

Ensemble methods for bioinformatics

Giorgio Valentini

e-mail: valenti@disi.unige.it



Istituto Nazionale per la Fisica della Materia

DISI Dipartimento di Informatica
e Scienze dell'Informazione

Università di Genova

Ensemble methods for bioinformatics and for gene expression data analysis

Applied in different bioinformatics domains: e.g.

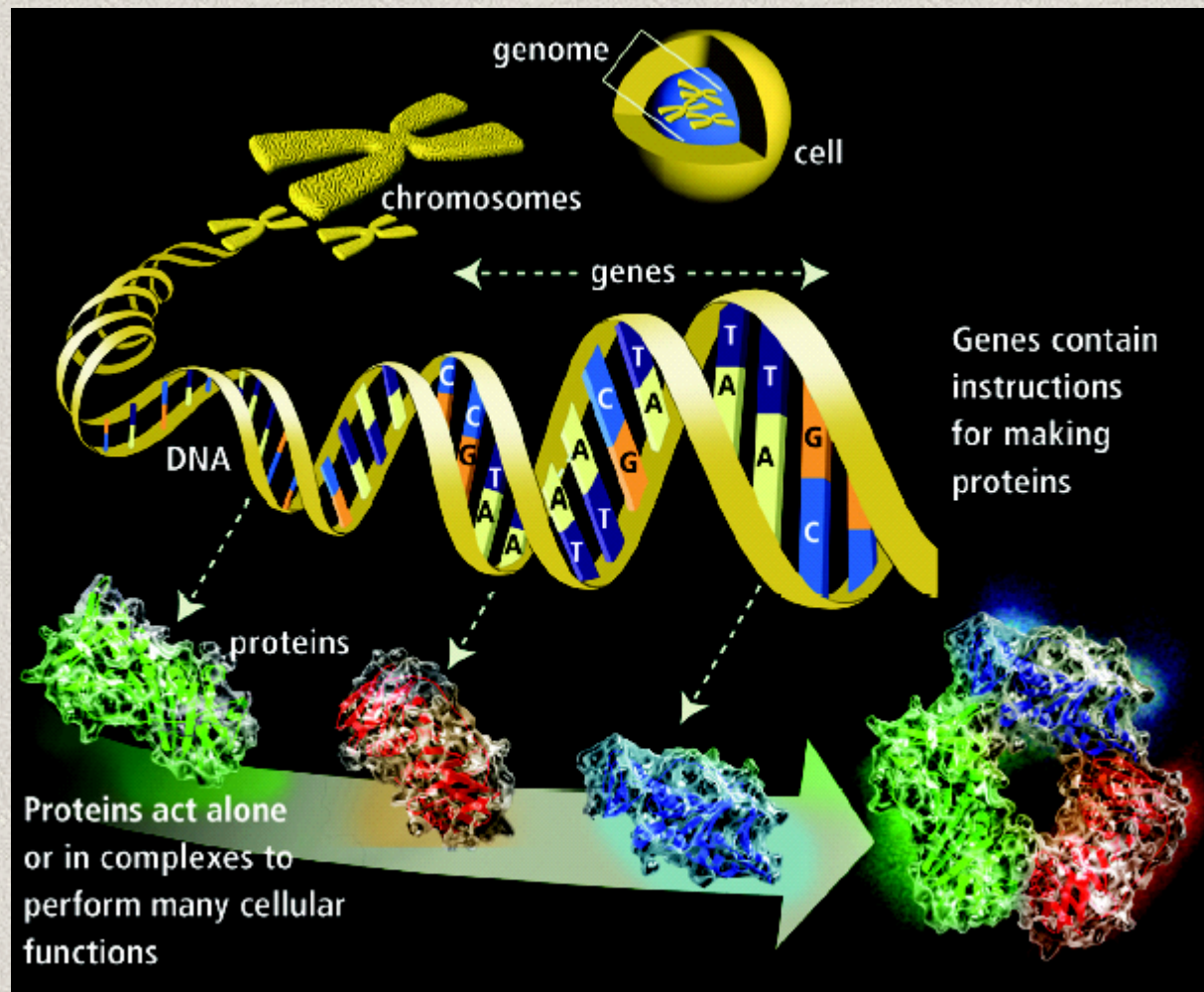
- Protein secondary structure predictions (*Riis and Krogh, 1996, Petersen et al., 2000*)
- Gene finding and intron splice site prediction (*Brunak et al., 1991*)

But we focus on **ensemble methods for gene expression analysis**.

Outline

- Gene expression
- cDNA microarray technology
- Ensemble methods for gene expression data analysis

Relationships between DNA and proteins



Expression of the genetic information

The **expression** of the genetic information stored in the DNA molecule occurs in two stages:

- (i) **transcription**, during which DNA is transcribed into mRNA;
- (ii) **translation**, during which mRNA is translated to produce a protein.

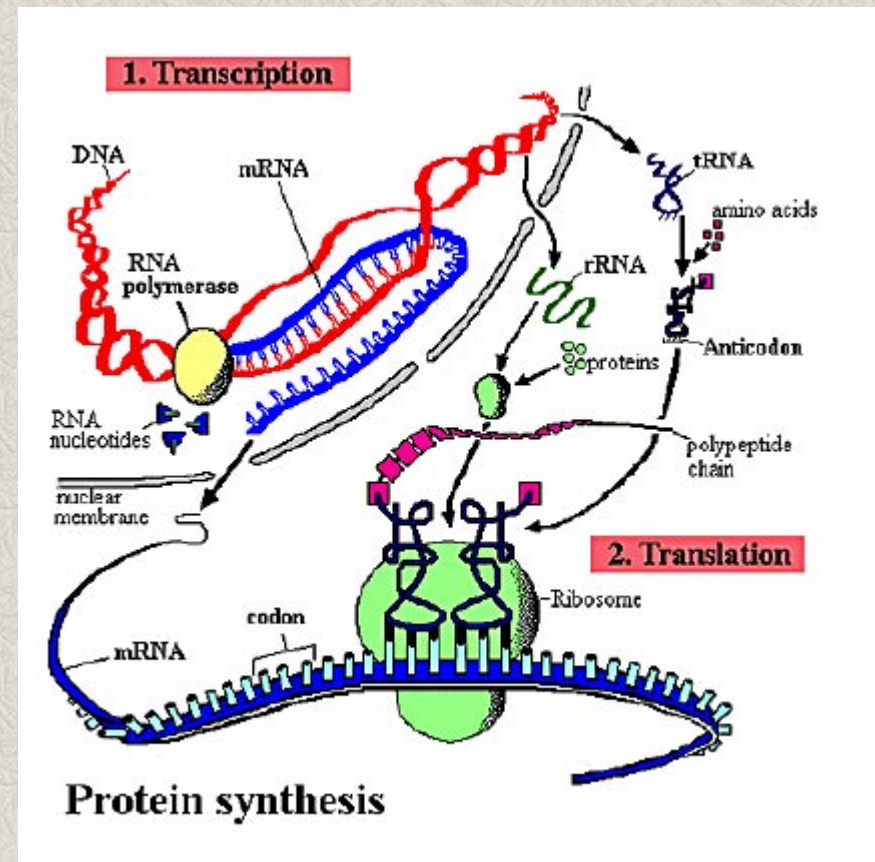
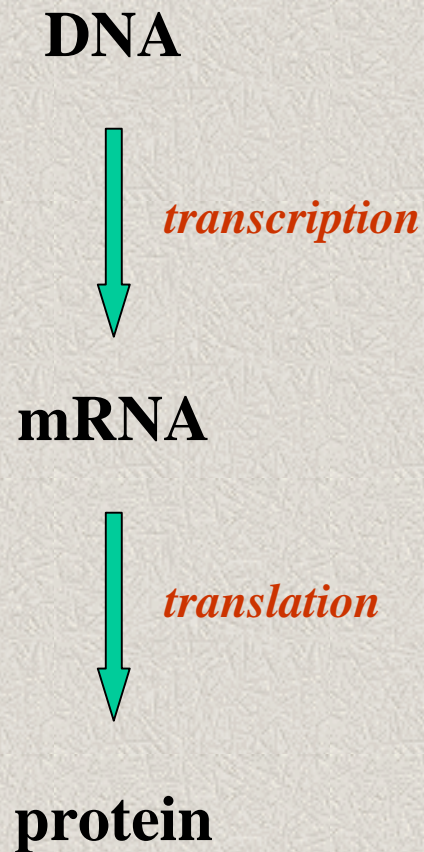
Differential expression

- Each cell contains a complete copy of the organism's genome (that is, the same genome).
- Cells are of many different types and states: blood, nerve, and skin cells, dividing cells, cancerous cells, etc.

What makes the cells different?

- **Differential gene expression**, i.e., **when**, **where**, and **how much** each gene is expressed.
- On average, 40% of our genes are expressed at any given time.

The “central dogma” of Molecular Biology



Transcriptome

- mRNA transcript levels reflect the functional status of a cell.

- Measuring protein levels (translation) would be more direct but more difficult.

The **transcriptome** reflects

- Tissue source: cell type, organ.
- Tissue activity and state:
 - Stage of development, growth, death.
 - Cell cycle.
 - Disease vs. healthy.
 - Response to therapy, stress.

Functional genomics

- The **genome projects** have yielded the complete DNA sequences of many organisms: human, mouse, yeast, fruitfly, etc.

Human: 3 billion base-pairs, 30-40 thousand genes.

- Challenge: **go from sequence to function**,

i.e., define the role of each gene, their interactions and understand how the genome functions as a whole.

- **Transcriptomics** involves large-scale analysis of messenger RNAs to follow when, where, and under what conditions genes are expressed.
- **Proteomics**—the study of protein expression, protein-protein interactions and functions
- **Structural genomics** studies the 3-D structures of one or more proteins from each protein family, thus offering clues to function and biological targets for drug design.
- **Comparative genomics**—analyzing DNA sequence patterns of humans and well-studied model organisms side-by-side for identifying human genes and interpreting their function.

DNA microarray: a technology for transcriptome analysis

- DNA hybridization microarrays supply information about gene expression through measurements of mRNA levels of large amounts of genes in a cell
- They offer a snapshot of the overall functional status of a cell: virtually all differences in cell type or state are related with changes in the mRNA levels of many genes.
- DNA microarrays have been used in mutational analyses, genetic mapping studies and in genome monitoring of gene expression

Applications of microarrays

- Molecular diagnosis of polygenic diseases
- Molecular characterization of tumors on a genomic scale: diagnosis and effective treatment of cancer.
- Prognostic tools for clinical use to predict the outcome or treatment response
- Pharmacogenomics: Identification of molecular targets for drugs
- Analysis of gene expression response to external stimuli (drugs, environment, hormones)
- Analysis of metabolic pathways
- Non-expression uses: assessing presence/absence of sequences in the genome, SNP and mutations analysis.

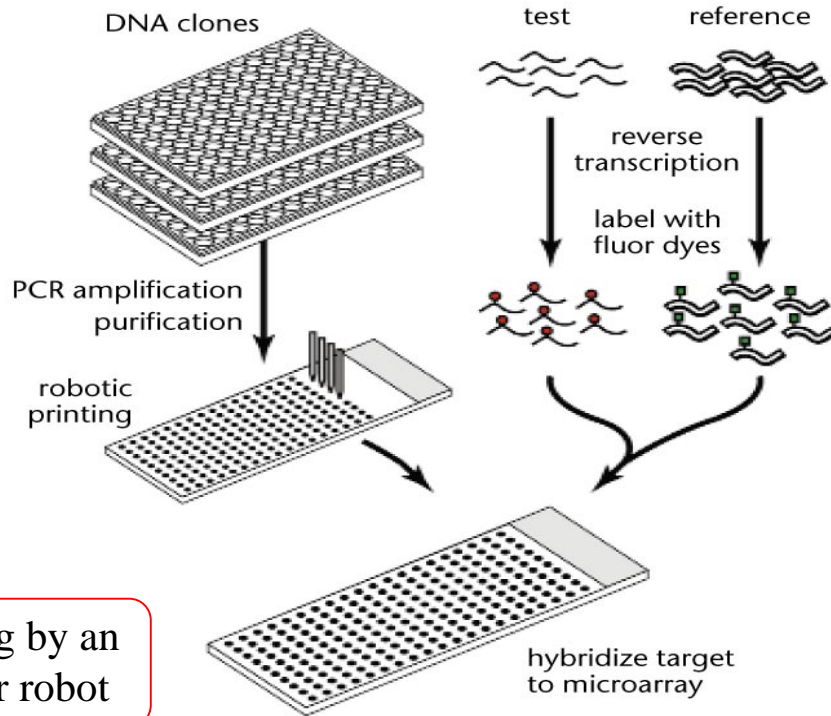
Gene expression assays

- Serial analysis of gene expression (SAGE);
- Short oligonucleotide arrays (Affymetrix);
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- **cDNA microarrays**

cDNA microarray hybridization experiments

Selection of DNA probes (cDNA clones from cDNA libraries)

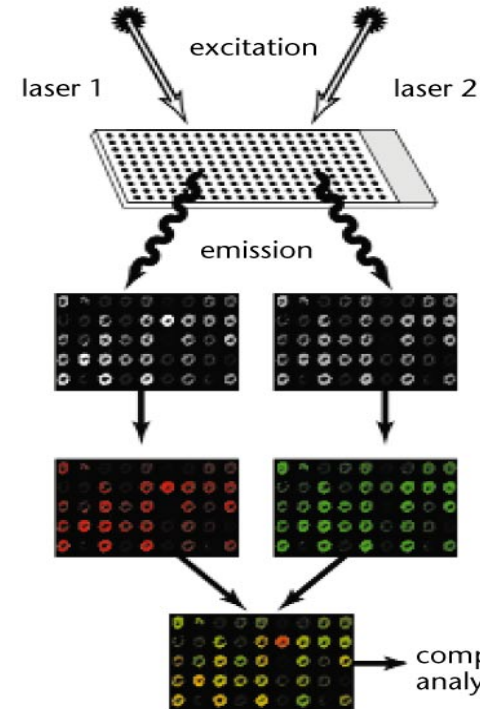
Preparation of mRNA and cDNA synthesis by reverse transcription



Printing by an arrayer robot

Hybridization of the cDNA sequences with the DNA samples of the microarray

Cy5: ~650 nm Cy3: ~550 nm

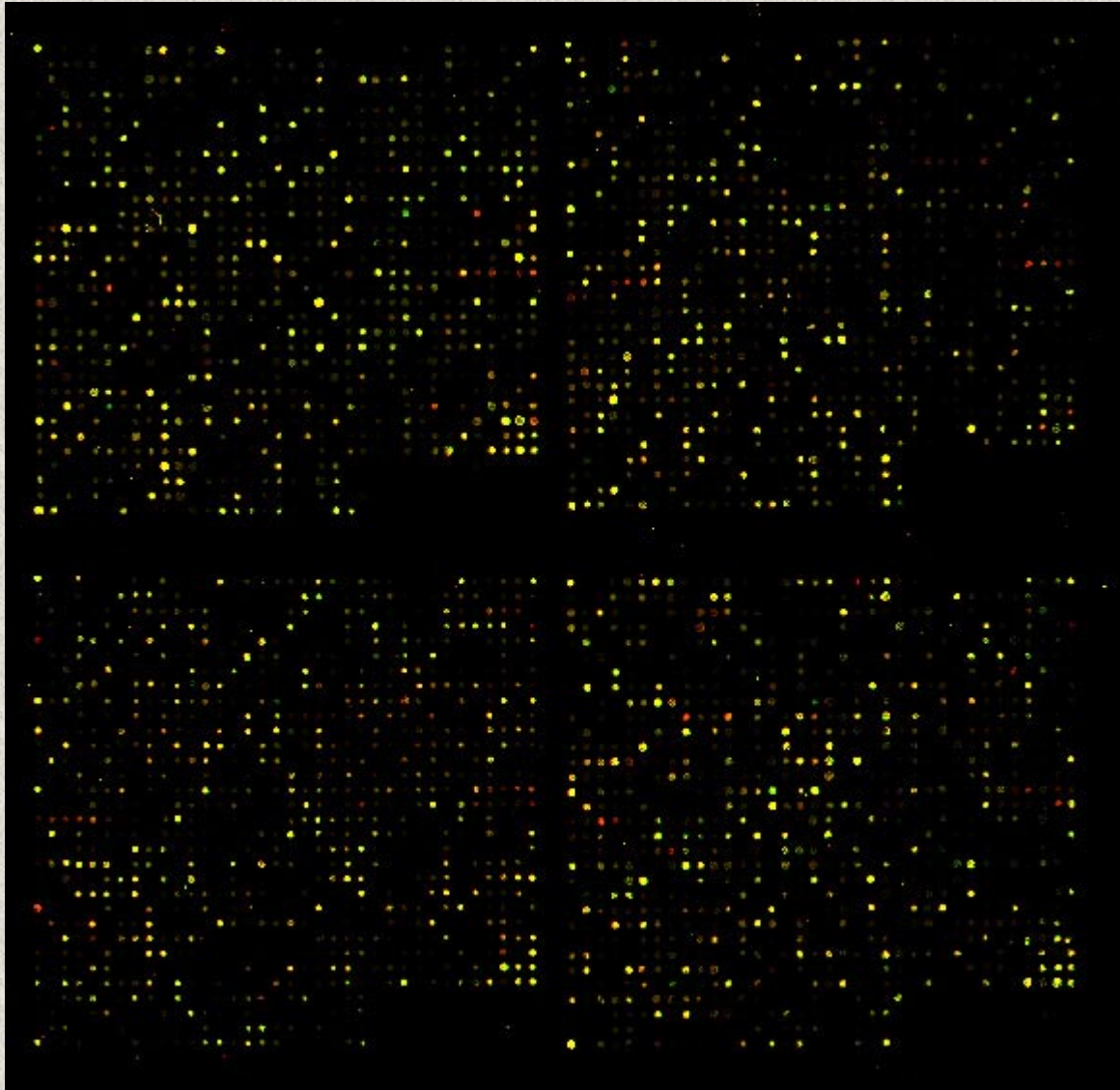


Scanning the slide to produce a raster image of the array

Fluorescent intensities → mRNA levels

Data preprocessing and data analysis

A DNA microarray image (E. coli)



- Each spot corresponds to the expression level of a particular gene
- Red spots correspond to over expressed genes
- Green spots to under expressed genes
- Yellow spots correspond to intermediate levels of gene expression

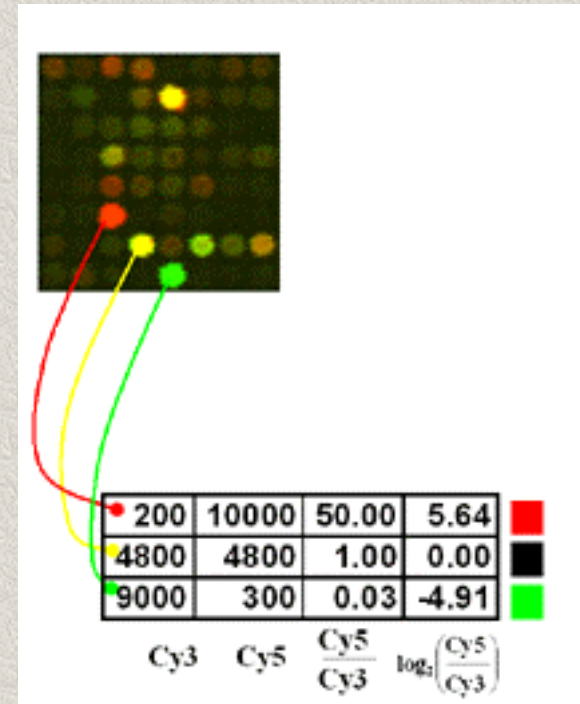
Preprocessing of cDNA microarray data

1. We do not know the exact number of clones in each DNA spot.
2. Length of the cDNA spotted sequences are not equal



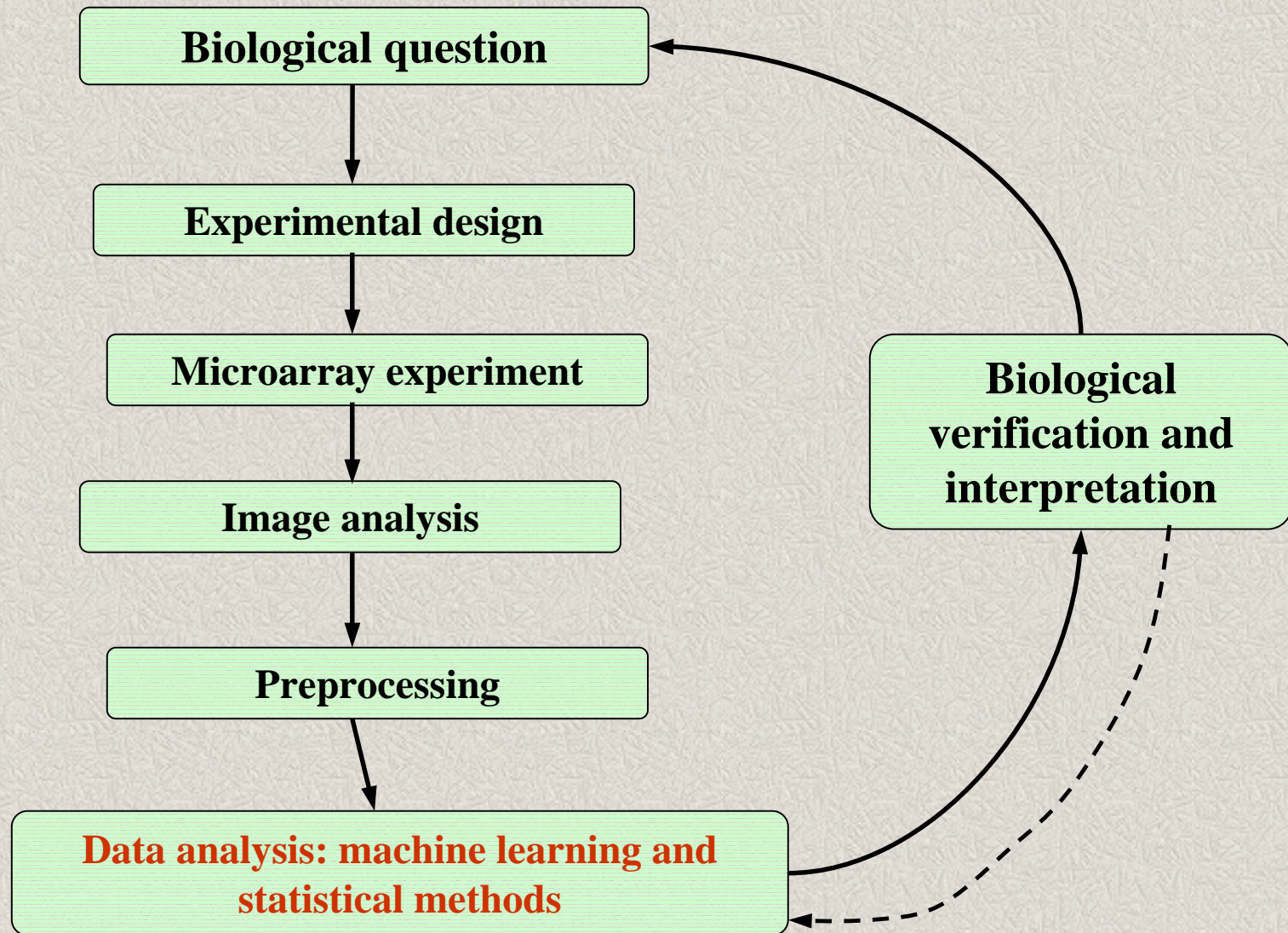
Both influence the amount of hybridization

Solution: Comparing the hybridization level in each spot with the level of hybridization under “control conditions”.



Many other problems related to preprocessing: see e.g.
http://globin.cse.psu.edu/courses/spring2002/3_Norm_miss.pdf

The cycle of gene expression experiments



Levels of analysis of DNA microarray data

0. **Image analysis**: Analysis of intensities levels of the fluorescent dyes
1. **Single genes analysis**: each gene in isolation behaves differently in a treatment vs. a control situation?
2. **Multiple genes**: analysis of interactions, common functionalities, co-regulations
3. **Pathway analysis** and exploration: analysis of the relationships between networks of interacting molecules

Analyzing microarray data by machine learning methods

The large amount of gene expression data requires machine learning methods to analyze and extract significant knowledge from DNA microarray data

Unsupervised approach

- No or limited a priori knowledge.
- Clustering algorithms are used to group together similar expression patterns :
 - grouping sets of genes
 - grouping different cells or different functional status of the cell.
- Example: hierarchical clustering, fuzzy or possibilistic clustering, self-organizing maps.

Supervised approach

- “A priori” biological and medical knowledge on the problem domain.
- Learning algorithms with labeled examples are used to associate gene expression data with classes:
 - separating normal from cancerous tissues
 - classifying different classes of cells on functional basis
 - Prediction of the functional class of unknown genes.
- Example: multi-layer perceptrons, support vector machines, decision trees.

Unsupervised approaches to gene expression data analysis

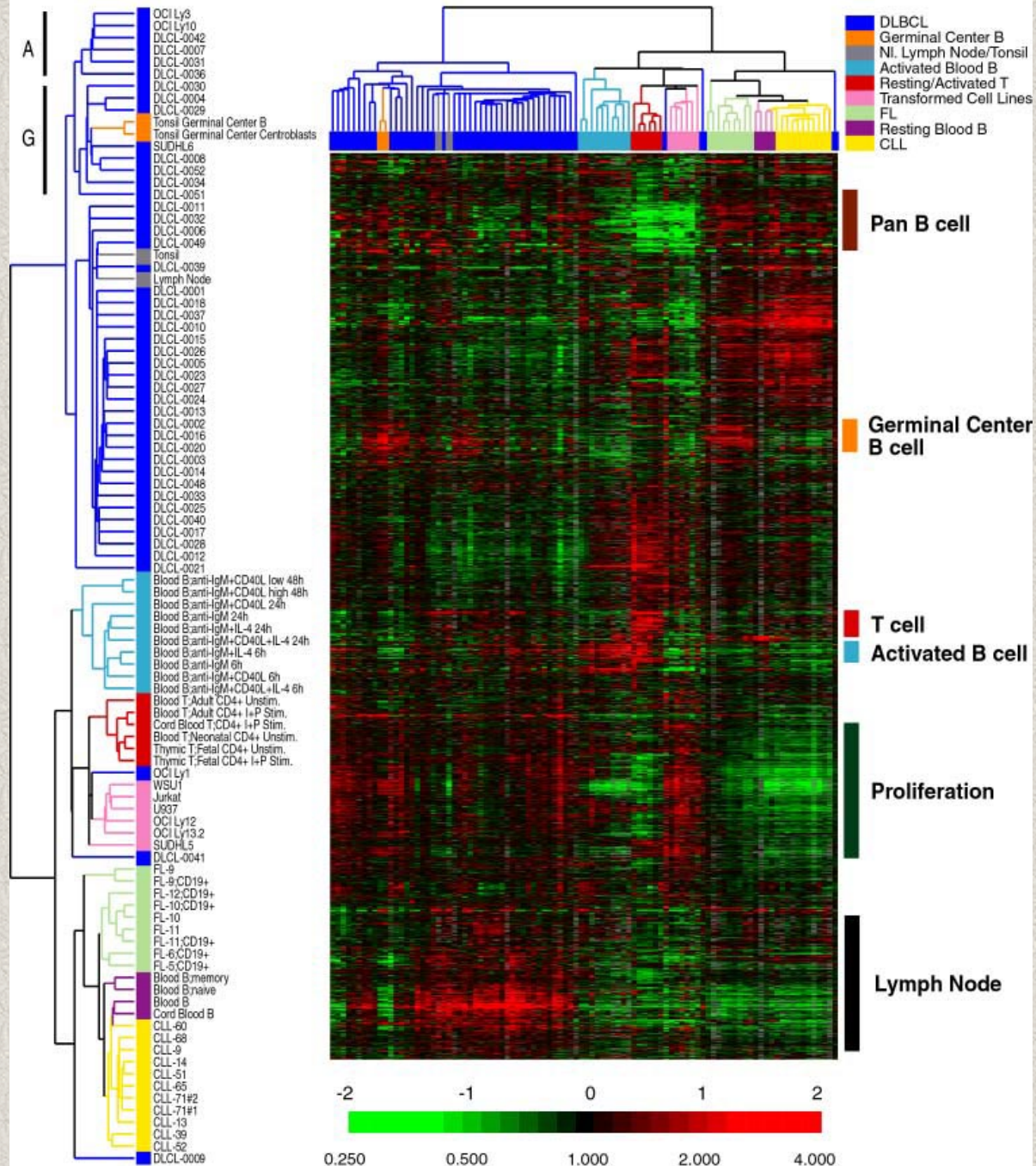
Biological problems

- Functional class discovery (e.g. discovery new diseases on molecular basis)
- Gene expression signature discovery
- Gene subset selection
- Exploratory analysis

Unsupervised methods

- Hierarchical clustering (Eisen & Brown, 1999)
- Self Organizing Maps (Tamayo et al., 1999)
- K-means (Tavazoie et al, 1999)
- Graph-based algorithms (Sharan and Shamir, 2000)
- Biclustering (Tanay et al, 2002)
- ...

Visualizing data with Tree View



Supervised approaches to gene expression data analysis

Biological problems

- Prediction of the functional status of tissues
- Prediction of the functional class of genes
- Diagnosis of tumours of molecular basis
- Prognostic tools to predict the outcome or treatment response
- Identification of molecular targets for drugs

Supervised methods

- Decision trees
- Fisher linear discriminant
- Multi-Layer Perceptrons
- Nearest-Neighbours classifiers
- Linear discriminant analysis
- Parzen windows
- Support Vector Machines

Proposed by different authors:

Golub et al. (1999), Pavlidis et al. (2001), Khan et al. (2001), Furey et al. (2000), Ramaswamy et al. (2001), Yeang et al. (2001), Dudoit et al. (2002).

Ensemble methods for gene expression data analysis

Why using ensemble methods for gene expression data analysis ?

- General motivations
- Domain specific motivations

Why should we use ensembles?

- From *empirical studies* : ensembles are often much more accurate than individual learning machines (Freund & Schapire (1995), Bauer & Kohavi (1999), Dietterich (2000), ...)
- Different *theoretical explanations* proposed to justify their effectiveness (Kittler (1998), Schapire et al. (1998), Kleinberg(2000), Allwein et al. (2000)).
- Very fast development of *computer technology*: availability of very fast computers and networks of workstations at a relatively low cost.

Reasons for combining multiple learners

- *Statistical*: data are limited
- *Representational*: learning algorithms cannot always represent all the functions
- *Algorithmical*: optimization techniques are not always optimal
- *Theoretical*: for instance, bias-variance reduction

Ensembles of learning machines for gene expression data analysis

Gene expression data are characterized by low cardinality:



Ensembles of learning machines can reduce bias and/or variance due to the low cardinality of the available training data.

Gene expression data are characterized by high dimensionality:



Dimensionality reduction ensemble methods (e.g. random subspace) or ensembles combined with feature selection

Different works showed that ensemble methods can be successfully applied to DNA microarray data analysis (Dudoit et al., 2000; Yeang et al., 2001; Ramaswamy et al., 2001; Valentini, 2001; Su et al. 2002).

Ensembles of learning machines for gene expression data analysis: examples

- **Output Coding methods** (*Dietterich and Bakiri, 1995*) for multiple-class cancer diagnosis.
- **Resampling methods** for gene expression based prediction of functional classes:
 - Bagging (*Breiman, 1996*)
 - Boosting (*Freund and Schapire, 1996*)
 - Cross validated committees (*Parmanto et al., 1996*)
 - Random forests (*Breiman, 2001*)

Output Coding methods for multiple-class cancer diagnosis.

1. OVA-WTA and AP ensemble approach (*Yeang et al, 2001; Ramaswamy et al, 2001*)
2. ECOC ensemble approach (*Valentini, 2002*)

Multiclass classification is hard in this context:

- **large dimensionality** of the datasets
- **small number** of examples
- small but significant **uncertainty** in the original labelings
- **noise** in the experimental and measurement processes
- intrinsic **biological variation** from specimen to specimen

Multiclass cancer diagnosis of 14 common tumor types using OVA and AP decomposition methods (Ramaswamy et al, 2001)

- Data set of about 200 tumor specimens spanning 14 different tumor classes obtained from the NCI, Memorial Sloan-Kettering Cancer Center (NY), and 3 hospitals in Boston.
- All tumors are biopsy specimens from primary sites
- Hybridization of targets to oligonucleotide microarrays (Affymetrix) provides expression data for 16063 genes.
- The largest study about gene expression based prediction of multiple tumor types

OVA-WTA ensemble approach

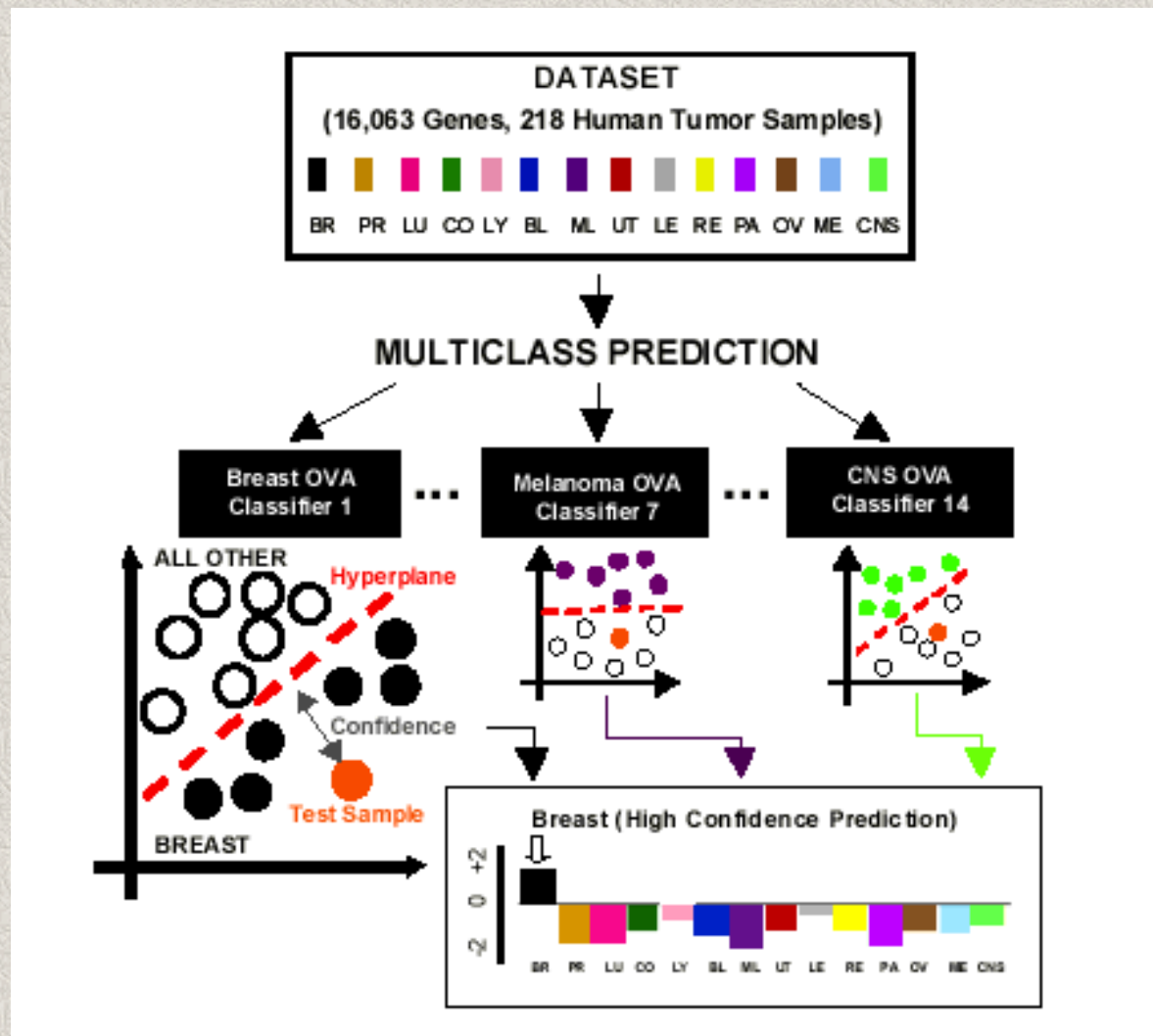
	codewords				
	c ₁	c ₂	c ₃	c ₄	c ₅
f ₁	+1	-1	-1	-1	-1
f ₂	-1	+1	-1	-1	-1
f ₃	-1	-1	+1	-1	-1
f ₄	-1	-1	-1	+1	-1
f ₅	-1	-1	-1	-1	+1

- Base classifiers: kNN, SVM
- Dichotomies: One class Versus All (OVA)
- Winner Takes All (WTA) decoding:

$$class = \arg \max_{i=1..k} \hat{f}_i$$

- “Simple” dichotomies to learn
- No error recovering capabilities

OVA-WTA multiclass prediction



All Pairs approach

All pairs decomposition

	codewords					
	c_1	c_2	c_3	c_4	c_5	
dichotomies	f_1	+1	-1	0	0	0
	f_2	+1	0	-1	0	0
	f_3	+1	0	0	-1	0
	f_4	+1	0	0	0	-1
	f_5	0	+1	-1	0	0
	f_6	0	+1	0	-1	0

• • • • •

- $k(k-1)/2$ dichotomies

- Dichotomies: class i versus class j : f_{ij}

- Vote decoding:

$$class = \arg \max_{i=1..k} \sum_{j=1}^k \hat{f}_{ij}$$

Limits:

- Binary classifiers trained with fewer examples

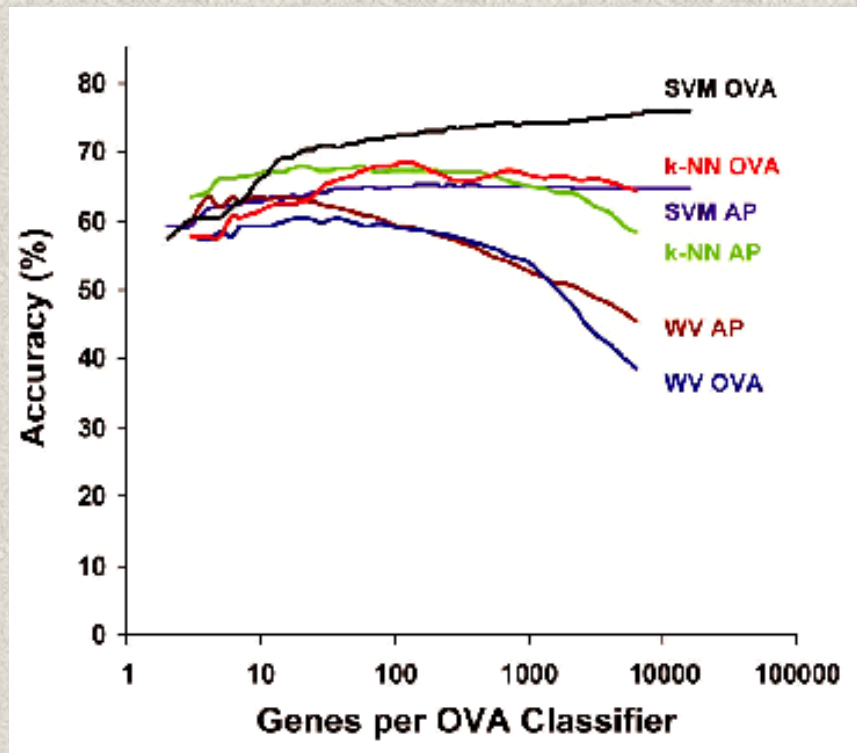
- Added noise: only k classifiers are relevant for a given sample

A better alternative:

Correcting Classifiers (Moreira and Mayoraz, 1998)

Results

Dataset	Method	Samples	Accuracy	Confidence			
				High		Low	
				Fraction	Accuracy	Fraction	Accuracy
Training	CV	144	78%	80%	90%	20%	28%
Test	Train / Test	54	78%	78%	83%	22%	58%



Genes selected through the RFE method (Guyon et al., 2002)

Gene expression based analysis of lymphoma using ECOC ensemble methods (Valentini, 2002)

Classification problems

1. Separating cancerous and normal tissues using the overall information available.
2. Identifying groups of genes specifically related to the expression of two different tumour phenotypes through *expression signatures*.
3. Classifying different types of lymphoma (a multiclass problem), using all the available gene expression data.

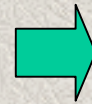
Methods

1. - Support Vector Machines (SVM) :
linear, RBF and polynomial kernels
 - Multi Layer Perceptron (MLP)
 - Linear Perceptron (LP)
2. Two step method:
 - A priori knowledge and unsupervised methods to select “candidate” subgroups
 - SVM or MLP identify the most correlated subgroups
3. - MLP
 - ECOC-MLP and OPC-MLP ensembles
 - ECOC-LP and OPC-LP ensembles

The data

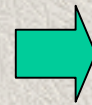
- Data of a specialized DNA microarray, named "Lymphochip", developed at the Stanford University School of Medicine:

4026 different genes
preferentially expressed in
lymphoid cells or with known
roles in processes important in
immunology or cancer



High dimensional data

96 tissue samples from normal
and cancerous populations of
human lymphocytes



Small sample size



A challenging machine learning problem

Types of lymphoma

Three main classes of lymphoma:

- *Diffuse Large B-Cell Lymphoma (DLBCL)*,
- *Follicular Lymphoma (FL)*
- *Chronic Lymphocytic Leukemia (CLL)*
- *Transformed Cell Lines (TCL)*

and normal lymphoid tissues

Type of tissue	Number of samples
<i>Normal lymphoid cells</i>	24
<i>DLBCL</i>	46
<i>FL</i>	9
<i>CLL</i>	11
<i>TCL</i>	6

Application of OC ensembles to the classification of lymphoma.

1. *Parallel Linear Dichotomizer ensembles (PLD)*
2. *One-Per-Class Parallel Non linear Dichotomizers (OPC-PND) ensembles*
3. *Error-Correcting-Output-Coding Parallel Non linear Dichotomizers (ECOC-PND) ensembles*

ECOC-PND, OPC-PND and PLD are OPC and ECOC ensembles of MLPs or LPs, where each LP or MLP is independently trained to learn a different bit of the codeword coding the classes.

Classifying different types of lymphoma with ECOOC methods

codewords

	c ₁	c ₂	c ₃	c ₄	c ₅
f ₁	-1	+1	+1	+1	-1
f ₂	-1	+1	+1	+1	+1
f ₃	+1	+1	+1	-1	-1
f ₄	-1	+1	+1	-1	+1
f ₅	-1	+1	-1	+1	-1
f ₆	-1	+1	-1	+1	+1
f ₇	+1	+1	-1	-1	-1
f ₈	+1	+1	-1	-1	+1
f ₉	-1	-1	+1	+1	-1
f ₁₀	+1	-1	+1	+1	+1
f ₁₁	+1	-1	+1	-1	-1
f ₁₂	+1	-1	+1	-1	+1
f ₁₃	+1	-1	-1	+1	-1
f ₁₄	-1	-1	-1	+1	+1
f ₁₅	+1	-1	-1	-1	-1

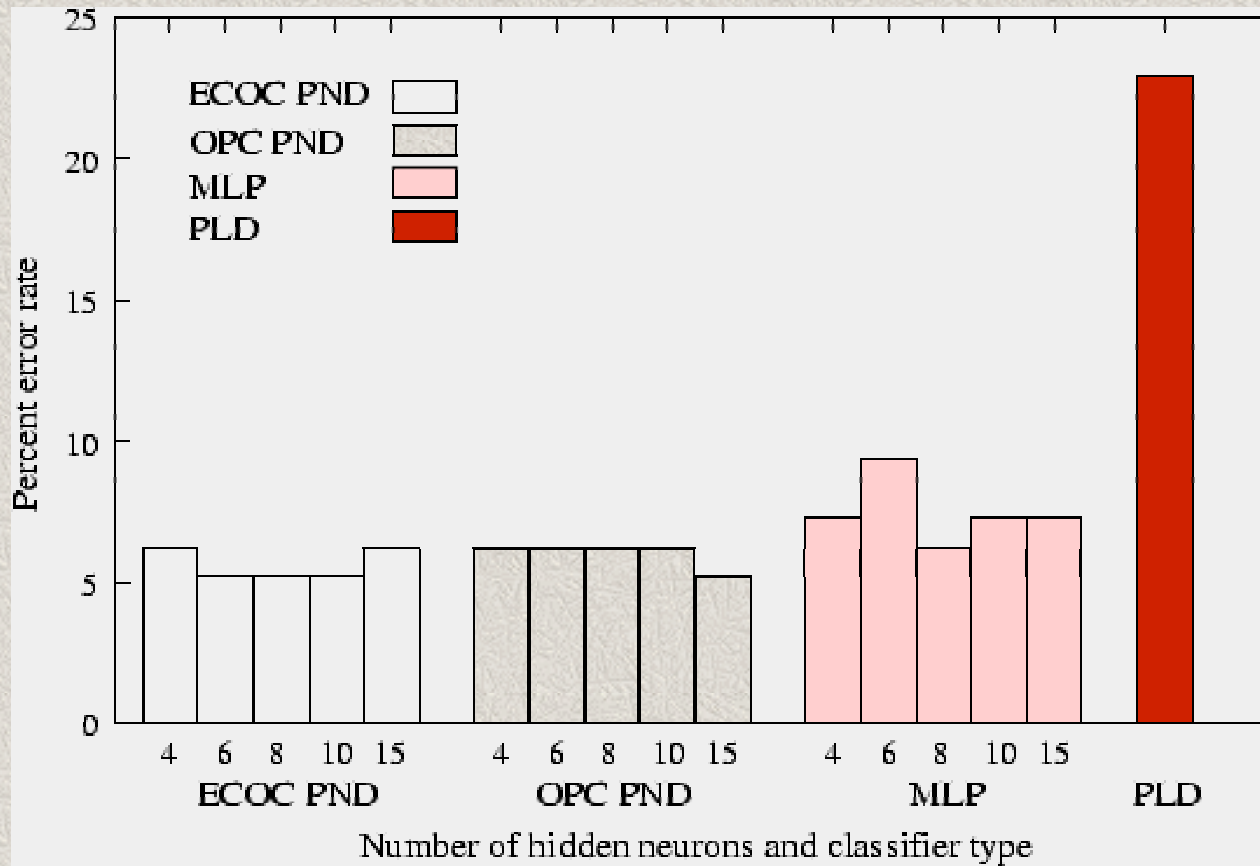
dichotomies

- Base learners: MLPs and LPs
- 15 bit ECOOC generated by exhaustive algorithms
- L₁ norm distance decoding:

$$\arg \min_{i \in C} L_1(y, D_i) = \arg \min_{i \in C} \sum_{j=1}^{15} |y_j - D_{ij}|$$

- Error recovering capabilities: at least 3 errors admissible in this case
- More complex dichotomies (in general) with respect to OPC decomposition
- ECOOC reduces both bias and variance

Results (1)



- ECOC and OPC PND achieve the best results
- PLD fail on this task
- MLP performs slightly worse than PND

- OPC and ECOC PND ensembles less sensitive to model parameters with respect to MLPs

Results (2)

Confusion matrix for the classification of different types of lymphoma.

		<i>Expected</i>				
		DLBCL	CLL	normal	FL	TCL
<i>Pre- dicted</i>	DLBCL	44	0	3	0	0
	CLL	0	11	0	0	0
	normal	0	0	21	0	0
	FL	1	0	0	9	0
	TCL	1	0	0	0	6

- Errors are due to:
 - false positives DLBCL
 - false positives TCL and FL
- Errors are the same in:
 - OPC ensembles
 - ECOC ensembles
- Similar genetic programs between GCB-like and normal lymphoid cells?

OC methods for multiple type cancer classification: results and perspectives

- OC decomposition methods for multiple type cancer classification using gene expression data obtained encouraging results:
 - Multiclass cancer classification of 14 common tumor types using OVA-SVM decomposition methods achieved an estimated accuracy of about 78% (Ramaswamy et al, 2001)
 - Multiclass cancer classification of 4 different lymphoma types and normal cells using ECOC-MLP achieved an estimated accuracy of about 95% (Valentini, 2002)
- **Open problems:**
 - Using SVMs as base learners we can use WTA or Hamming decoding, but there are problems in using other more reliable similarity measure that can exploit the strength of the prediction (especially for ECOC). We need to experiment with “normalized” SVMs or with SVMs whose output estimates probabilities.
 - Is feature selection useful for multiclass cancer classification, or we may lose information ?
 - Is it useful in this context applying ensembles of ensembles? (e.g. in the style of Adaboost.OC)

Gene expression-based classification of normal and heterogeneous malignant tissues using bagged SVMs: some preliminary results

- Data set of 300 normal and tumor specimens spanning 14 different tumor classes obtained from the Whitehead Institute – Massachusetts Institute of Technology Center for Genome Research.
- Hybridization of targets with oligonucleotide microarrays (Affymetrix) provides expression data for 16063 genes.
- Preprocessing of raw data using standard thresholding, filtering and normalization methods for oligonucleotide microarray data.
- Stratified-random splitting of the data in a separated learning and test set (1:1).

SVMs for classification of normal and heterogeneous malignant tissues

SVMs using all the genes (16063) achieved a low accuracy (~30% error on the test set).

Problems due to:

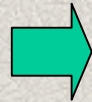
- heterogeneity of the data?
- too high dimension and low cardinality of the data?
- noise?



Feature selection methods can enhance accuracy ?

Gene selection

Selecting subsets of genes mostly related to carcinogenic processes



- *Genomic diagnosis* of tumors
- *Genomic therapy* of tumors
- Insights into *genetic networks* correlated to carcinogenic processes

From a machine learning standpoint, it is a *feature selection* problem

A filter approach to gene selection: Gene-specific neighborhood analysis

It is a method for gene selection applied before and independently of the induction algorithm (filter method).

It is a variant of the classic neighborhood analysis proposed by *Golub et al.* (1999)

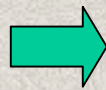
1. For each gene the S2N ratio c_i is calculated:
$$c_i = \frac{(m_i^+ - m_i^-)}{(\sigma_i^+ + \sigma_i^-)}$$
2. A gene-specific random permutation test is performed:
 - i. Generate n random permutations of the class labels computing each time the S2N ratio for each gene.
 - ii. Select a p significance level (e.g. $0 < p < 0.1$)
 - iii. If the randomized S2N c_{rand_i} is larger than the actual S2N c_i in less than $p * n$ random permutations, select the i^{th} gene as significant for tumor discrimination at p significance level.

Gene-specific neighborhood analysis

- It is a simple method $O(n \times d)$, n = number of examples, d = number of features (genes) to assess the correlation of genes with tumors.
- It estimates the significance of the matching of a given phenotype to a particular set of marker genes
- The permutation test is distribution independent: no assumptions about the functional form of the gene distribution.

Limits:

It assumes that the expression patterns of each gene are independent



It fails in detecting the role of coordinately expressed genes in carcinogenic processes

Gene-specific neighborhood analysis enhances SVM accuracy

1. Gene selection by *gene-specific neighborhood analysis*.

- Selected 592 genes correlated with tumoral examples ($p=0.01$) (set A)
- Selected > 3000 genes correlated with normal examples ($p=0.01$) (set B)
- Data set composed by set A and the 592 genes most correlated (with higher S2N ratio values) with normal examples in the set B
- As a result, the selected set of genes is composed by 1182 genes .

2. *Results*: SVMs with the selected subset of genes achieve a significant reduction of the prediction error with respect to the SVMs trained using all the available genes: about 12% of relative error reduction with linear SVMs, 27% with Gaussian SVMs, and even better with polynomial kernels.

Can bagged ensembles of SVMs enhances accuracy ?

Gene expression data are characterized by low cardinality:

Ensembles of learning machines can reduce variance due to the low cardinality of the available training data.

Gene expression data are characterized by high dimensionality:

SVMs can manage high dimensionality data and have “good” theoretical and practical properties


Bagged ensembles of SVMs enhance accuracy

- The low cardinality of the available data and the large degree of biological variability in gene expression suggested to apply variance-reduction methods (bagging) for this task.
- As in other works (*Dudoit et al.*, 2000) we can observe a slight reduction of the error bagging unstable base learners.
- Can we enhance the accuracy of the prediction exploiting the specific learning characteristics of SVMs ?

Bias-variance analysis based ensemble methods

Bias-variance decomposition of the error (*Domingos, 2000*) recently proposed as a tool to properly design ensemble methods well-tuned to the characteristics of a specific base learner (*Valentini, Dietterich 2002*).

Bias variance decomposition of the error as a tool to:



study the properties
of learning
algorithms

design ensemble
methods base
learner specific

An example of the application of bias-variance analysis of the error to SVMs: *bagged ensemble of selected low-biased SVM (Lobag)*

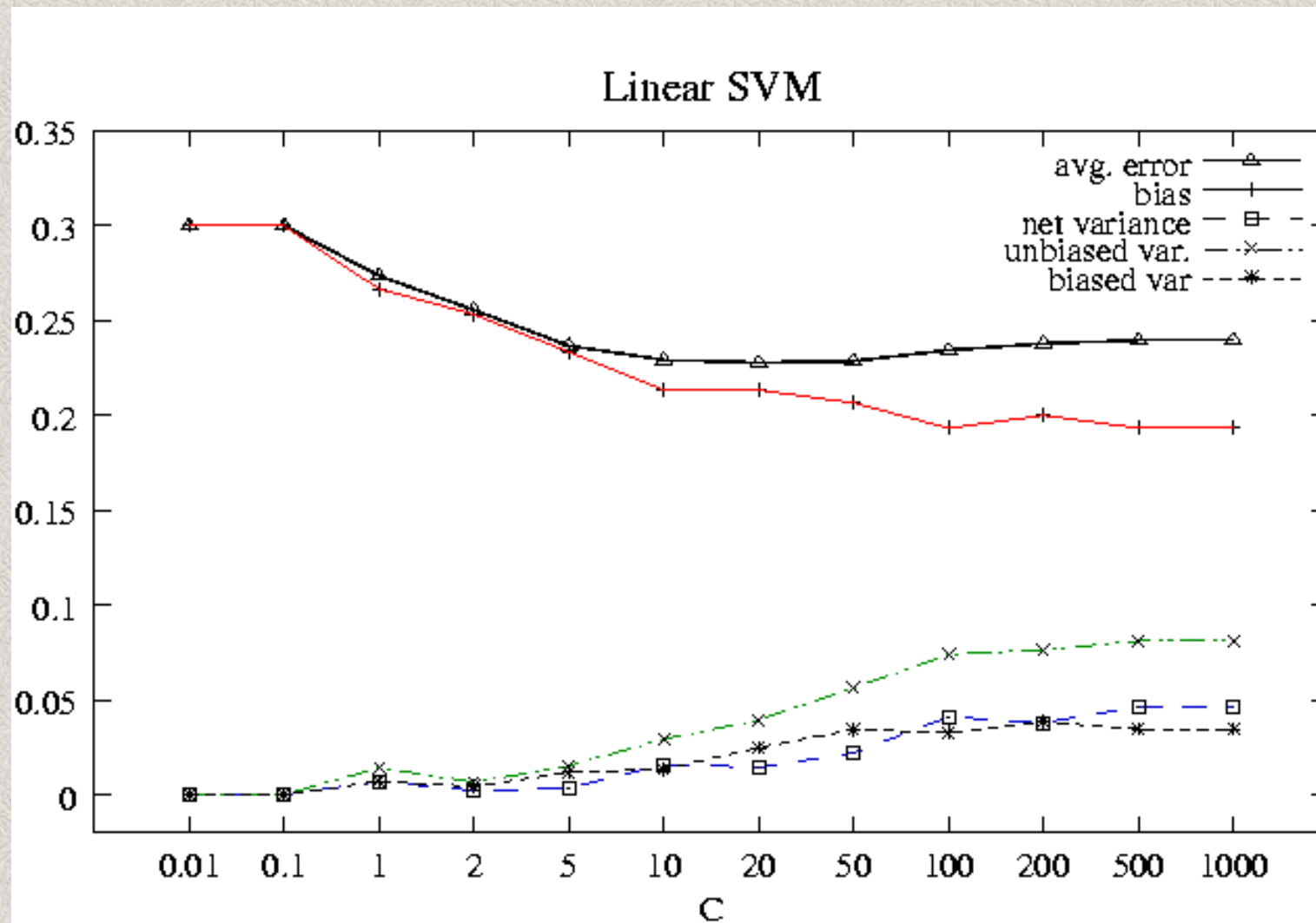
- We know that bagging lowers variance, but not bias.
- SVM are “strong” low-biased learners, but this property depends on the proper selection of the kernel and its parameters.
- If we can identify low-biased base learners with a relatively high unbiased variance, bagging can lower the error.
- Bias-variance analysis can identify SVM with low bias.



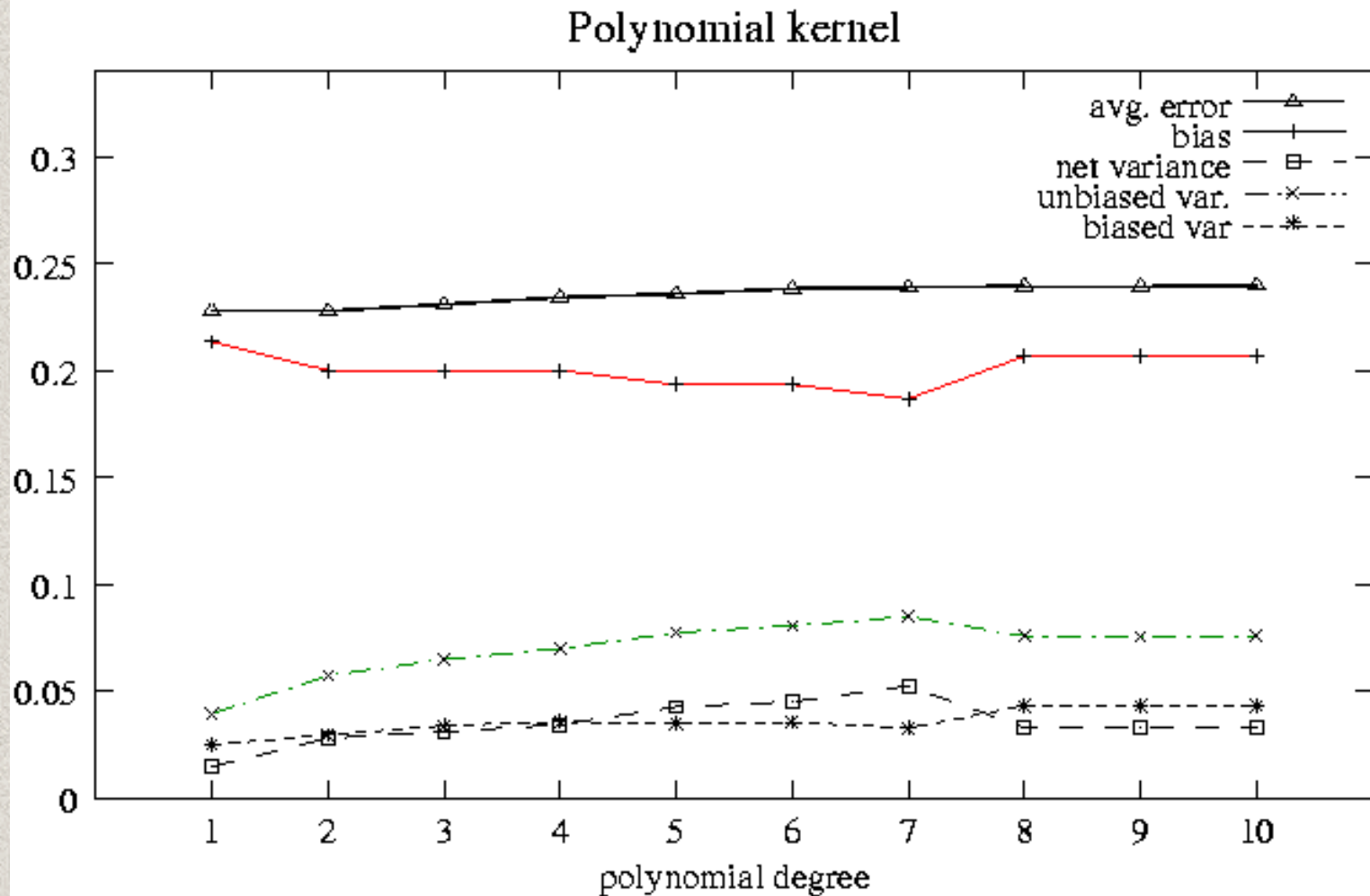
A basic high-level algorithm for a general *Bagged ensemble of selected low-biased SVM* could be:

1. Estimate bias-variance decomposition of the error for different SVM models
2. Select the SVM model with the lowest bias
3. Perform bagging using as base learner the SVM with the estimated lowest bias.

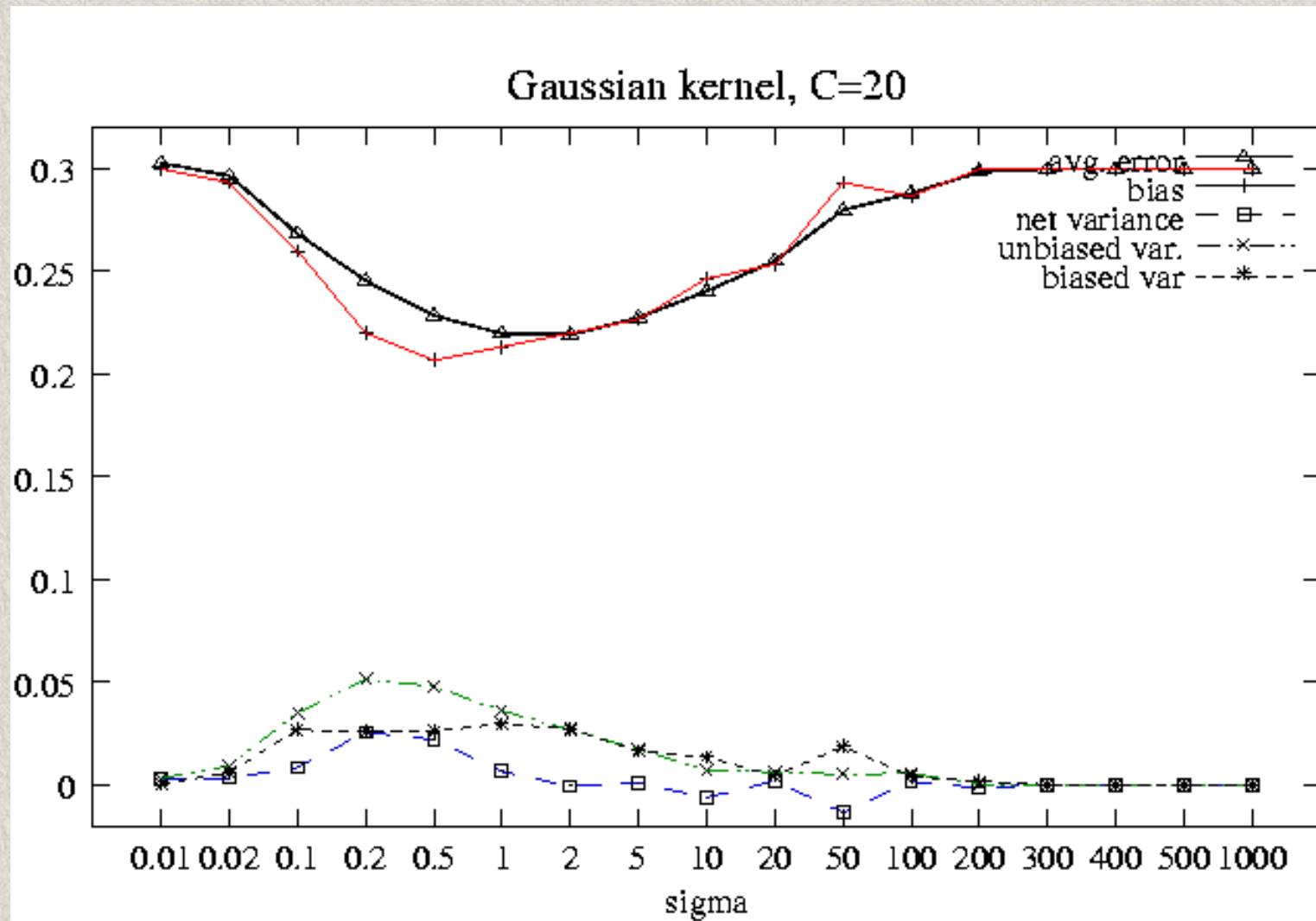
Bias-variance analysis in dot-product SVMs



Bias-variance analysis in polynomial SVMs



Bias-variance analysis in gaussian SVMs



Lobag can enhance further accuracy

Lobag ensembles of SVMs, such as bagged ensembles of SVMs work if the unbiased variance is relatively high w.r.t. the bias.

For this specific dichotomic classification problem Lobag enhances further accuracy w.r.t. both single SVMs and bagged SVM ensembles (preliminary results)

Open problems:

- Estimating bias-variance decomposition of the error is computationally expensive.
- Can we estimate bias-variance decomposition without an “exhaustive” search ?

Bagging and boosting for the classification of DNA microarray data (*Dudoit et al. 2002*)

5 data sets:

- 3 classes **Lymphoma**. Alizadeh et al. (2000) (cDNA arrays).
- **Leukemia**. Golub et al. (1999) (oligonucleotide arrays): 72 samples, 3571 genes, 3 classes (Bcell ALL, Tcell ALL, AML).
- **NCI 60**. Ross et al. (2000) (cDNA arrays): 64 samples, 5244 genes, 8 classes.
- **Brain cancer**. Pomeroy et al. (2002) (oligonucleotide arrays): 34 samples, 5893 genes, Classic vs. desmoplastic medulloblastoma
- **Breast cancer**. West et al. (2001) (oligonucleotide arrays): 49 samples, 7129 genes, ER + vs. ER -.

Methods:

- Bagging
 - Parametric bagging
 - Convex pseudo-data (Breiman, 1996)
 - Adaboost (Freund and Schapire, 1997)
 - LogitBoost (Friedman et al., 2000)
- (Dettling and Buhlmann, 2002 also applied an OVA-LogitBoost to gene expression data).

Main results:

- Aggregation improves performance of unstable classifiers
- Gene selection sometimes improves accuracy
- Simple classifiers can outperform more complex ones

Random forests (*Breiman, 2001*) for gene expression data analysis

Gene expression data characteristics:

Small and noisy sample size

High dimensional data

Random resampling of the learning data (e.g. Bagging)

Random resampling of the features (e.g. “Random subspace” method, Ho, 1998)

Random forests combine both

This general approach can be extended to other base learners

Thank you