Feature Vector Denoising with Prior Network Structures

(with Y. Fan, L. Raphael)
NESS 2015, University of Connecticut

Summary:

I. General idea: denoising functions on Euclidean space ---> denoising in index/gene space

II. Applications in computational biology: cancer classification

III. Extensions: pathway methods

IV. Extensions: arbitrary structure on the index space - conceptual structures

1. Motivation:

We live in a time of massive feature vectors for classification.

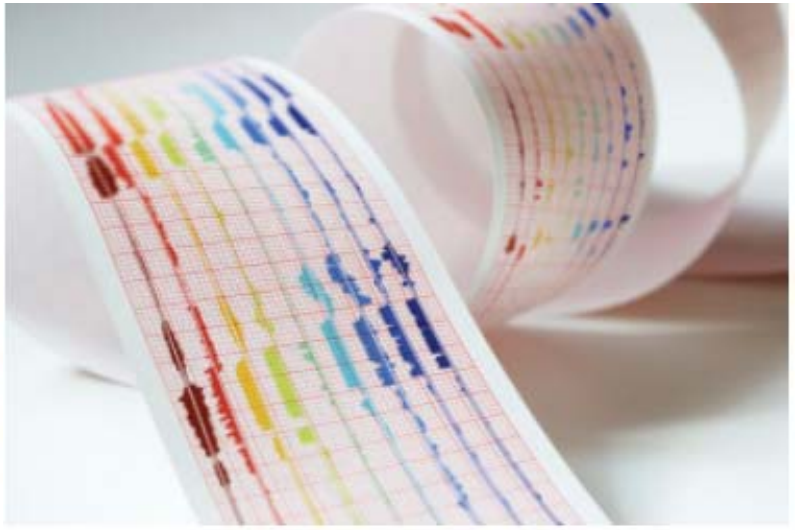
In computational biology, these include (per tissue sample):

Features	Cardinality
Gene expression array values	20K
Single Nucleotide Polymorphism (Illumina)	500K
Methylation and Phosphorylation data	200K
Gene copy number data (Agilent)	250K
TOTAL:	~1,000K

Total numbers of DNA biomarkers available are approaching 3 x 10^9, since each DNA base will be a biomarker once full genome sequencing is common.

Example: Gene expression arrays and inference

An RNA-seq gene expression array produces approximately 20k gene-level biomarkers describing a tissue sample:



http://www.polyomics.gla.ac.uk/event-rnaseq2014.html; Broad Institute

Result: for each subject tissue sample s, obtain feature vector:

$$\Phi(s) = \mathbf{x} = (x_1, \ldots, x_{20,000})$$

$=$ feature vector of gene expression levels

Can we classify tissues this way?

If this is an ovarian cancer tissue sample:

Questions:

(a) What type of cancer is it?

(b) What is prognosis if untreated?

(c) What will be the reaction to standard chemotherapies?

Goals:

1. Differentiate two different but similar cancers.

2. Determine the future course of a cancer

3. Determine what chemical agents the cancer will respond to

4. Understand genetic origins and pathways of cancer

Basic difficulties: few samples (e.g., 30-200); high dimension (e.g., 10^4 - 10^6).

Curse of dimensionality - too few samples and too many parameters (dimensions) to fit them.

Primary Problem:

- Problems in machine learning (ML) often involve noisy input data $\mathbf{x} = (x_1, \ldots, x_p)$ (particularly in computational biology).

- ML classification methods have in some cases reached limiting accuracies on 'standard' ML datasets

An approach:

- An important step to greater accuracy in ML requires incorporation of prior structural information on data

- A potentially important regularization involves denoising of feature vectors *alone* using Tikhonov and related regularization methods usually used on functions $f : \mathbb{R}^p \longrightarrow \mathbb{R}$.

- These are denoted as *unsupervised regularization methods* -- they use Lagrangian optimization functionals like in supervised learning.

- This viewpoint takes feature vectors $\{x_q\}_{q=1}^n$ as functions on their indices q, and requires continuity with respect to graph or proximity structures on the q.

- Two standard function regularzation methods on $\mathbb{R}^p$, local averaging and kernel regression, are adapted here to unsupervised regularization.

- Result is improved feature vector recovery and thus subsequent improved classification/regression done with improved feature vectors.

- An example in gene expression analysis for cancer classification with the genome as index space for gene expression feature vectors.

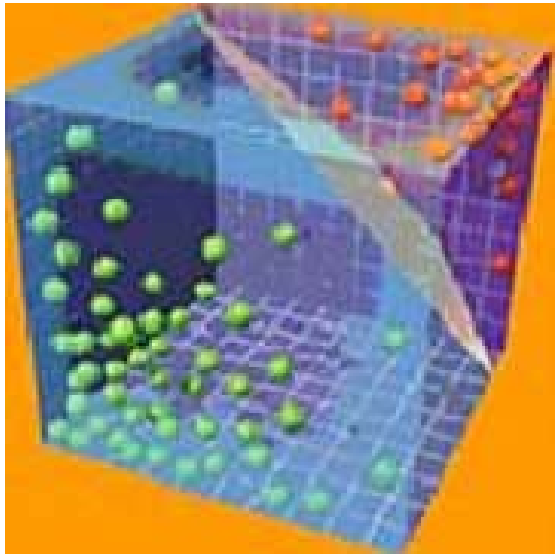
- Here noise in data is viewed as a source of complexity; feature vector regularization denoises by seeking less complex data forms.

2. SVM as a tool

Method: Support vector machine (SVM)

Procedure: look at feature space F in which $\Phi(s)$ lives, and differentiate examples of one and the other cancer with a hyperplane:

SVM as a tool



Train machine: take $n = 50$ subjects with different responses to therapy T, locate their feature vectors in F, labeling them as unresponsive or responsive.

SVM as a tool

Other machine learning methods can also discriminate feature vectors with respect to prognosis, response to therapies, etc.

Our data is obtained in collaboration with TCGA (the Cancer Genome Atlas).

3. The principle: more is more

Past: too many variables spoil the statistics; < 50 variables
was typical requirement

Present: more is better

Machine learning allows massive integration of relationship information:

On a gene level:

Machine learning vs. classical statistics

- protein-protein interactions
- coexpression
- gene ontology
- pathway connections

Machine learning allows seamless combination of many different data types using kernel matrices

Kernel trick: incorporate relational information into a kernel matrix: for genes g_i and g_j:

$$\mathbf{K}_{ij} = K(g_i, g_j) = \text{'closeness' of } g_i \text{ and } g_j$$

as measured by above relationship information.

Machine learning vs. classical statistics

- Each type of gene relation gives a different kernel matrix.

Machine learning vs. classical statistics

To integrate information in kernel matrices $\mathbf{K}^{(1)}$, $\mathbf{K}^{(2)}, \ldots, \mathbf{K}^{(n)}$, we form the sum,

$$\mathbf{K} = \mathbf{K}^{(1)} + \ldots + \mathbf{K}^{(n)}.$$

which incorporates all these measures into one.

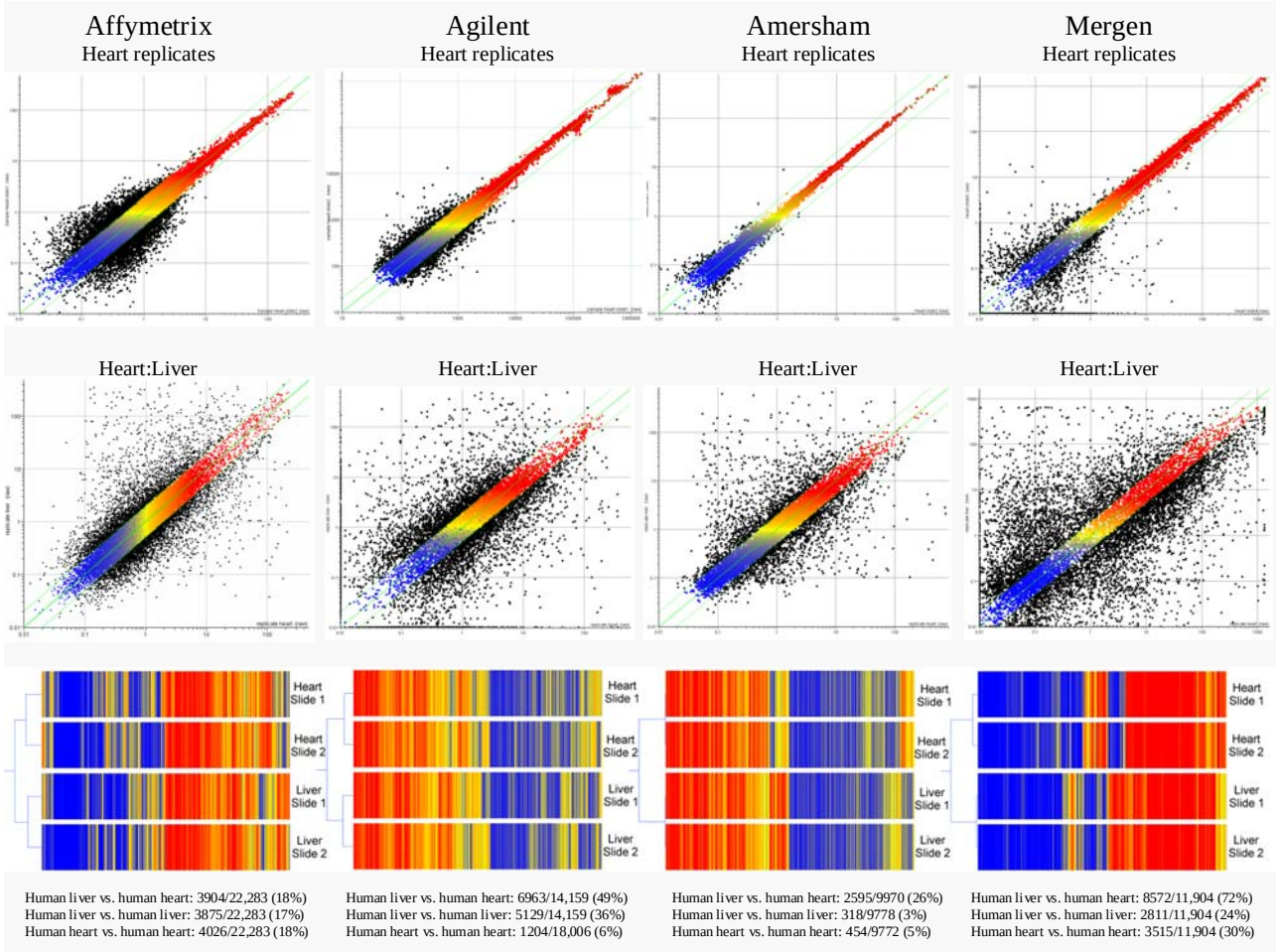
Information types (for example, in The Cancer Genome Atlas, TCGA) contain:

- Gene expression (microarray)
- Single nucleotide polymorphism (SNP) information
- Methylation, epigenetic information
- Gene copy numbers
- micro-RNA (miRNA) data

4. Problem: biomarkers are noisy!

Gene expression is non-self-replicating (microarray example):

Noisy biomarkers



How to clean up the noise? Use the same methods as denoising functions in Euclidean space.

Noisy biomarkers

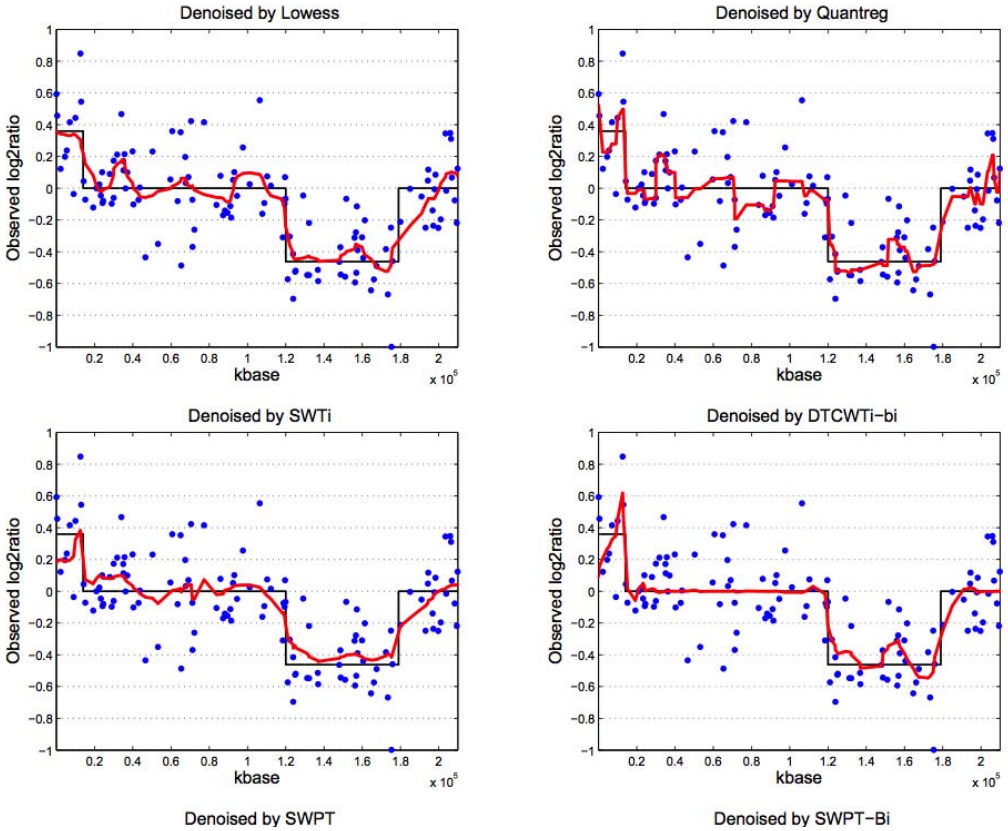


Figure: smoothing of gene copy number arrays using wavelet denoising.
Huang, et al. http://www.biomedcentral.com/content/pdf/1471-2164-9-S2-S17.pdf

What methods help in denoising functions on Euclidean spaces?

Noisy biomarkers

1. Local averaging (Haar wavelet denoising) - above

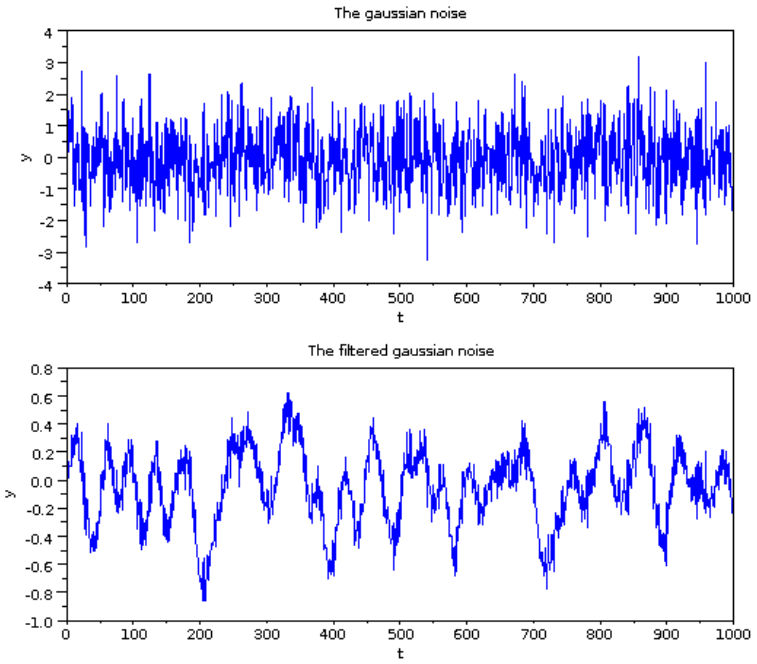
2. Smoothing using convolutions $f(x) \to f * g(x)$, where $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is say a Gaussian kernel.

3. More generally, smoothing using kernel regression:

$$f(x) \to \sum_i \alpha_i K(x, x_i) + b$$

4. Spectral smoothing - filtering high spectral components of a function:

Noisy biomarkers



http://www.scilab.org/product/man/DesignEllipticFilter.html

and many other modes.

Rapaport, Vert, et al. (2007) have used spectral methods for denoising gene expression arrays.

5. How to transfer Euclidean space methods to gene space?

One can use similar methods for denoising gene expression arrays, and more generally machine learning (ML) feature vectors.

Gene expression arrays: Given a gene expression feature vector **x** $= (x_1, x_2, \ldots, x_p)$, we can *view it as a function on its indices* $G = \{1, 2, \ldots, p\}$ or equivalently the genes $g_1, \ldots, g_p$.

Euclidean denoising on gene space

Purpose: if index set G has a distance measure (e.g. a metric or network structure), and thus a notion of when two points i, j in G are 'close', then we will try to use this metric structure similarly to Euclidean metric to eliminate noise.

In Euclidean space denoising of a function $f(\mathbf{x})$ is done using continuity, i.e.,

$|f(\mathbf{x}) - f(\mathbf{y})|$ small when $d(\mathbf{x}, \mathbf{y})$ is small.

Euclidean denoising on gene space

In ML denoising can be done when we expect

$$|f(i) - f(j)| \text{ small when } d(i,j) \text{ is small,}$$

where d is a distance measure on indices i, j (e.g. genes)

Genes in a network: if index i represents gene g_i and j represents gene g_j, and if nodes g_i and g_j are close in the gene network, we believe their expressions x_i and x_j should be close to each other.

Note this is an unsupervised method which can regularize feature vectors for any classifier (e.g., SVM, random forest, k-nearest neighbors, etc.)

6. More formally:

Given distance structure (e.g., metric or network) on the index set G (e.g. genes) of a basis for a feature space F, so that $\mathbf{x} \in F$ is

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

with index $G = \{1, 2, \ldots, p\}$.

View features $x_q = f(q) = f(g_q)$ as a function on the indices q (the *feature function*).

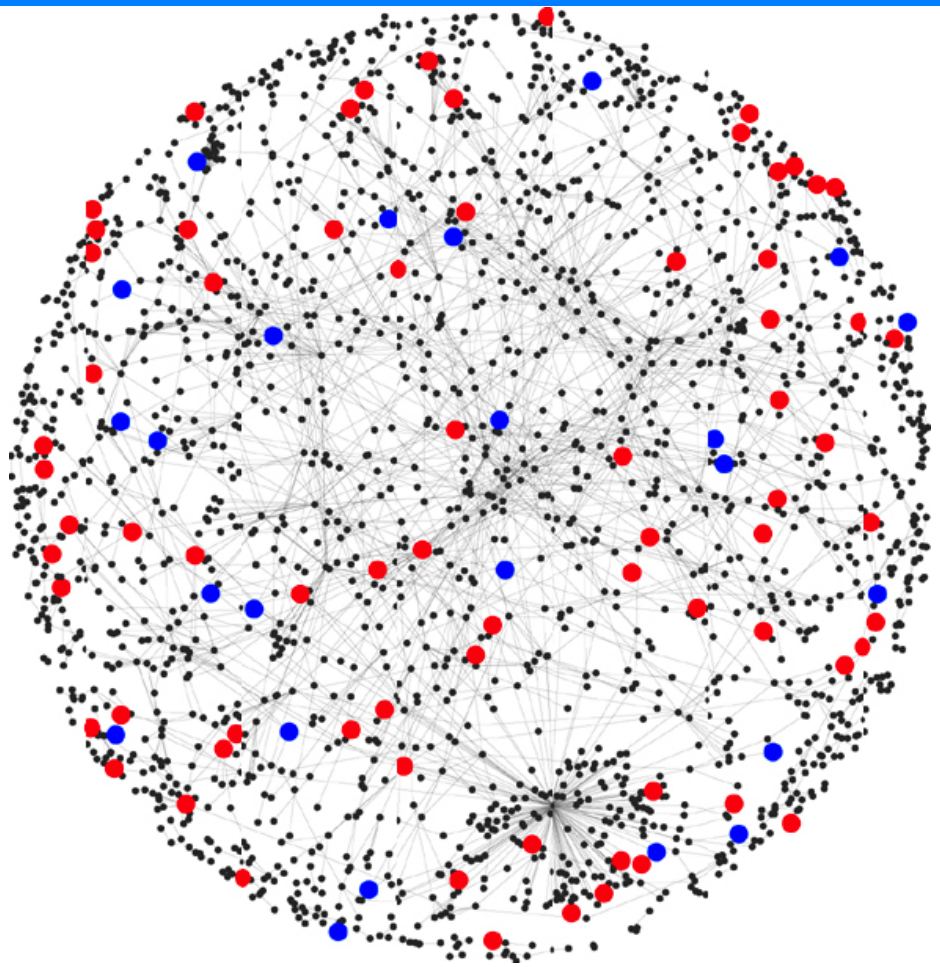
Formalities

Model features (e.g. gene expressions)

$$x_q = f(g_q) = f_1(g_q) + \epsilon(g_q),$$

where ϵ_i represents noise, and $f_1(g_q)$ is the 'true' expression signal.

Formalities



Base space G for feature vector $\mathbf{x} = f(g)$ (gene network)

Lu, et al. http://www.nature.com/msb/journal/v3/n1/full/msb4100138.html

Formalities

Now consider a smoothing transformation T on $f(g)$ to smooth out noise:

Mapping $f(g) \rightarrow T(f(g))$ gives

$$T(f(g)) = T(f_1(g)) + T(\epsilon(g)).$$

Transformation $T(f_1(g))$ will differ from the true expression $f_1(g)$, so we have introduced bias

If k $=$ regularization parameter (cluster size)

Formalities

loss of signal through bias increase:

(*) $f_1 - T(f_1)$ (increases with k)

However, the smoothing $T(\epsilon(g))$ of noise ϵ will quench it: averaging over k genes will reduce $\epsilon \longrightarrow \frac{1}{\sqrt{k}}\epsilon$

Gain in signal through variance decrease

(**) $\epsilon - \frac{1}{\sqrt{k}}\epsilon$ (decreases with k)

For some value of regularization parameter k the bias loss in (*) is balanced by the variance gain in (**).

Formalities

This is the usual bias-variance dilemma - when do we quench so much noise (**) that the increase (*) in bias is overcome?

Principle: local averaging eliminates noise.

7. Some theorems -

Theorem. (a) Let F be a space of feature vectors with basis $\{b_q\}_{q \in G}$ whose indices q form a graph structure. Let $f_1(q) = f(q) + \eta(q)$ be a noisy feature vector with η independent Gaussian noise. Let $\{\mathcal{F}_t : 0 \leq t \leq T\}$ be a filter (an family of increasingly refined partititions of G based on graph clustering.) Then the regularization $f_{1t}(q)$ of f_1 obtained by averaging the noisy function $f_1(q)$ over the clusters in $\mathcal{F}_t$ has an error that decreases and then increases, so there is a $t > 0$ for which the averaging regularization is optimal.

(b) This holds also if the above averaging regularization is replaced by support vector

regression using a Gaussian graph kernel, i.e., local kernel averaging helps regularize feature vectors.

8. Example: local averaging noise reduction

Using the protein-protein interaction (PPI) gene network as an example: consider differentiation of metastatic and non-metastatic breast cancer (Wang; van de Vijver).

Example: Local averaging

Wang data set:

93 metastatic
183 non-metastatic

van de Vijver data set:

79 metastatic
216 non-metastatic

How to predict metastasis?

Strategy - regularize the feature vectors before the classification begins.

Example: Local averaging

Regularizer for feature vectors: clustering using PPI network and then averaging over clusters

Classifier: SVM

Results:

Area under ROC curve improved by 5% to 20%

Example: Local averaging

No. Clusters	Wang			van de Vijver		
	AUROC	AUPRC	ACC90	AUROC	AUPRC	ACC90
64	0.658 (0.014)	0.450 (0.019)	0.470 (0.027)	0.687 (0.014)	0.371 (0.015)	0.472 (0.024)
128	0.680 (0.015)	0.462 (0.021)	0.477 (0.023)	0.705 (0.013)	0.399 (0.019)	0.520 (0.023)
256	0.692 (0.019)	0.475 (0.026)	0.526 (0.029)	0.689 (0.016)	0.398 (0.022)	0.490 (0.030)
512	0.684 (0.019)	0.487 (0.031)	0.502 (0.029)	0.686 (0.021)	0.375 (0.023)	0.489 (0.038)
1024	0.708 (0.019)	0.500 (0.032)	0.527 (0.029)	0.712 (0.019)	0.403 (0.026)	0.520 (0.026)
2048	0.730 (0.017)	0.522 (0.029)	0.567 (0.024)	0.500 (0.038)	0.270 (0.026)	0.311 (0.026)
RAW	0.534 (0.044)	0.362 (0.032)	0.430 (0.035)	0.660 (0.027)	0.346 (0.028)	0.535 (0.020)

Performance of local averaging of microarray data locally averaged in PPI network

9. Performance using support vector regression:

In Euclidean space: replace noisy gene expression function by a regularized one based on support vector regression (here $x = g$ represents a variable gene in base space gene network)

$$f(x) \rightarrow f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b$$

for selected points (centers) $\{x_i\}_i$, where $K(x, y)$ is a kernel which gives a metric between genes x and y.

Support vector regression

Example: in our gene space kernel

$$K(x, y) = \text{graph diffusion kernel}$$

(heat kernel on gene network graph).

Support vector regression

Regularized function $f(x) = f(g)$ is optimizer of objective fn.

$$f = \underset{f(g)}{\operatorname{argmin}} \sum_j L(f(g_j), z_j) + \lambda \| f \|_K^2,$$

where $g_j = j^{th}$ gene

$f(g) = f(x) =$ fn. on genes (regularized expressions),

$z_j =$ original measured expression on gene j

$L(f(g_j), z_j) = (|f(g_j) - z_j| - \epsilon)^+$

$=$ loss function (difference between measured and regularized gene expression)

$\lambda =$ regularization parameter

Support vector regression

$\|f\|_K$ = norm of f with respect to kernel K

Regularization done within clusters of genes, grouped by similar expressions in the training set

Support vector regression

No. Clusters	Wang			van de Vijver		
	AUROC	AUPRC	ACC90	AUROC	AUPRC	ACC90
1	0.618 (0.013)	0.405 (0.014)	0.456 (0.024)			
64	0.672 (0.018)	0.503 (0.025)	0.476 (0.033)	0.706 (0.017)	0.441 (0.025)	0.468 (0.032)
128	0.698 (0.018)	0.519 (0.026)	0.52 (0.032)	0.738 (0.017)	0.456 (0.024)	0.527 (0.035)
256	0.716 (0.017)	0.526 (0.024)	0.565 (0.030)	0.741 (0.017)	0.465 (0.025)	0.536 (0.033)
512	0.71 (0.016)	0.515 (0.023)	0.567 (0.027)	0.746 (0.015)	0.478 (0.026)	0.552 (0.030)
1024	0.701 (0.015)	0.494 (0.022)	0.558 (0.027)	0.74 (0.013)	0.48 (0.025)	0.536 (0.022)
2048	0.676 (0.019)	0.47 (0.026)	0.521 (0.031)	0.718 (0.015)	0.441 (0.026)	0.532 (0.018)
RAW	0.54 (0.040)	0.364 (0.030)	0.434 (0.035)	0.661 (0.023)	0.351 (0.026)	0.535 (0.020)

Support vector regression

Support vector regression performance (expression clustering followed by regression in each cluster)

There are other potential gene metrics based on gene networks derived gene ontology (GO), gene copy number information (in cancer), etc.

Summary:

- Regularization of classification functions $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^p$ is standardly done using *Tikhonov regularization functionals:*

$$\mathcal{L}(f) = \sum_i |f(\mathbf{x}_i) - y_i|^2 + \alpha ||f||^2.$$

- The success of such methods suggests that similar methods might be used in a different stage of the

machine learning process, the formation of the feature vector $\mathbf{x} = (x_1, \ldots, x_p)$ itself, which is now viewed as a *feature function* $x_q = f(q)$.

- The spatial structure of $\mathbb{R}^p$ in regularization is replaced by a structure on the index space G (the set of indices q, e.g. the set of genes).

- We denote this as *unsupervised regularization.*

- This regularization denoising of feature vectors is a pre-processing step *prior* to any supervised learning of the data.

Questions

Some further questions:

1. Are separations into the above biomarkers together with regularization the best way to structure the index sets of feature vectors?

a *hierarchical SVM* (feature vectors are based on a tree structure for feature indices)

- parses feature information like the brain;

- leads to both better predictability and better clinically
applicable biomarker sets e.g., in cancer analysis

Questions

2. What is the best way to translate detailed biomarker-based algorithms into clinical practice?

There is a lack of standard individual gene biomarkers in microarrays

Pathway-based biomarkers work toward the model of hierarchical SVM and have proved to be much more stable in classyfing cancers.

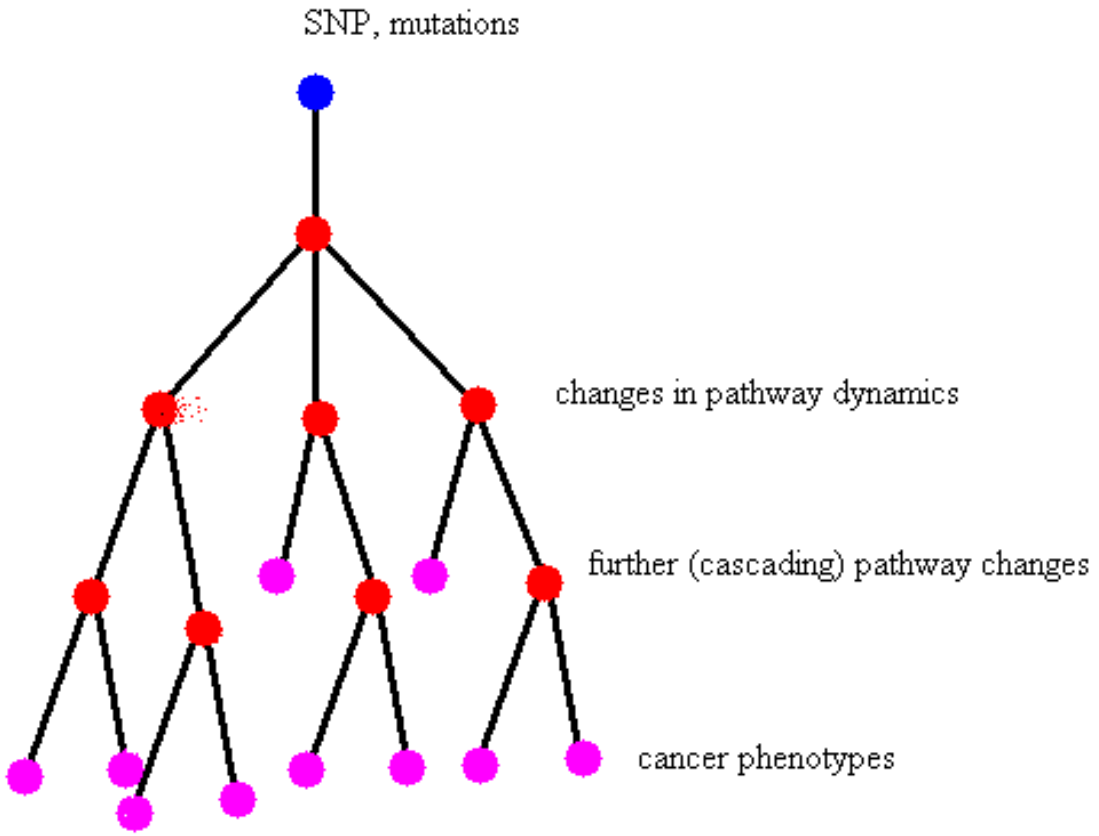
3. How should we integrate genomic (SNP, mutation, copy number) information with expression and epigenetic information to classify cancer prognoses and therapeutics?

Cancer causes a cascade of changes:

- starts with inherited SNPs/ mutations
- augmented by one or more somatic mutations
- result is a causal tree of subsequent cell changes
- SNP/mutation data provide information on the root of this tree

Questions

- expression information reveals changes in the branches and leaves



Questions

Challenge: to integrate information which propagates from the causes at top to effects below.

10. Better cancer signatures: gene copy number

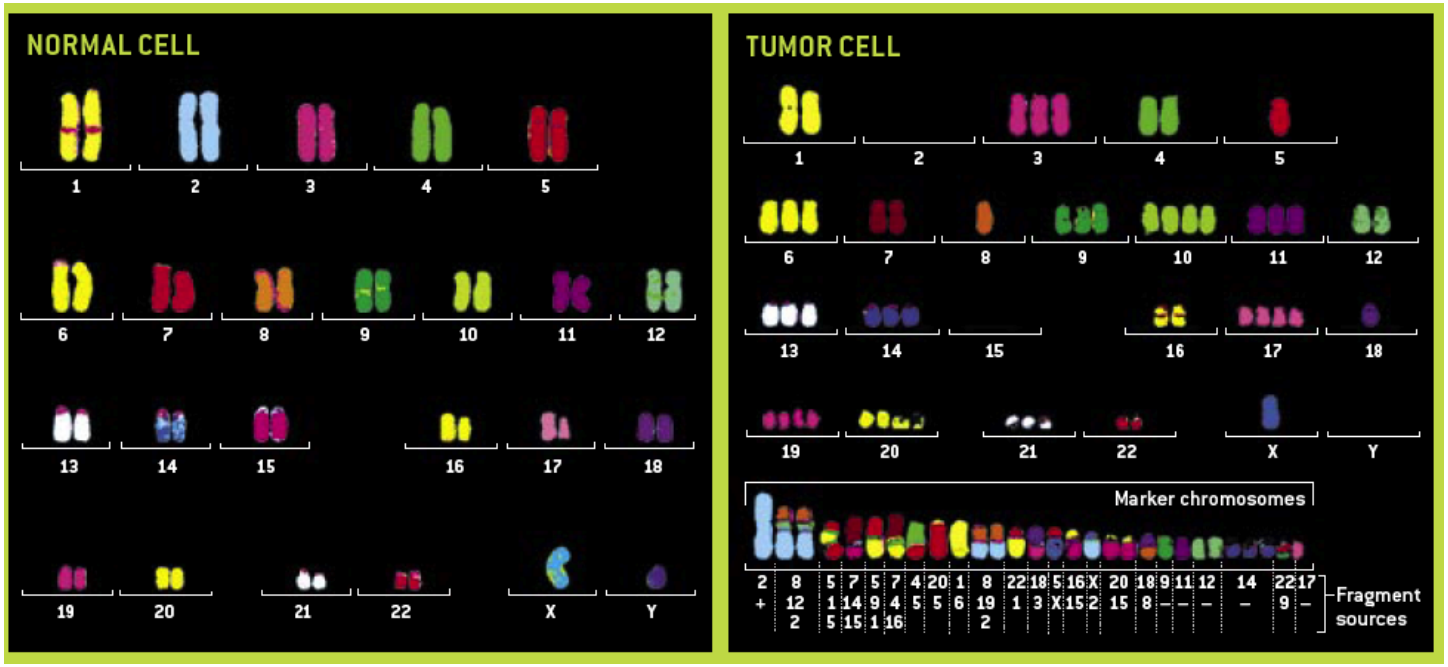
Method: Agilent copy number arrays from TCGA (The Cancer Genome Atlas) for Glioma patients:

250 k DNA markers with local copy signals

CopyArray algorithm translates 250k Agilent copy signals to 15k gene-level copy signals - a pseudo-microarray

Allows analysis of cancer copy number arrays using microarray software (e.g., GSEA, clustering software, etc.)

Copy number signatures



Scientific American

Copy number signatures

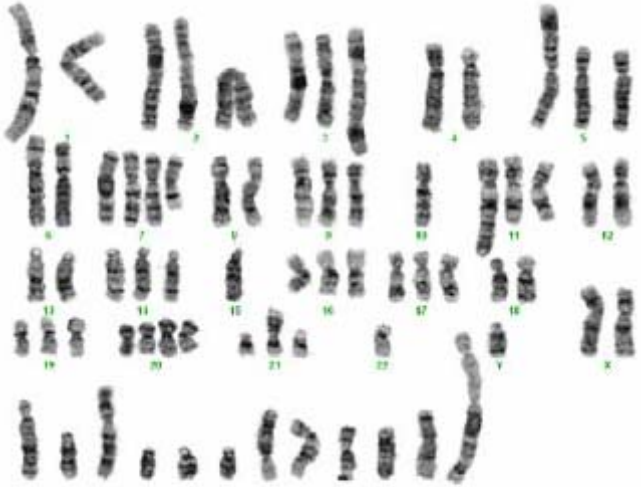


Fig 1: *Conventional cytogenetic analysis of glioblastoma cell line LN-428 by GTG-banding reveals a complex karyotype and multiple marker chromosomes (bottom row) that cannot be unequivocally identified by this method.*

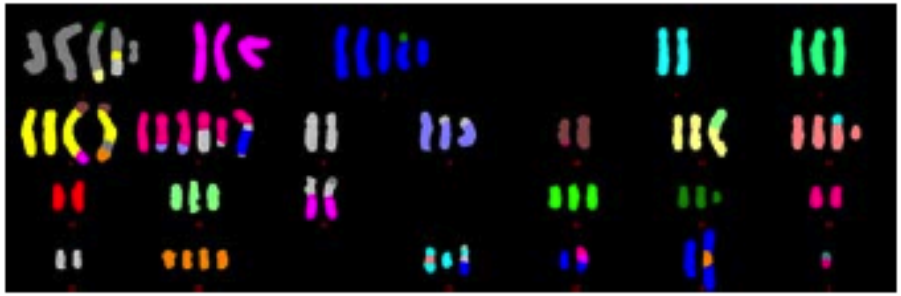


Fig 2: *24-color-FISH-analysis of glioblastoma cell line LN-428 allows identification of the different components of marker chromosomes by painting material from each chromosome in a different color. Breakpoints in marker chromosomes can be identified by comparison with GTG-banded chromosomes.*

Copy number signatures

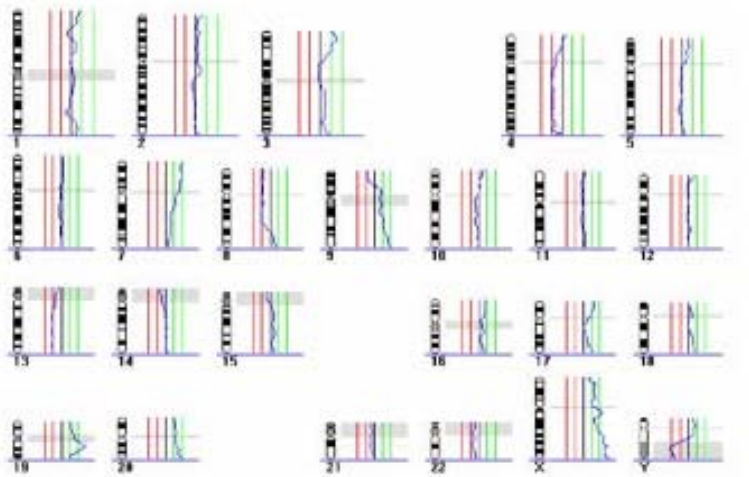


Fig 3: *Chromosomal comparative genomic hybridization (CGH)-analysis of glioblastoma cell line LN-428 allows detection of net gains (profile deviation beyond the green line (1.25) next to the central black line (1.0)) and losses (profile deviation beyond the red line (0.75) next to the central black line (1.0)) of chromosomal material in the tumor cells.*

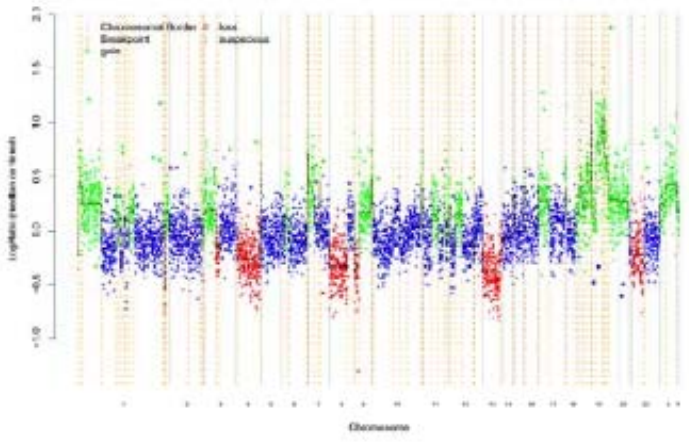
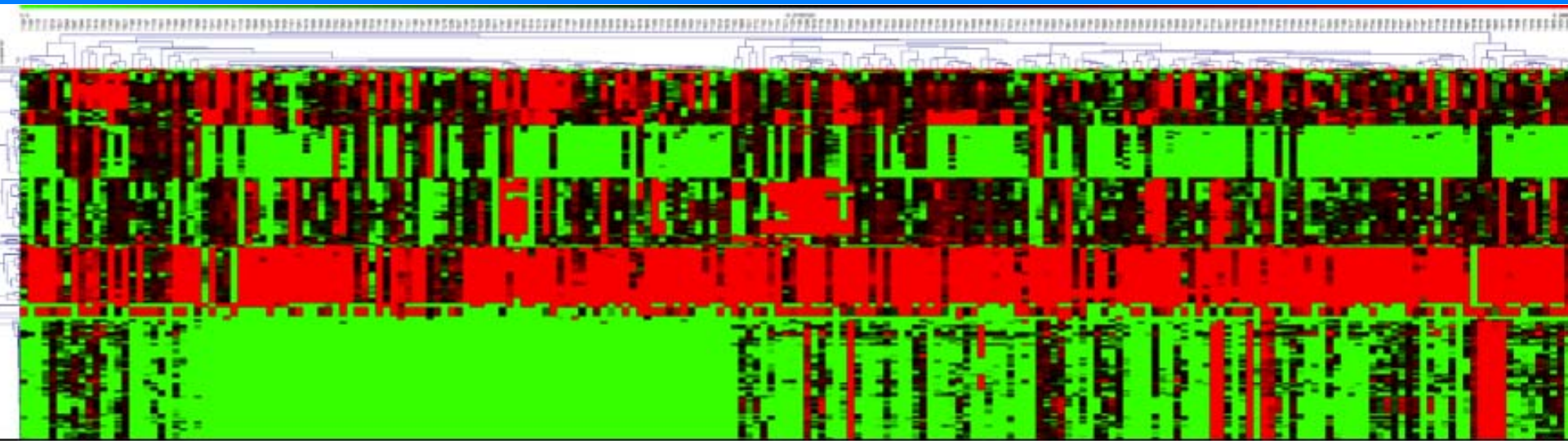


Fig 4: *Array-based comparative genomic hybridization (aCGH)-analysis of glioblastoma cell line LN-428 allows an at least 10-fold increase in the resolution of detected net gains (green data points) and losses (red data points) of chromosomal material in the tumor cells compared to CGH. Clones are plotted in genomic order (from 1p to Yq) against their normalized log2 tumor-DNA to reference-DNA fluorescence ratio.*

http://www.science.ngfn.de/dateien/N3KR-S04T04_Weber.pdf

Copy number signatures



204 glioma tissue samples (horizontal) / 200 genes (vertical)

CopyArray: Natural clustering: 5 groups of genes which co-varyn within subjects --> reduced CopyArray dimensionality can summarize copy number information in 5 dimensional feature vector for phenotype prediction/clustering.

Copy number signatures

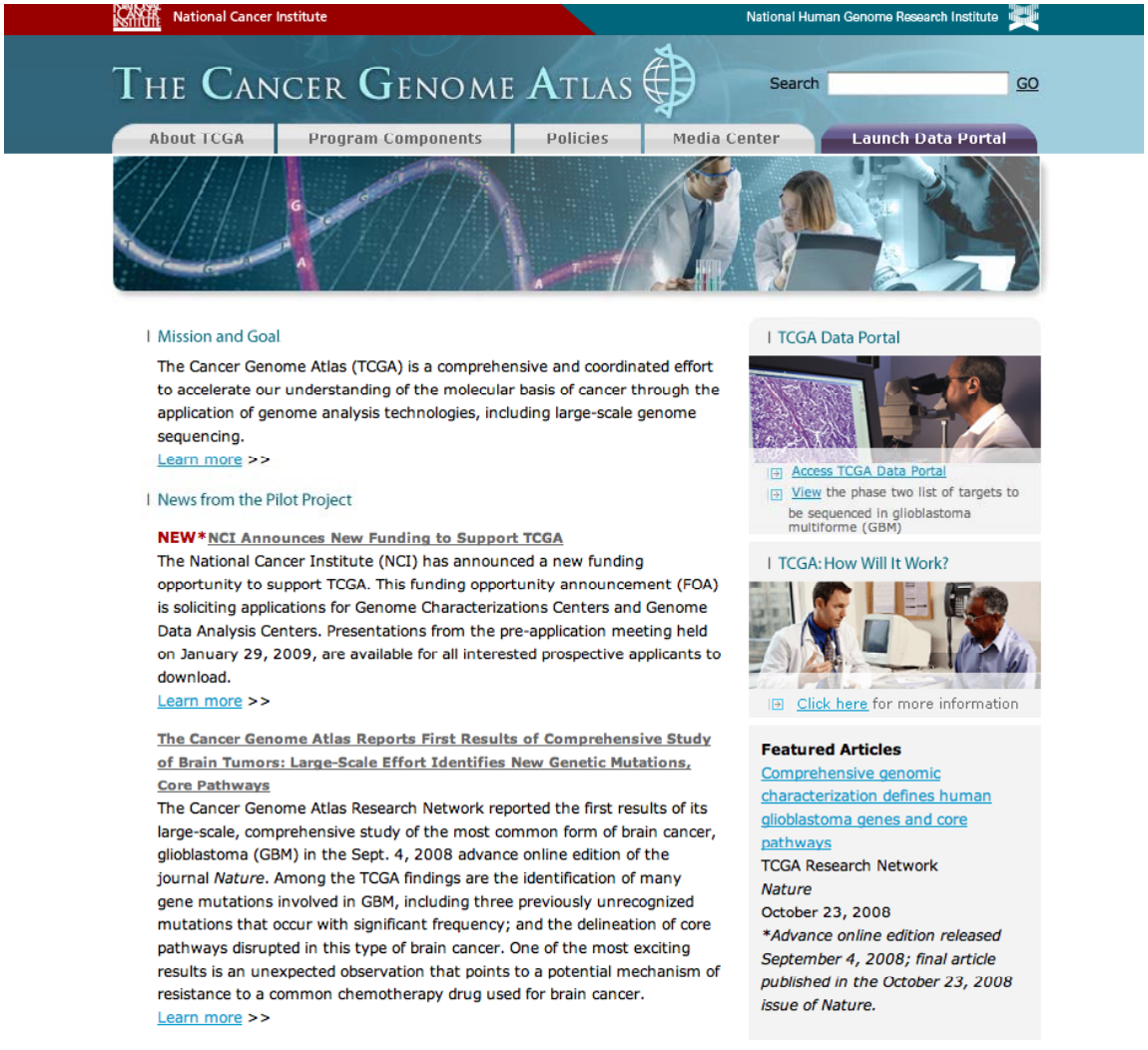
(Dimensionality reduction tool)

Ongoing: use of cluster-based reduced features for prediction of phenotypes (e.g. survival time)

11. Cancer genomics: TCGA

The cancer genome atlas (TCGA) provides high-quality cancer data for large scale analysis by many groups:

Cancer genomics: TCGA



Cancer genomics: TCGA

About TCGA Data | Portal Help | Data Access | Browse Data | Analyze TCGA Data

Overview | Types of Data

TCGA Data Portal

Welcome to The Cancer Genome Atlas (TCGA) Data Portal.

TCGA Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. This portal contains all TCGA data pertaining to clinical information associated with cancer tumors and human subjects, genomic characterization, and high-throughput sequencing analysis of the tumor genomes.

New data is derived on an ongoing basis from TCGA analyses and is deposited into databases. The Data Portal offers access to download these data sets.

Click here to access and download TCGA data.

In addition, the Cancer Molecular Analysis Portal provides the ability for researchers to use analytical tools designed to integrate, visualize, and explore genome characterization from TCGA data.

TCGA Data Portal

Application Help

For more information about how to search the Data Portal for TCGA data, click here.

TCGA Updates

Click here to read more about the latest progress of TCGA pilot project.

View the phase two list of targets to be sequenced in glioblastoma multiforme (GBM).

For more information about initiatives related to TCGA, click here.

Click here to learn more about the new TCGA data use policy and publication guidelines.

Cancer genomics: TCGA

About TCGA Data | Portal Help | Data Access | Browse Data | Analyze TCGA Data

Get TCGA Data

The **Data Access Matrix** allows you to select results of individual samples from multiple centers, platforms and data types, thereby creating a custom archive with your customized data. Simply choose the disease type and data type(s) you would like to work with and proceed to the Data Access Matrix.

Disease Type

GBM – Glioblastoma multiforme

Data Types

All
Clinical
Copy Number Results
DNA Methylation
Expression–Exon
Expression–Genes
Expression–miRNA
SNP

Go to the Data Access Matrix

Alternatively, you can **search by archive** to search for and download complete data archives as submitted by the TCGA research centers.

If you prefer to access the downloads directly you may do so from either FTP (open access) or SFTP (controlled access).

TCGA Sample Counts

	CN			Methyl		Exp-Exon			Exp-Gene			Exp-miRNA			SNP		
	L1	L2	L3	L1	L2	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
GBM	458	460	460	29	247	249	249	232	287	287	287	279	250	250	471	470	470
OV	159	159	159	86	86				49	49	49				86	86	86

TCGA Related Resources

- GBM Publication Site
- Somatic Mutation Data
- Analytical Views of TCGA data
- Sequence Data from NCBI Trace Archive
- TCGA-Data Listserv

DCC Resources:

- BCR Biospecimen Barcodes Table
- Sample-to-file Association Matrix

+ Portal News

01/29/09 - Public Clinical Data File
All current public GBM clinical data is available in tab-delimited format here.

10/03/08 - Tier 1 Clinical Data Spreadsheet
The Tier 1 Clinical Data as of the 10/01/08 update of the BCR Data is available here

09/09/08 - GBM Publication Data Freeze
A list of the archives that comprise the GBM Publication Data Freeze is available here.

09/04/08 - TCGA Reports First Results
In a paper published Sept. 4, 2008, in the advance online edition of the journal Nature, the TCGA team describes the discovery of new genetic mutations and other types of DNA alterations with potential implications for the diagnosis and treatment of

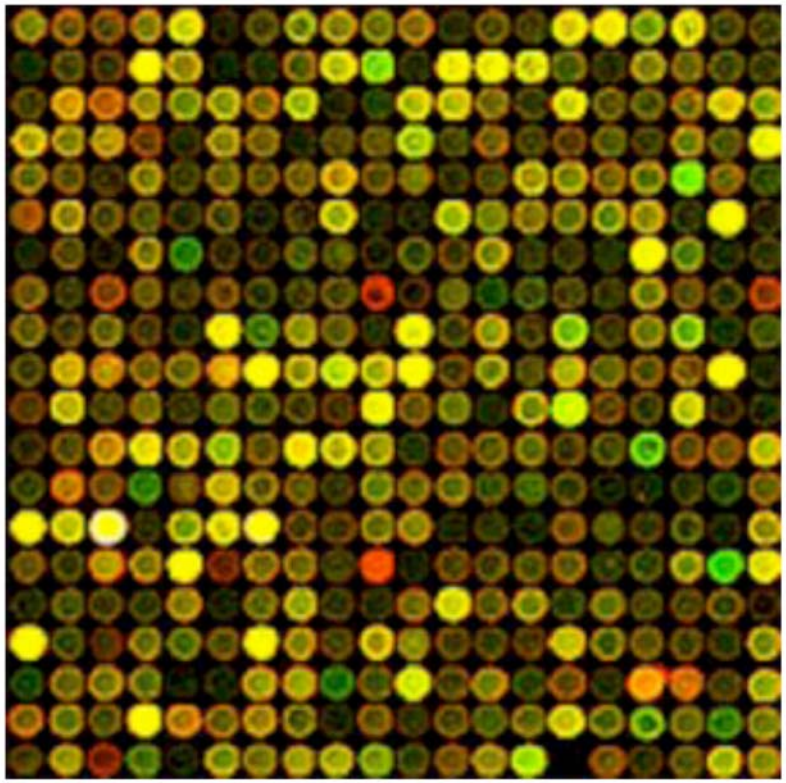
... Show More

12. Some results:

Discrimination of survival time in ovarian cancer: better biomarkers.

Imagine a microarray which measures metabolic pathway activation instead of gene activation:

Some results



Problem with genes: there are too many of them!

Gene expression microarrays are massively redundant; feature vector **x** will often discriminate

Some results

cancer from normal tissue with 100% accuracy (e.g. as in ovarian cancer, glioma).

Complete discriminative set of genes from one TCGA database (e.g. UNC) is sometimes completely different from complete set for another (e.g. Broad institute).

Stable biomarkers: hierarchical SVM

Pathway biomarkers are much more robust and non-redundant than gene biomarkers - they are *canonical.*

Most significant pathways can be extracted from most significant gene biomarkers

Stable biomarkers: hierarchical SVM

Enriched pathway in stage IV (BI)	Enriched pathway in stage IV (UNC)
CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION[1,2]	ANTIGEN_PROCESSING_AND_PRESENTATION[2]
TYPE_I_DIABETES_MELLITUS	TYPE_I_DIABETES_MELLITUS
ANTIGEN_PROCESSING_AND_PRESENTATION[2]	NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY[1,2]
COMPLEMENT_AND_COAGULATION_CASCADES[2]	CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION[1,2]
TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY[2]	RIBOSOME[1,2]
GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIOSERIES[2]	JAK_STAT_SIGNALING_PATHWAY[1,2]
CELL_COMMUNICATION[1,2]	HEMATOPOIETIC_CELL_LINEAGE[2]
NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY[1,2]	CELL_COMMUNICATION[1,2]
HEMATOPOIETIC_CELL_LINEAGE[2]	CELL_ADHESION_MOLECULES[1,2]
ALZHEIMERS_DISEASE	PPAR_SIGNALING_PATHWAY[2]
BLADDER_CANCER	TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY[2]
	LINOLEIC_ACID_METABOLISM
	COMPLEMENT_AND_COAGULATION_CASCADES[2]
	TYPE_II_DIABETES_MELLITUS
	T_CELL_RECEPTOR_SIGNALING_PATHWAY
	NEUROACTIVE LIGAND RECEPTOR INTERACTION[2]

Stable biomarkers: hierarchical SVM

Clinical applications - standardized biomarkers:

Form SVM from feature vectors with e.g. 120 pathway strength biomarkers:

$$\mathbf{x} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_{120} \end{bmatrix}$$

with pathway strengths p_i based on microexpression gene activities in p_i.

Stable biomarkers: hierarchical SVM

Use machine learning methods on standardized biomarkers.

Expectation: in near future these will form a standard portion of clincal data.

Further stable biomarkers: genes selected by pathway membership

Given pathway p_i, select most important genes $\{g_{ij}\}_j \subset p_i$.

(a) These can be the 'leading edge' genes chosen using GSEA (gene set enrichment analysis), or genes g_{ij} in path p_i with highest SVM weights w_{ij}.

Stable biomarkers: hierarchical SVM

These genes again form a set of biomarkers which is stable under change of data source.

(b) The genes can be selected from SVM weights. If SVM decision function is

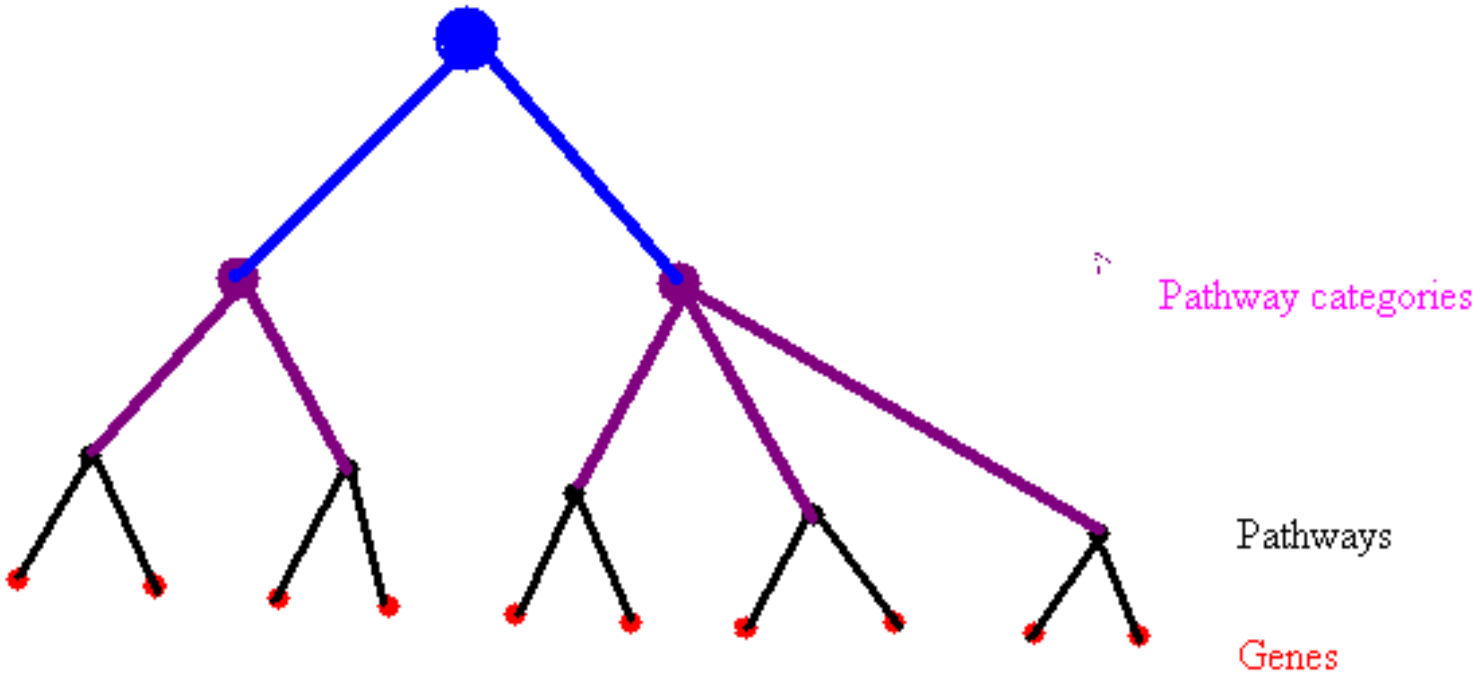
$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b,$$

where **x** is microarray feature vector and **w** $=$ vector $\perp$ to separating hyperplane H.

Large components w_i of **w** represent the gene components in feature space which are most important.

Stable biomarkers: hierarchical SVM

Directions i with largest values w_i represent genes g_i in a fixed pathway p that form a canonical set.



Stable biomarkers: hierarchical SVM

In small dimensional spaces (higher on the tree) the ordering of features is much more consistent than in high dimensional spaces.

This yields consistent feature (pathway) biomarkers.

In addition, in each pathway we have *canonical genes* (most significant) in addition to above *canonical pathways.*

Application to SVM discrimination:

TCGA: Consider discrimination between metastatic and non-metastatic cancer.

Example: Breast cancer metastasis data

Wang:

93 metastatic
183 non-metastatic

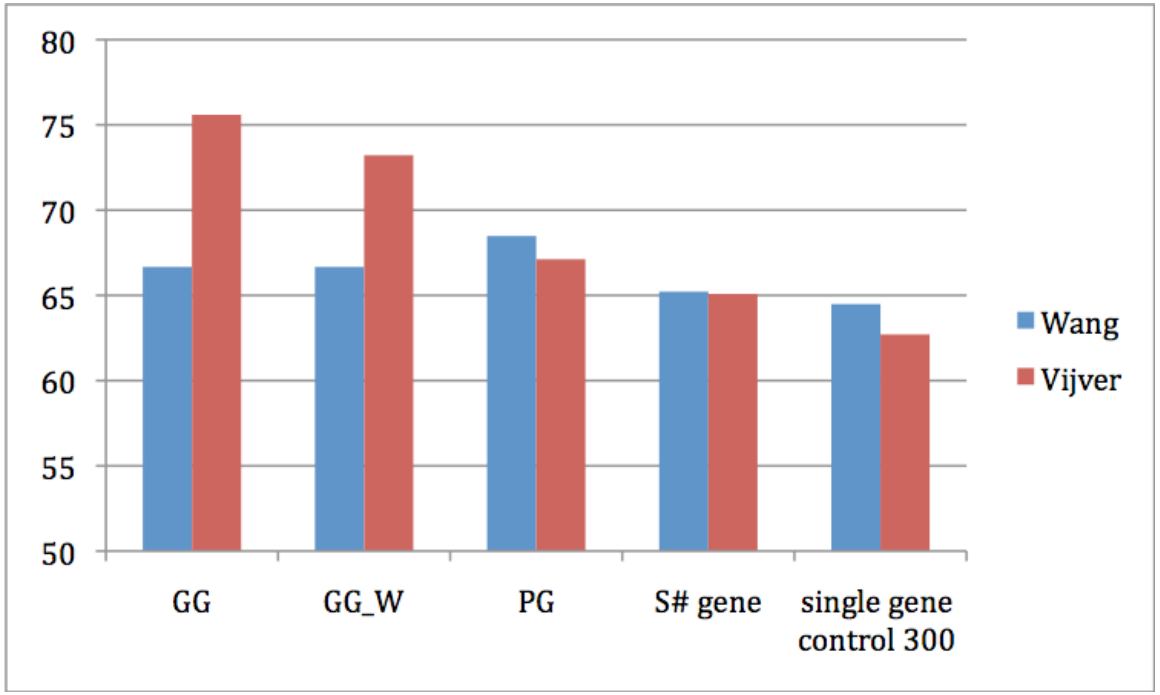
van de Vijver:

79 metastatic

Pathway applications to SVM discrimination

216 non-metastatic

Here we show that our stable biomarker methods give better performance than standard single gene control methods:



Pathway applications to SVM discrimination

Canonical gene methods Pathway method single gene co

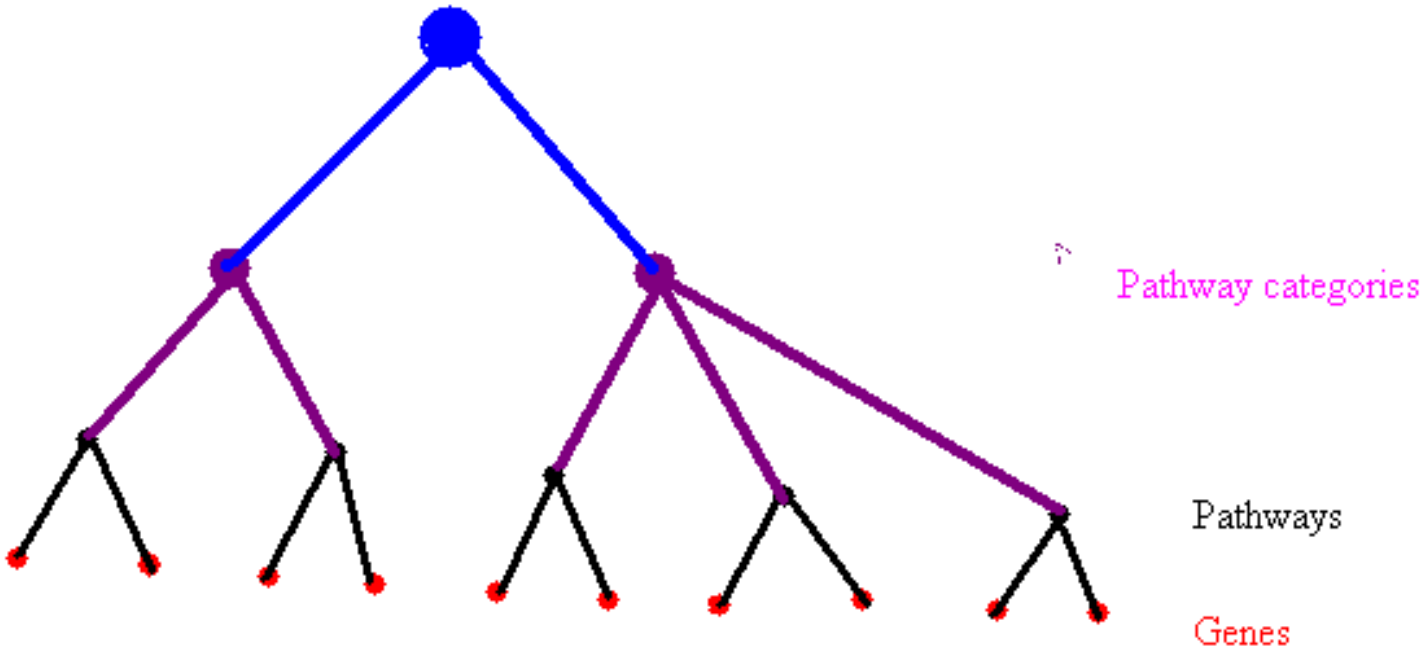
GG: leading edge genes; GG_W: leading edge genes weighted

PG: Pathway biomarkers; S#: control single gene classification

Single gene control 300: control single gene classification using top 300 Fisher selected genes

Significantly, they provide more stability in choice of significant aggregate and non-aggregate biomarkers:

Pathway applications to SVM discrimination



Important pathways p_i in Wang study overlap significantly with those in the van de Vijver study:

Pathway applications to SVM discrimination

Common pathways for discriminating metastatic & non-metastatic BC between Van de Vijver and Wang data
HSA04940_TYPE_I_DIABETES_MELLITUS
HSA04640_HEMATOPOIETIC_CELL_LINEAGE
HSA04610_COMPLEMENT_AND_COAGULATION_CASCADES
HSA00590_ARACHIDONIC_ACID_METABOLISM
HSA00071_FATTY_ACID_METABOLISM
HSA04110_CELL_CYCLE
HSA00100_BIOSYNTHESIS_OF_STEROIDS
HSA03050_PROTEASOME
HSA03030_DNA_POLYMERASE
HSA00240_PYRIMIDINE_METABOLISM
HSA00970_AMINOACYL_TRNA_BIOSYNTHESIS
HSA03020_RNA_POLYMERASE
HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM
HSA05219_BLADDER_CANCER

This is a significant overlap of biomarkers for selecting 40 pathways out of 200.

Pathway applications to SVM discrimination

p-value: with Poisson approximation we have

$$\lambda = np = 40 \cdot \frac{40}{200} = 8.$$

Thus have: $P(\geq 14 \text{ overlapping}) = P(S \geq 14)$ where S is Poisson with parameter $\lambda = 8$, getting $p = .0342$.

There is a clear signal here, obscured by sample/condition fluctuations, which indicates pathways worthy of biological investigation for roles in metastasis.

Some suspects in this role

Pathway applications to SVM discrimination

Complement and coagulation cascades: proteolytic cascade in blood plasma and a mediator of innate immunity, a nonspecific defense mechanism against pathogens.

Arachodinic acid metabolism: Arachodinic acid is metabolized to both pro- and anti-inflammatory elements
http://en.wikipedia.org/wiki/Arachidonic_acid

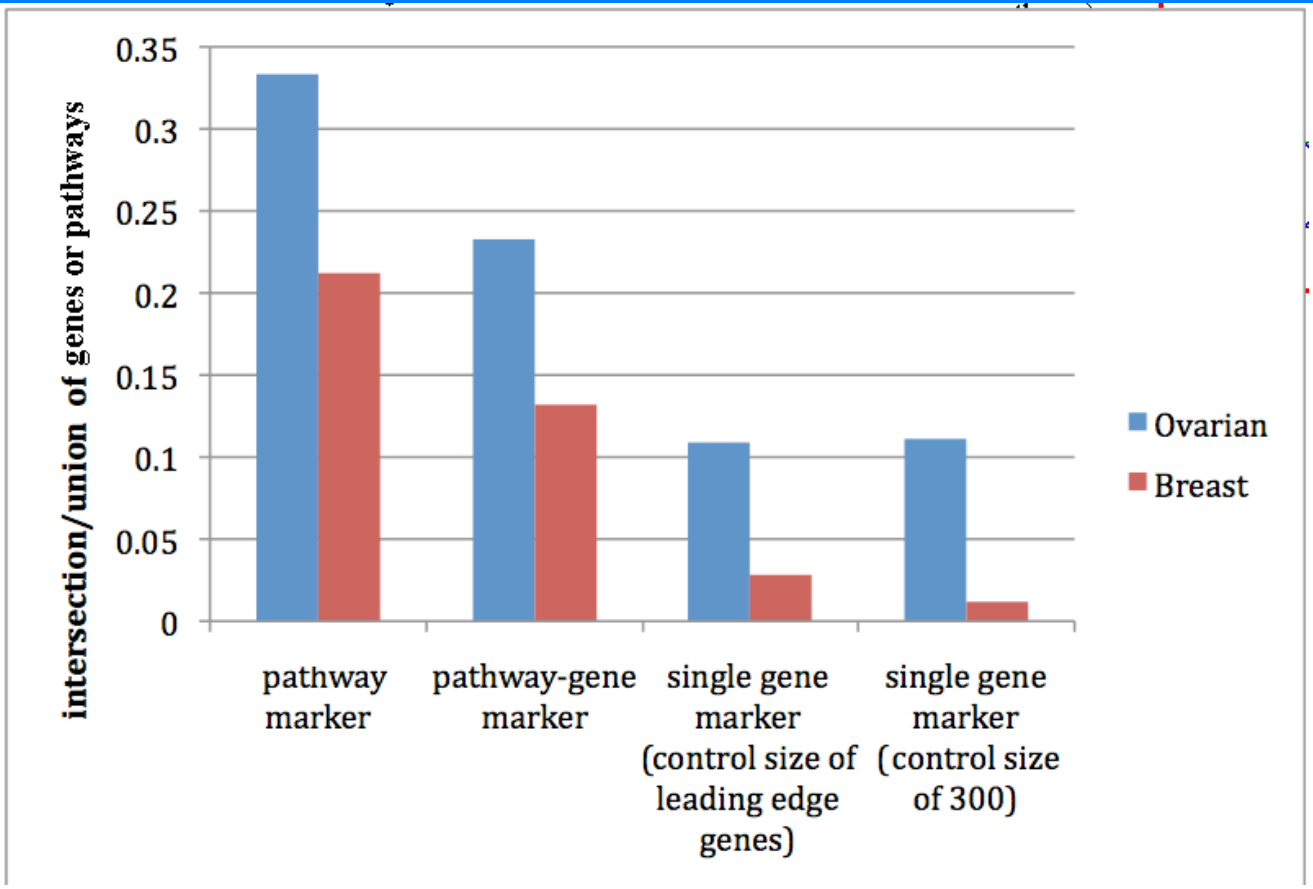
Cell cycle pathway: well known links to cancer

DNA polymerase pathway: strongly linked to DNA repair

Bladder cancer pathway

Stability of gene/hierarchical biomarkers:

Pathway applications to SVM discrimination



hierarchical biomarkers gene biomarkers

Pathway stability:

Pathway applications to SVM discrimination

Blue is ovarian (TCGA data - UNC vs. BI)

Tan is breast: Wang vs. van de Vijver

Left two are based on hierarchically based biomarkers

Right two based on individual gene biomarkers

13. MicroRNA variations

Machine learning (using SVM) has shown that miRNA levels turn out to be crucial in predicting cancer outcomes:

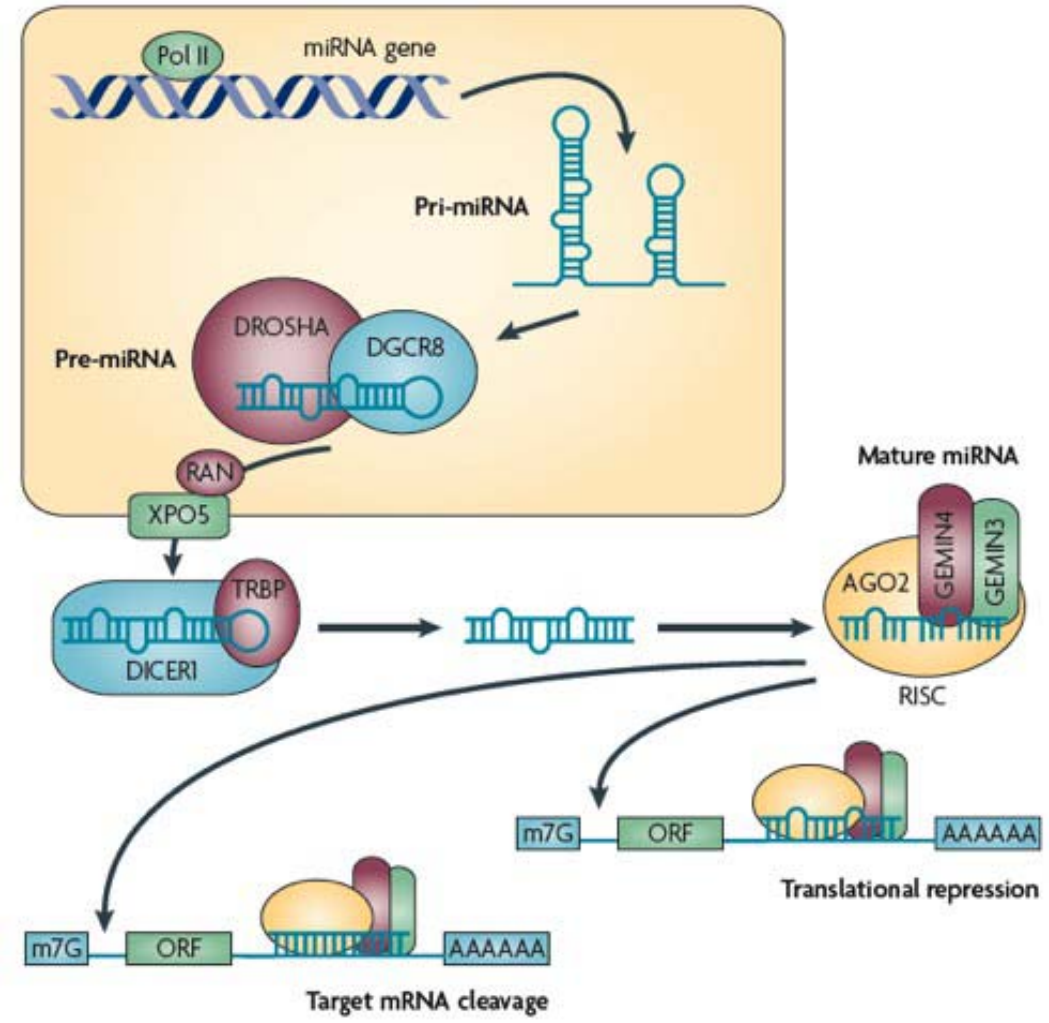
1)miRNA is more accurate method of classifying cancer subtype than using the expression profiles (ref: Calin and Crosce; Volinia)

the performance of TCGA ovarian SURVIAL:
82% miRNAs, 60% mRNA, 84% miRNA+mRNA.

Cancer: microRNA variations

2) miRNAs regulate their target mRNA controlling biological functions such as cellular proliferation, differentiation, and apoptosis (ref : Calin and Crosce).

Cancer: microRNA variations

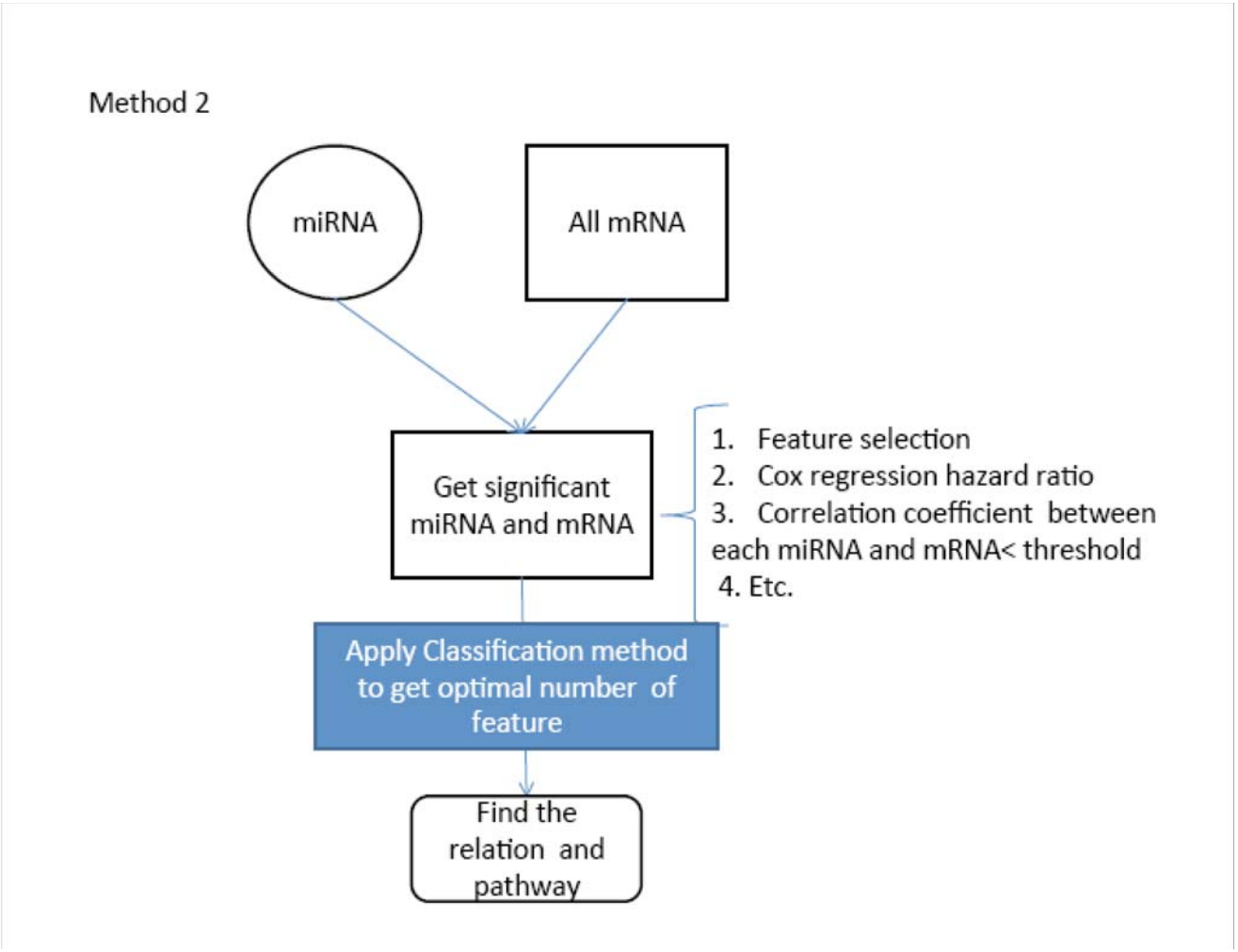


- Ovarian cancer data sets:

Cancer: microRNA variations

- 22 short survival patients (less than 1 year)
- 22 long survival patients (greater than 5 years)
- 799 miRNA
- 17,814 mRNA

Cancer: microRNA variations



Cancer: microRNA variations

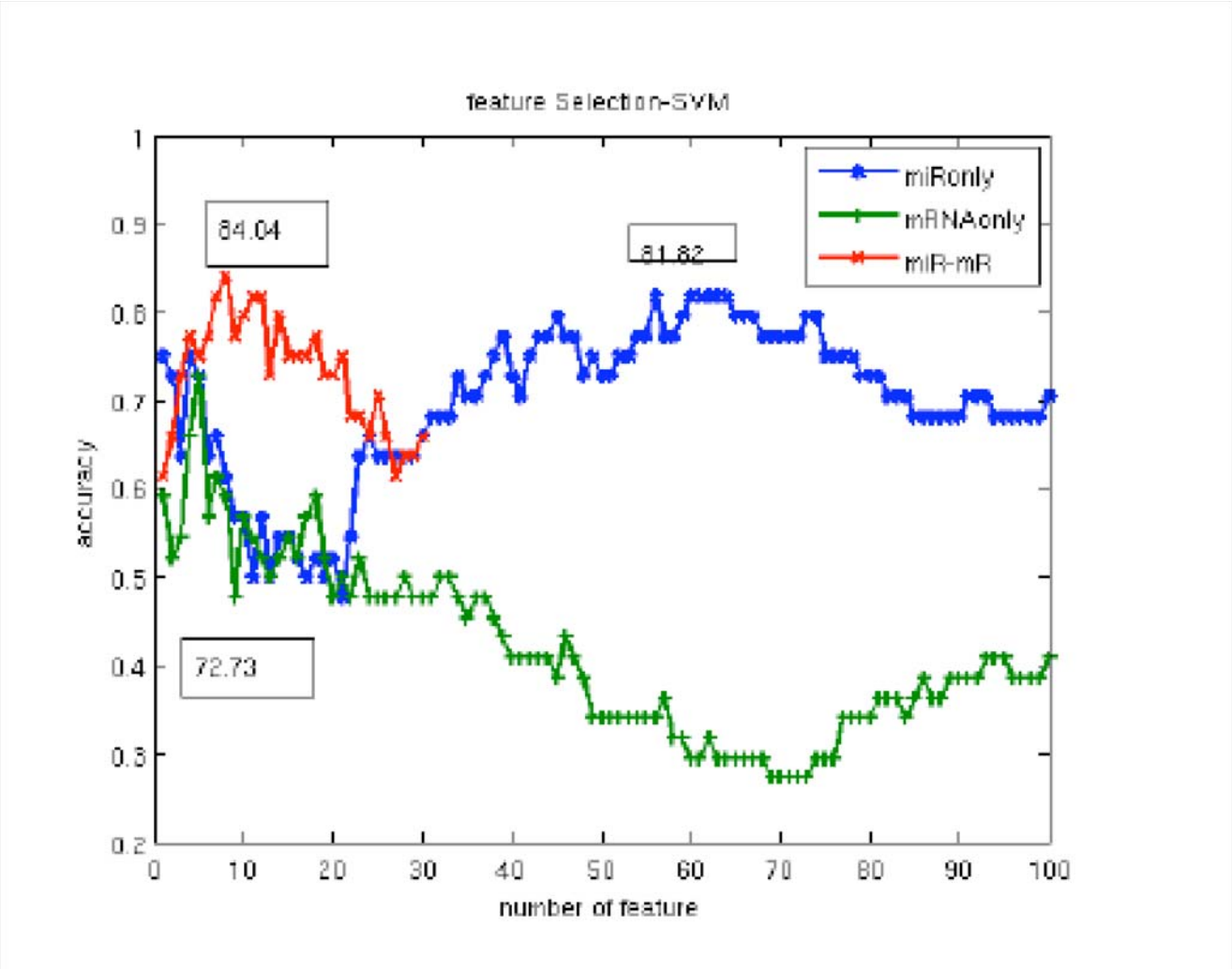
More importantly, joint information from miRNA and mRNA (regular microarrays) gives better information than either alone - add the kernel matrices

$$\mathbf{K}^{(1)} + \mathbf{K}^{(2)} = \mathbf{K}_{\text{full}}$$

and use machine learning.

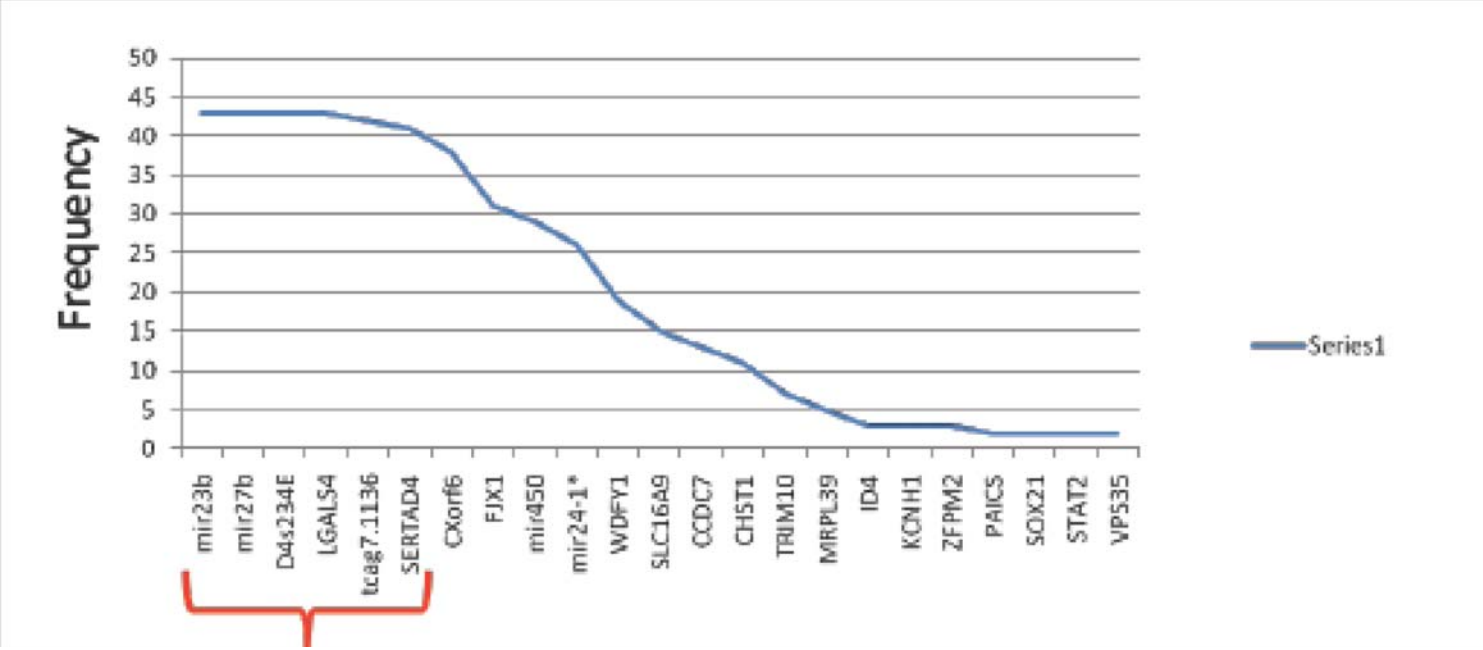
SVM:

Cancer: microRNA variations



Most important genes in machine learning runs:

Cancer: microRNA variations



Glioblastoma

- Data sets
- 24 short survival patients (less than 150 days)
- 23 long survival patients (greater than 700 days)
- 534 miRNA
- 17,530 mRNA

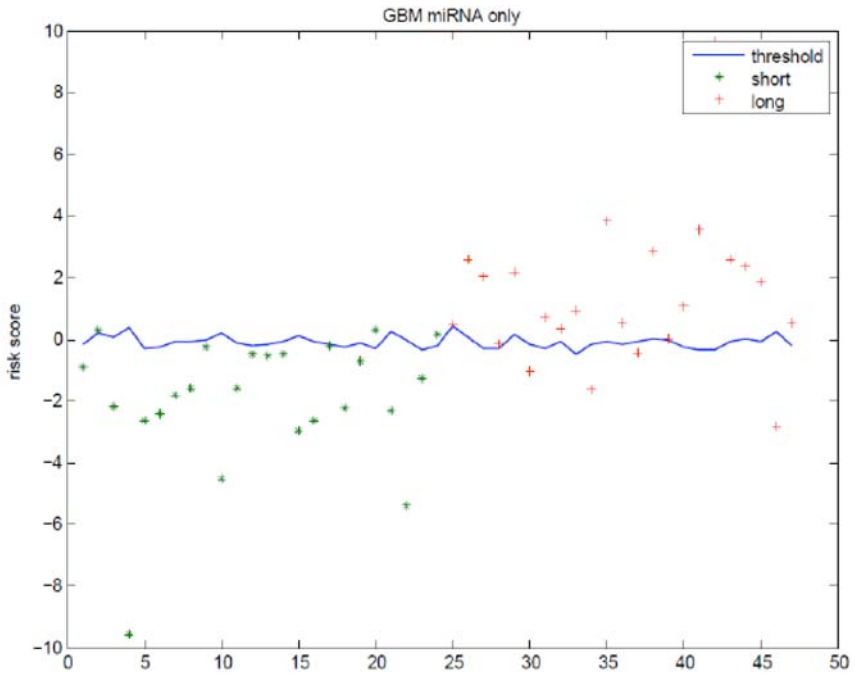
Cancer: microRNA variations

Result for GBM

- SVM :
  1) miRNA :72.34%(13 genes selected)
  2) mRNA : 72.34% (76~77 genes selected)

- Cox Regression
  1) miRNA : 85.11% (19 genes selected)
  2) mRNA : 74.47% (17 genes)

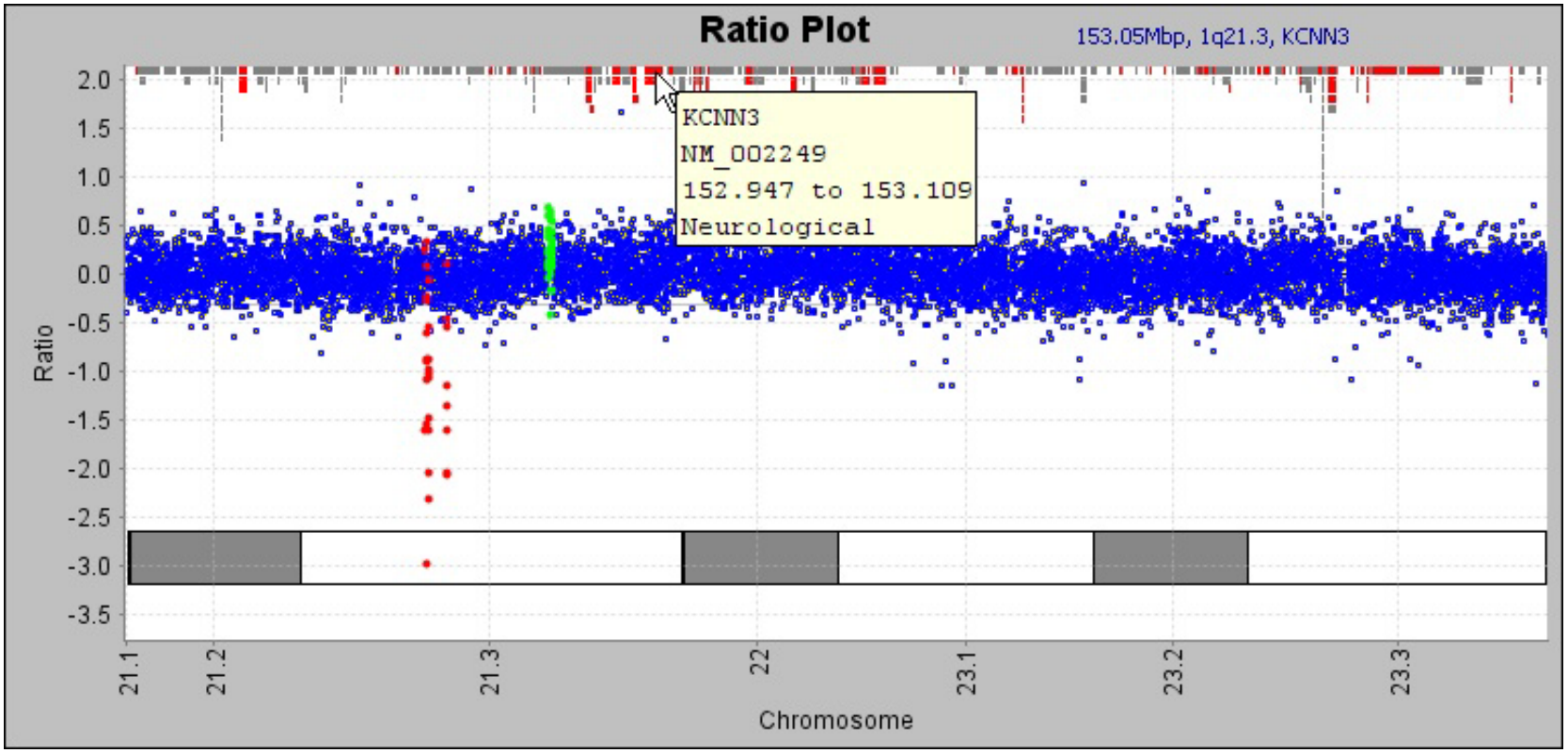
Cancer: microRNA variations



14. Copy number variations in cancer

Cancer: copy number variations

Agilent copy number array



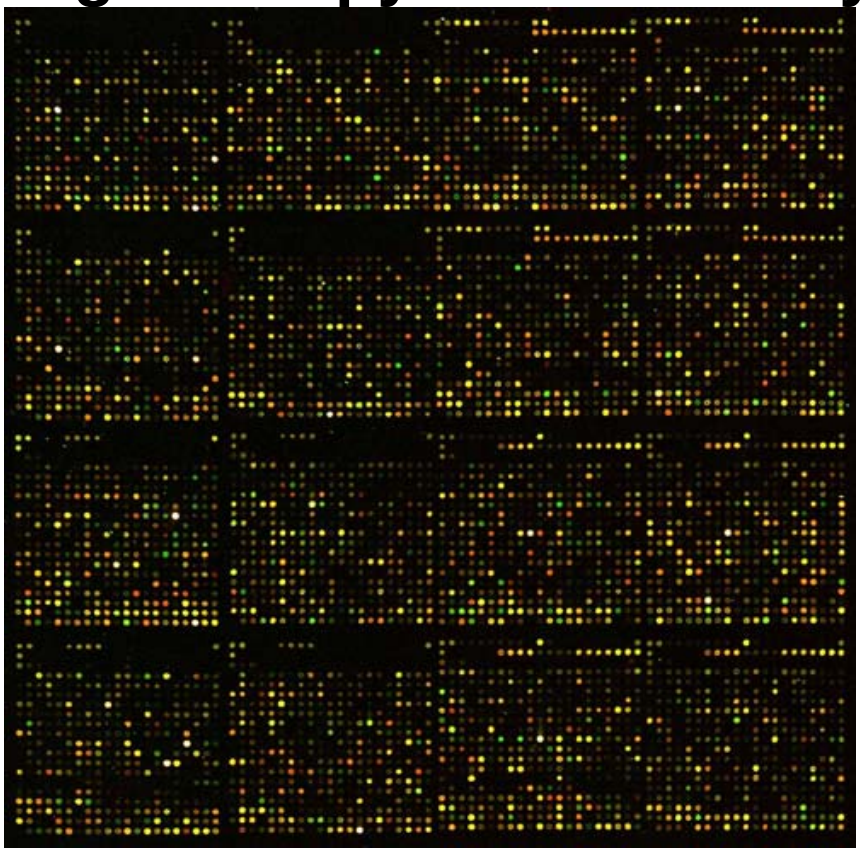
Cancer: copy number variations

http://www.infoquant.com/index/platform-highres

Roche Nimblegen HD2 (2.1M) Whole Genome Tiling Array

Cancer: copy number variations

Idea: convert 260,000 local DNA copy number markers into gene copy number array:



http://ki.se/ki/jsp/polopoly.jsp?d=3833&a=1349&l=en

Cancer: copy number variations

Replace a microarray with a gene copy number array

Each spot now represents gene copy number in cancer - this is the DNA version of an RNA microarray.

High correlation with microarray - gene expression is proportional to gene copy number.

More authentic information for prediction of e.g. survival time or drug response - genomic signature of the cancer.

15. Next step: full sequencing

The $1000 genome will yield clinical as well as research applications. Imagine 3 x 10^9 biomarkers - the full genome.

Extraction of relevant features now becomes crucial - an important task for data mining and machine learning.

Cancer genome: broken chromosomes, multiple copies, deletions of chromosome regions.

Construction of the cancer genome will start with precise copy number data and algorithms which determine maximal connected components of the genome.

Integration of large quantities of data with RNA and phenomic observations will have to become standard.