Some contributions of machine learning to bioinformatics

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Mines ParisTech / Institut Curie / Inserm

Ecole normale supérieure, Paris, October 13, 2009.

Outline

1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

Outline

1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

Outline

1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

Outline

1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

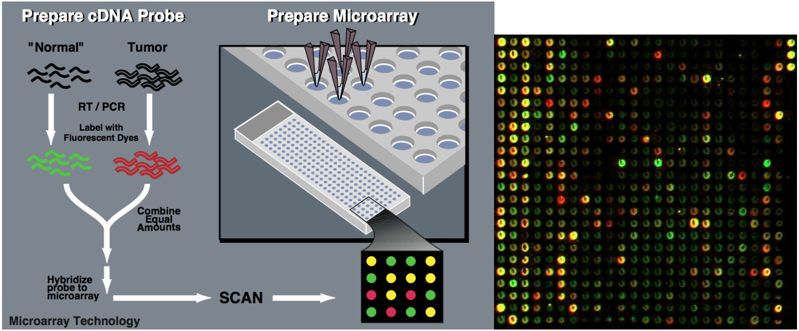
Outline

1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

Outline

1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

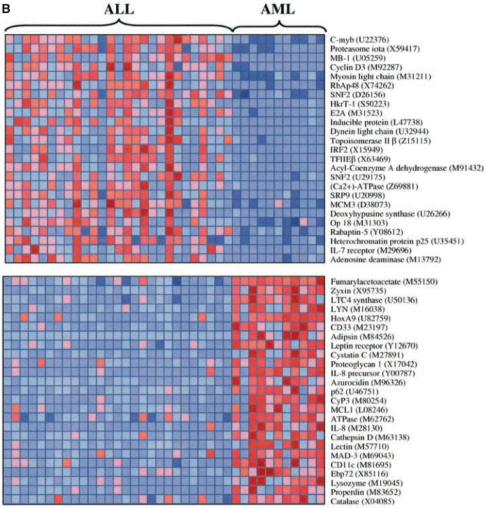
Tissue profiling with DNA chips



Data

- Gene expression measures for more than 10*k* genes
- Measured typically on less than 100 samples of two (or more) different classes (e.g., different tumors)

Tissue classification from microarray data



Goal

- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes

Difficulty

- Large dimension
- Few samples

Linear classifiers

The approach

- Each sample is represented by a vector $x = (x_1, \ldots, x_p)$ where $p > 10^5$ is the number of probes
- Classification: given the set of labeled samples, learn a linear decision function:

  $$f_\beta(x) = \sum_{i=1}^{p} \beta_i x_i + \beta_0 \,,$$

  that is positive for one class, negative for the other
- Interpretation: the weight β_i quantifies the influence of gene i for the classification
- We must use prior knowledge for this small n large p problem.

Outline

1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data

2. Inference on biological networks

3. Virtual screening and chemogenomics

4. Conclusion

Motivation

- In feature selection, we look for a linear function $f(\mathbf{x}) = \mathbf{x}^\top \beta$, where only a limited number of coefficients in β are non-zero.
- Motivations
  - Accuracy: by restricting $\mathcal{F}$, we increase the bias but decrease the variance. This should be helpful in particular in high dimension, where bias is low and variance is large.
  - Interpretation: with a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects.
- Of course, this is particularly relevant if we believe that there exist good predictors which are sparse (prior knowledge).

Best subset selection

- In best subset selection, we must solve the problem:

$$\min R(f_\beta) \quad \text{s.t.} \quad \| \beta \|_0 \leq k$$

  for $k = 1, \ldots, p$.

- The state-of-the-art is branch-and-bound optimization, known as *leaps and bound* for least squares (Furnival and Wilson, 1974).
- This is usually a NP-hard problem, feasible for p as large as 30 or 40

Efficient feature selection

To work with more variables, we must use different methods. The state-of-the-art is split among

- Filter methods : the predictors are preprocessed and ranked from the most relevant to the less relevant. The subsets are then obtained from this list, starting from the top.
- Wrapper method: here the feature selection is iterative, and uses the ERM algorithm in the inner loop
- Embedded methods : here the feature selection is part of the ERM algorithm itself (see later the shrinkage estimators).

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

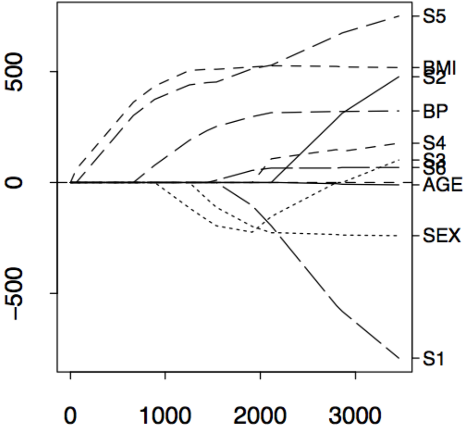
- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

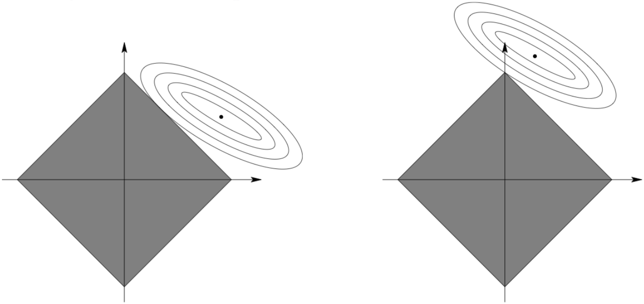
Embedded methods (LASSO)

$$\min_{\beta} R(\beta) + \sum_{i=1}^{p} |\beta_i|$$

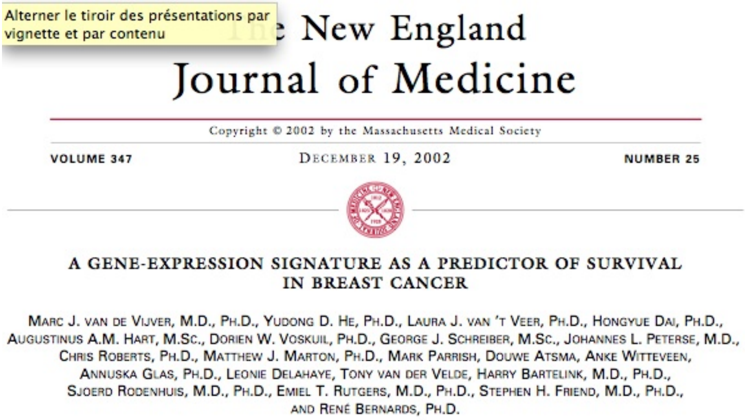


Why LASSO leads to sparse solutions

Geometric interpretation with $p = 2$



Example: MAMMAPRINT



Alterner le tiroir des présentations par vignette et par contenu

The New England Journal of Medicine

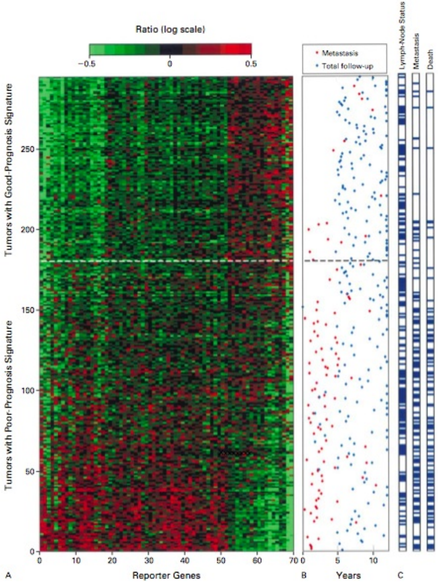
Copyright © 2002 by the Massachusetts Medical Society

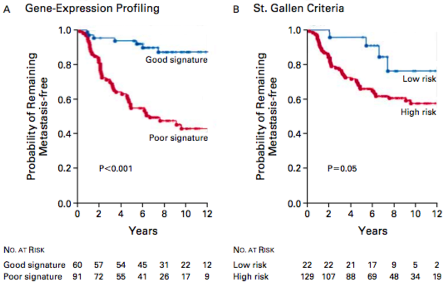
VOLUME 347 DECEMBER 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D., AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D., CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE ATSMA, ANKE WITTEVEEN, ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D., SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D., AND RENÉ BERNARDS, PH.D.

Example: MAMMAPRINT





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

Outline

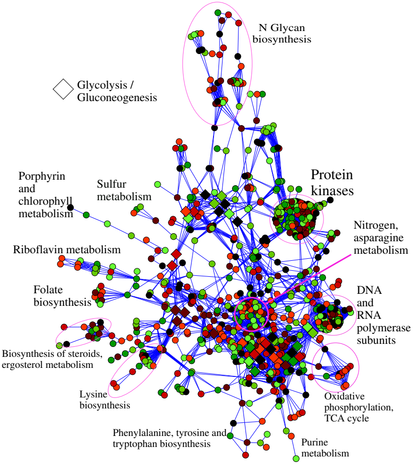
1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data

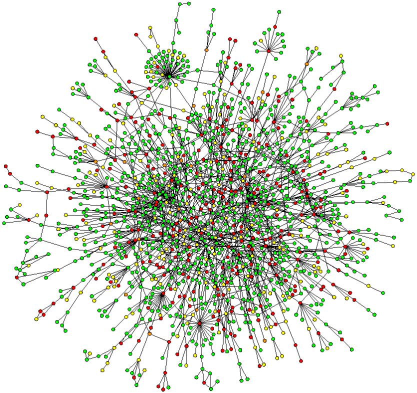
2. Inference on biological networks

3. Virtual screening and chemogenomics

4. Conclusion

Gene networks

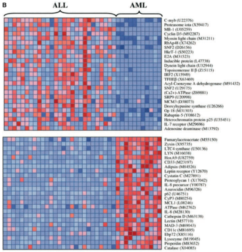

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative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

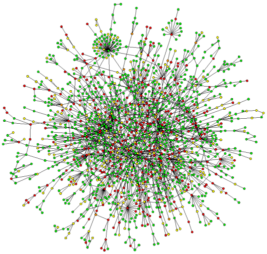


Gene networks and expression data

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





The fused LASSO

- The LASSO performs gene selection by solving

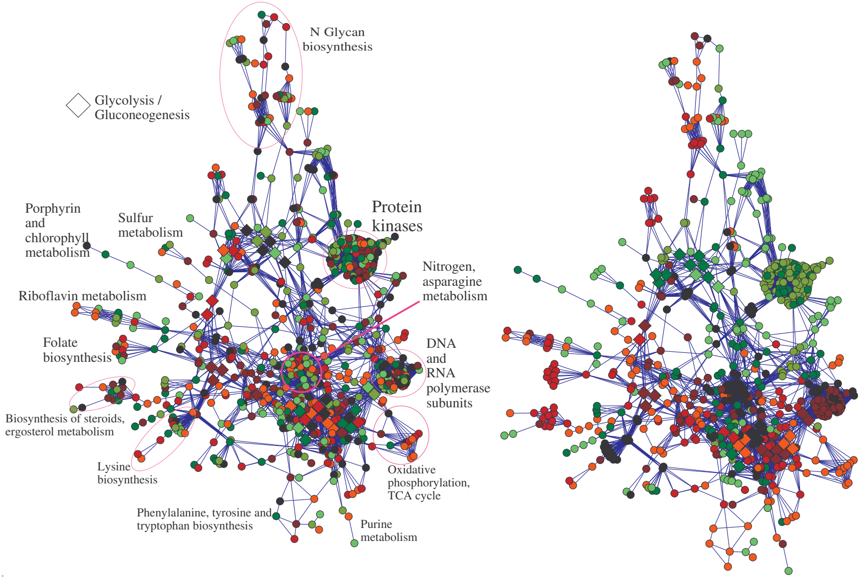
$$\min_{\beta} R(\beta) + \sum_{i=1}^{p} |\beta_i| \,.$$

- Here we want instead to enforce connected genes to have similar weights
- We can try the following embedded methods:

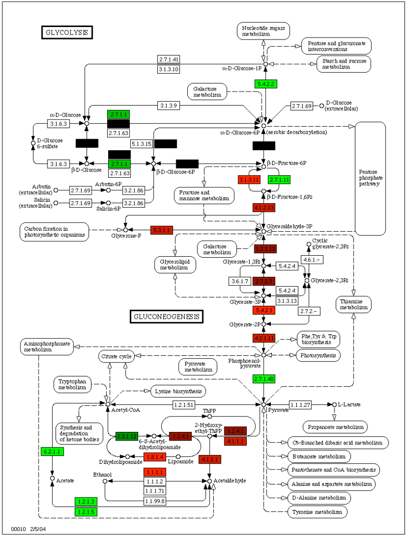
$$\min_{\beta} R(\beta) + \sum_{i \sim j} (\beta_i - \beta_i)^2 \,, \tag{1}$$

$$\min_{\beta} R(\beta) + \sum_{i \sim j} |\beta_i - \beta_i| \,. \tag{2}$$

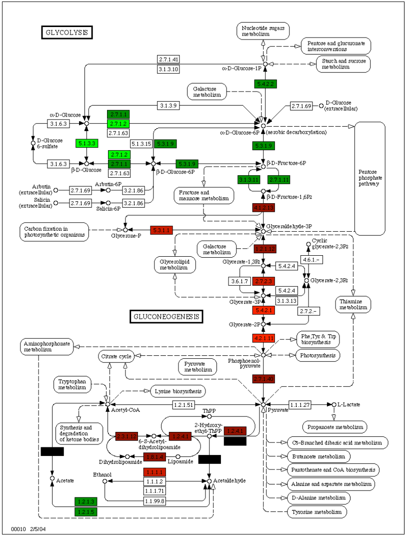
Classifier


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



a)

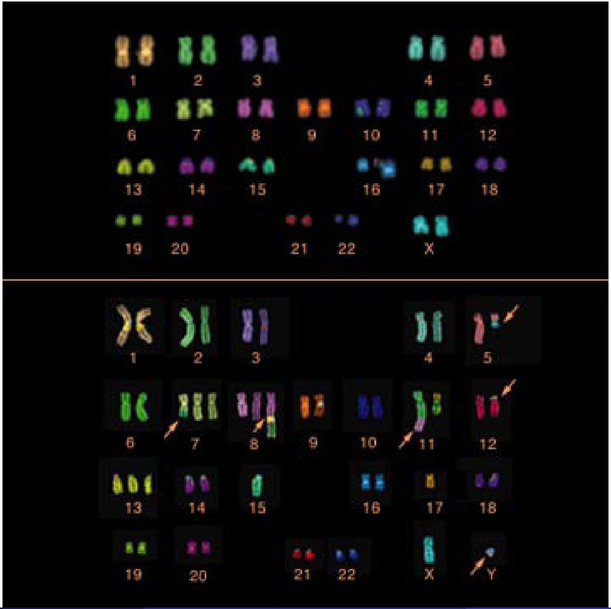


b)

Outline

1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

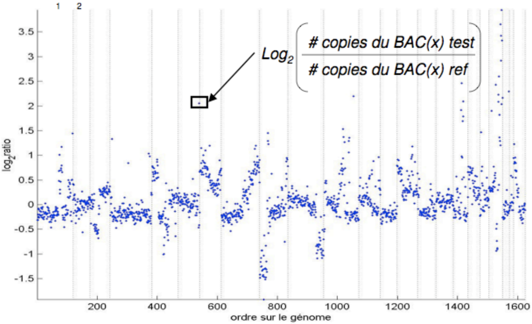
Chromosomic aberrations in cancer



Comparative Genomic Hybridization (CGH)

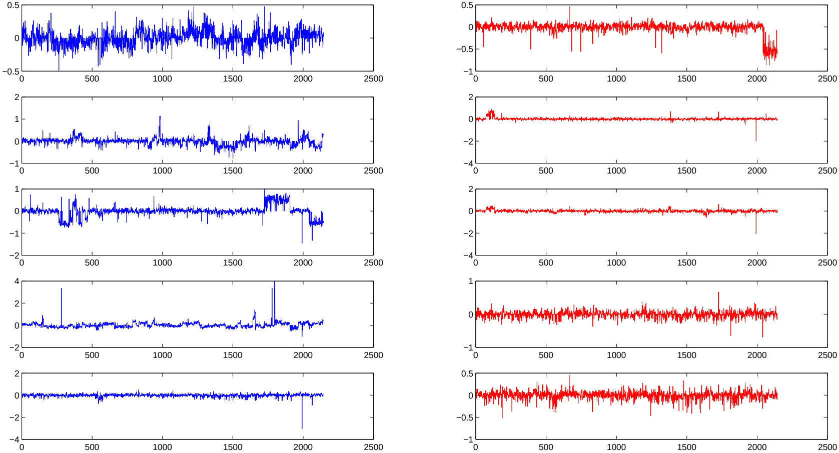
Motivation

- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research
- Can we classify CGH arrays for diagnosis or prognosis purpose?



Jain et al. Genome research 2002 12:325-332

Aggressive vs non-aggressive melanoma



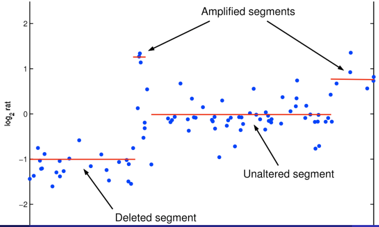
Classification of array CGH

Prior knowledge

- Let **x** be a CGH profile
- We focus on linear classifiers, i.e., the sign of :

$$f(\mathbf{x}) = \mathbf{x}^\top \beta \,.$$

- We expect β to be
  - sparse : only a few positions should be discriminative
  - piecewise constant : within a region, all probes should contribute equally



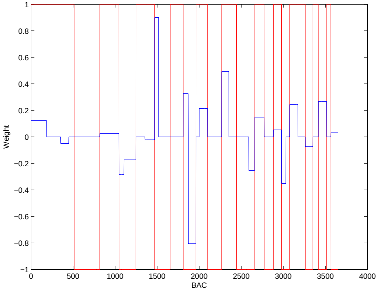
A penalty for CGH array classification

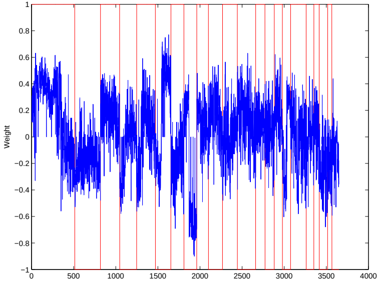
The fused LASSO penalty (Tibshirani et al., 2005)

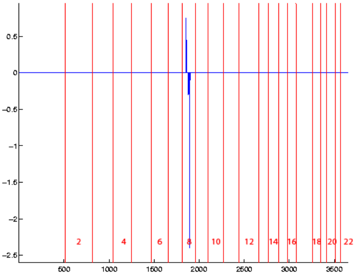
$$\Omega_{fusedlasso}(\beta) = \sum_i |\beta_i| + \sum_{i \sim j} |\beta_i - \beta_j| \,.$$

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions
- Combined with a hinge loss leads to a fused SVM (Rapaport et al., 2008);

Application: metastasis prognosis in melanoma







Example: finding discriminant modules in gene networks

The problem

- Classification of gene expression: too many genes
- A gene network is given (PPI, metabolic, regulatory, signaling, co-expression...)
- We expect that "clusters of genes" (modules) in the network contribute similarly to the classification

Two solutions (Rapaport et al., 2007, 2008)

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

$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Example: finding discriminant modules in gene networks

The problem

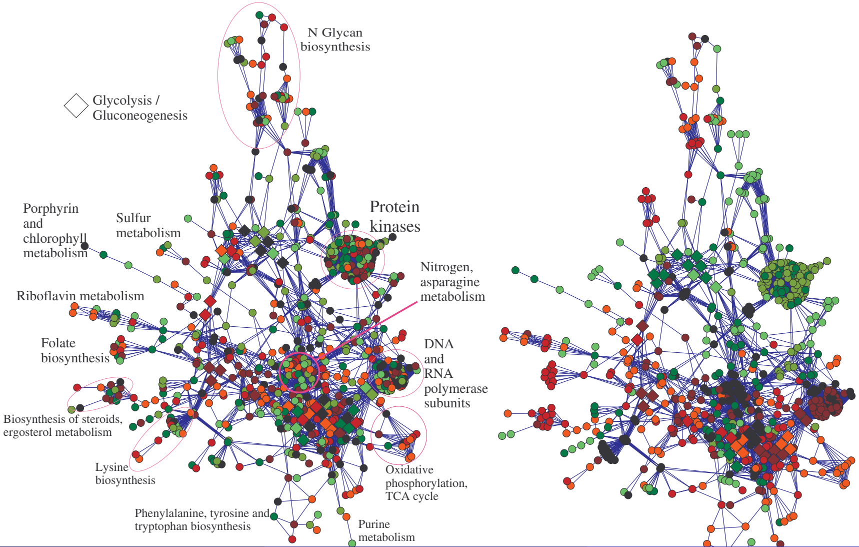
- Classification of gene expression: too many genes
- A gene network is given (PPI, metabolic, regulatory, signaling, co-expression...)
- We expect that "clusters of genes" (modules) in the network contribute similarly to the classification

Two solutions (Rapaport et al., 2007, 2008)

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

$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Example: finding discriminant modules in 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

Outline

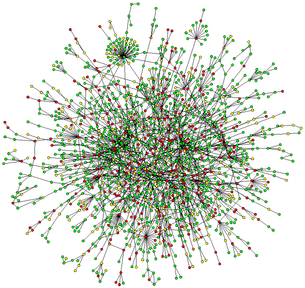
1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data

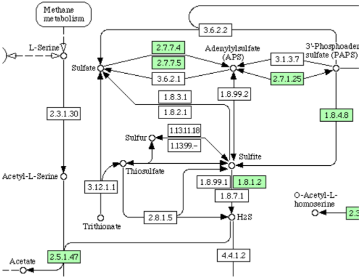
2. Inference on biological networks

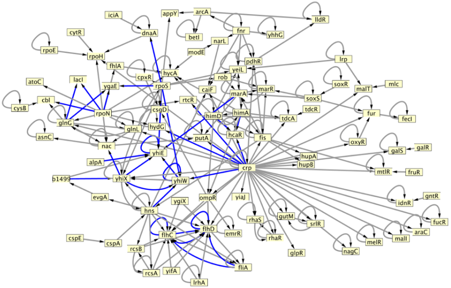
3. Virtual screening and chemogenomics

4. Conclusion

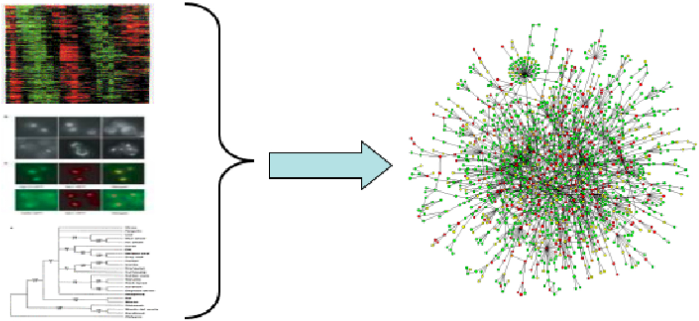
Biological networks







Our goal



Data

- Gene expression,
- Gene sequence,
- Protein localization, ...

Graph

- Protein-protein interactions,
- Metabolic pathways,
- Signaling pathways, ...

More precisely

“De novo” inference

- Given data about individual genes and proteins
- Infer the edges between genes and proteins

“Supervised” inference

- Given data about individual genes and proteins
- and given some known interactions
- infer unknown interactions

More precisely

"De novo" inference

- Given data about individual genes and proteins
- Infer the edges between genes and proteins

"Supervised" inference

- Given data about individual genes and proteins
- and given some known interactions
- infer unknown interactions

Main messages

1. Most methods developed so far are “de novo” (e.g., co-expression, Bayesian networks, mutual information nets, dynamical systems...)
2. However most real-world application fit the “supervised” framework
3. Solving the “supervised” problem is much easier (and more efficient) than the “de novo” problem. It requires less hypothesis.

De novo methods

Typical strategies

- Fit a dynamical system to time series (e.g., PDE, boolean networks, state-space models)
- Detect statistical conditional indenpence or dependency (Bayesian netwok, mutual information networks, co-expression)

Pros

- Excellent approach if the model is correct and enough data are available
- Interpretability of the model
- Inclusion of prior knowledge

Cons

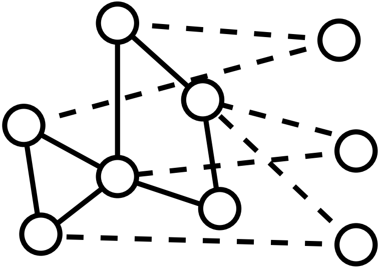
- Specific to particular data and networks
- Needs a correct model!
- Difficult integration of heterogeneous data
- Often needs a lot of data and long computation time

Supervised methods

Motivation

In actual applications,

- we know in advance parts of the network to be inferred
- the problem is to add/remove nodes and edges using genomic data as side information



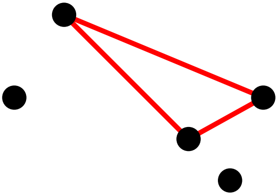
Supervised method

- Given genomic data and the currently known network...
- Infer missing edges between current nodes and additional nodes.

Supervised approach by Metric learning

Idea

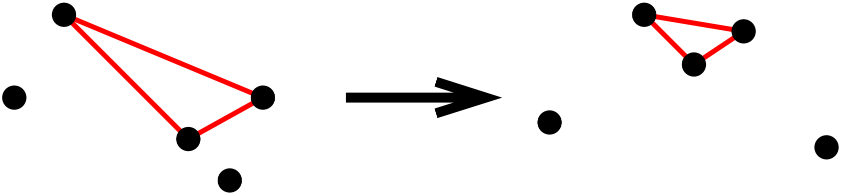
- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)



Supervised approach by Metric learning

Idea

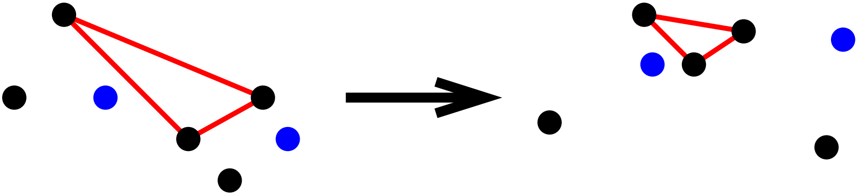
- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)



Supervised approach by Metric learning

Idea

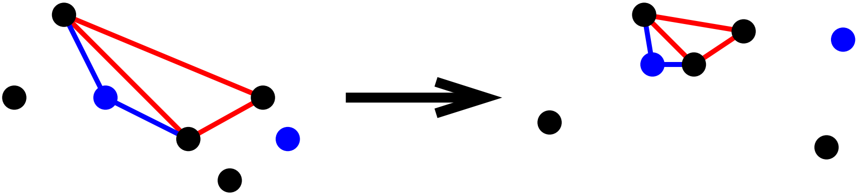
- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)



Supervised approach by Metric learning

Idea

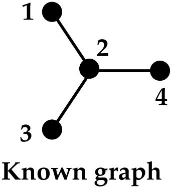
- The direct similarity-based method fails because the distance metric used might not be adapted to the inference of the targeted protein network.
- Solution: use the known subnetwork to refine the distance measure, before applying the similarity-based method
- Examples: kernels CCA (Yamanishi et al. 2004), kernel metric learning (V and Yamanishi, 2005)

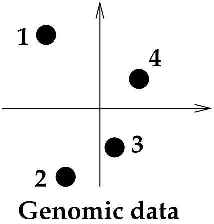


Supervised inference by pattern recognition

Formulation and basic issue

- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!

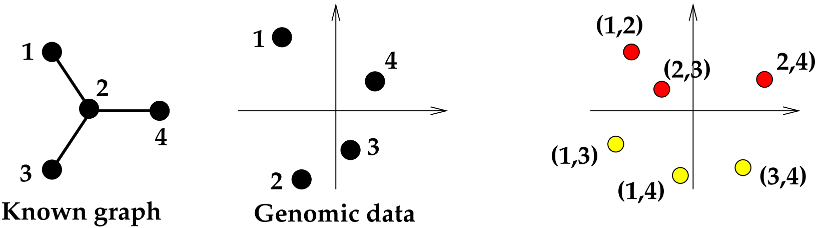




Supervised inference by pattern recognition

Formulation and basic issue

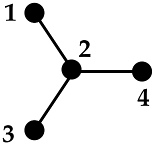
- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!



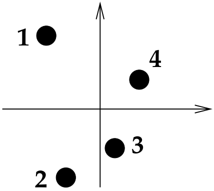
Supervised inference by pattern recognition

Formulation and basic issue

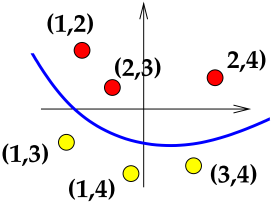
- A pair can be connected (1) or not connected (-1)
- From the known subgraph we can extract examples of connected and non-connected pairs
- However the genomic data characterize individual proteins; we need to work with pairs of proteins instead!



Known graph



Genomic data



Tensor product SVM (Ben-Hur and Noble, 2006)

- Intuition: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C and B is similar to D, or...
  - A is similar to D and B is similar to C
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}((a, b), (c, d)) = K(a, c)K(b, d) + K(a, d)K(b, c) .$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized tensor product:

$$(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a) .$$

Tensor product SVM (Ben-Hur and Noble, 2006)

- Intuition: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C and B is similar to D, or...
  - A is similar to D and B is similar to C
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}\left((a, b), (c, d)\right) = K(a, c)K(b, d) + K(a, d)K(b, c) \,.$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized tensor product:

$$(a, b) \rightarrow (a \otimes b) \oplus (b \otimes a) \,.$$

Tensor product SVM (Ben-Hur and Noble, 2006)

- Intuition: a pair (A, B) is similar to a pair (C, D) if:
  - A is similar to C and B is similar to D, or...
  - A is similar to D and B is similar to C
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{TPPK}((a,b),(c,d)) = K(a,c)K(b,d) + K(a,d)K(b,c) \,.$$

- If K is a positive definite kernel for individuals then K_{TPPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized tensor product:

$$(a,b) \rightarrow (a \otimes b) \oplus (b \otimes a) \,.$$

Metric learning pairwise SVM (V. et al, 2007)

- Intuition: a pair (A,B) is similar to a pair (C,D) if:
  - $A-B$ is similar to $C-D$, or...
  - $A-B$ is similar to $D-C$.
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}\left((a,b),(c,d)\right) = \left(K(a,c)+K(b,d)-K(a,c)-K(b,d)\right)^2 \,.$$

- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a,b) by the symmetrized difference:

$$(a,b) \rightarrow (a-b)^{\otimes 2} \,.$$

Metric learning pairwise SVM (V. et al, 2007)

- Intuition: a pair (A, B) is similar to a pair (C, D) if:
  - $A - B$ is similar to $C - D$, or...
  - $A - B$ is similar to $D - C$.
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}((a,b),(c,d)) = (K(a,c) + K(b,d) - K(a,c) - K(b,d))^2 \ .$$

- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized difference:

$$(a,b) \rightarrow (a-b)^{\otimes 2} \ .$$

Metric learning pairwise SVM (V. et al, 2007)

- Intuition: a pair (A, B) is similar to a pair (C, D) if:
  - $A - B$ is similar to $C - D$, or...
  - $A - B$ is similar to $D - C$.
- Formally, define a similarity between pairs from a similarity between individuals by

$$K_{MLPK}\left((a,b),(c,d)\right) = \left(K(a,c) + K(b,d) - K(a,c) - K(b,d)\right)^2 \ .$$

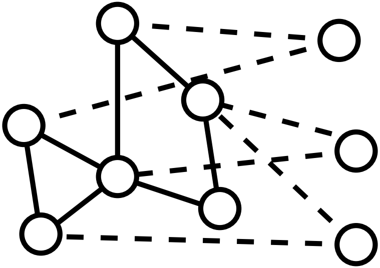
- If K is a positive definite kernel for individuals then K_{MLPK} is a p.d. kernel for pairs which can be used by SVM
- This amounts to representing a pair (a, b) by the symmetrized difference:

$$(a,b) \rightarrow (a-b)^{\otimes 2} \ .$$

Supervised inference with local models

The idea (Bleakley et al., 2007)

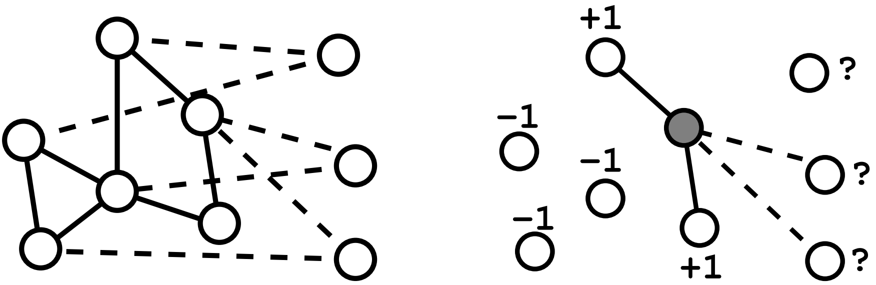
- Motivation: define specific models for each target node to discriminate between its neighbors and the others
- Treat each node independently from the other. Then combine predictions for ranking candidate edges.



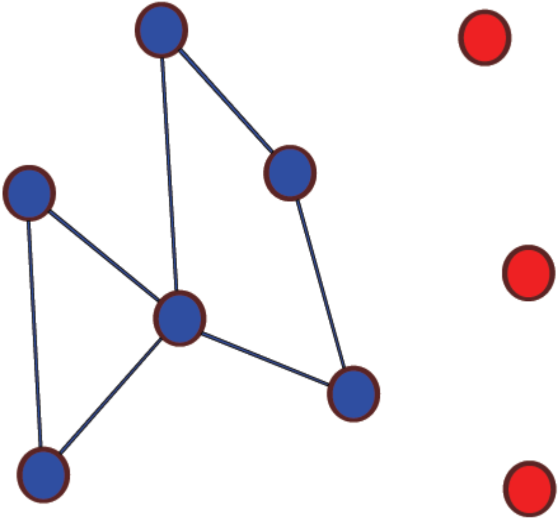
Supervised inference with local models

The idea (Bleakley et al., 2007)

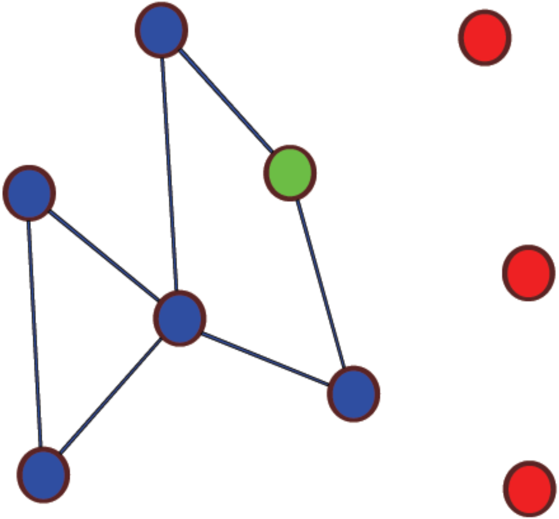
- Motivation: define specific models for each target node to discriminate between its neighbors and the others
- Treat each node independently from the other. Then combine predictions for ranking candidate edges.



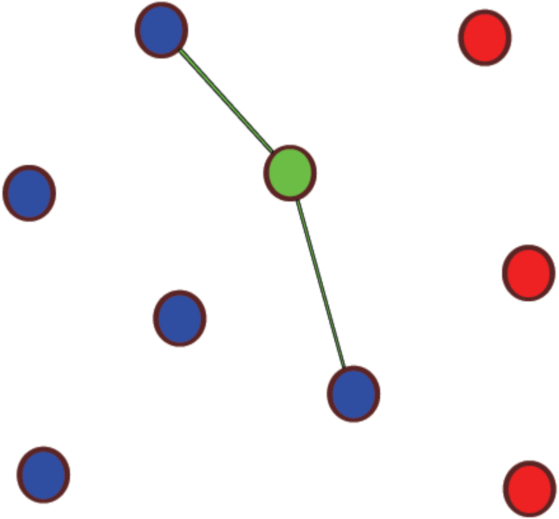
The LOCAL model



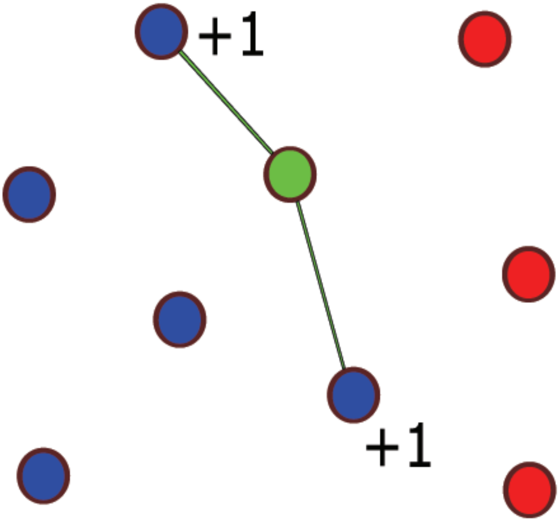
The LOCAL model



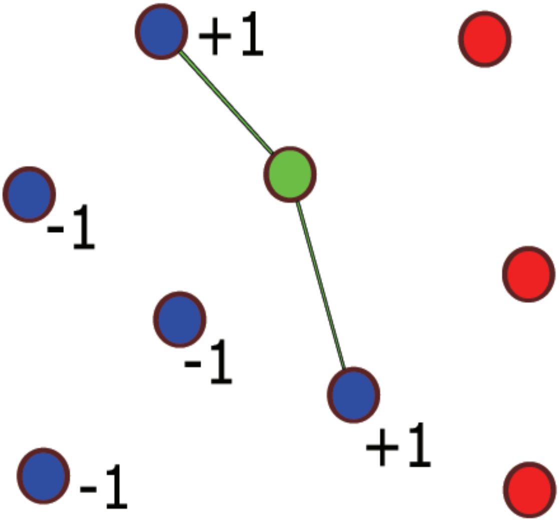
The LOCAL model



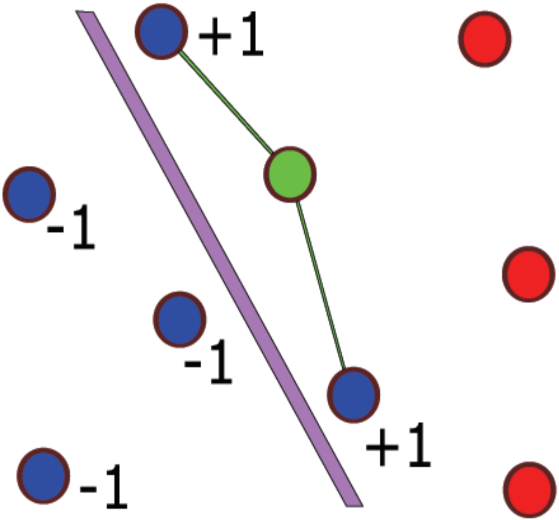
The LOCAL model



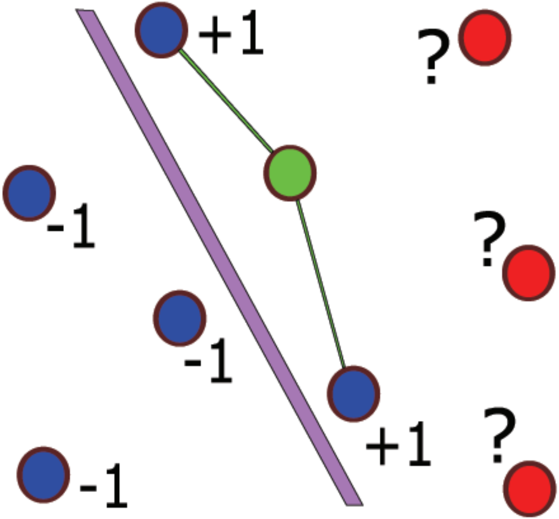
The LOCAL model



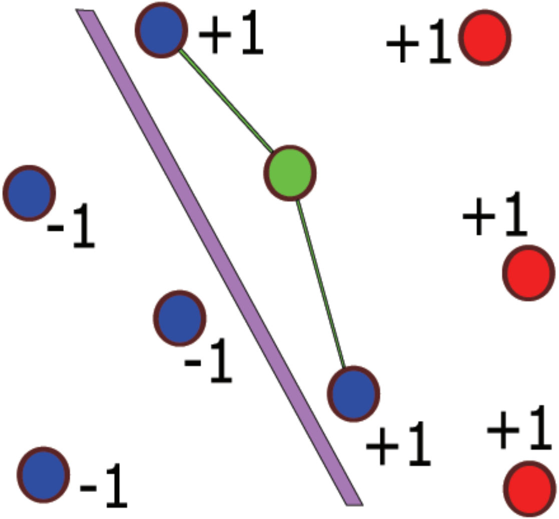
The LOCAL model



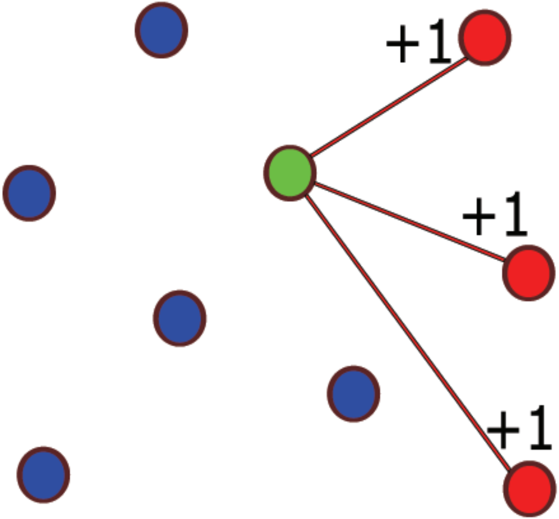
The LOCAL model



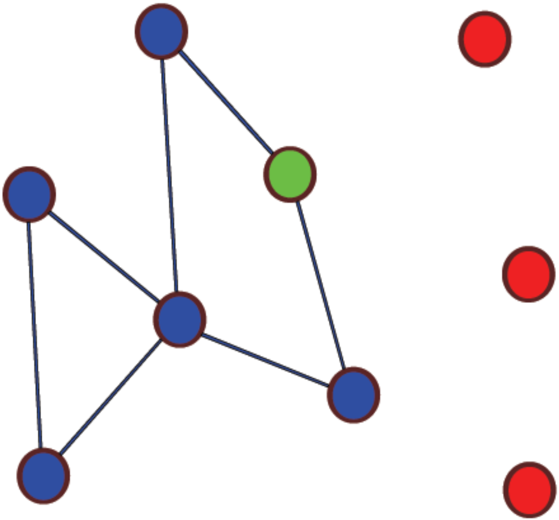
The LOCAL model



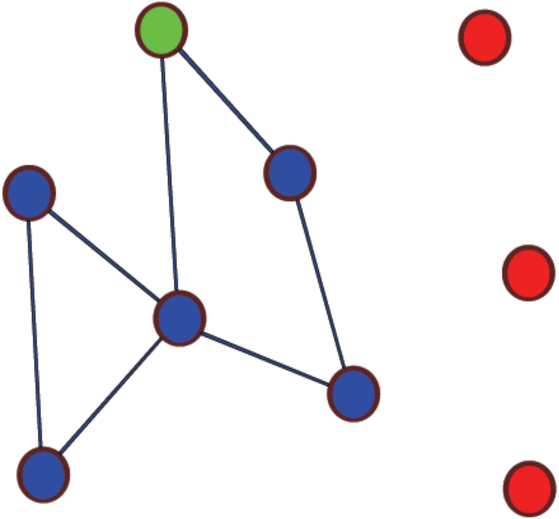
The LOCAL model



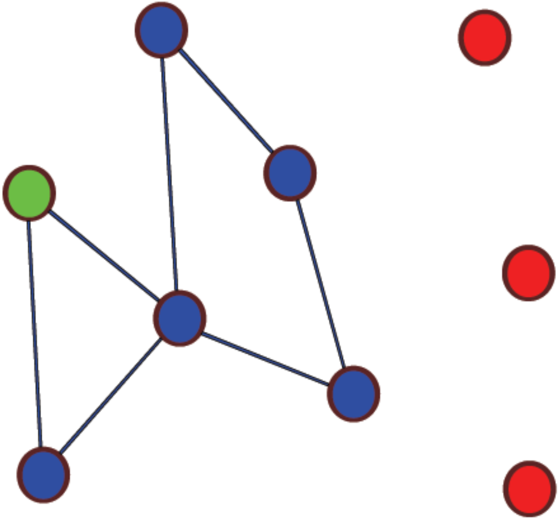
The LOCAL model



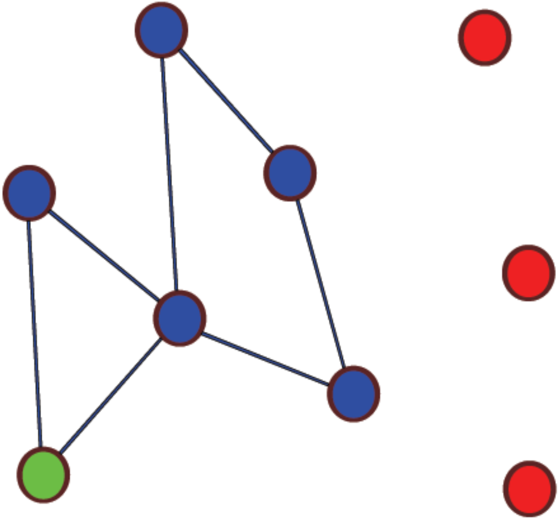
The LOCAL model



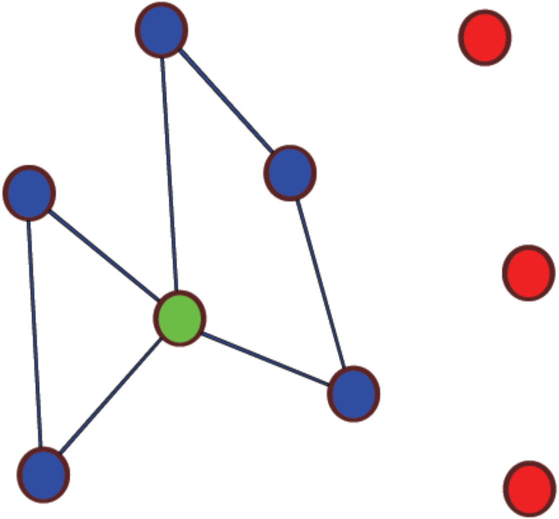
The LOCAL model



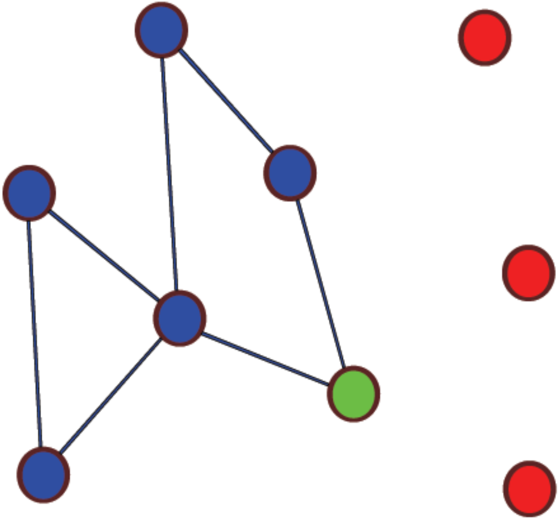
The LOCAL model



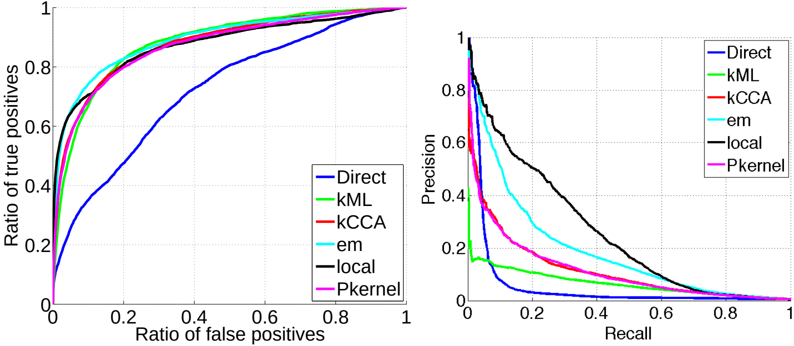
The LOCAL model



The LOCAL model

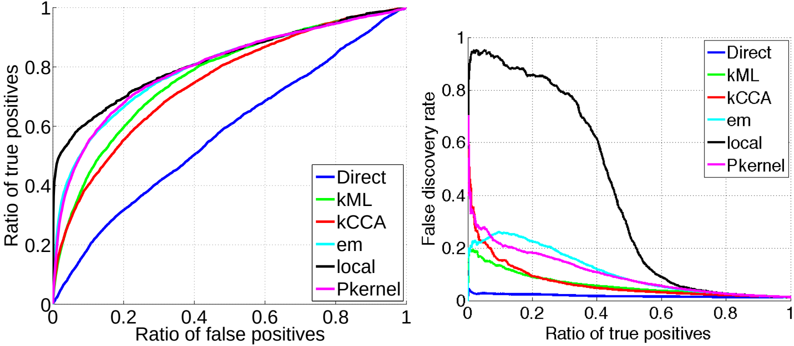


Results: protein-protein interaction (yeast)



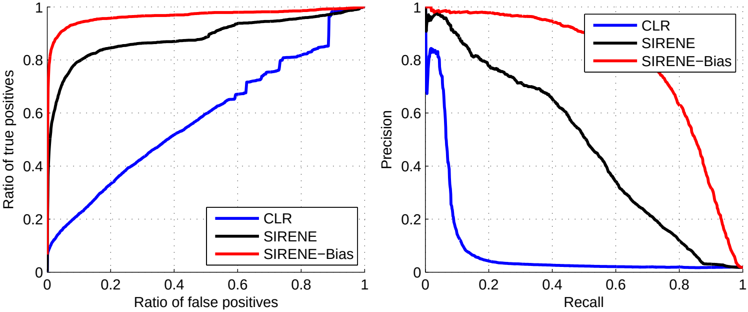
(from Bleakley et al., 2007)

Results: metabolic gene network (yeast)



(from Bleakley et al., 2007)

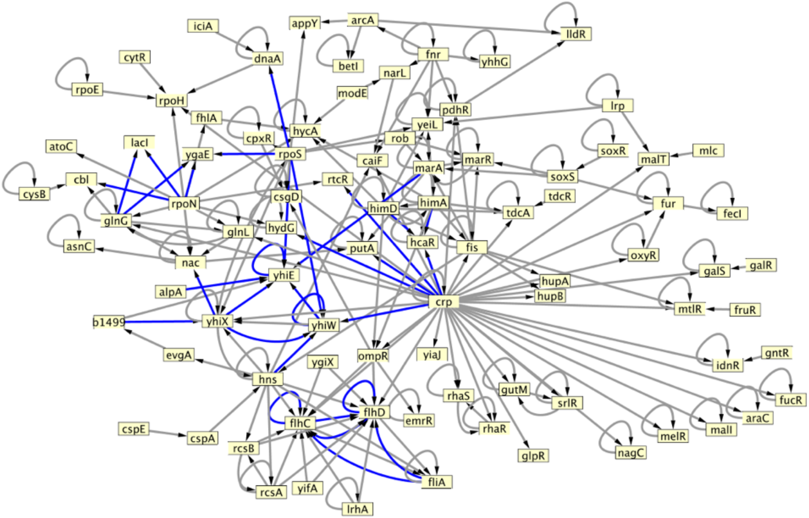
Results: regulatory network (E. coli)



Method	Recall at 60%	Recall at 80%
SIRENE	**44.5%**	**17.6%**
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)

Results: predicted regulatory network (E. coli)



Prediction at 60% precision, restricted to transcription factors (from Mordelet and V., 2008).

Outline

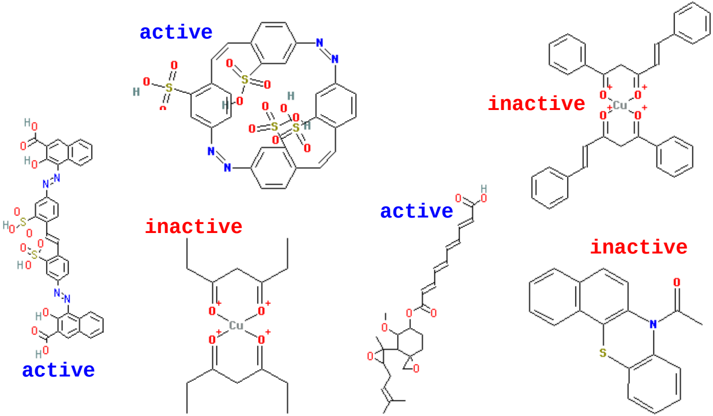
1. Supervised classification of genomic data
    - Context
    - Gene selection for transcriptomic predictive signatures
    - Pathway signatures
    - Predictive chromosomic aberrations with CGH data

2. Inference on biological networks

3. Virtual screening and chemogenomics

4. Conclusion

Ligand-Based Virtual Screening and QSAR



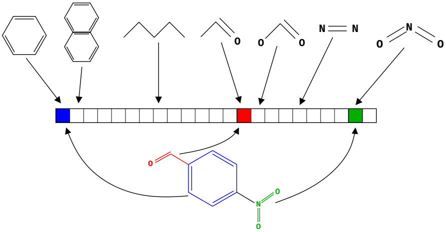
NCI AIDS screen results (from http://cactus.nci.nih.gov).

Classical approaches

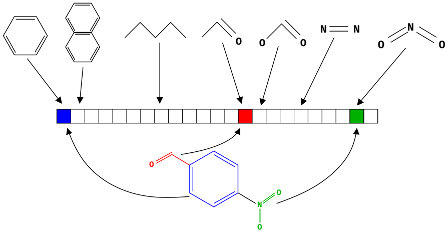
Two steps

1. Map each molecule to a vector of fixed dimension using molecular descriptors
   - Global properties of the molecules (mass, logP...)
   - 2D and 3D descriptors (substructures, fragments, ...)
2. Apply an algorithm for regression or pattern recognition.
   - PLS, ANN, ...

Example: 2D structural keys



Which descriptors?



Difficulties

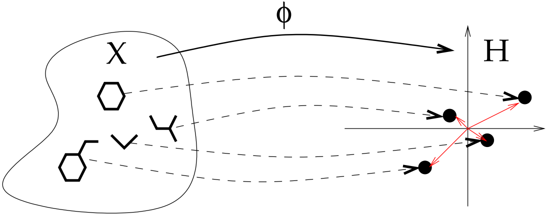
- Many descriptors are needed to characterize various features (in particular for 2D and 3D descriptors)
- But too many descriptors are harmful for memory storage, computation speed, statistical estimation

Kernels

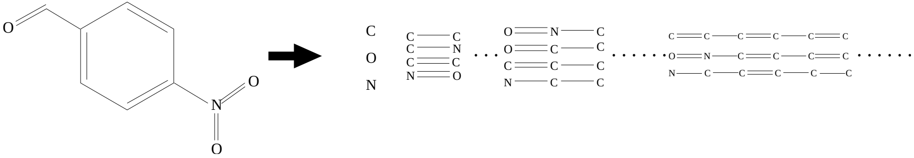
Definition

- Let $\Phi(x) = (\Phi_1(x), \ldots, \Phi_p(x))$ be a vector representation of the molecule x
- The kernel between two molecules is defined by:

$$K(x, x') = \Phi(x)^\top \Phi(x') = \sum_{i=1}^{p} \Phi_i(x)\Phi_i(x') .$$



Example: 2D fragment kernel



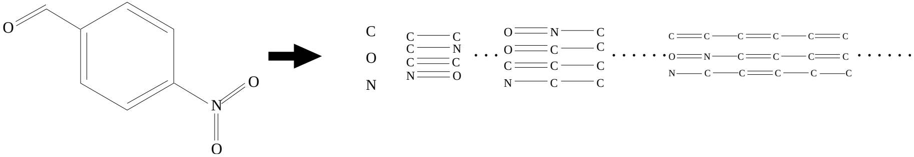
- $\phi_d(x)$ is the vector of counts of all fragments of length d:

$$\phi_1(x) = (\quad \texttt{\#(C),\#(O),\#(N),}\quad \ldots)^\top$$
$$\phi_2(x) = (\quad \texttt{\#(C-C),\#(C=O),\#(C-N),}\quad \ldots)^\top \quad \text{etc...}$$

- The 2D fragment kernel is defined, for $\lambda < 1$, by

$$K_{fragment}(x, x') = \sum_{d=1}^{\infty} r(\lambda) \phi_d(x)^\top \phi_d(x') \,.$$

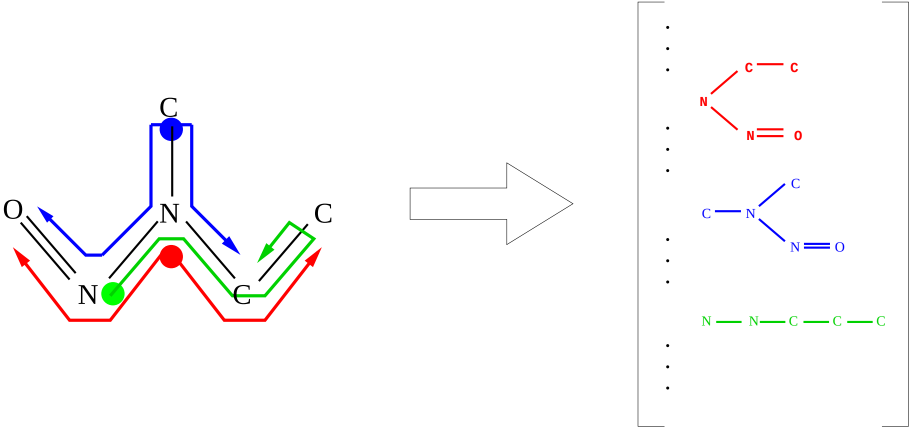
Example: 2D fragment kernel



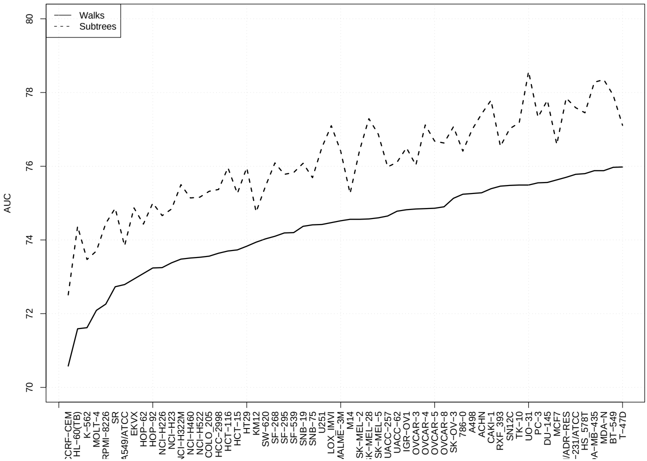
In practice

- $K_{fragment}$ can be computed efficiently (geometric kernel, random walk kernel...) although the feature space has infinite dimension.
- Increasing the specificity of atom labels improves performance
- Selecting only “non-tottering” fragments can be done efficiently and improves performance.

Example: 2D subtree kernel

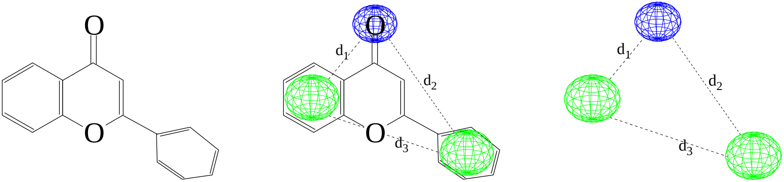


2D Subtree vs fragment kernels (Mahé and V, 2007)



Screening of inhibitors for 60 cancer cell lines (from Mahé and V., 2008)

Example: 3D pharmacophore kernel (Mahé et al., 2005)



$$K(x,y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \exp\left(-\gamma d\left(p_x, p_y\right)\right) \, .$$

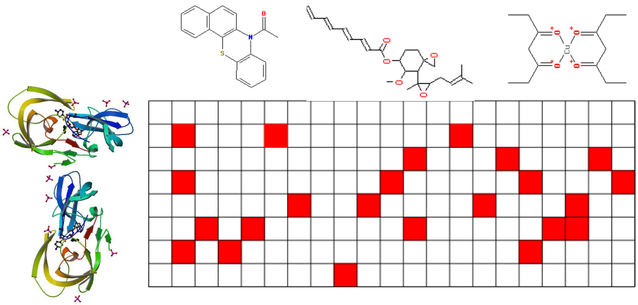
Results (accuracy)

Kernel	BZR	COX	DHFR	ER
2D (Tanimoto)	71.2	63.0	76.9	77.1
3D fingerprint	75.4	67.0	76.9	78.6
3D not discretized	**76.4**	**69.8**	**81.9**	**79.8**

Chemogenomics

The problem

- Similar targets bind similar ligands
- Instead of focusing on each target individually, can we screen the biological space (target families) vs the chemical space (ligands)?
- Mathematically, learn $f(target, ligand) \in \{bind, notbind\}$



Chemogenomics with SVM

Tensor product SVM

- Take the kernel:

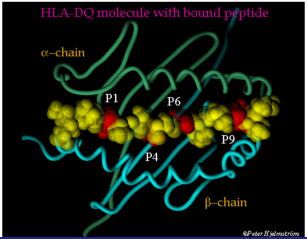
$$K\left((t,l),(t',l')\right) = K_t(t,t')K_l(l,l')\,.$$

- Equivalently, represent a pair (t,l) by the vector $\phi_t(t) \otimes \phi_l(l)$
- Allows to use any kernel for proteins K_t with any kernel for small molecules K_l
- When K_t is the Dirac kernel, we recover the classical paradigm: each target is treated independently from the others.
- Otherwise, information is shared across targets. The more similar the targets, the more they share information.

Example: MHC-I epitope prediction across different alleles

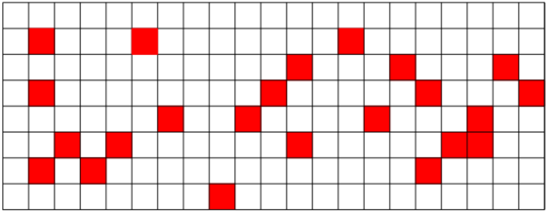
The approach (Jacob and V., 2007)

- take a kernel to compare different MHC-I alleles (e.g., based on the amino-acids in the paptide recognition pocket)
- take a kernel to compare different epitopes (9-mer peptides)
- Combine them to learn the $f(allele, epitope)$ function
- State-of-the-art performance
- Available at `http://cbio.ensmp.fr/kiss`



Generalization: collaborative filtering with attributes

- General problem: learn $f(x, y)$ with a kernel K_x for x and a kernel K_y for y.
- SVM with a tensor product kernel $K_x \otimes K_y$ is a particular case of something more general: estimating an operator with a spectral regularization.
- Other spectral regularization are possible (e.g., trace norm) and lead to efficient algorithms
- More details in Abernethy et al. (2008).



Outline

1. Supervised classification of genomic data
   - Context
   - Gene selection for transcriptomic predictive signatures
   - Pathway signatures
   - Predictive chromosomic aberrations with CGH data
2. Inference on biological networks
3. Virtual screening and chemogenomics
4. Conclusion

What we saw

- Modern machine learning methods for regression / classification lend themselves well to the integration of prior knowledge in the penalization / regularization function, in particular for feature selection / grouping. Applications in array CGH classification, siRNA design, microarray classification with gene networks
- Inference of biological networks can be formulated as a supervised problem if the graph is partly known, and powerful methods can be applied. Application in PPI, metabolic and regulatory networks inference.
- Kernel methods (eg SVM) allow to manipulate complex objects (eg molecules, biological sequences) as soon as kernels can be defined and computed. Applications in virtual screening, QSAR, chemogenomics.

People I need to thank

Including prior knowledge in penalization

Franck Rapaport, Emmanuel Barillot, Andrei Zynoviev, Christian Lajaunie, Yves Vandenbrouck, Nicolas Foveau...

Virtual screening, kernels etc..

Pierre Mahé, Laurent Jacob, Liva Ralaivola, Véronique Stoven, Brice Hoffman, Martial Hue, Francis Bach, Jacob Abernethy, Theos Evgeniou...

Network inference

Kevin Bleakley, Fantine Mordelet, Yoshihiro Yamanihi, Gérard Biau, Minoru Kanehisa, William Noble, Jian Qiu...