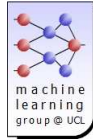


# Biomarker selection from microarray data: a transfer learning approach

Pierre.Dupont@uclouvain.be



Computing Science and Engineering Department  
Université catholique de Louvain – Belgium

October 02, 2009

## Outline

- 1 Context and motivations
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 Gene selection by transfer learning across datasets
- 5 Conclusion and some open issues

A joined work with the "Microarray Team", especially with T. Helleputte

## 1 Context and motivations

- DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint

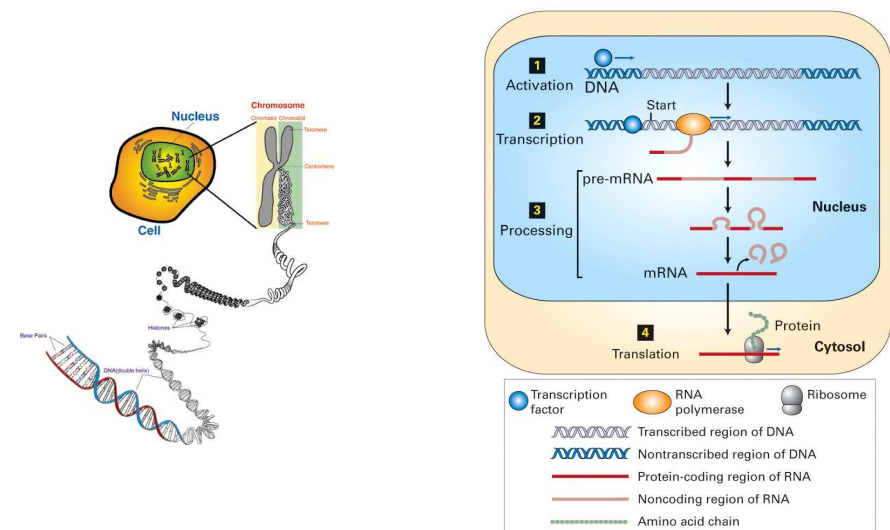
## 2 Embedded SVM-based feature selection

## 3 Gene selection with partial supervision from prior knowledge

## 4 Gene selection by transfer learning across datasets

## 5 Conclusion and some open issues

## Molecular biology in one slide!

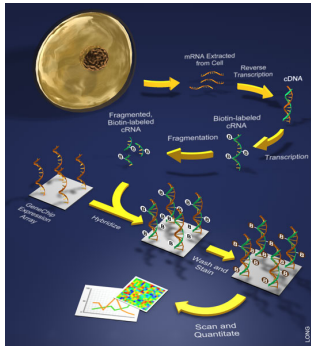


# Gene expression is **highly variable**

## Gene expression

- A **gene** is said to be **expressed** when it is actually translated to a protein
- Gene expressions **vary** due to many different factors
  - ▶ growth of the organism
  - ▶ cell types
  - ▶ chemical and physical environment of the cell
  - ▶ a pathology of the organism
  - ▶ a genetic susceptibility to develop a pathology
  - ▶ a response to a treatment
  - ▶ ...

## DNA Microarrays



- DNA Microarrays measure the level of expression of **all genes** in a **single experiment**
- This technology is **much faster** than previous technologies (e.g. RT-PCR) which are working gene by gene
- This technology is also a **bit expensive**: **500 ... 1,000 €**/chip
- ... and **not perfect** (noisy data, many possible confounding factors)

- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 Gene selection by transfer learning across datasets
- 5 Conclusion and some open issues

## CRISTALL: predicting the risk of allergies of newborns



- More than **30%** of children are affected by allergies in industrial countries
- Collaboration with **Pediatric Gastroenterology Hepatology Unit** at *Cliniques Universitaires St-Luc*
- First microarray data of  $\approx 50$  children are currently available.

# GSK MAGE

## MAGE-A3 immunotherapeutic

Significant clinical activity of MAGE-A3 antigen specific cancer immunotherapeutic treatment has been reported in a Phase II study in metastatic melanoma.



## Objectives

- Gene expression profiling using microarrays to identify markers predictive of the clinical activity **before** treatment
- Predict responders versus non-responders from gene expression **before** treatment

# BIOWIN: Rheumagene

## Objectives

- **Early diagnosis** of the type of rheumatoid infection
  - ▶ Rheumatoid arthritis
  - ▶ Psoriatic rheumatism
  - ▶ Microcrystalline arthritis
  - ▶ Inflammatory osteoarthritis
  - ▶ ...
- Gene expression data (low density microarray + real time PCR) and protein levels (ELISA tests)
- Collaboration with **rheumatology clinical departments** of UCL, Univ. Liège, CHU Brugmann, U. Ghent + **EUROGENTEC, Eppendorf Array Technologies**
- Expected outcome: diagnosis clinical kit

- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 Gene selection by transfer learning across datasets
- 5 Conclusion and some open issues

# Microarray data: a machine learning viewpoint

	gene 1	gene 2	...	gene d	class label
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,d}$	$y_1$
sample 2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,d}$	$y_2$
...	...	...	...	...	...
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,d}$	$y_n$

- The number  $d$  of **genes** or **probe sets**  $\approx 55,000$
- The number  $n$  of **samples** (tissues, patients)  $\leq 100$
- Each sample is characterized by a vector  $\mathbf{x}$  of expression values
- The class labels  $y$  are not measured on the chip but come from clinical status: **diagnosis? survival chance? responder to treatment?**

## Biomarker Selection and Classifier Estimation

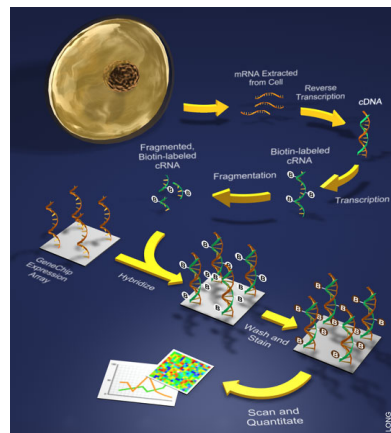
Find a small subset of ( $\approx 50$ ) genes to predict the outcome  $y$  of **new samples**

## Learning/Estimation Challenges

- The dimension  $d$  of the input space (= number of genes) is very large ( $\approx 55,000$ ) with respect to the number of samples ( $n \leq 100$ )
  - ▶ the classification problem is mathematically strongly undetermined
  - ▶ many “apparently equivalent” solutions
- The measurements are **noisy**
  - ▶ experimental protocol variability
  - ▶ non linearities of laser sensitivity
  - ▶ image processing
  - ▶ **intrinsic variability of RNA level of expression**
  - ▶ variability in expression level **not** related to the problem under study (age, sex, cell types, other pathologies, external factors, ...)

## Microarray data analysis workflow

- 1 Data extraction
- 2 Summarization
- 3 Sample normalization
- 4 Non-specific filtering
- 5 Gene expression normalization
- 6 **Biomarker selection and supervised classification**

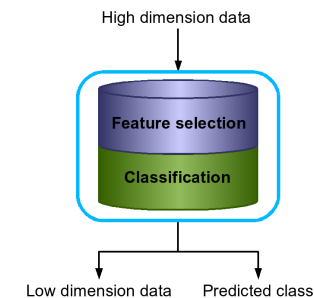


## Gene selection: ranking methods



- Use training data + class labels only during gene selection
- Often **univariate** techniques
  - ▶ measure to which extent a **single** gene is **differently expressed** between responders and non-responders
- Standard techniques: **Golub's S/N ratio**, **t-Test**, mutual information, information gain, ... to **rank** genes
- Train a single classifier taking the selected genes as inputs
- The **simplest** and **less computing intensive** approach

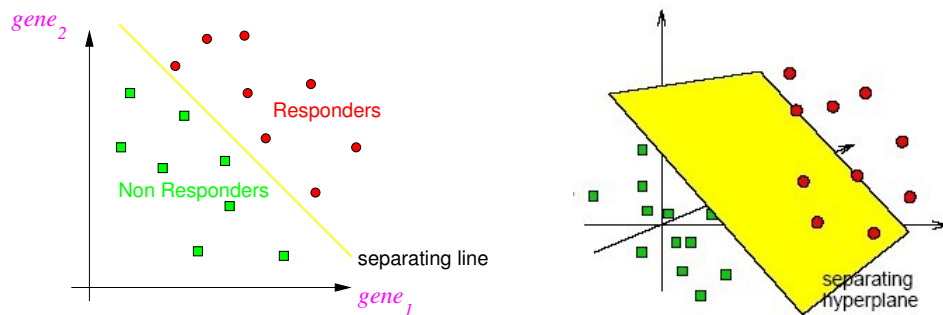
## Gene selection: embedded approaches



- Define the gene selection and the classifier estimation as a **combined optimization process**
- Include possible regularization and classifier optimization in the gene selection process
- **Multivariate** methods: find genes that are **jointly predictive**
- More elegant mathematically but also more computing intensive
- **Support vector machines** have been extended along these lines

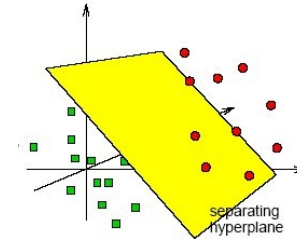
- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 Gene selection by transfer learning across datasets
- 5 Conclusion and some open issues

## Linear Discriminants



- The actual number of dimensions is  $\geq 10,000$  = the number of genes after non-specific filtering
- The linear discriminant is a **hyperplane** in  $\mathbb{R}^{\geq 10,000}$

## Linear Separability

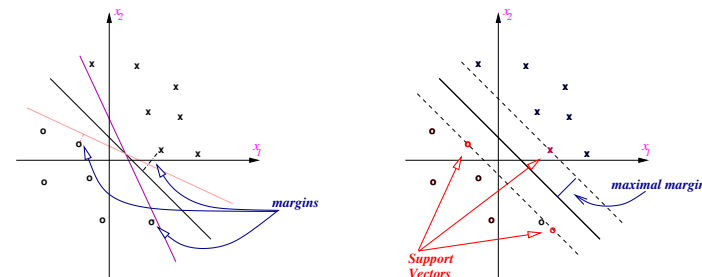


### Facts

- The data is **linearly separable** if the two classes can be perfectly separated by a hyperplane
- A hyperplane in  $\mathbb{R}^{10,000}$  can **separate perfectly** at least **10,001** (unaligned) points, given any possible 2 class labeling
- There is no problem to find a perfect linear separator of less than **100** points in  $\mathbb{R}^{\geq 10,000}$
- The problem is that there are many *apparently perfect* models

## Support Vector Machines in a nutshell

- When the data is linearly separable the separating hyperplane is not unique but the **maximal margin hyperplane** separates the data with the largest margin
- **Statistical learning theory [Vapnik, 1995]** maximizing the margin is a form of **regularization** which is relevant for generalizing well from a **finite sample**
- For each separating hyperplane, there is an associated set of **support vectors**



## SVM in a mathematical nutshell

## SVM Optimization Problem

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 \text{ subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1, \forall i$$

with  $\mathbf{x}_i$  the  $i^{\text{th}}$  sample and  $y_i \in \{+1, -1\}$  its class label

## SVM solution

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \text{ and } f(\mathbf{x}) = \text{sign}[\langle \mathbf{w}, \mathbf{x} \rangle + w_0] = \text{sign}[\sum_{j=1}^d w_j x_j + w_0]$$

- $n$  values  $\alpha_1 \dots \alpha_n$  (the "weight" of each sample) define  $d$  components of the vector  $\mathbf{w}$  (the weight of each gene)
- Very convenient and appropriate when  $n \ll d$

Recursive Feature Elimination [Guyon *et al.*, 2002]

## Embedded Backward Selection

- 1 Estimate a SVM on a given gene set (initially  $d$  genes)
- 2 Consider  $|w_j|$  as the relevance of the  $j^{\text{th}}$  gene
- 3 Remove the least relevant gene(s)
- 4 Iterate 1 to 3 on a reduced gene set

AROM methods [Weston *et al.*, 2003]

Find the separating hyperplane with the fewest nonzero elements satisfying the margin constraints

## Zero-Norm Minimization

$$\min_{\mathbf{w}} \|\mathbf{w}\|_0 \text{ subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$$

This **NP-hard** problem can be approximated and solved iteratively

## L1-AROM objective

$$\min_{\mathbf{w}} \sum_{j=1}^d \log(|w_j| + \epsilon) \text{ subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$$

- 1  $\min_{\mathbf{w}} \|\mathbf{w}\|_1$  subject to  $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$
- 2 Rescale the input data  $\mathbf{x}_1, \dots, \mathbf{x}_n$  component-wise by  $\mathbf{w}$
- 3 Iterate 1 to 2 on the rescaled inputs

## L1-AROM versus L2-AROM

## Iterative procedures

- 1  $\min_{\mathbf{w}} \|\mathbf{w}\|_1$  subject to  $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$   
versus  
 $\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2$  subject to  $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$
- 2 Rescale the input data  $\mathbf{x}_1, \dots, \mathbf{x}_n$  component-wise by  $\mathbf{w}$
- 3 Iterate 1 to 2 on the rescaled inputs

- L1-AROM better enforces sparsity
- L2-AROM is computationally faster

## L2-AROM versus RFE

### L2-AROM

- Gene expression values are rescaled by multiplying them iteratively by  $|w_j| \Rightarrow$  some gene values will gradually decrease towards 0 and eventually vanish
- The number of genes kept depends on the optimization process

### RFE

- At each iteration, genes are either kept (= multiplied by 1) or eliminated (= multiplied by 0)
- RFE can be seen as a hard thresholding version of L2-AROM
- The number of genes kept depends on the user choice

- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 Gene selection by transfer learning across datasets
- 5 Conclusion and some open issues

## Partially supervised gene selection

### Usual approach

- Select genes from microarray with no prior preference between genes
- Check the predictive performance of the signature
- Interpret the signature biologically

### Novel approach

- Make use of an *a priori proposed list of reporters*
- Favor these genes in the selection process
- Let the data confirm or infirm the prior selection + complement with additional genes

## Partially Supervised AROM [Helleputte & Dupont, 2009a]

### L1-AROM objective

$$\min_{\mathbf{w}} \sum_{j=1}^d \log(|w_j| + \epsilon) \text{ subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$$

### PS-L1-AROM objective

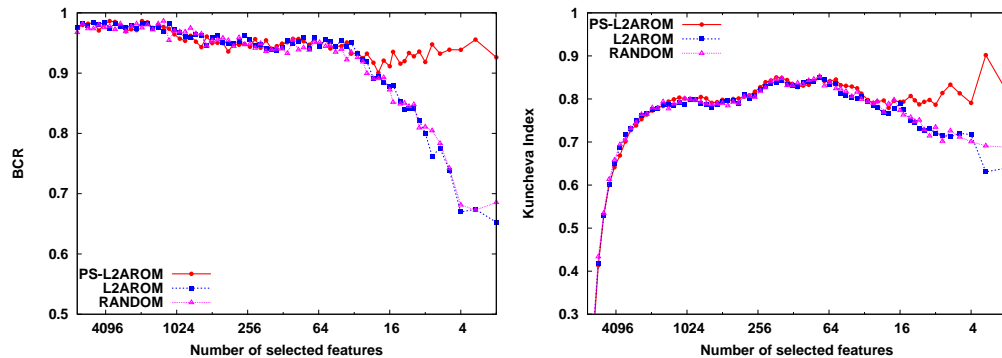
$$\min_{\mathbf{w}} \sum_{j=1}^d \frac{1}{\beta_j} \log(|w_j| + \epsilon) \text{ subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$$

- $\beta_j \geq 1$ : the higher  $\beta_j$  the more relevant feature  $j$  is *a priori assumed*
- $\beta_j = 1$  when no prior information is known about feature  $j$

- 1  $\min_{\mathbf{w}} \|\mathbf{w}\|_1$  subject to  $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1$
- 2 Rescale the input data  $\mathbf{x}_1, \dots, \mathbf{x}_n$  component-wise by  $\mathbf{w} * \beta$
- 3 Iterate 1 to 2 on the rescaled inputs

# Partial Supervision Results

**LEUKEMIA data set:** discrimination between two subtypes of leukemia while favoring 3 genes (out of 7,129 genes) known as clinical markers



## PS-AROM properties

- **Positive effect** on **classification** and **stability** by only favoring **very few genes** ( $\beta_j = 10$ ) versus many other genes ( $\beta_j = 1$ )
- **Multivariate selection:** the genes a priori favored influence the selection of other genes
- The genes a priori favored are not necessarily selected
  - ▶  $|w_j|\beta_j$  matters more than  $\beta_j$  alone!
- Favoring genes **at random** does **not degrade performance**
  - ▶ *Prior knowledge need not be fully accurate, nor complete!*
- $\beta_j = 10$  for favored genes is an arbitrary value, but this is not an issue (see later)

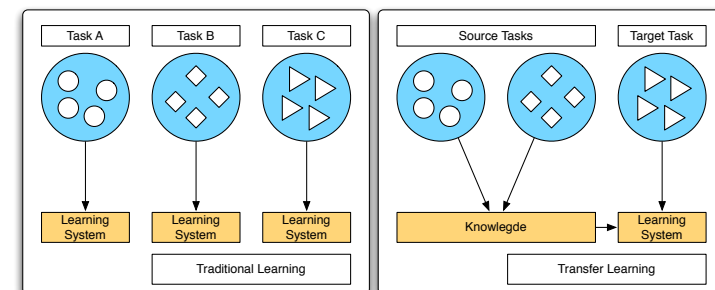
- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint
- 2 Embedded SVM-based feature selection
- 3 Gene selection with partial supervision from prior knowledge
- 4 **Gene selection by transfer learning across datasets**
- 5 Conclusion and some open issues

## Transfer feature relevance across datasets

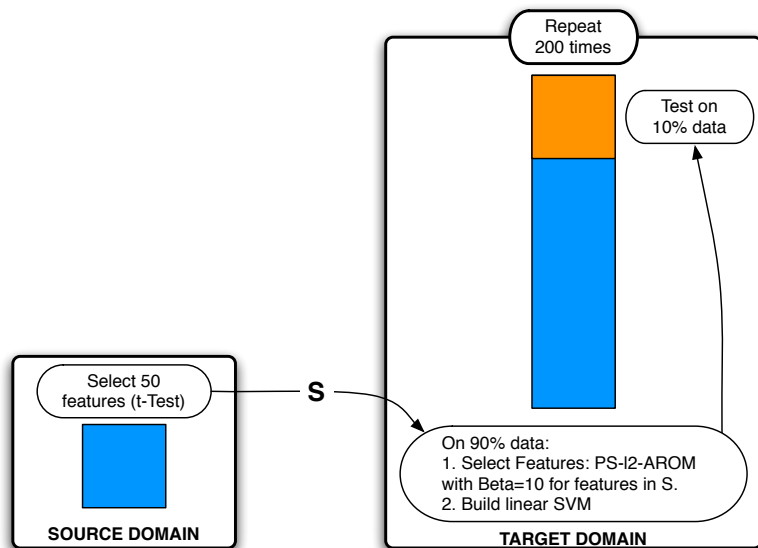
[Helleputte & Dupont, 2009b]

### Facts

- Not many reliable markers are already known (this is part of our research!)
- Microarray datasets are generally very small but many datasets are publicly available



# Protocol 1: Single Transfer of Features

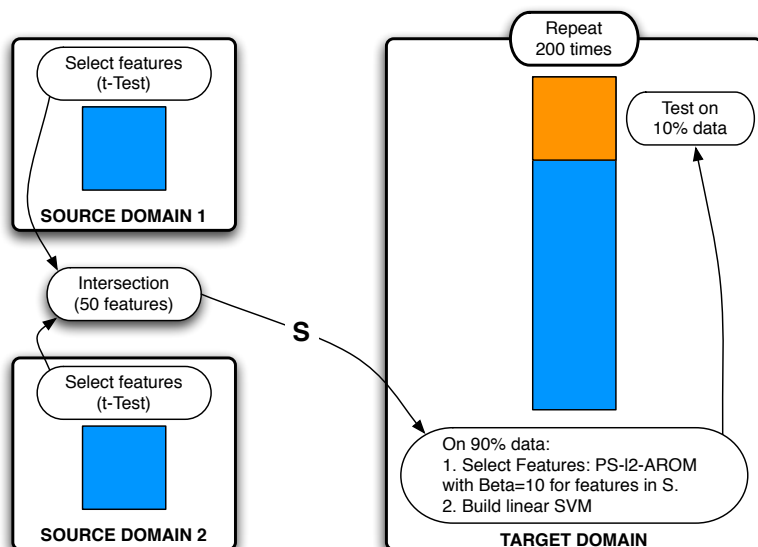


# Datasets

Data Set	Normal/Tumor	Feat.	Technology
Singh	50/52	12625	HGU95Av2
Chandran	18/86	12625	HGU95Av2
Welsh	9/25	12626	HGU95A

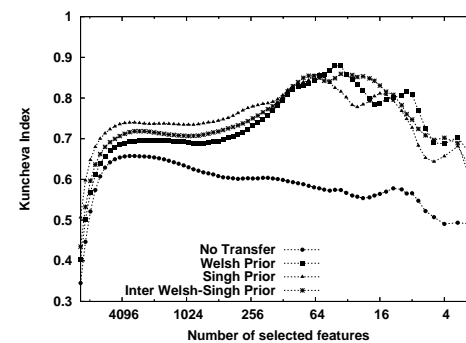
- 3 prostate cancer "related" datasets
- 12,600 probesets ( $\approx$  genes) in common

# Protocol 2: Multiple Transfer of Features

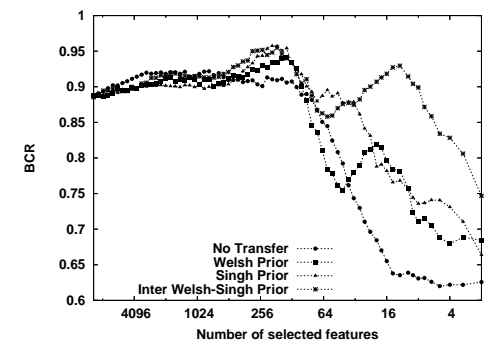


# Transfer results

## Stability



## Classification Performance

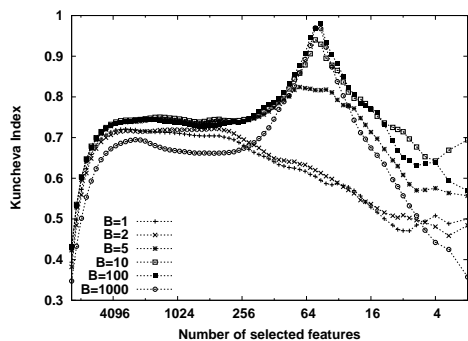


## Notes

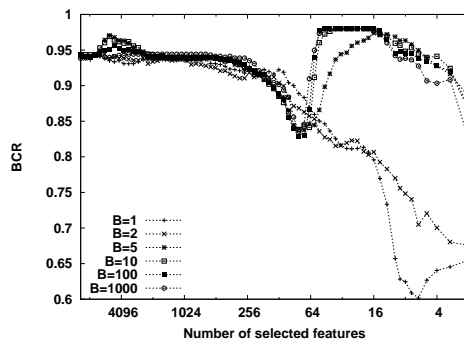
- performance differences are **statistically significant** for small signature sizes
- multiple transfer does not always improve over single transfer

Sensitivity to  $\beta > 1$  values

## Stability



## Classification Performance



## Take home message

- Biomarker discovery from microarray data can be tackled through **multivariate embedded partially supervised feature selection**
- There is a benefit to favor some genes **a priori assumed more relevant or apparently relevant on related datasets**
- *Prior/Transferred knowledge need not be fully accurate nor complete* (soft constraint)
- The actual “strength” of the prior knowledge does not matter much

- 1 Context and motivations
  - DNA microarrays: a technology to measure gene expression
  - Some ongoing research projects
  - Data analysis/ML viewpoint

## 2 Embedded SVM-based feature selection

## 3 Gene selection with partial supervision from prior knowledge

## 4 Gene selection by transfer learning across datasets

## 5 Conclusion and some open issues

## Perspectives and open issues

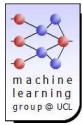
## Practical issues

- Fine tune which genes to transfer or how many
- Relate strengths with confidence in prior belief
- Iterate with a molecular biologist in the loop
  - further link with functional annotations of genes
- Draw relevant medical conclusions

## Theoretical issues

- extend those ideas to closely related approaches
  - LARS/LASSO
  - L1-penalized logistic regression (Generalized LASSO)
  - Elastic Net
- define a multi-objective for jointly optimizing **classification**, **sparsity** and **stability** and design a solver accordingly

# Thanks to a growing team



- $\approx 30$  researchers at UCL

- The **Microarray Team** (in order of arrival in the team)



1 Prof. P. Dupont



4 Roman Zakharov



2 Thibault Helleputte



5 Dr. Daniel  
Hernández-Lobato



3 Dr. Nizar Touleimat



6 ...

# Further Reading

- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. 2002.  
 Gene selection for cancer classification using support vector machines.  
*Machine learning*, 46, 389–422.
- Helleputte, T., & Dupont, P. 2009a.  
 Partially supervised feature selection with regularized linear models.  
*In: International conference on machine learning*.
- Helleputte, T., & Dupont, P. 2009b.  
 Feature selection by transfer learning with linear regularized models.  
*Pages 533–547 of: European conference on machine learning*.  
 Lecture Notes in Artificial Intelligence, no. 5781.
- Vapnik, V.N. 1995.  
*The nature of statistical learning theory*.  
 Springer-Verlag.
- Weston, J., Elisseeff, A., Schölkopf, B., & Tipping, M. 2003.  
 Use of the zero-norm with linear models and kernel methods.  
*Journal of machine learning research*, 3, 1439–1461.