Data Mining Algorithms for Cancer Detection

Nirmalya Bandhopadhay, Jun Liu, Sanjay Ranka, Tamer Kahveci

http://www.cise.ufl.edu

UF UNIVERSITY of FLORIDA

Outline

- Cancer Datasets are growing
  - CGH, Microarray, Microarray time course
- Datasets are High Dimensional
  - 1000 to 20000 dimensions
- Maximum Influence Feature Selection
- Biological Pathway Feature Selection
- Cancer Progression Modeling

Gene copy number

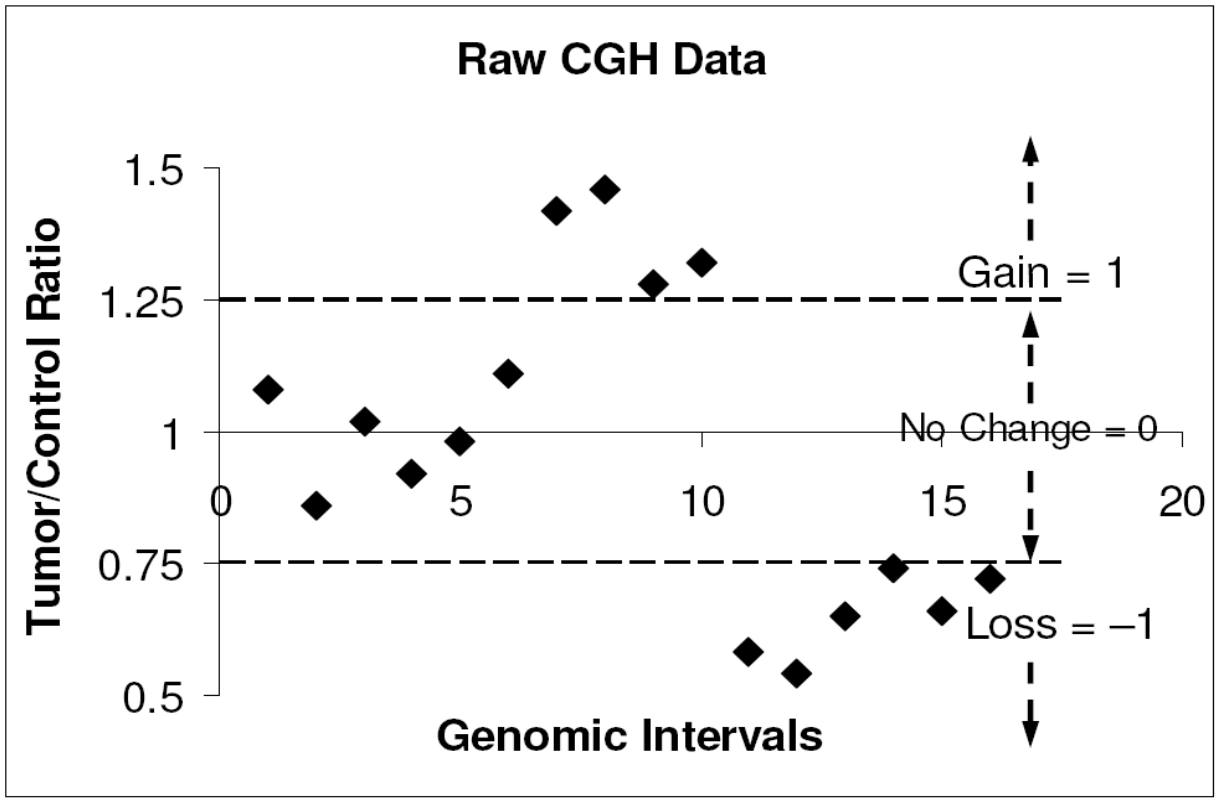
- The number of copies of genes can vary from person to person.
  - ~0.4% of the gene copy numbers are different for pairs of people.
- Variations in copy numbers can alter resistance to disease
  - EGFR copy number can be higher than normal in Non-small cell lung cancer.

Lung images (ALA)

Healthy

Cancer

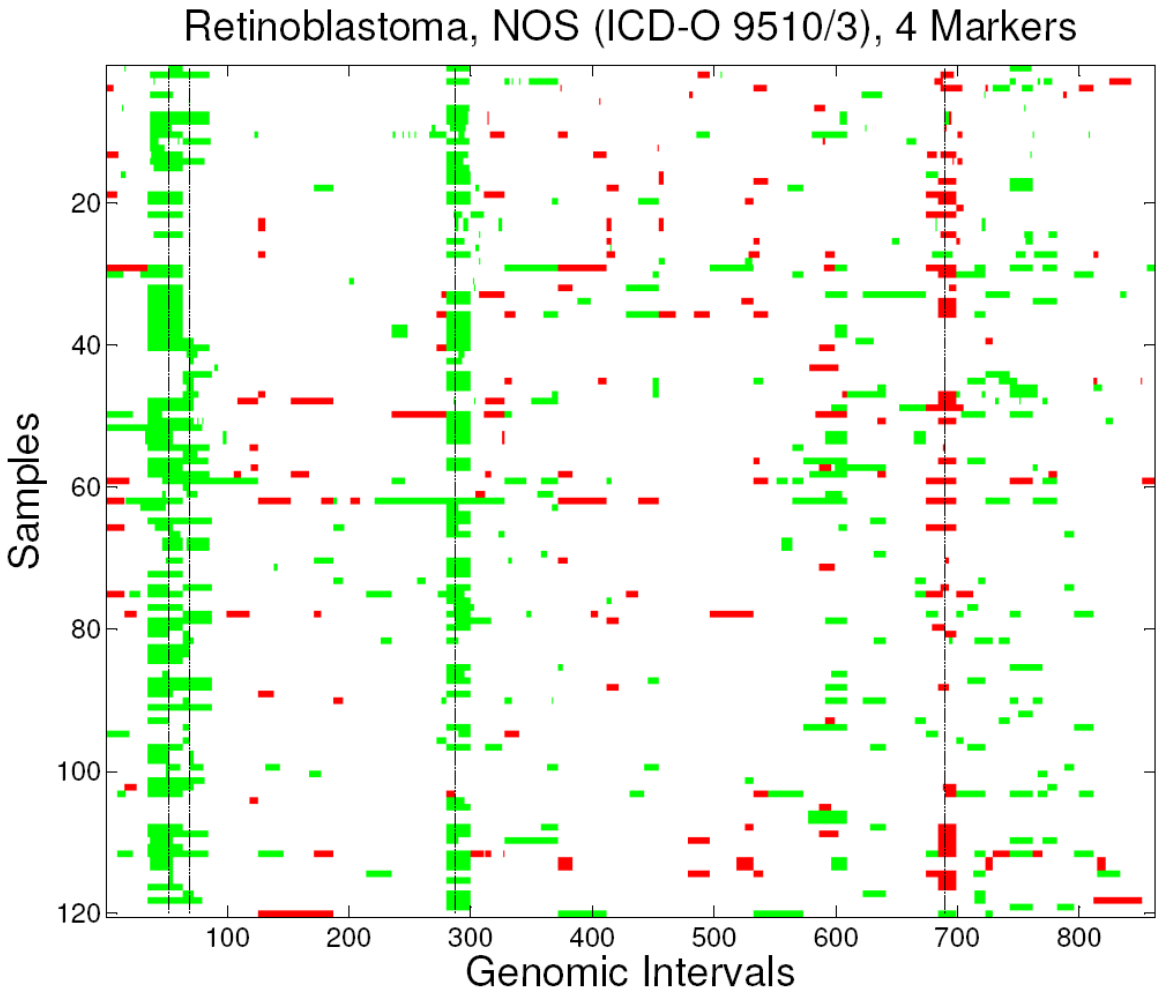
Raw and smoothed CGH data



0 0 0 0 0 0 1 1 1 1 -1 -1 -1 -1 -1 -1

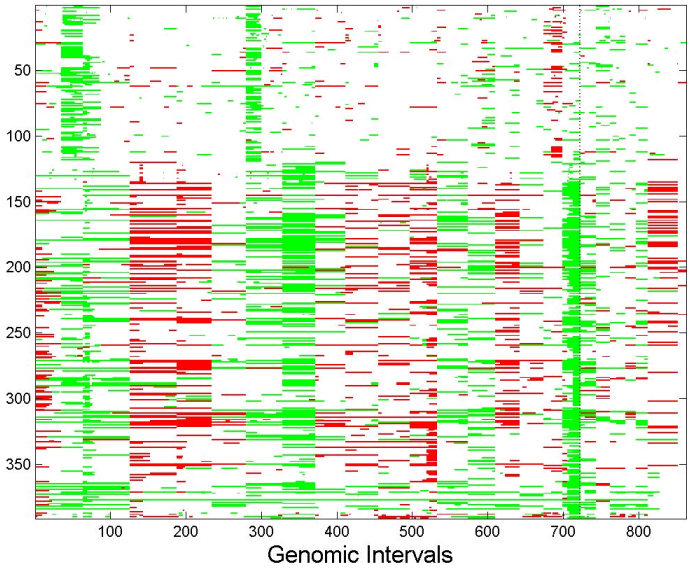
Smoothed CGH Data

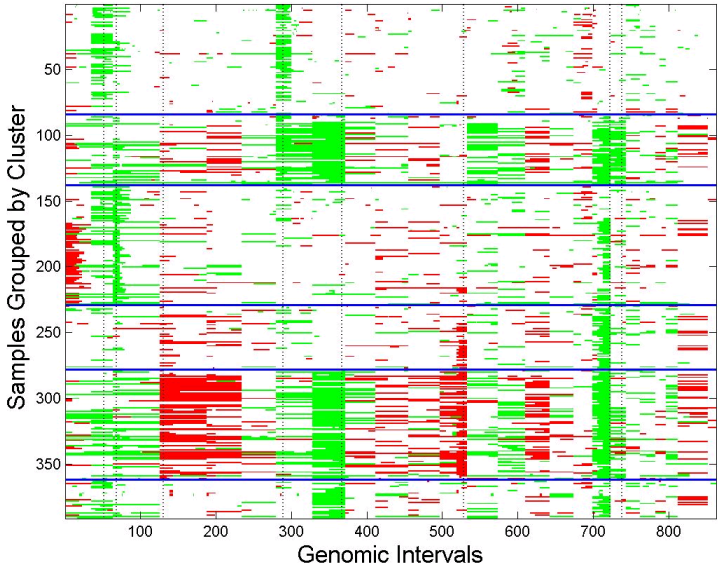
Example CGH dataset



862 genomic intervals in the Progenetix database

Problem description

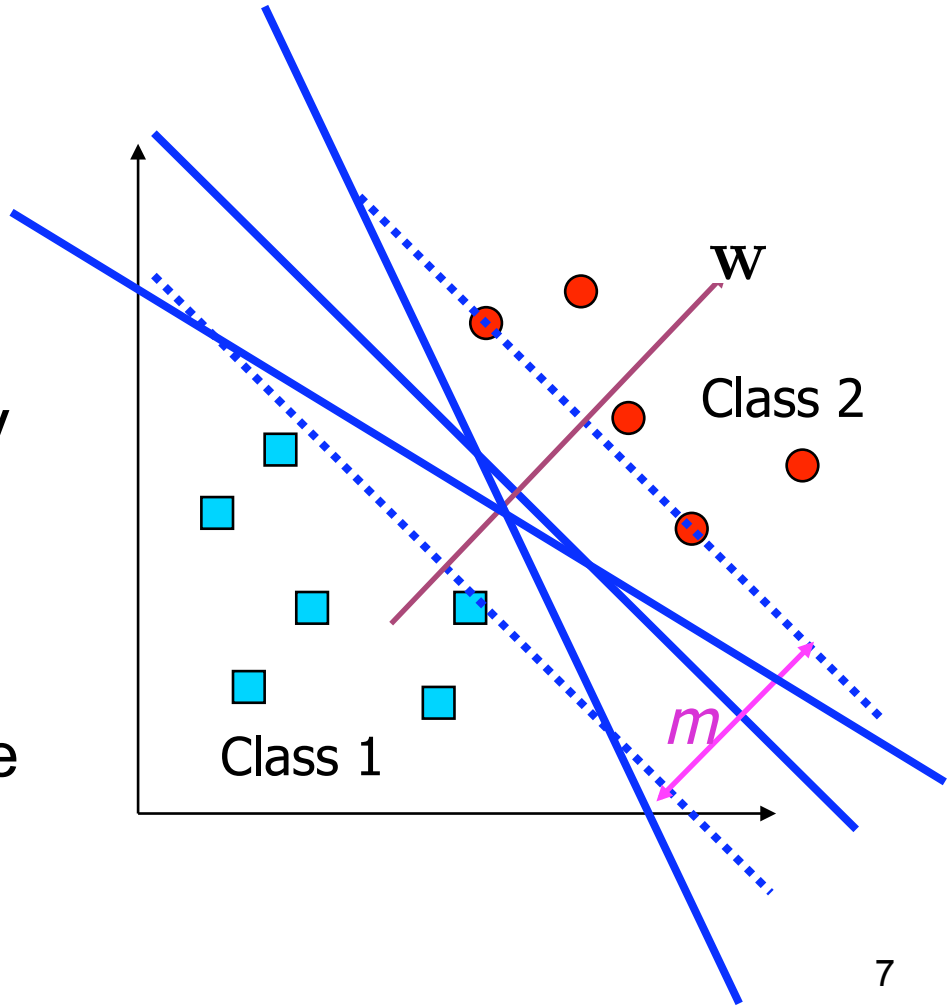




- Given a new sample, which class does this sample belong to?
- Which features should we use to make this decision?

Classification with SVM

- Consider a two-class, linearly separable classification problem
- Many decision boundaries!
- The decision boundary should be as far away from the data of both classes as possible
  - We should maximize the margin, *m*



SVM Formulation

- Let $\{x_1, ..., x_n\}$ be our data set and let $y_i \in \{1,-1\}$ be the class label of x_i
- Maximize J over α_i

$$J = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

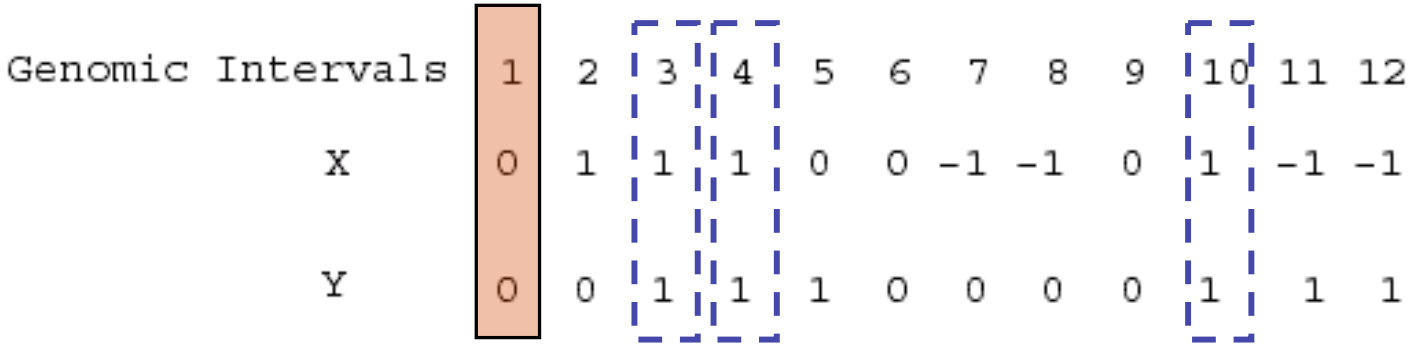
Similarity between x_i and x_j

- The decision boundary can be constructed as

$$D(z) = w \cdot z + b = \sum_i \alpha_i y_i (x_i^T z) + b$$

Pairwise similarity measures

- Raw measure
  - Count the number of genomic intervals that both samples have gain (or loss) at that position.



Raw = 3

SVM based on Raw kernel

- Using SVM with the Raw kernel amounts to solving the following quadratic program

  Maximize *J* over α_i :

$$J = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j Raw(x_i, x_j)$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

Use Raw kernel to repla$x_i^T x_j$

- The resulting decision function is

$$D(z) = \sum_{i} \alpha_i y_i Raw(x_i, z) + b$$

Use Raw kernel to repla$x_i^T z$

Is this cool?

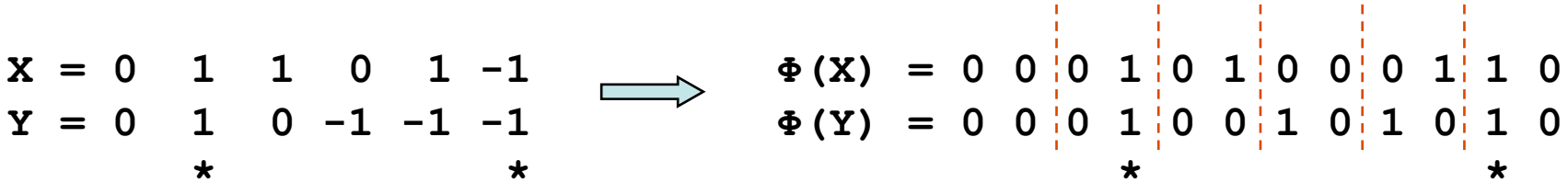
Is Raw kernel valid?

- Not all similarity function can serve as kernel. This requires the underlying kernel matrix M is “positive semi-definite”.
- *M* is positive semi-definite if for all vectors v, $v^T Mv \geq 0$

$$M = \begin{bmatrix} Raw(x_1, x_1) & Raw(x_1, x_2) & \cdots \\ Raw(x_2, x_1) & Raw(x_2, x_2) & \cdots \\ \cdots & & \end{bmatrix}$$

Is Raw kernel valid?

- Proof: define a function Φ() where
  - Φ: a $\in\{1, 0, -1\}^m \rightarrow$ b $\in\{1, 0\}^{2m}$,where
    - Φ(gain) = Φ(1) = 01
    - Φ(no-change) = Φ(0) = 00
    - Φ(loss) = Φ(-1) = 10
  - Raw(X, Y) =$\Phi(X)^T \Phi(Y)$



Raw(X, Y) = 2

$\Phi(X)^T \Phi(Y)$ = 2

Raw Kernel is valid!

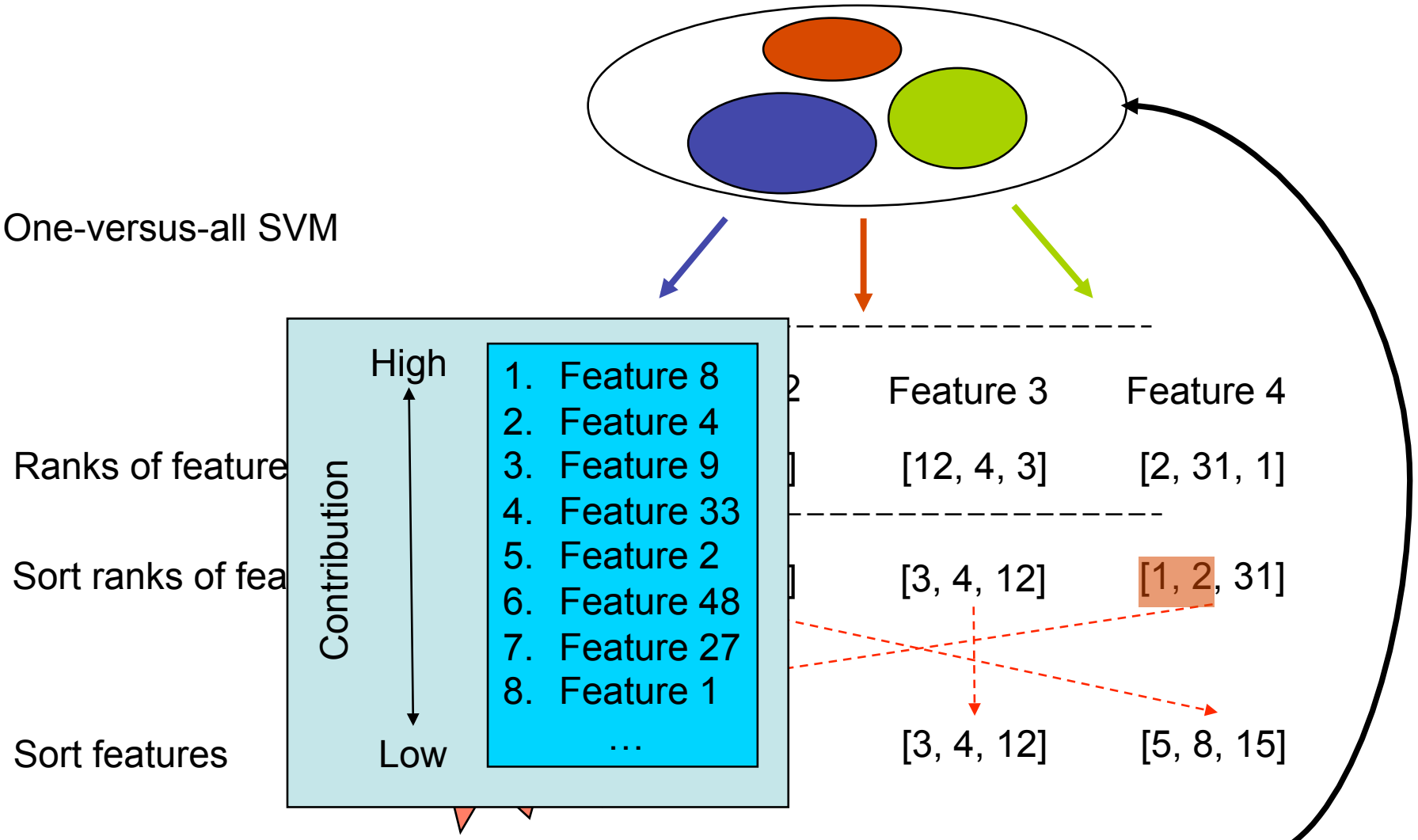
- Raw kernel can be written as $\text{Raw}(X, Y) = \Phi(X)^T \Phi(Y)$
- Define a *2m* by *n* matrix $u := \begin{bmatrix} \Phi(x_1) & \Phi(x_2) & \cdots \end{bmatrix}$

Let *M* denote the Kernel matrix of Raw

$$
\begin{aligned}
M &= \begin{bmatrix} Raw(x_1, x_1) & Raw(x_1, x_2) & \cdots \\ Raw(x_2, x_1) & Raw(x_2, x_2) & \cdots \\ \cdots & & \end{bmatrix} \\
&= \begin{bmatrix} \Phi(x_1)^T \cdot \Phi(x_1) & \Phi(x_1)^T \cdot \Phi(x_2) & \cdots \\ \Phi(x_2)^T \cdot \Phi(x_1) & \Phi(x_2)^T \cdot \Phi(x_2) & \cdots \\ \cdots & & \end{bmatrix} \\
&= \begin{bmatrix} \Phi(x_1)^T \\ \Phi(x_2)^T \\ \cdots \end{bmatrix} \begin{bmatrix} \Phi(x_1) & \Phi(x_2) & \cdots \end{bmatrix} \\
&= u^T \cdot u
\end{aligned}
$$

- Therefore, $v^T M v = v^T (u^T u) v = (uv)^T uv = \|uv\|^2 \geq 0, \forall v \in \mathbf{R}^n$

MIFS for multi-class data



Dataset Details

Data taken from
Progenetix database

#cases	Code translation
310	Infiltrating duct mixed with carcinoma
323	Diffuse large B-cell lymphoma, NOS
346	B-cell chronic/small lymphocytic leukemia
1057	Adenocarcinoma, NOS
657	Squamous cell carcinoma, NOS
209	Adenoma, NOS
110	non-neoplastic or benign
286	Hepatocellular carcinoma, NOS
120	Retinoblastoma, NOS
171	Mantle cell lymphoma
180	Carcinoma, NOS
190	Multiple myeloma
141	Precursor B-cell lymphoblastic leukemia
133	Osteosarcoma, NOS
144	Adenocarcinoma, intestinal type
118	Leiomyosarcoma, NOS
126	Ependymoma, NOS
271	Neuroblastoma, NOS

Time requirements:
$O(N^3 R^2 C)$

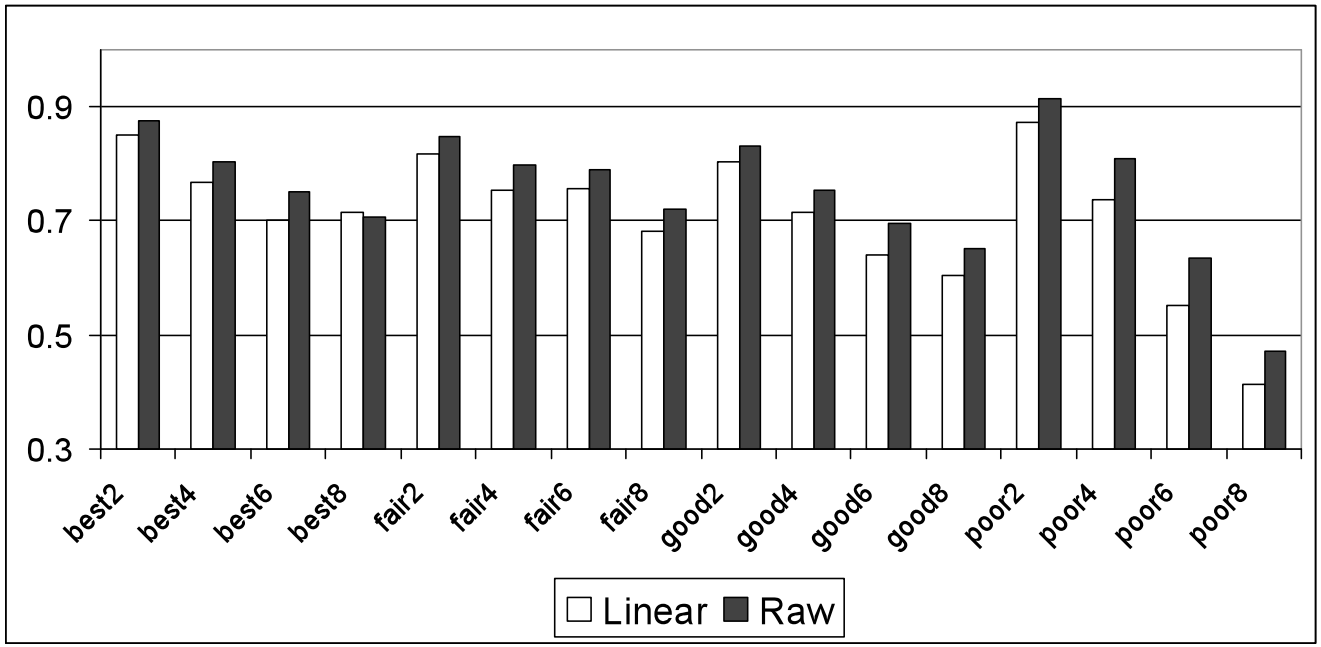
Datasets

	Similarity level			
#cancers	best	good	fair	poor
2	478	466	351	373
4	1160	790	800	800
6	1100	850	880	810
8	1000	830	750	760

Dataset size

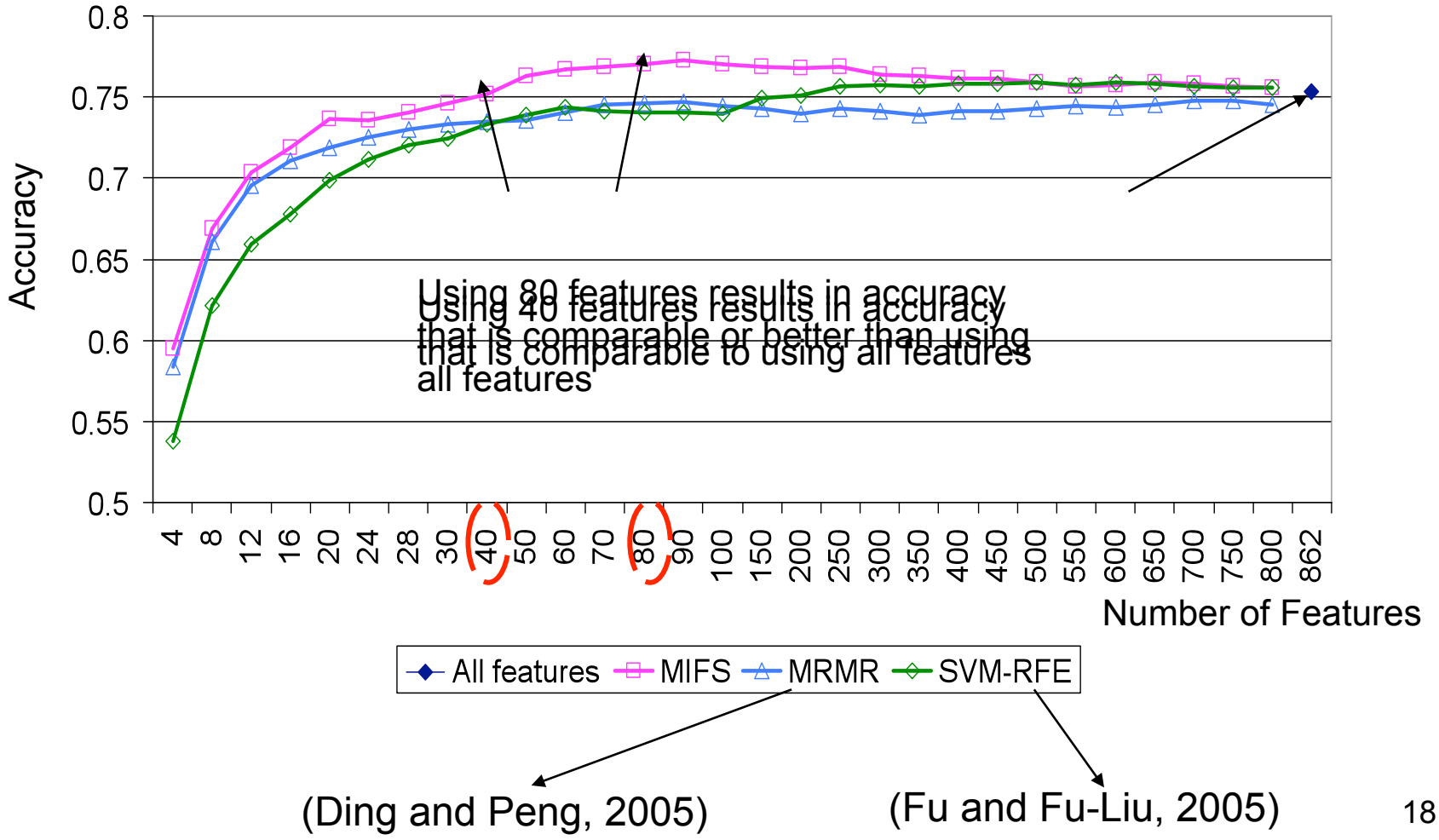
Experimental results

- Comparison of linear and Raw kernel



On average, Raw kernel improves the predictive accuracy by 6.4% over sixteen datasets compared to linear kernel.

Experimental results



Using MIFS for feature selection

- Result to test the hypothesis that 40 features are enough and 80 features are better

Dataset	Number of Features		
	40	80	862
newds1	0.801	0.792	0.799
newds2	0.803	0.819	0.8
newds3	0.629	0.67	0.637
newds4	0.706	0.748	0.719
Average	0.735	0.757	0.739

Feature Selection with Microarray Data

- Microarray is widely used to record gene expression.
- Due to thousands of genes, standard classification algorithms perform poorly on microarray data.
- Most existing feature selection algorithms depends on the microarray data only and ignores underlying biological process.
- So, the selected features are relatively redundant and prone to overfitting.
- Also, biologically, they may not be meaningful.

Biological Pathway based Feature Selection (BPFS)

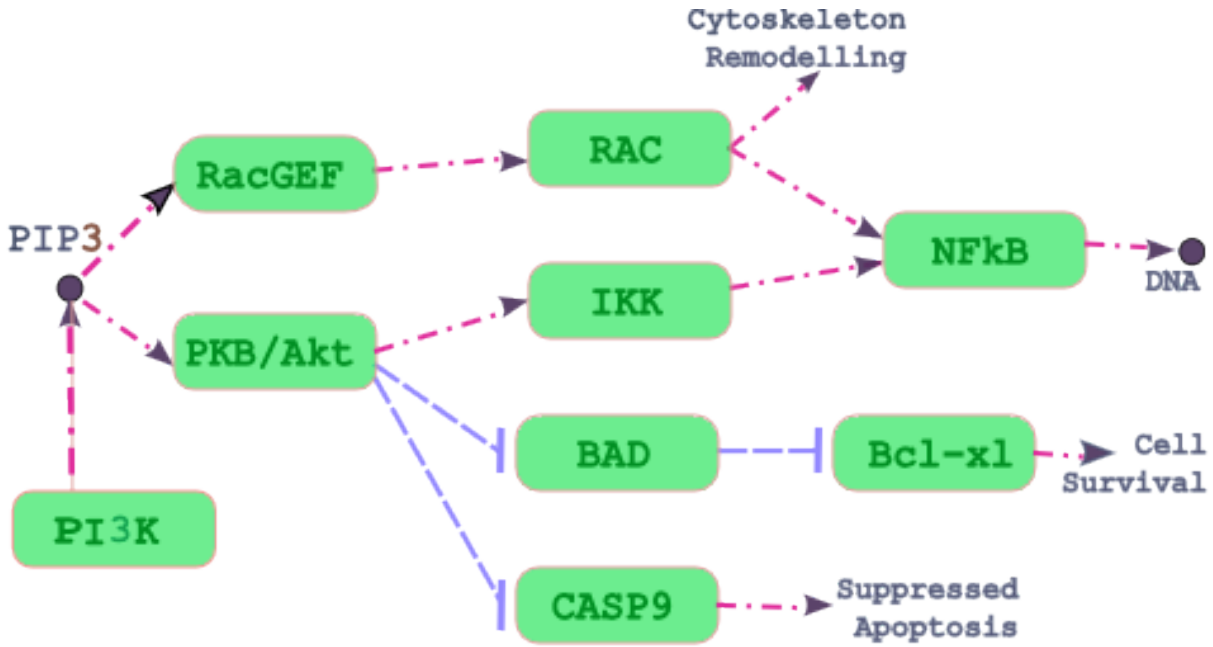
If two genes interacts on the pathway, they have functional dependency and correlated expression.

Selecting both of the is relatively redundant and doesn't improve accuracy significantly.

Algorithm:

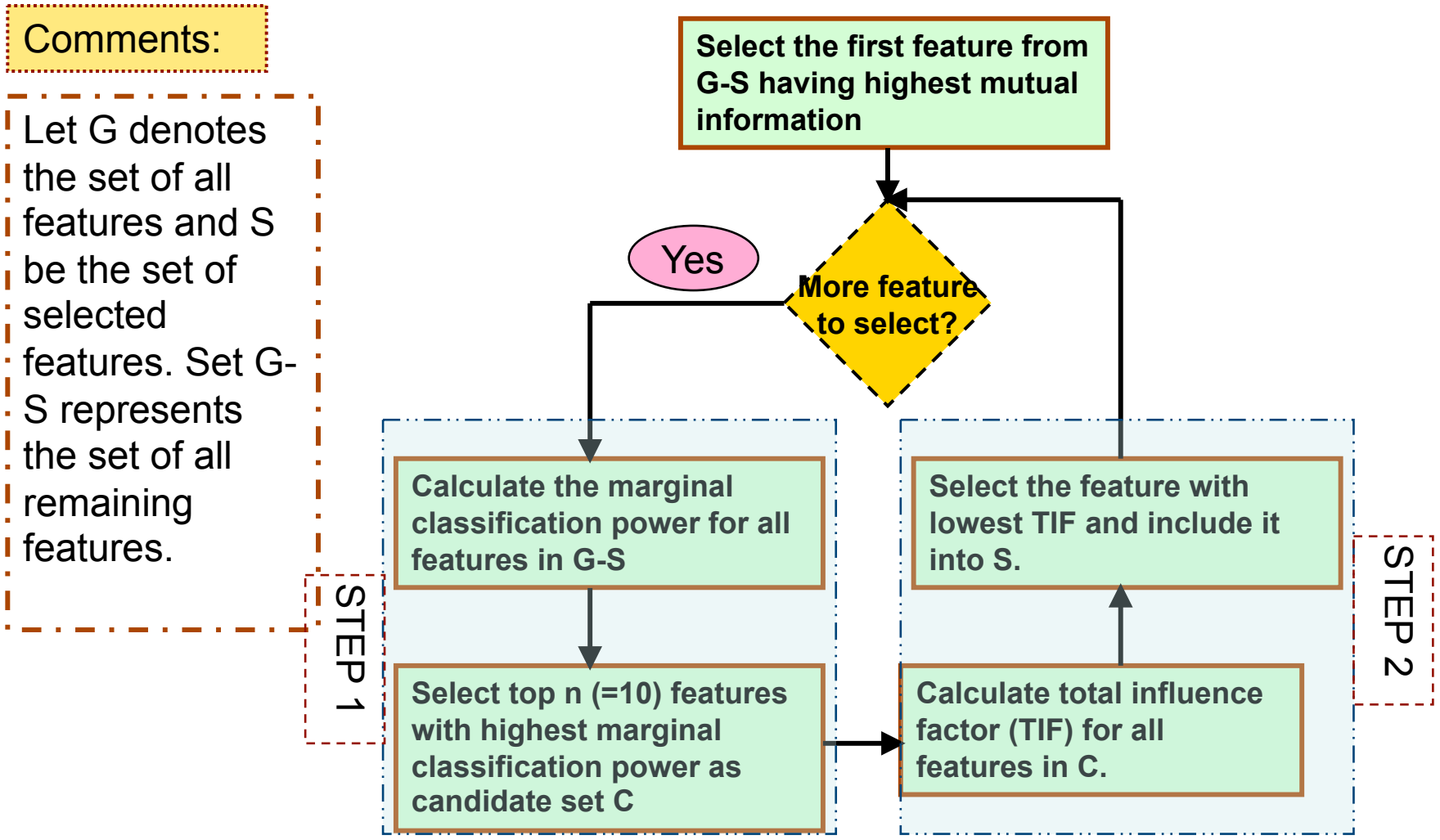
- Selects features that have high discriminating power and are relatively non redundant.
- Feature selection considers both microarray data and pathways.
- Encompasses the non-annotated genes in the microarray.

A Gene Regulatory Pathway



Part of Pancreatic Cancer Pathway adapted from KEGG showing regulatory relationships among genes. The green rectangles represent genes. The purple pointed arrows and blue blunt arrows represent activation and inhibition relationships respectively. For example, RacGEF activates RAC and BAD inhibits Bcl-xl.

High Level Overview of BPFS



Unresolved genes and feature set refinement

- Unresolved genes:
  - A good number of microarray entries (nearly 70%) don't have corresponding genes in the pathway.
  - As pathway interaction between unresolved genes is yet to be discovered, we modify TIF to encompass this.
- Feature Set refinement:
  - We repeat this for m (=50) times and get m (=50) individual rankings.
  - We consolidate these m rankings to produce the final feature set from the dataset.

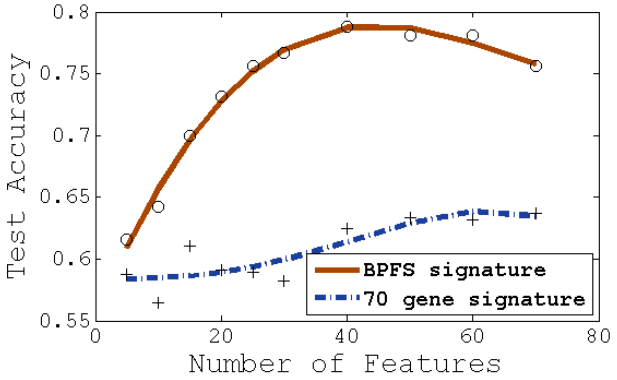
Experiments..

- BPFS was tested on five breast cancer dataset: Nature, JNCI, BCR, CCR and Lancet.
  - Biological Significance: Relevance of first twenty features from BCR was validated from literature. All of them are related to some kind of cancer.
  - Comparison with van ‘t Veer’s 70-gene signature: Improvement up to 18% over own dataset.
  - Classification accuracy: BPFS outperforms I-RELIEF with up to 8% more accuracy.
  - Cross Validation across datasets.

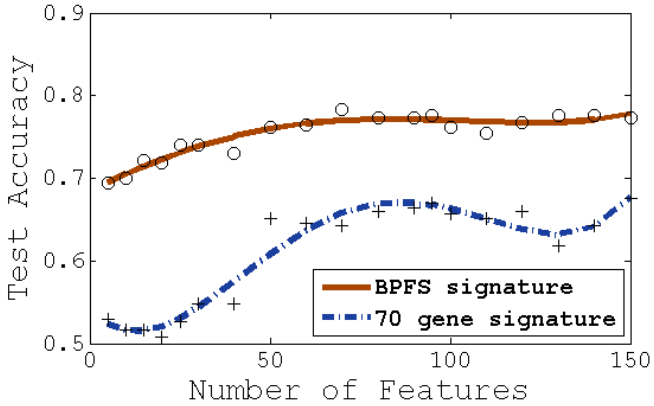
Biological Significance of Selected Features (First 10 features from BCR)

- KCNK2- Acute Lymphoblastic Leukemia
- ZNF222- Breast cancer
- P2RY2- Human lung epithelial tumor, Non-melanoma skin cancer, Thyroid cancer
- SLC2A6- Human Leukemia
- CD163- Breast Cancer, Human colorectal cancer
- HOXC13- Acute myeloid leukemia
- PCSK6- Breast Cancer, Ovarian cancer
- AQP9- Leukemia
- RUNX1- Gastric cancer, Ovarian cancer, Classical tumor suppressor gene
- KLRC4- KLRC4 is a member of the NKG2 group that are expressed primarily in natural killer (NK) cells

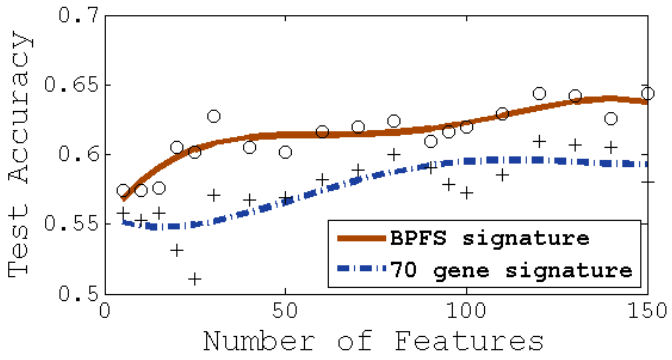
Comparison of Classification Accuracy to that of van ‘t Veer’s 70-gene Signature



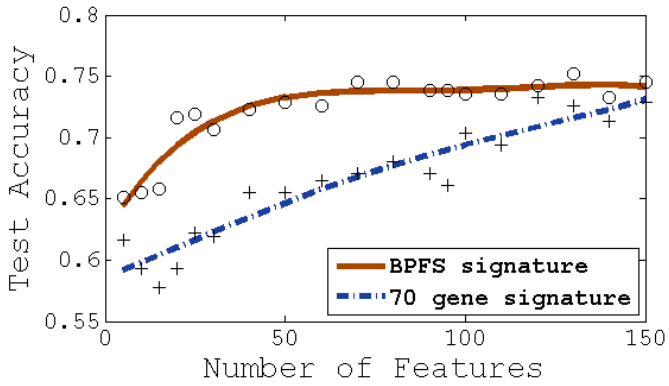
JNCI



CCR

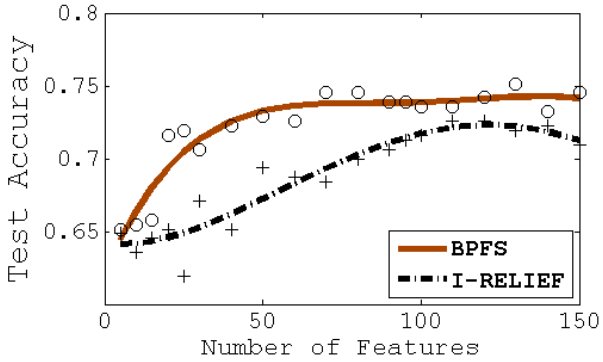


Lancet

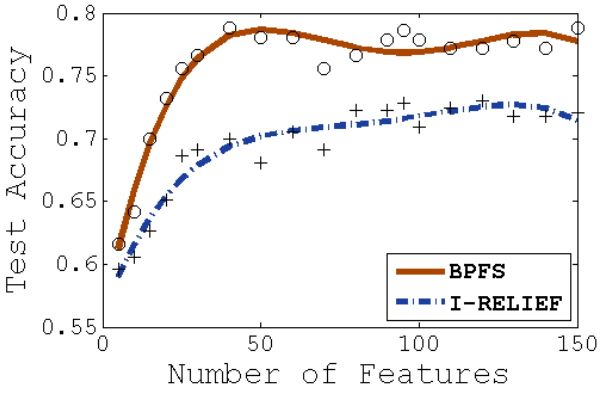


BCR

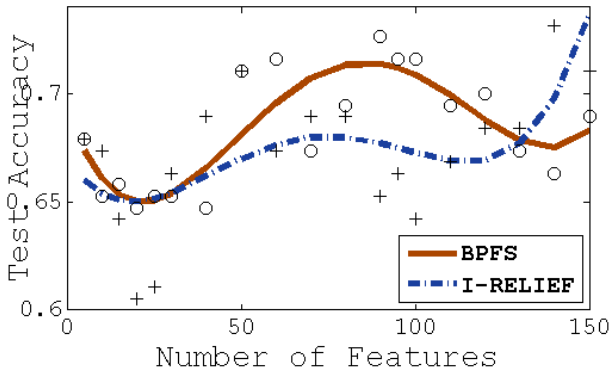
Comparison of Classification Accuracy to that of I-RELIEF



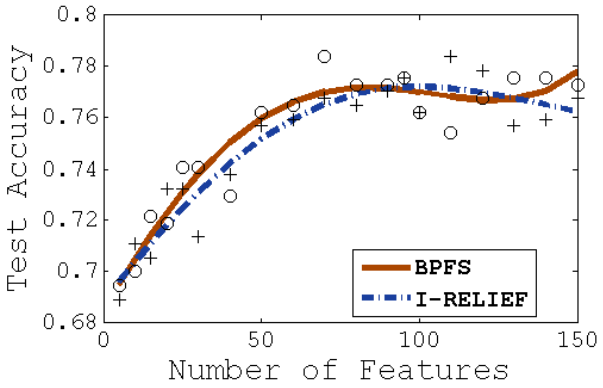
BCR



JNCI



Nature



CCR

Time Complexity

- Nunber of genes be p, the number of samples be n. X is the data matrix.
- Training with Support Vector Machine: SMO is an iterative optimization method. So calculating an average case bound is difficult. Let it be O(f(n,p)).
- Thus, for a single gene selection, time complexity is $O(f(n,p) + |C|*p^2)$ provided p > n. Assuming m features are selected, the overall complexity of BPFS is $O(m(f(n,p) +|C|* p^2))$.

Progression model for cancers

- Motivation
  - Helps explain known clinical and molecular evidence of cancers
  - Existing works focus on individual genetic event, not the pattern of aberrations
- Goal: Develop progression models for multiple cancers based on pattern of aberrations

Progression model for markers

O(DNRK) For each one of K cancers, identify R markers

C_1: (+53, -402, +576, +701)
C_2: (+54, -402, +612, -748)
C_3: (+53, -399, +576, -748)

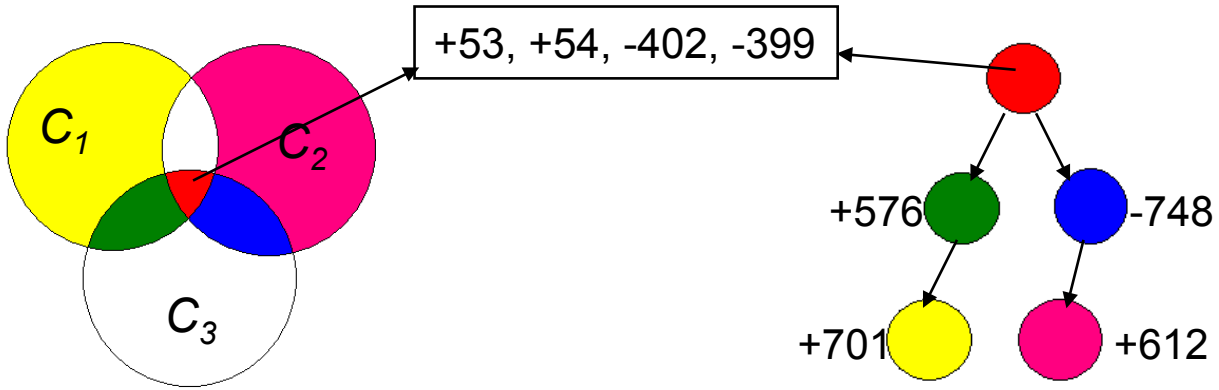
Step 1

O(NR²K²) *Shared Status*

markers	C_1	C_2	C_3
+53,+54	X	X	X
-402,-399	X	X	X
+576	X		X
-748		X	X
+612		X	
+701	X		

Step 2

Worst case *O(R²K²)*

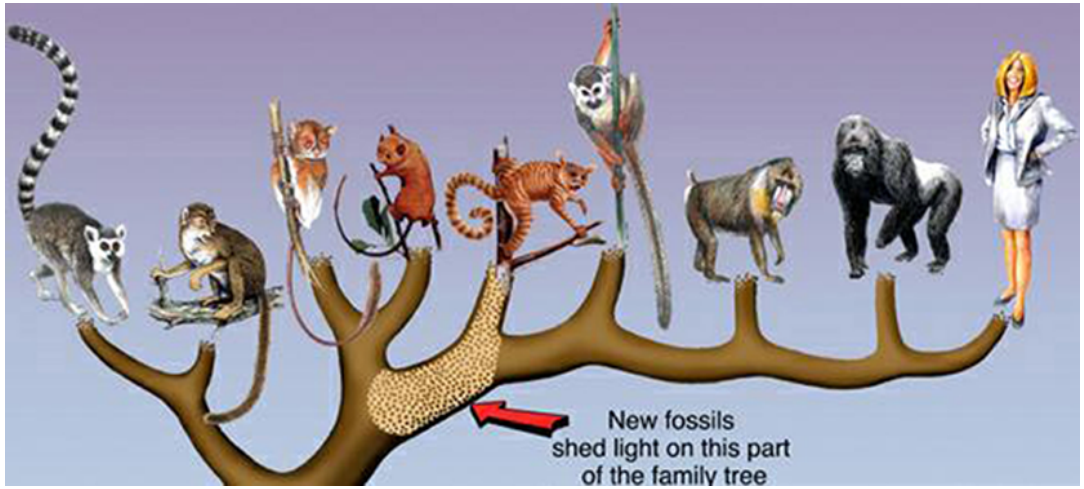


Step 3

Venn diagram Graph model

Progression model for cancers

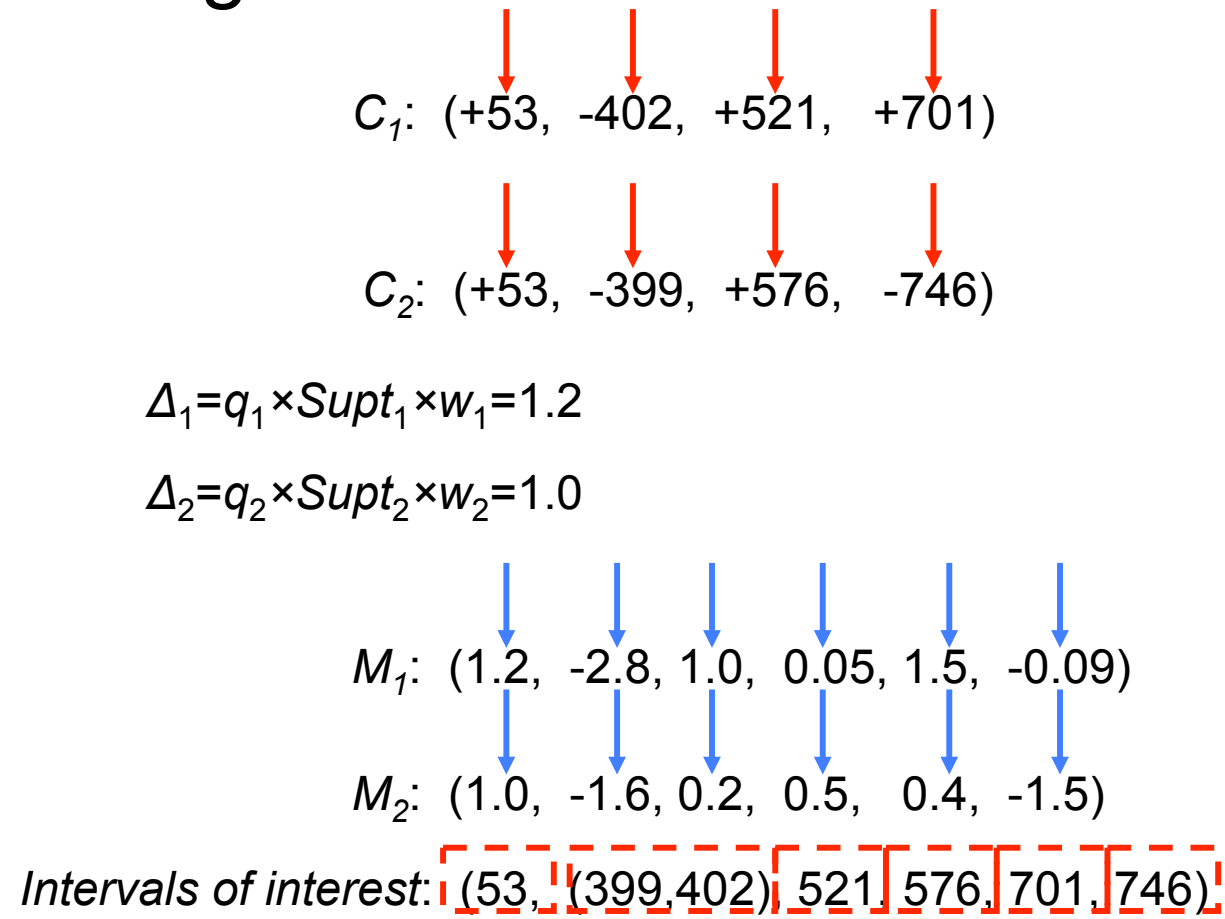
- What is a phylogenetic tree?



- Advantages
  - Make use of well-defined, pre-existing software packages
  - A large number of literature can guide us

Distance measure for cancers

- The alignment of markers



Distance measure of cancers

- Use extended Jaccard (EJ) similarity

$$EJ(M_1, M_2) = \frac{M_1 \cdot M_2}{\|M_1\|^2 + \|M_2\|^2 - M_1 \cdot M_2}$$

- Difference between EJ and cosine

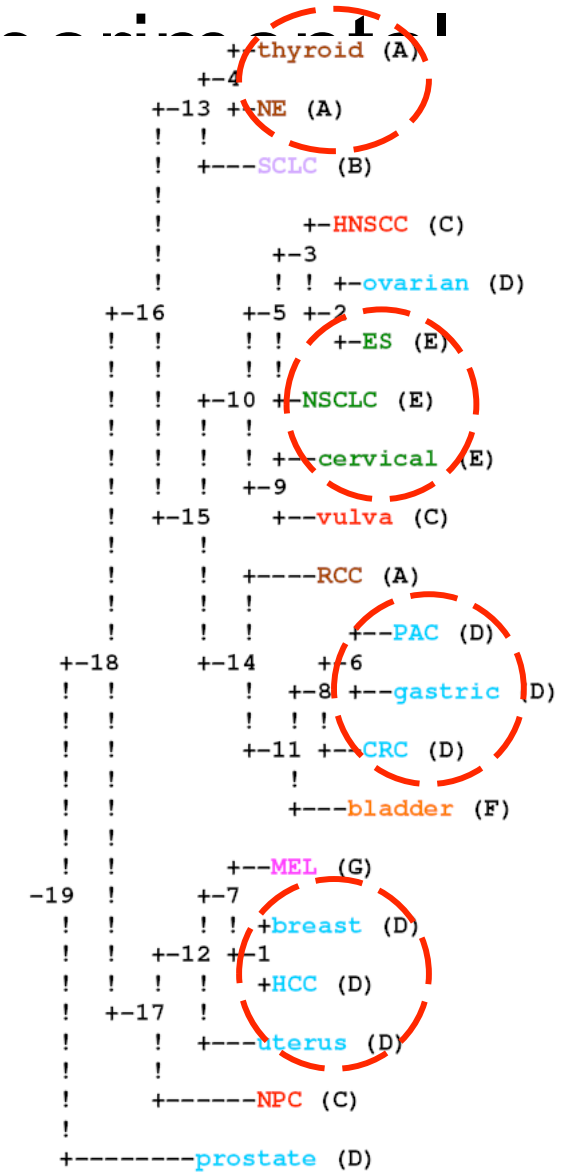
M_1: (0.1, 0.3) ⟶ *Cosine(M_1, M_2) = 1*

M_2: (0.2, 0.6) ⟶ *EJ(M_1, M_2) = 0.667*

- The distance between C_1 and C_2 is computed as
  - *$D(C_1, C_2) = 1 - EJ(M_1, M_2)$*

Ex

- Phylogenetic tree for cancers
  - 20 markers each c
- Use existing distance matrix based method UPGMA to generate the phylogenetic tree
- Cancers with the sam histological compositi are placed close toge in the tree.



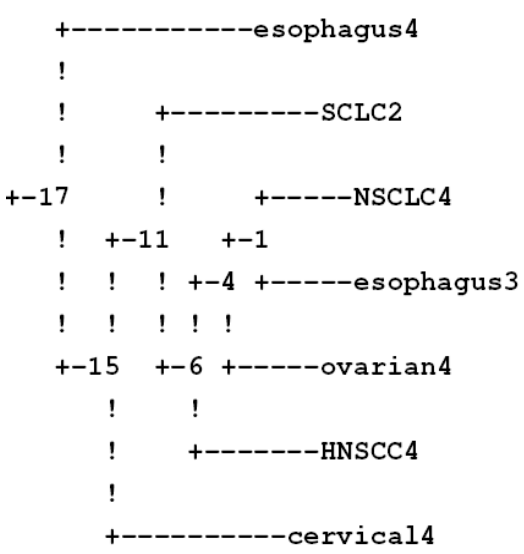
Coding and Legend

A: endocrine_and_clear
B: small_cell_neuroendo
C: squamous
D: adenocarcinomas
E: mixed squamous/adeno
F: transitional
G: melanoma

Experimental results

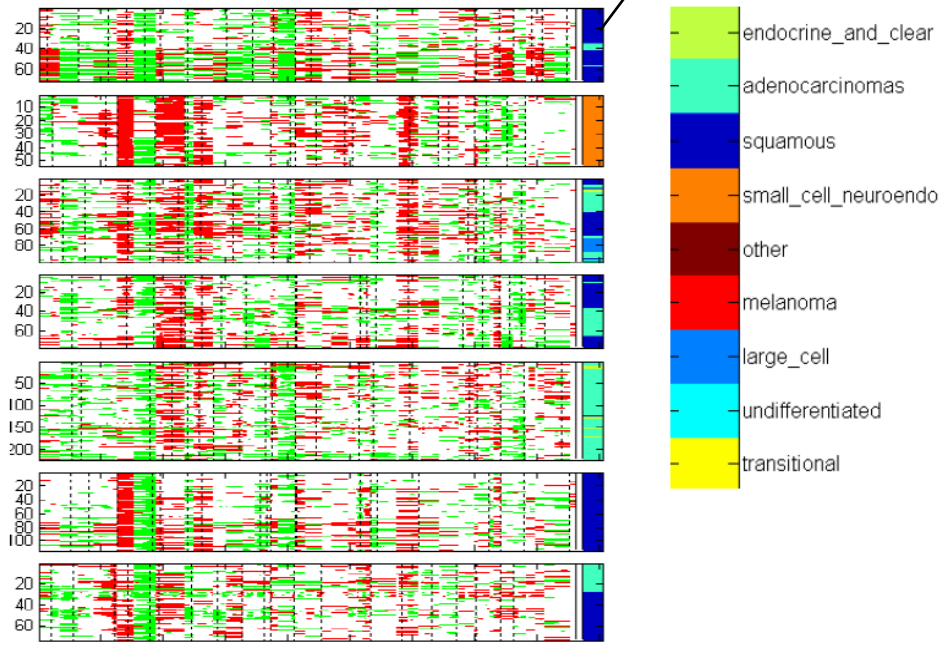
- Each cancer may contain multiple subtypes with heterogeneous aberration patterns
- Build a phylogenetic tree for 58 clusters of 20 cancers

A subtree of 7 clusters



Plots of aberrations

Color bar



Performance Issues and Future Work

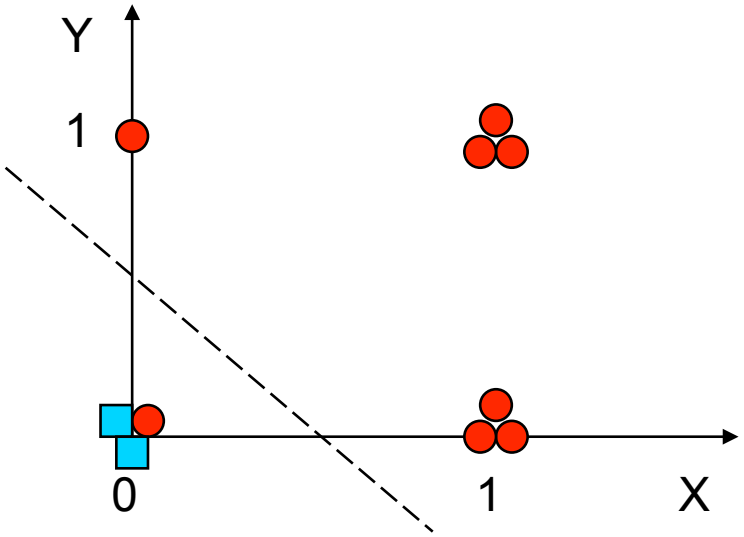
- Each run can take days for a single dataset (1000 samples)
- Scaling will require use of parallel machines
  - Increase the size of the datasets
  - Modeling temporal components

Thank you

Appendix

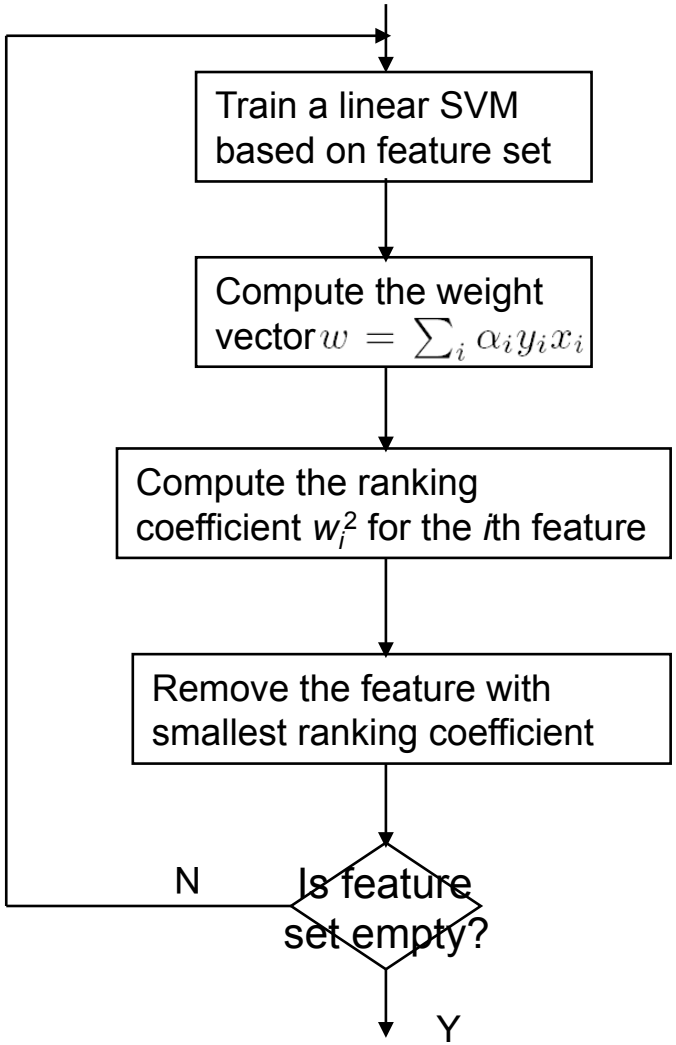
Minimum Redundancy and Maximum Relevance (MRMR)

- Relevance *V* is defined as the average mutual information between features and class labels
- Redundancy *W* is defined as the average mutual information between all pairs of features
- Incrementally select features by maximizing (*V* / *W*) or (*V* – *W*)



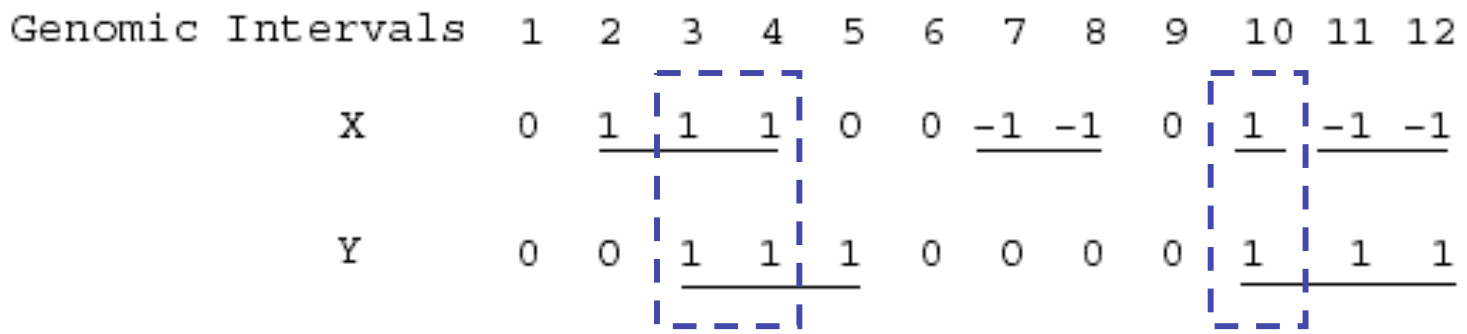
	Features 1	2	3	4	Class
x_1	0	1	1	0	1
x_2	0	1	1	0	1
x_3	0	1	1	0	1
x_4	0	0	0	1	1
x_5	0	0	0	0	-1
x_6	0	0	0	0	-1

Support Vector Machine Recursive Feature Elimination (SVM-RFE)



Pairwise similarity measures

- Sim measure
  - Segment is a contiguous block of aberrations of the same type.
  - Count the number of overlapping segment pairs.



Non-linear Decision Boundary

- How to generalize SVM when the two class classification problem is not linearly separable?
- Key idea: transform $\mathbf{x}_i$ to a higher dimensional space to "make life easier"
  - Input space: the space the point $\mathbf{x}_i$ are located
  - Feature space: the space of $\phi(\mathbf{x}_i)$ after transformation