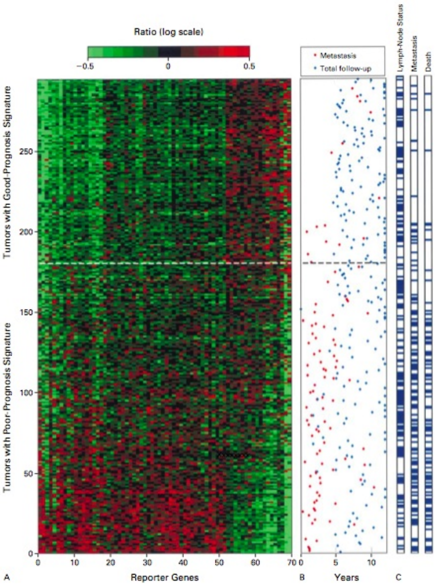
2. Learning molecular classifiers with network information

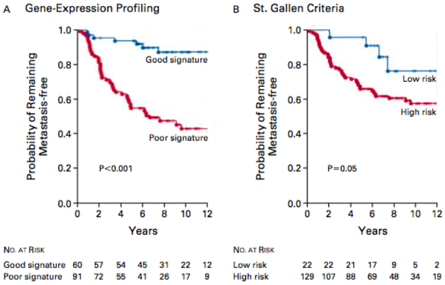
3. Kernel bilinear regression for toxicogenomics

Joint work with...

Franck Rapaport, Emmanuel Barillot, Andrei Zinovyev, Anne-Claire Haury, Laurent Jacob, Guillaume Obozinski

Breast cancer prognosis

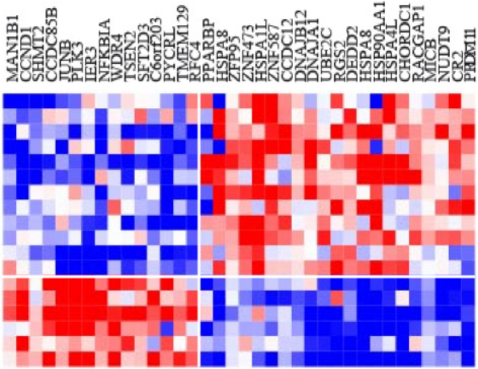




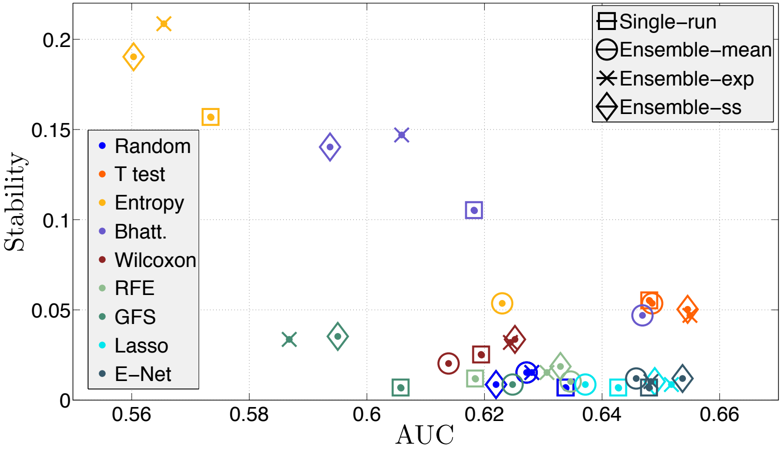
Gene selection, molecular signature

The idea

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

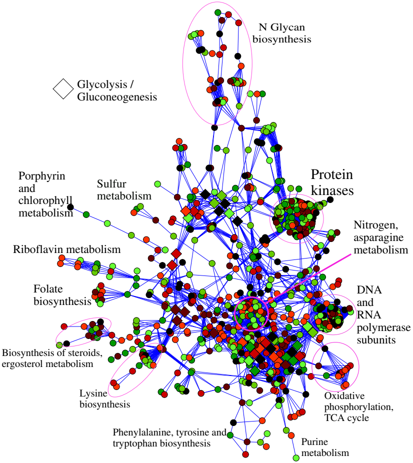


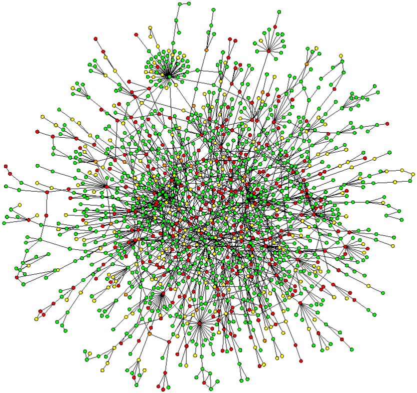
Lack of stability of signatures



Haury et al. (2011)

Gene networks

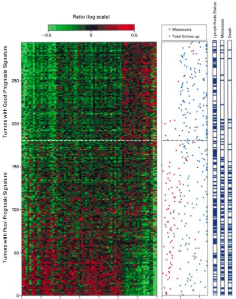

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Protein kinases
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative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

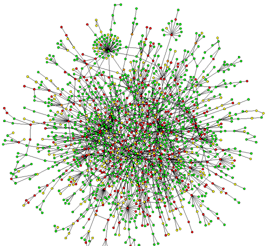


Gene networks and expression data

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

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

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

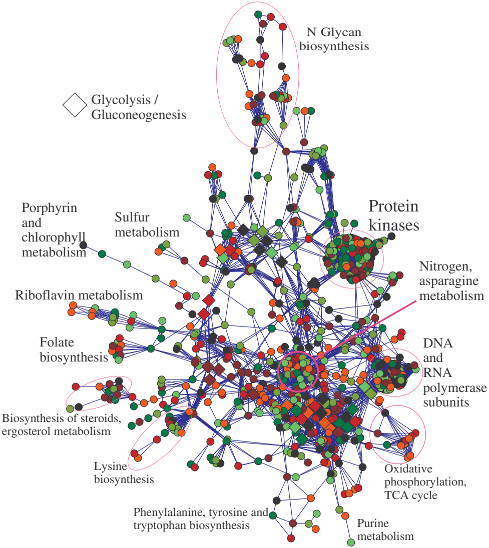
Genes near each other on the graph should have similar weigths.

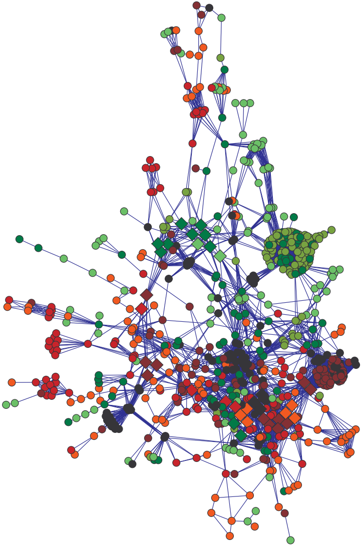
An idea (Rapaport et al., 2007)

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

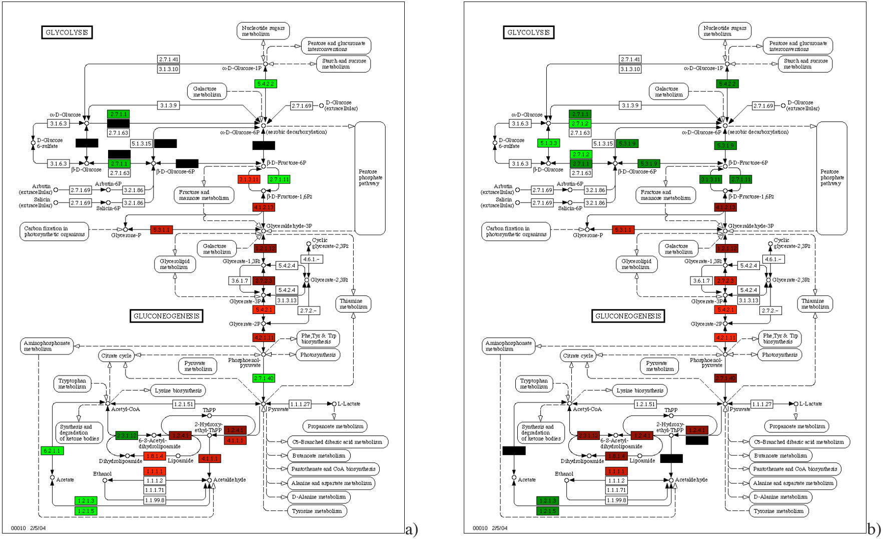
$$\min_{\beta \in \mathbb{R}^p} R(f_\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2 \,.$$

Classifiers


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



Classifier


GLYCOLYSIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-1P
5.4.2.2
Galactose metabolism
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
α-D-Glucose
2.7.1.1
3.1.6.3
2.7.1.63
α-D-Glucose-6P
(aerobic decarboxylation)
D-Glucose 6-sulfate
5.1.3.15
2.7.1.1
2.7.1.63
3.1.6.3
β-D-Glucose
β-D-Glucose-6P
β-D-Fructose-6P
Pentose phosphate pathway
3.1.3.11
2.7.1.11
Arbutin (extracellular)
2.7.1.69
Arbutin-6P
3.2.1.86
Salicin (extracellular)
2.7.1.69
3.2.1.86
Salicin-6P
Fructose and mannose metabolism
β-D-Fructose-1,6P2
4.1.2.13
Carbon fixation in photosynthetic organisms
5.3.1.1
Glycerone-P
Glyceraldehyde-3P
Galactose metabolism
1.2.1.12
Cyclic glycerate-2,3P2
Glycerolipid metabolism
Glycerate-1,3P2
5.4.2.4
4.6.1.-
3.6.1.7
2.7.2.3
Glycerate-2,3P2
5.4.2.4
3.1.3.13
Glycerate-3P
Thiamine metabolism
GLUCONEOGENESIS
5.4.2.1
2.7.2.-
Glycerate-2P
4.2.1.11
Phe,Tyr & Trp biosynthesis
Aminophosphonate metabolism
Citrate cycle
Phosphoenol-pyruvate
Photosynthesis
Pyruvate metabolism
Tryptophan metabolism
2.7.1.40
Lysine biosynthesis
1.2.1.51
1.1.1.27
L-Lactate
Acetyl-CoA
ThPP
Pyruvate
Synthesis and degradation of ketone bodies
2-Hydroxy-ethyl-ThPP
Propanoate metabolism
2.3.1.12
1.2.4.1
1.2.4.1
6-S-Acetyl-dihydrolipoamide
4.1.1.1
C5-Branched dibasic acid metabolism
6.2.1.1
1.8.1.4
Dihydrolipoamide
Lipoamide
4.1.1.1
Butanoate metabolism
Ethanol
1.1.1.1
Pantothenate and CoA biosynthesis
Acetate
1.1.1.2
Acetaldehyde
Alanine and aspartate metabolism
1.1.1.71
1.2.1.3
1.1.99.8
D-Alanine metabolism
1.2.1.5
Tyrosine metabolism
00010 2/5/04
a)
GLYCOLYSIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-1P
5.4.2.2
Galactose metabolism
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
α-D-Glucose
2.7.1.1
2.7.1.2
3.1.6.3
2.7.1.63
α-D-Glucose-6P
(aerobic decarboxylation)
D-Glucose 6-sulfate
5.1.3.3
5.1.3.15
5.3.1.9
5.3.1.9
2.7.1.2
2.7.1.1
2.7.1.63
3.1.6.3
5.3.1.9
β-D-Glucose
β-D-Glucose-6P
β-D-Fructose-6P
Pentose phosphate pathway
3.1.3.11
2.7.1.11
Arbutin (extracellular)
2.7.1.69
Arbutin-6P
3.2.1.86
Salicin (extracellular)
2.7.1.69
3.2.1.86
Salicin-6P
Fructose and mannose metabolism
β-D-Fructose-1,6P2
4.1.2.13
Carbon fixation in photosynthetic organisms
5.3.1.1
Glycerone-P
Glyceraldehyde-3P
Galactose metabolism
1.2.1.12
Cyclic glycerate-2,3P2
Glycerolipid metabolism
Glycerate-1,3P2
5.4.2.4
4.6.1.-
3.6.1.7
2.7.2.3
Glycerate-2,3P2
5.4.2.4
3.1.3.13
Glycerate-3P
Thiamine metabolism
GLUCONEOGENESIS
5.4.2.1
2.7.2.-
Glycerate-2P
4.2.1.11
Phe,Tyr & Trp biosynthesis
Aminophosphonate metabolism
Citrate cycle
Phosphoenol-pyruvate
Photosynthesis
Pyruvate metabolism
Tryptophan metabolism
2.7.1.40
Lysine biosynthesis
1.2.1.51
1.1.1.27
L-Lactate
Acetyl-CoA
ThPP
Pyruvate
Synthesis and degradation of ketone bodies
2-Hydroxy-ethyl-ThPP
Propanoate metabolism
2.3.1.12
1.2.4.1
1.2.4.1
6-S-Acetyl-dihydrolipoamide
C5-Branched dibasic acid metabolism
1.8.1.4
Dihydrolipoamide
Lipoamide
Butanoate metabolism
Ethanol
1.1.1.1
Pantothenate and CoA biosynthesis
Acetate
1.1.1.2
Acetaldehyde
Alanine and aspartate metabolism
1.1.1.71
1.2.1.3
1.1.99.8
D-Alanine metabolism
1.2.1.5
Tyrosine metabolism
00010 2/5/04
b)

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

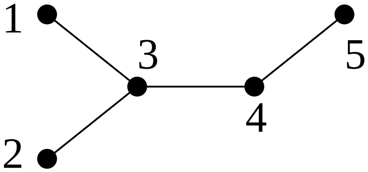
$$\Phi(x)^\top \Phi(x') = x^\top K_G x'$$

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

Graph Laplacian

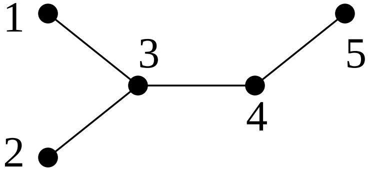
Definition

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



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

Pseufo-inverse of the Laplacian



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

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.

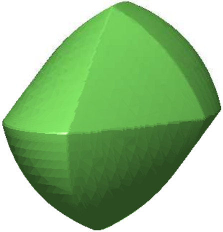
Other penalties without kernels

- Gene selection + Piecewise constant on the graph

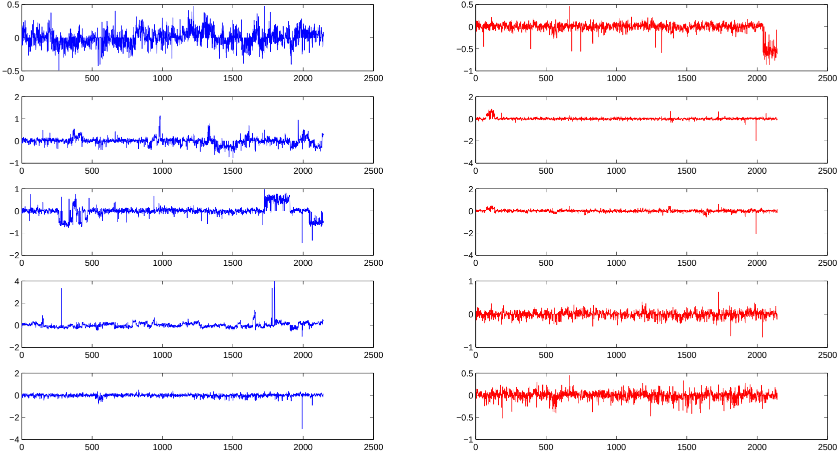
$$\Omega(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$

- Gene selection + smooth on the graph

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



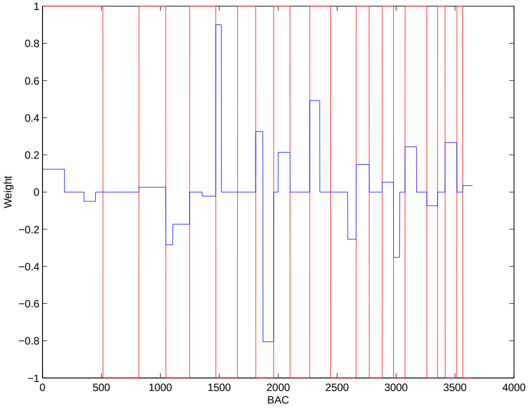
Example: classification of DNA copy number profiles



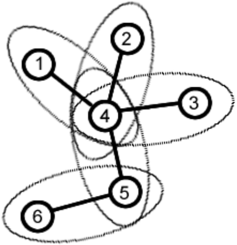
Aggressive (left) vs non-aggressive (right) melanoma

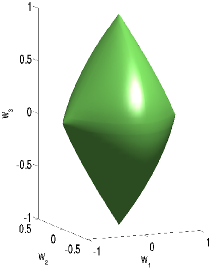
Fused lasso solution (Rapaport et al., 2008)

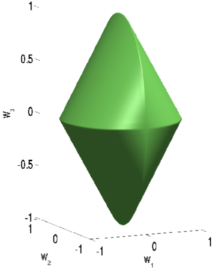
$$\Omega(\beta) = \sum_{i \sim j} |\, \beta_i - \beta_j \,| + \sum_{i=1}^{p} |\, \beta_i \,|$$



Graph-based structured feature selection





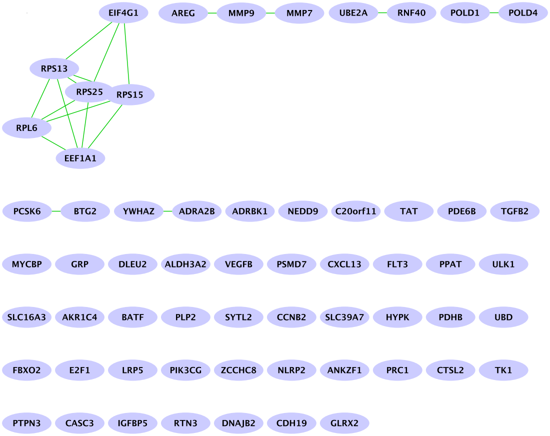


Graph lasso(s)

$$\Omega_1(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,, \quad \text{(Jenatton et al., 2009)}$$

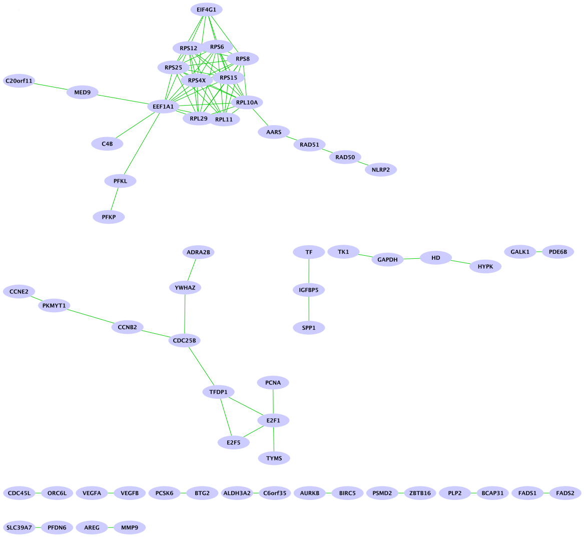
$$\Omega_2(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,. \quad \text{(Jacob et al., 2008)}$$

Lasso signature (accuracy 0.61)



Breast cancer prognosis

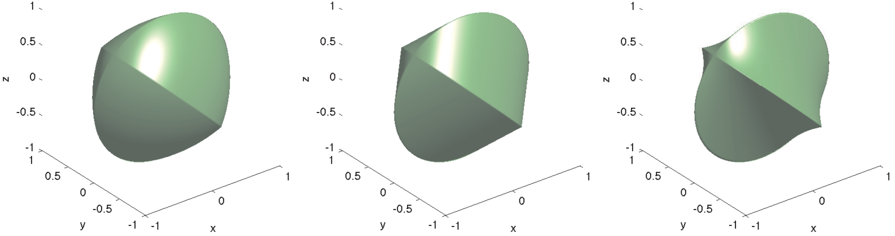
Graph Lasso signature (accuracy 0.64)



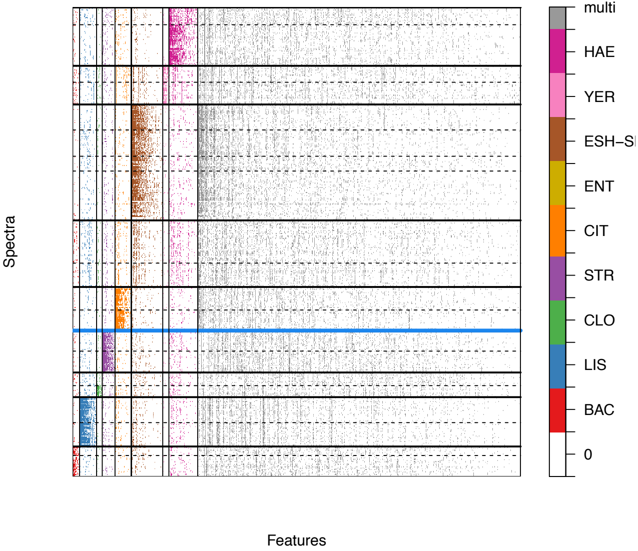
Breast cancer prognosis

Disjoint feature selection
(Vervier, d'Aspremont, Mahé, Veyrieras and V., 2014)

$$W = (w_i)_{i \in V} \in \mathbb{R}^{p \times V} \qquad \Omega(W) = \min_{-H \leq W \leq H} \sum_{i \sim j} K_{ij} \left| h_i^\top h_j \right|$$



Example: multiclass classification of MS spectra



(Vervier et al, 2013, unpublished)

Outline

1. FlipFlop: fast isoform prediction from RNA-seq data

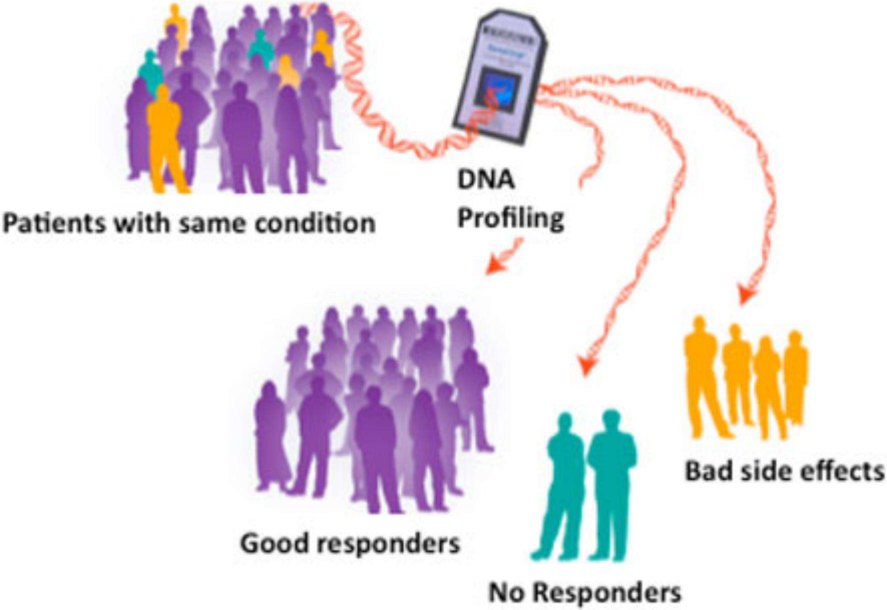
2. Learning molecular classifiers with network information

3. Kernel bilinear regression for toxicogenomics

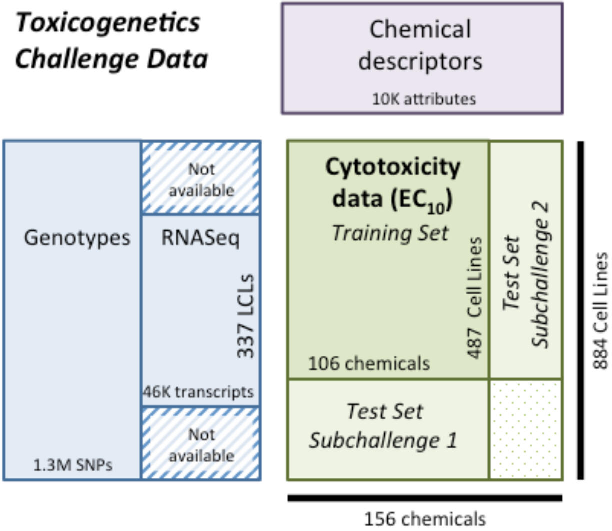
Joint work with...

Elsa Bernard, Erwan Scornet, Yunlong Jiao, Véronique Stoven, Thomas Walter

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*

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

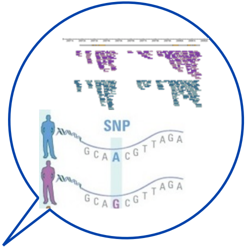
and

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

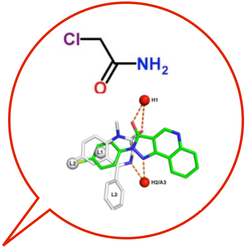
where

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

Kernel Trick

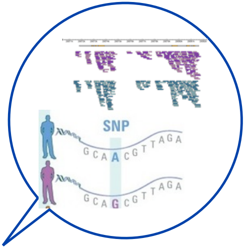


cell line descriptors

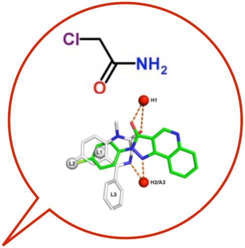


drug descriptors

Kernel Trick



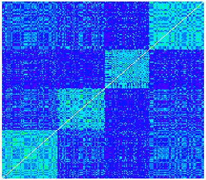
cell line descriptors



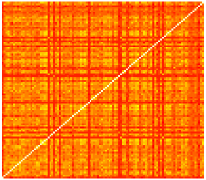
drug descriptors

kernelized →

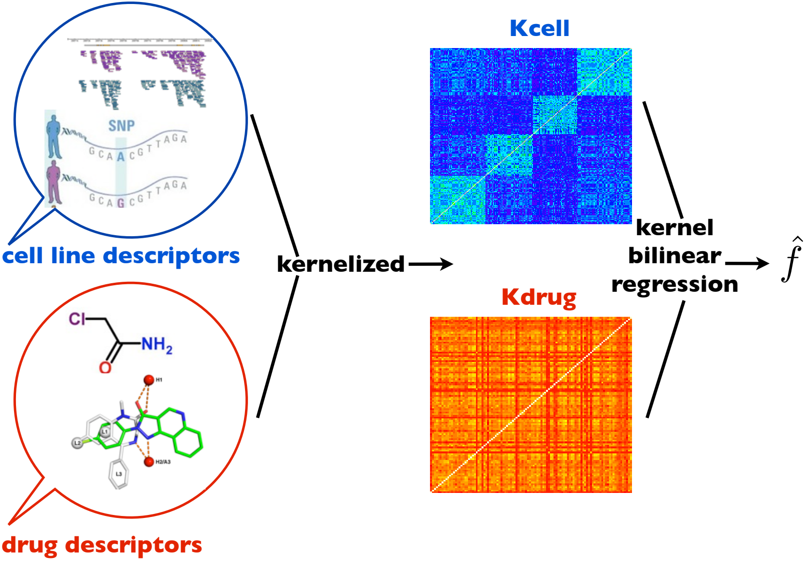
Kcell



Kdrug



Kernel Trick



cell line descriptors

drug descriptors

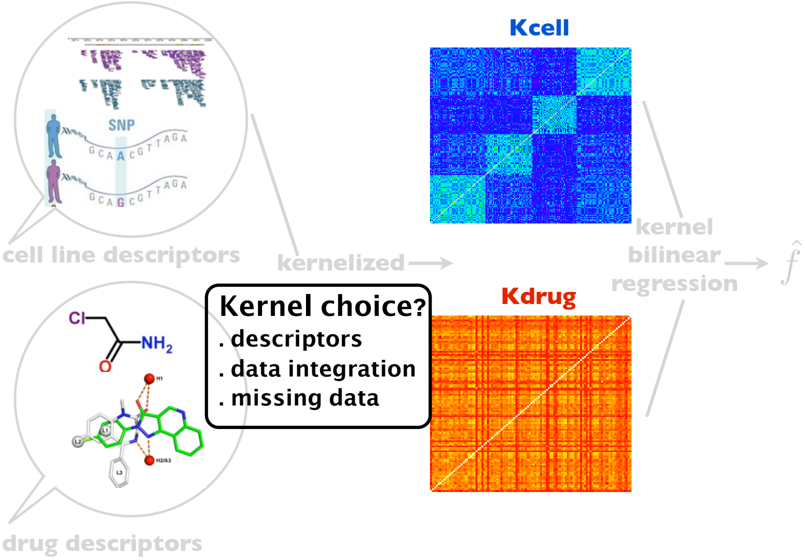
kernelized →

Kcell

Kdrug

kernel bilinear regression → $\hat{f}$

Kernel Trick



Kernel choice?

- . descriptors
- . data integration
- . missing data

Kernel choice

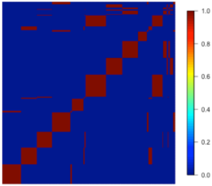
1. $\mathbf{K}_{\text{cell}}$:
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. $\mathbf{K}_{\text{drug}}$:
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Kernel choice

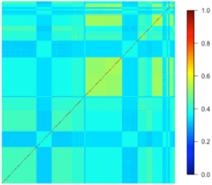
1. $\mathbf{K}_{\text{cell}}$:
   $\Longrightarrow$ 29 cell line kernels tested
   $\Longrightarrow$ 1 kernel that *integrate all information*
   $\Longrightarrow$ deal with missing data
2. $\mathbf{K}_{\text{drug}}$:
   $\Longrightarrow$ 48 drug kernels tested
   $\Longrightarrow$ multi-task kernels

Cell line data integration

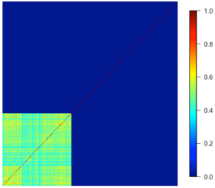
Covariates
. linear kernel



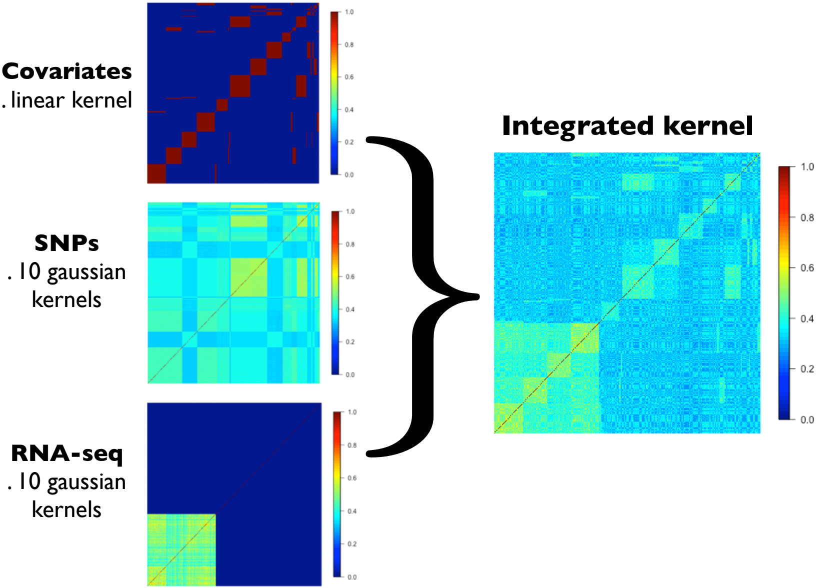
SNPs
. 10 gaussian kernels



RNA-seq
. 10 gaussian kernels



Cell line data integration



Covariates
. linear kernel

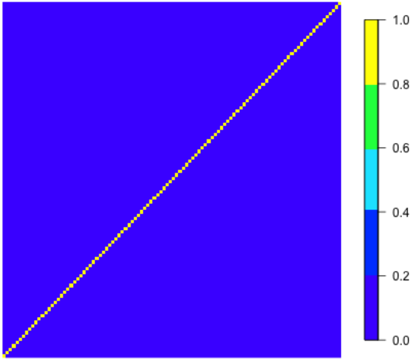
SNPs
. 10 gaussian kernels

RNA-seq
. 10 gaussian kernels

Integrated kernel

Multi-task drug kernels

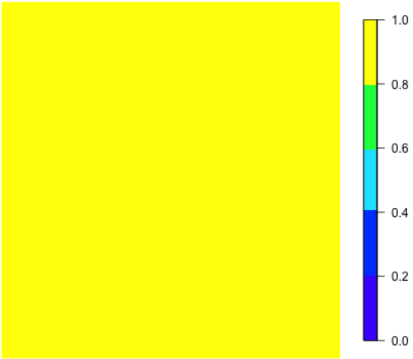
1. **Dirac**
2. Multi-Task
3. Feature-based
4. Empirical
5. Integrated



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

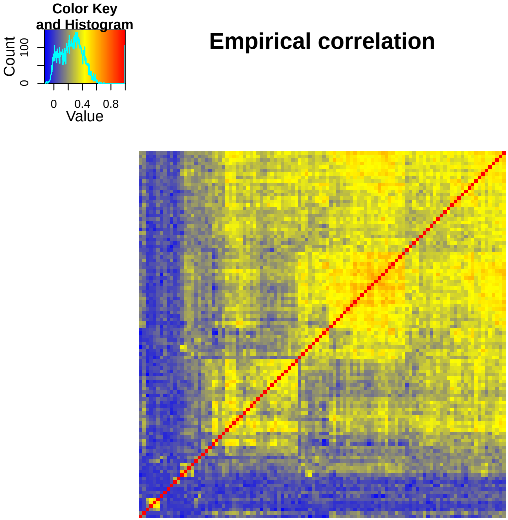
1. Dirac
2. Multi-Task
3. **Feature-based**
4. Empirical
5. Integrated

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



Empirical correlation

Multi-task drug kernels

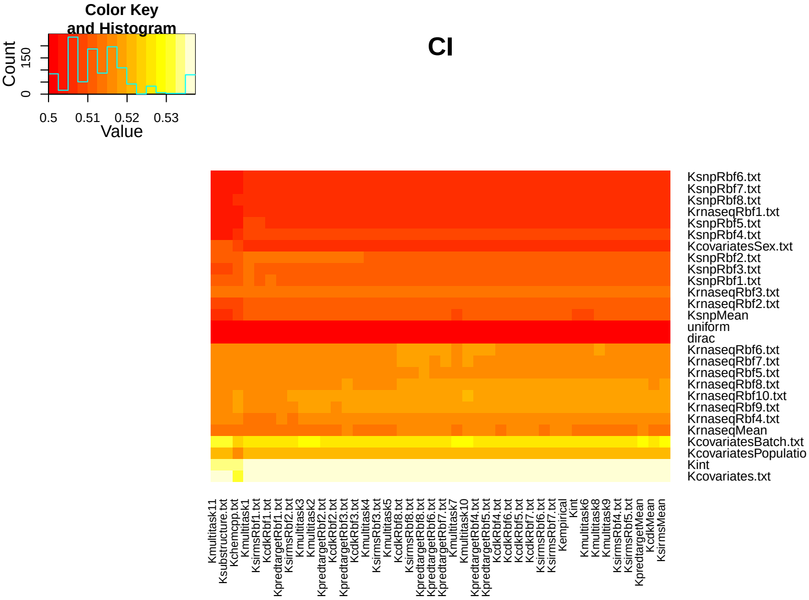
1. Dirac
2. Multi-Task
3. Feature-based
4. Empirical
5. **Integrated**

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

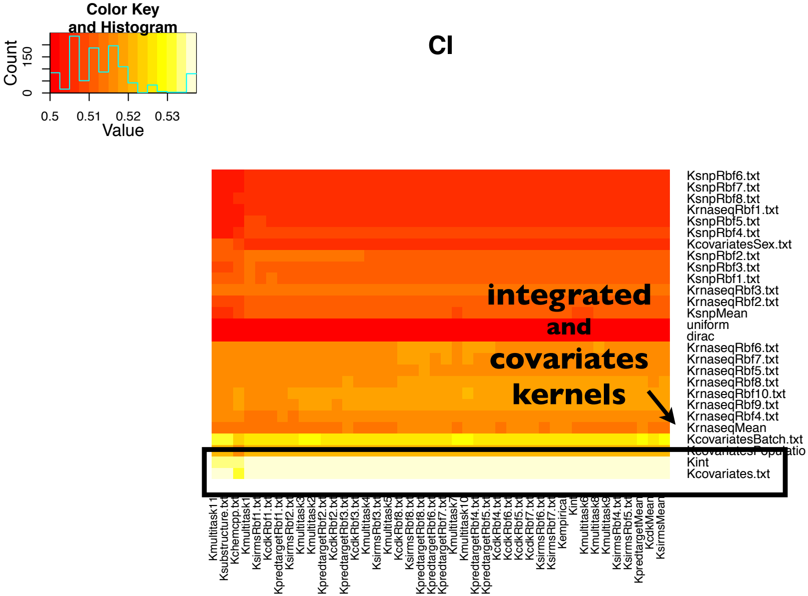
Integrated kernel:

- Combine all information on drugs

29x48 kernel combinations: CV results

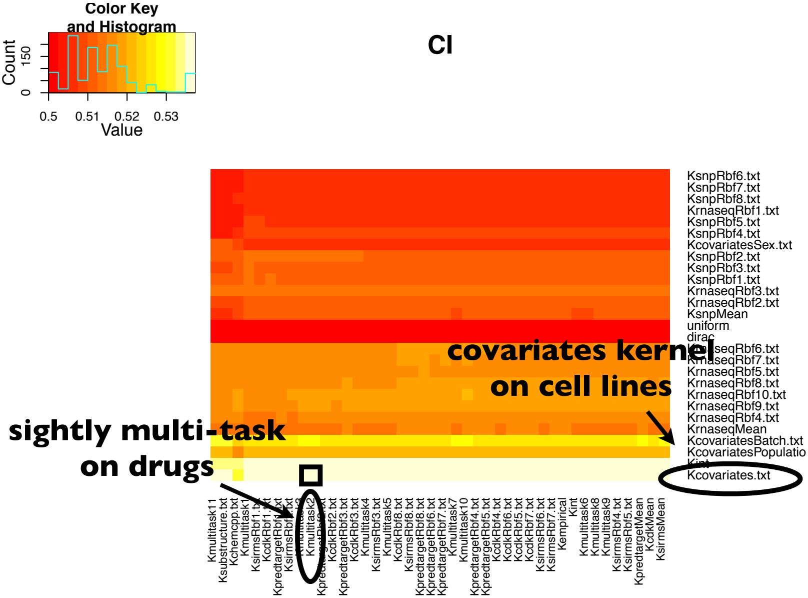


29x48 kernel combinations: CV results



integrated and covariates kernels

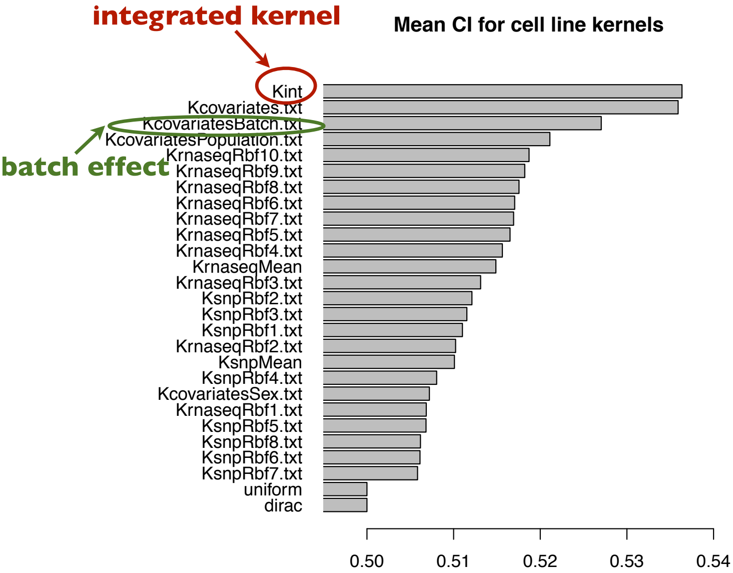
29x48 kernel combinations: CV results



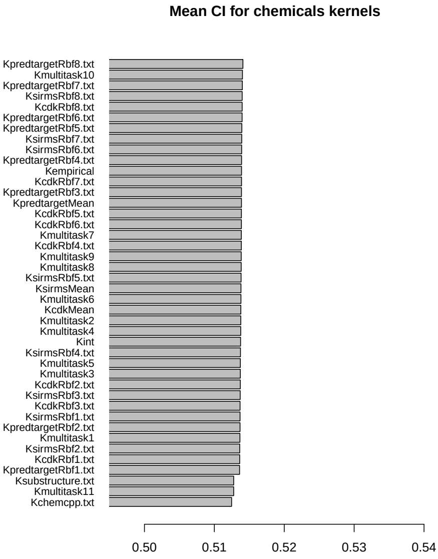
covariates kernel on cell lines

sightly multi-task on drugs

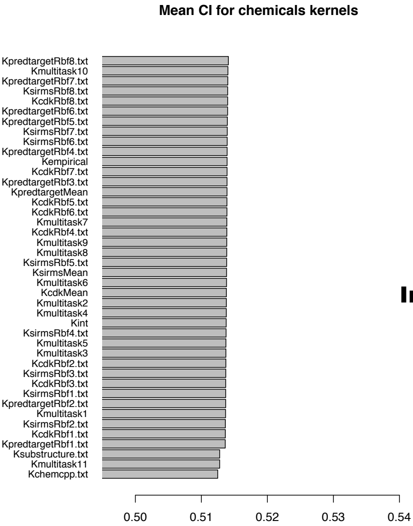
Kernel on cell lines: CV results



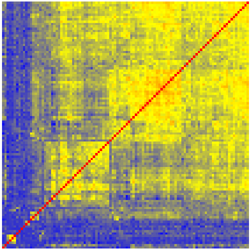
Kernel on drugs: CV results



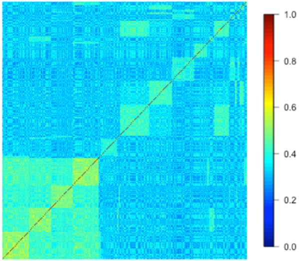
Final Submission (ranked 2nd)



Empirical kernel on drugs



Integrated kernel on cell lines



Conclusion

- Many new problems and lots of data in computational genomics
- Computational constraints $\Longrightarrow$ fast sparse models (FlipFlop)
- Small n large p $\Longrightarrow$ regularized models with prior knowledge
- Heterogeneous data integration $\Longrightarrow$ kernel methods
- Personalized medicine promising but difficult!

Thanks

Alexandre d'Aspremont, Emmanuel Barillot, Anne-Claire Haury, Laurent Jacob, Pierre Mahé, Julien Mairal, Guillaume Obozinski, Franck Rapaport, Jean-Baptiste Veyrieras, Andrei Zynoviev
... and all CBIO@Mines

institutCurie
European Research Council erc
Inserm Institut national de la santé et de la recherche médicale
MINES ParisTech