Design of Clinical Decision Support Systems for Cancer based upon Clinical and Molecular Data

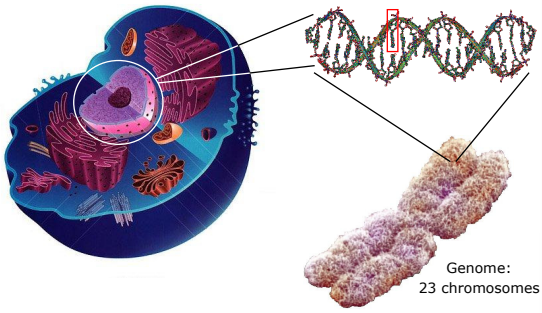
Anneleen Daemen
ESAT, Department of Electrical Engineering
Katholieke Universiteit Leuven, Belgium

PhD Defense
May 31, 2010
Leuven, Belgium

Cell biology

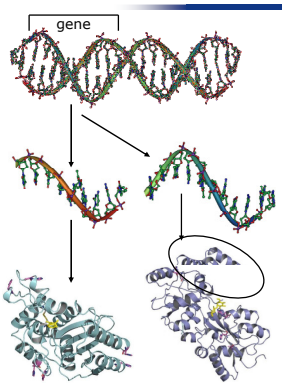
Human body: 100 x 10^{12} cells

DNA: 3.2 x 10^{9} nucleotides



Genome:
23 chromosomes

Central dogma



DNA – genomics (25.000 genes)

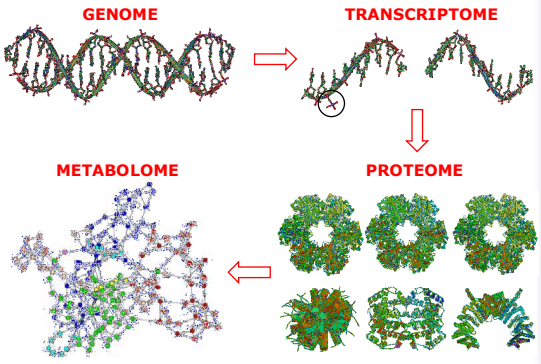
TRANSCRIPTION
alternative splicing

mRNA - transcriptomics

TRANSLATION
post-translational modifications

Protein – proteomics (10^{6} proteins)

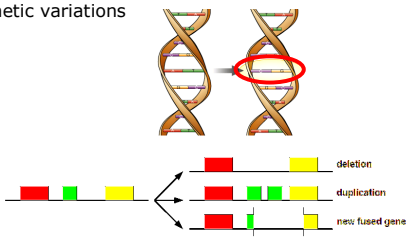
(epi)Genetics



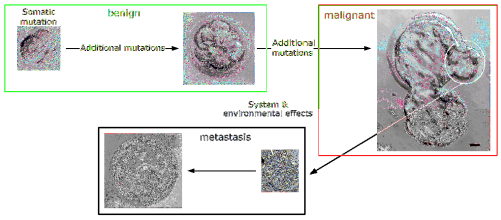
Cancer

- Genetic and epigenetic disease
- Incidence in Europe: 3.2 million
- Mortality in Europe: 1.7 million
- Responsible for 10% of medical care cost

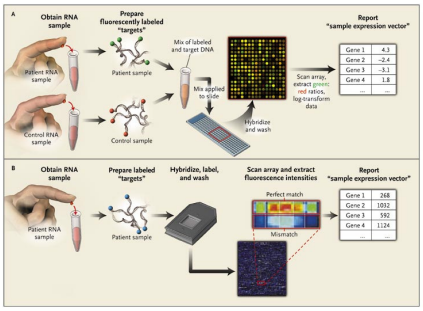
- Genetic variations



Cancer

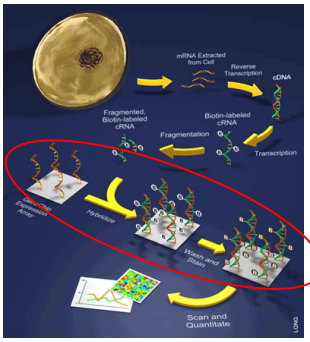


Microarray



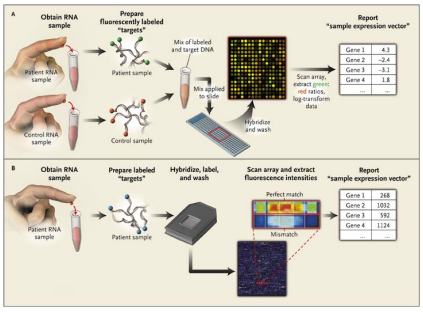
Quackenbush *et al.* (2006)

Microarray



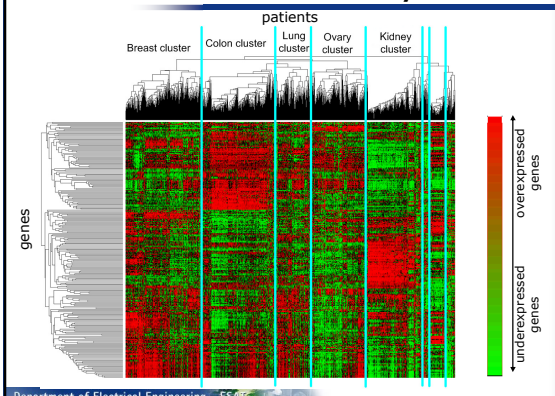
Coe *et al.* (2004)

Microarray

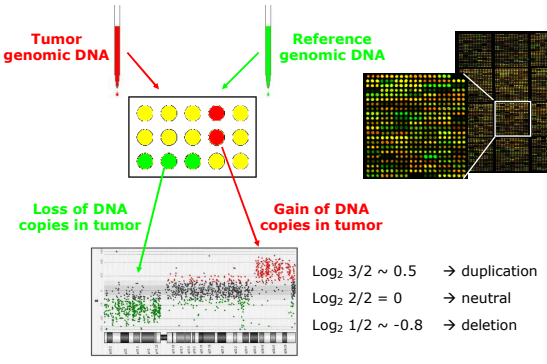


Quackenbush *et al.* (2006)

Microarray



Array CGH

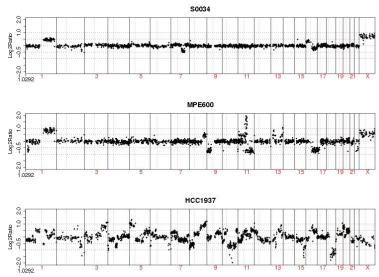


Log_2 3/2 ~ 0.5 → duplication

Log_2 2/2 = 0 → neutral

Log_2 1/2 ~ -0.8 → deletion

Array CGH



Select alterations in gene expression that favor tumor development

Clinical decision support

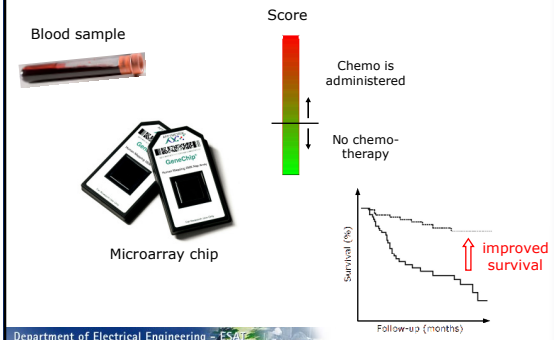
- Tsunami of data (multiple –ome levels)
- 4P medicine
  - Preventive
  - Predictive
  - Personalized
  - Participatory
- Decreasing cost-effectiveness of the health care system

➔ Clinical decision support systems

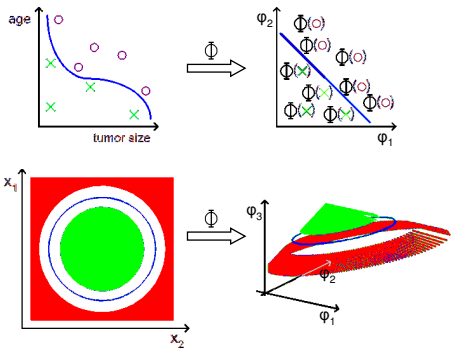
- To automate decisions based on domain knowledge and training data
- To improve speed, accuracy and reliability of diagnostic and prognostic tools
- To better select patients for therapy

Clinical decision support

- Example: gene signature



Kernel methods



Least Squares SVM

$$\min_{w,b,e} J_P(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} \varsigma_k e_k^2$$

subject to $y_k \left[w^T \varphi(x^k) + b \right] = 1 - e_k, \quad k = 1 \ldots N$

with $\varsigma_k = \begin{cases} N / 2N_P & \text{if } y_k = +1 \\ N / 2N_N & \text{if } y_k = -1 \end{cases}$

Kernel function

$$k(x_k, x_l) = \left\langle \Phi(x_k), \Phi(x_l) \right\rangle$$

$$k(x_k, x_l) = x_k^T x_l$$

$$k(x_k, x_l) = \left(x_k^T x_l + \tau \right)^d$$

$$k(x_k, x_l) = \exp\left(-\left\| x_k - x_l \right\|_2^2 / \sigma^2 \right)$$

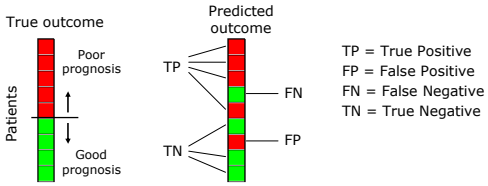
Feature selection

- Objectives
  - Exclusion of redundant & non-discriminatory features
  - Avoid overfitting
  - Improve model performance
  - Faster, more cost-effective models
- Additional layer of complexity

➔ Differential Expression via Distance Synthesis (DEDS)

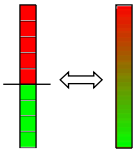
Yang *et al.* (2005)

Model selection



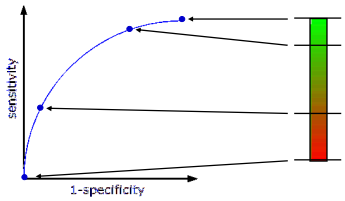
$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$



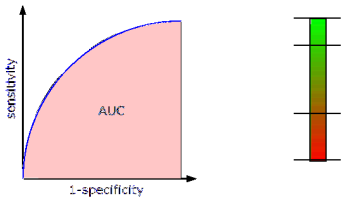
Model selection

Receiver Operating Characteristic Curve

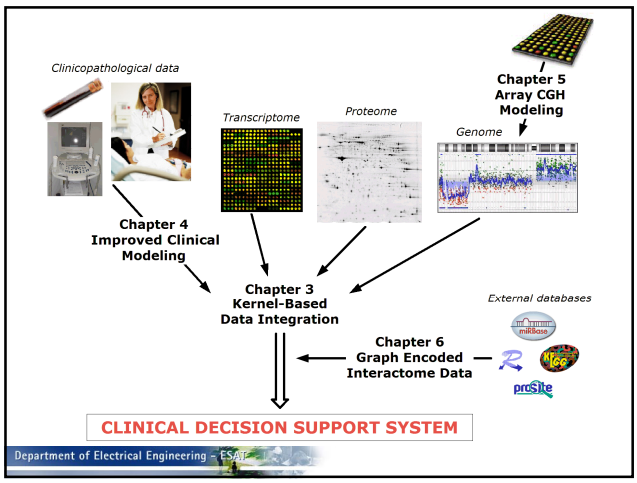


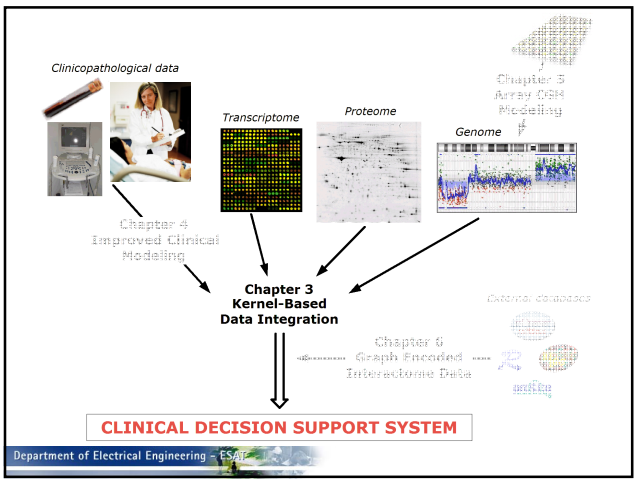
Model selection

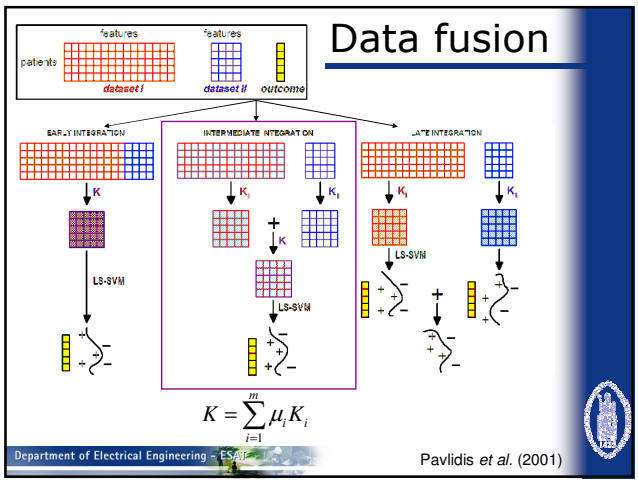
Receiver Operating Characteristic Curve

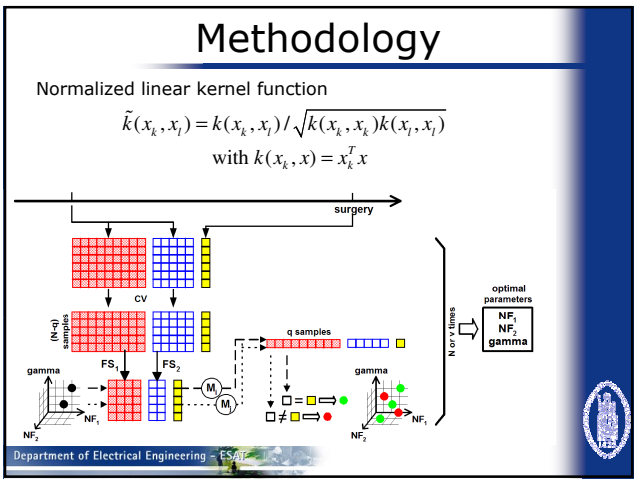


AUC = Area under the ROC curve


Clinicopathological data
Transcriptome
Proteome
Genome
Chapter 5
Array CGH
Modeling
Chapter 4
Improved Clinical
Modeling
Chapter 3
Kernel-Based
Data Integration
External databases
Chapter 6
Graph Encoded
Interactome Data
CLINICAL DECISION SUPPORT SYSTEM
Department of Electrical Engineering - ESAT

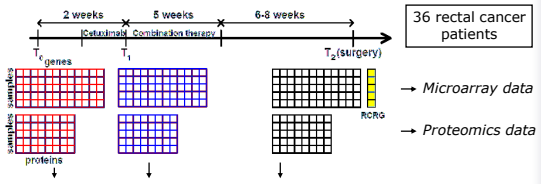

Clinicopathological data
Transcriptome
Proteome
Genome
Chapter 3
Kernel-Based
Data Integration
CLINICAL DECISION SUPPORT SYSTEM
Department of Electrical Engineering - ESAT


Data fusion
features
features
patients
dataset I
dataset II
outcome
EARLY INTEGRATION
INTERMEDIATE INTEGRATION
LATE INTEGRATION
LS-SVM
$K = \sum_{i=1}^{m} \mu_i K_i$
Department of Electrical Engineering - ESAT
Pavlidis et al. (2001)


Methodology
Normalized linear kernel function
$\tilde{k}(x_k, x_l) = k(x_k, x_l) / \sqrt{k(x_k, x_k) k(x_l, x_l)}$
with $k(x_k, x) = x_k^T x$
surgery
CV
q samples
gamma
optimal parameters
Department of Electrical Engineering - ESAT

Data Rectal cancer

Study Investigate the combination of cetuximab, capecitabine and radiotherapy in preoperative treatment of rectal cancer patients (Machiels *et al.* Ann Oncol 2007)



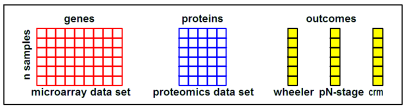
36 rectal cancer patients

→ *Microarray data*

→ *Proteomics data*

At 3 timepoints during therapy

Data Rectal cancer



Wheeler = tumor regression grade

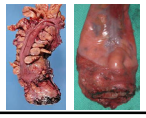
- Responder (26): good or total regression
- Nonresponder (10): no, minimal or moderate regression

pN-stage = number of lymph nodes found at surgery

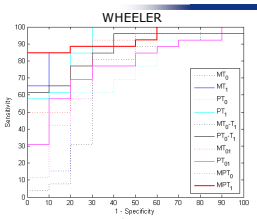
- Responder (22): no lymph nodes
- Nonresponder (14): ≥ 1 lymph node

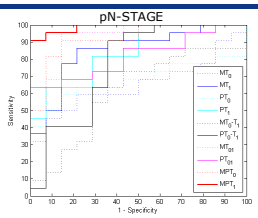
CRM (circumferential resection margin) = distance between tumor and mesorectal fascia

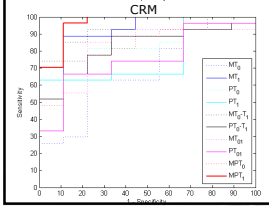
- Responder (27): > 2mm
- Nonresponder (9): ≤ 2mm



Results



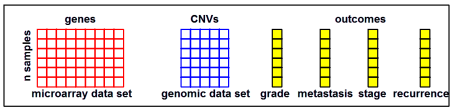




Both microarray and proteomics necessary due to complementarity

Limited number of genes and proteins, of which many related to (rectal) cancer

Data Prostate cancer



Publicly available data set on 55 primary prostate tumors
(Lapointe *et al.* PNAS 2004; Cancer Res 2007)

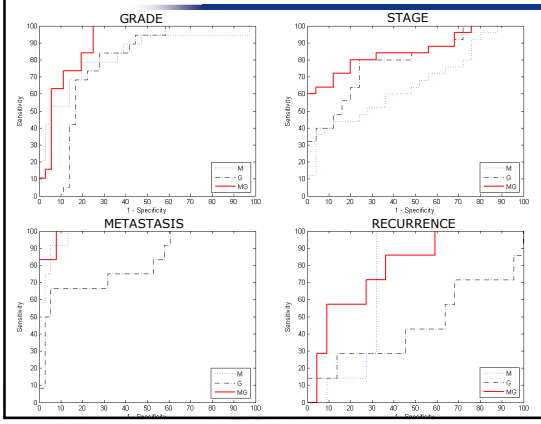
Data sources

- Microarray data (26.260 genes)
- DNA copy number variation data (22.279 CNVs)

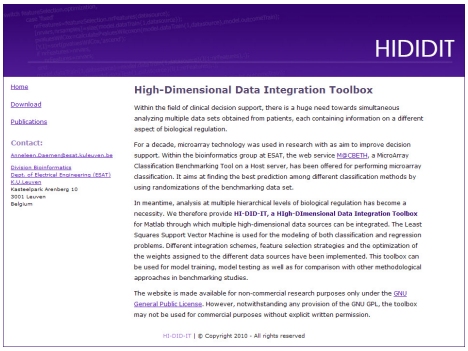
Outcomes

- Grade (36/19)
- Stage (25/25)
- Metastasis (38/12)
- Recurrence (22/7)

Results



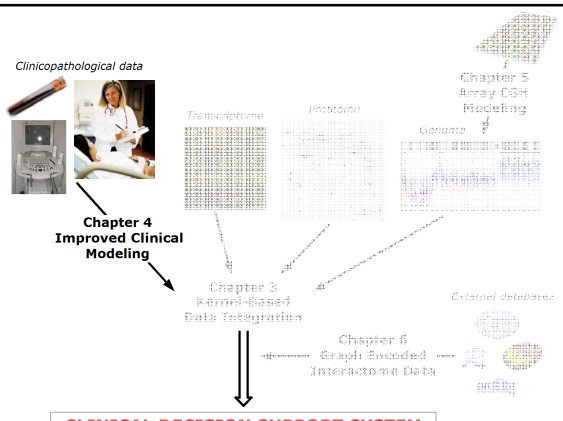
Toolbox



http://homes.esat.kuleuven.be/~bioiuser/HIDIDIT/index.php

Conclusions

- Integration of complementary data in the patient domain using kernel methods
- Improved decision support in cancer with limited number of variables
- Many features related to rectal cancer (e.g. *EGF-R, Cox-2, TGFa, MMP-2, TNFa*) or prostate cancer (e.g. *CXCL14, ERG, VAV3*)
- Multi-modal data should be gathered to ultimately obtain cost-efficient models
- Publications
  - Daemen *et al.* (2007) Integration of clinical and microarray data with kernel methods. *EMBC*, Lyon, France, 5411-5415 (6 citations).
  - Daemen *et al.* (2008) Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *PSB*, Kohala Coast, Hawaii, 166-177 (7 citations).
  - Daemen *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support. *Genome Med* **1**:39 (5 citations).
  - Debucquoy *et al.* (2009) Molecular response to cetuximab and efficacy of preoperative cetuximab-based chemoradiation in rectal cancer. *J Clin Oncol* **27**:2751-57 (12 citations).
  - Daemen *et al.* HI-DID-IT, a High-Dimensional Data Integration Toolbox for clinical applications. Submitted to *BMC Bioinf.*



Clinical kernel function

Linear kernel function: $k(i,j) = x^{iT} x^j$ with $x \in \mathbb{R}^p$

- variable type not taken into account
- inner product depends on the variable range
- different influence of variables on patient similarity
- dummy variables required for each nominal variable

Clinical additive kernel function:

- specifically developed for clinical data
- type and range of each variable taken into account
- only zero for most dissimilar patients

Clinical kernel function

Continuous & Ordinal variables:

$$k_z(i,j) = \frac{r - |z_i - z_j|}{r}$$

Nominal variables:

$$k_z(i,j) = \begin{cases} 1 & \text{if } z_i = z_j \\ 0 & \text{if } z_i \neq z_j \end{cases}$$

Final kernel for clinical data:

$$k(i,j) = \frac{1}{p}\sum_{z=1}^{p} k_z(i,j)$$

Polynomial kernel:

$$\left(x^{iT} x^j + \tau\right)^d \rightarrow \left(\frac{1}{p}\sum_{z=1}^{p} k_z(i,j) + \tau\right)^d$$

Gynecological data

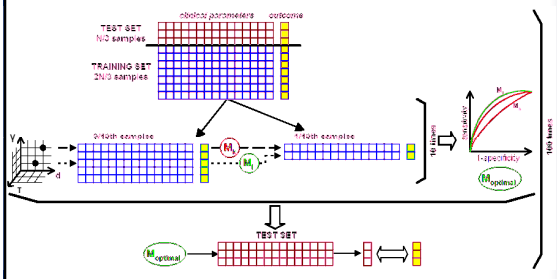
I. Endometrial disease: abnormal vs. normal
- 339 patients: 163/176
- 22 variables: 5C, 4O, 13N

II. First trimester pregnancy: miscarriage vs. normal
- 2356 pregnancies: 898/1458
- 18 variables: 1C, 8O, 9N

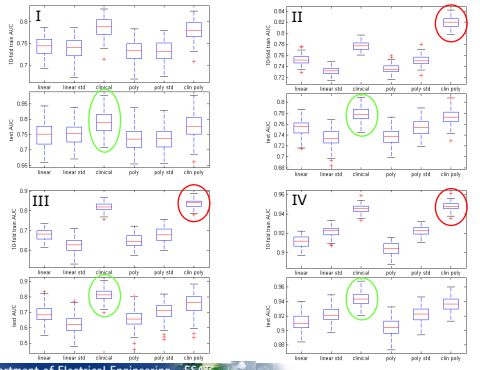
III. Pregnancy of unknown location: EP vs. failing PUL & IUP
- 856 PULs: 66/790
- 12 variables: 5C, 7N

IV. Adnexal mass: malignant vs. benign
- 1573 patients: 409/1164
- 15 variables: 3C, 2O, 10N

Methodology



Results



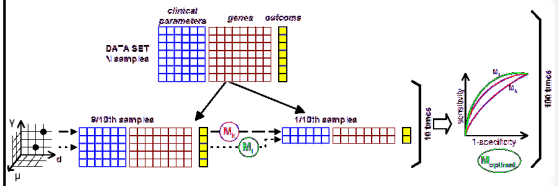
Breast cancer data

V. Recurrence: yes vs. no
- 110 patients: 25/85
- 12 variables: 2C, 3O, 7N

VI. Treatment response: residual vs. complete
- 129 patients: 96/33
- 8 variables: 1C, 3O, 4N

VII. Relapse: yes vs. no
- 177 patients: 65/112
- 5 variables: 2C, 3N

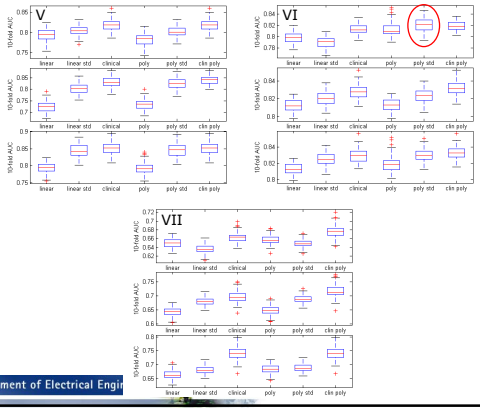
Methodology



Limited sample size: $\tau=1$

Three settings:
- 1 CL + 0 MA
- 0.5 CL + 0.5 MA
- μ CL + (1-μ) MA

Results

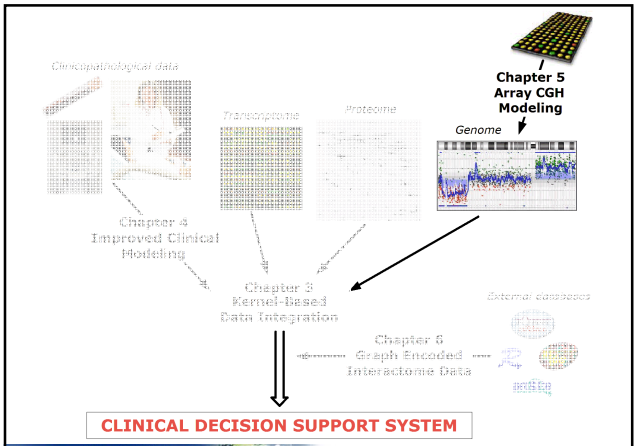


Conclusions

- Development of a clinical additive kernel function (both linear and non-linear)
- Type and range of each variable taken into account
- Each variable with same influence on patient similarity
- More accurate representation of patient similarity

- Improved results for clinical data and their combination with microarray data
- Similar results with SVM

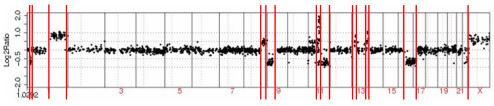
- Publications
  - Daemen *et al.* (2009) Development of a kernel function for clinical data. *EMBC*, Minneapolis, USA, 5913-5917 (1 citation).
  - Daemen *et al.* (2010) Improved modeling of clinical data with kernel methods. Revised manuscript submitted to *Artif Intell Med*.



Department of Electrical Engineering – ESAT

Hidden Markov Model

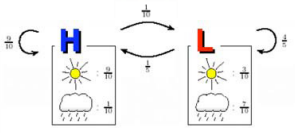
- Segmentation
  - Partition copy number profile into genomic regions of constant copy number
- Identification
  - Determine regions of copy number gain and loss
- Combination of both tasks
  - Hidden Markov Model



Department of Electrical Engineering – ESAT

Hidden Markov Model

- Hidden Markov Model
  - Hidden states
  - Observations
  - Initial probability of being in a state
  - Transition probabilities from 1 state to all the others

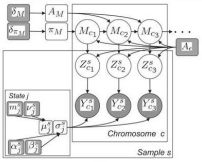


Department of Electrical Engineering – ESAT

Hidden Markov Model

- Hidden Markov Model
  - Hidden states = underlying copy number (loss, neutral, gain)
  - Observations = observed $\log_2$ ratio

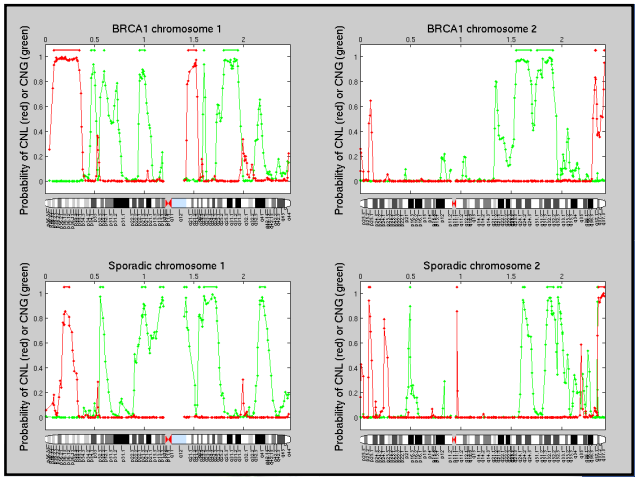
➔ Recurrent HMM of Shah *et al.* (2007)
  - Modeling of a group of samples
  - Statistical strength
  - Influence of noise
  - Individual clones

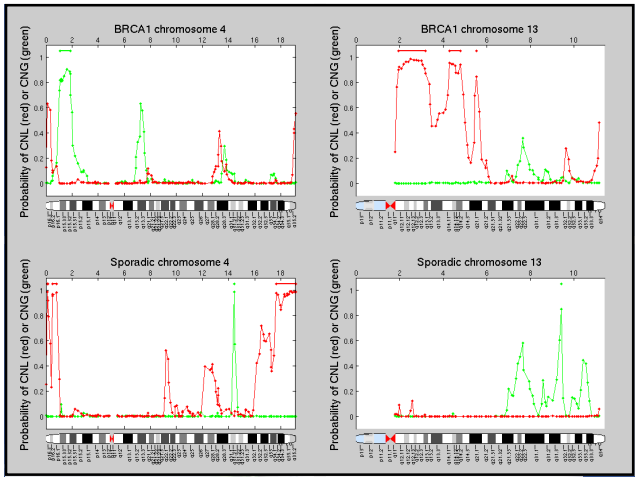


Data array CGH

Data set I: patients treated for ovarian cancer at University Hospital Leuven, Belgium (Leunen *et al.*, Hum Mut 2009)

- 8 sporadic samples
- 5 BRCA1 mutated samples
- 3.593 unique clones (CGH-SANGER 3K 7, Flanders Institute for Biotechnology, Leuven, Belgium)





Data array CGH

Data set I: patients treated for ovarian cancer at University Hospital Leuven, Belgium (Leunen *et al.*, Hum Mut 2009)

- 8 sporadic samples
- 5 BRCA1 mutated samples
- 3.593 unique clones (CGH-SANGER 3K 7, Flanders Institute for Biotechnology, Leuven, Belgium)

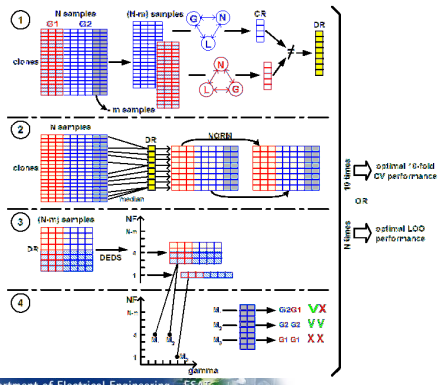
Data set II: oral squamous cell carcinoma Snijders *et al.* (2005)

- 59 samples wildtype for TP53
- 16 samples with a mutation for TP53
- 2.056 unique clones (HumArray2.0)

Data set III: non-small cell lung carcinoma Garnis *et al.* (2006)

- 13 adenocarcinoma
- 9 squamous cell carcinoma
- 29.781 unique clones (submegabase tiling array)

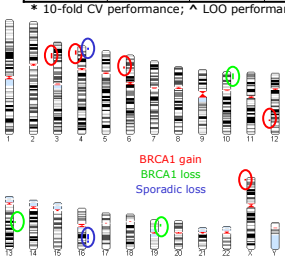
Methodology



Results

Data set	Nb regions	Accuracy	Sensitivity	Specificity	AUC
own data^	11	92.3 (12/13)	100 (5/5)	87.5 (7/8)	0.875
Snijders*	10	88 (66/75)	93.2 (55/59)	68.8 (11/16)	0.840
Garnis^	8	95.5 (21/22)	92.3 (12/13)	100 (9/9)	0.983

* 10-fold CV performance; ^ LOO performance



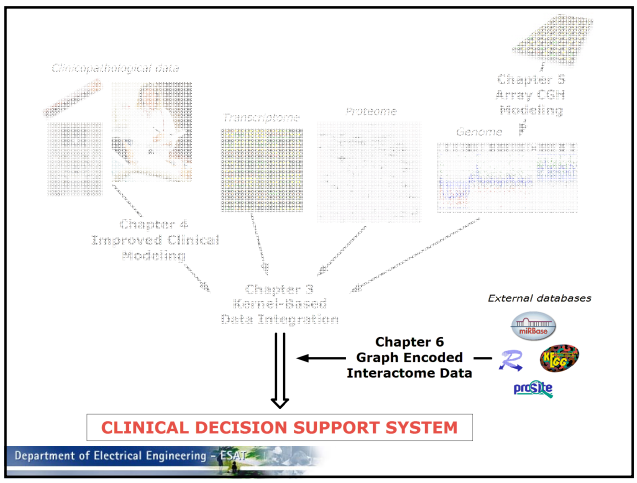
Region	Nb genes	Nb LOO iterations
1	0	8
2	5	11
3	0	9
4	22	11
5	24	10
6	32	5
7	66	7
8	81	4
9	39	4
10	86	4
11	36	6

Conclusions

- Many cancer studies: array CGH data for exploratory analysis
- Novel methodological approach: recurrent HMM and feature selection within classification setting
- Identification of class-specific aberrations
- Stability of the regions → robust

- Functional annotation analysis → oncogenes or tumor suppressor genes (*BAF57, HOXA5, LAMA3, CUTL1, FGF-10*)

- Publications
  - Daemen *et al.* (2008) Classification of sporadic and BRCA1 ovarian cancer based on a genome-wide study of copy number variations. *KES (Lecture Notes Comp Science)*, Zagreb, Croatia, 165-172.
  - Daemen *et al.* (2009) A genome-wide computational study of copy number variations: an example on ovarian cancer. Chapter 9 of *Investigating human cancer with computational intelligence techniques* (Vellido A, Lisboa P eds), 107-118.
  - Daemen *et al.* (2009) Supervised classification of array CGH data with HMM-based feature selection. *PSB*, Kohala Coast, Hawaii, 468-479.
  - Leunen et al. (2009) Recurrent copy number alterations in BRCA1-mutated ovarian tumors alter biological pathways. *Hum Mut* **30**:1693-1702.



Spectral graph theory

Prior biological knowledge → list of gene pairs
→ undirected graph G=(V,E)
- V = {genes}
- E = {gene regulation, protein interactions, etc.}

Adjacency matrix

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix}$$

Degree matrix

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

d_i = #neighbor nodes

Spectral graph theory

Laplacian matrix

$$L = D - A = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

$L^+ = G$ = Moore-Penrose pseudoinverse of L (Fouss *et al*, 2007)
= f (similarity between pairs of genes in the network)

$$L^+ = \begin{bmatrix} 0.55 & 0.21 & 0.08 & -0.32 & -0.52 \\ 0.21 & 0.54 & 0.08 & -0.32 & -0.52 \\ 0.08 & 0.08 & 0.28 & -0.12 & -0.32 \\ -0.32 & -0.32 & -0.12 & 0.48 & 0.28 \\ -0.52 & -0.52 & -0.32 & 0.28 & 1.08 \end{bmatrix}$$

→ For each gene, its neighborhood in the human interactome is taken into account

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways

Protein-protein interactions

Domain-domain interactions

Protein domains and families

Transcription factors

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways
- edge = genes/proteins belonging to same pathway

Protein-protein interactions

Domain-domain interactions

Protein domains and families

Transcription factors

KEGG HUMANCYC EHMN

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways

Protein-protein interactions
- edge = interacting proteins

Domain-domain interactions

Protein domains and families

Transcription factors

OPHID BioGRID STRING

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways

Protein-protein interactions

Domain-domain interactions
- edge = proteins interacting via a domain-domain interaction

Protein domains and families

Transcription factors

DOMINE Database of Protein Domain Interactions UniDomInt

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways

Protein-protein interactions

Domain-domain interactions

Protein domains and families
- edge = proteins with domains or families in common

Transcription factors

prosite Pfam

Secondary data sources

= knowledge in databases on different aspects of biological systems

Metabolic pathways

Protein-protein interactions

Domain-domain interactions

Protein domains and families

Transcription factors
- edge = genes targeted by the same miRNA

miRBase

microRNA.org

miRNAMap

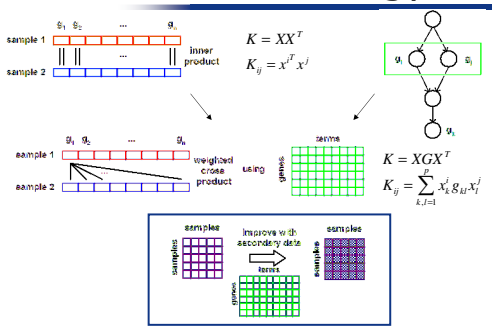
TargetScan Prediction of microRNA targets

Microarray data sets

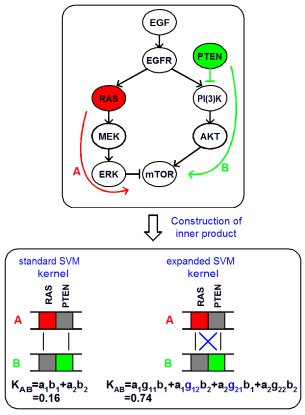
	Data set	Cancer type	Outcome	#samples (-/+)
T	Berchuck	ovarian	survival	53 (29/24)
	Hess	breast	pathologic response	133 (99/34)
	Ivshina	breast	local, regional or distant recurrence	249 (160/89)
	Pittman 1	breast	relapse	158 (95/63)
	Pittman 3	breast	distant metastasis	158 (108/50)
	Rosenwald	DLBCL	survival	220 (118/102)
	Singh	prostate	tumor status	102 (50/52)
	Sotiriou 1	breast	relapse	187 (139/40)
	Sotiriou 2	breast	distant metastasis	179 (139/40)
	Wang	breast	metastasis within 5 yrs	276 (183/93)
V	Bild	ovarian	survival	133 (88/45)
	Chin	breast	distant recurrence	129 (102/27)
	Huang 1	breast	disease recurrence	52 (34/18)
	Huang 2	breast	relapse	80 (53/27)
	Miller	breast	death from breast cancer	236 (181/55)
	Pittman 2	breast	loco-regional recurrence	158 (132/26)

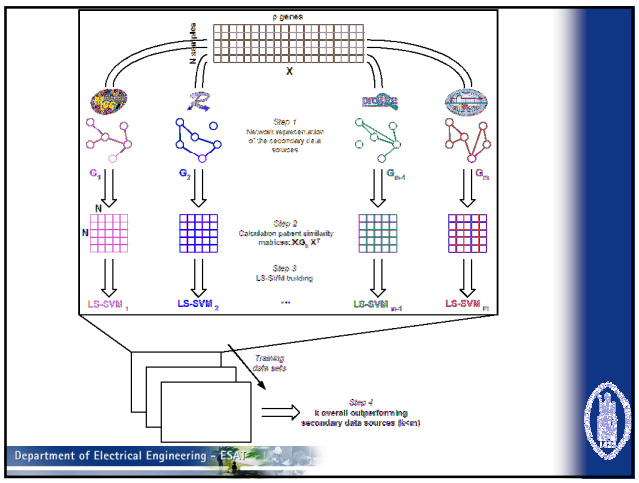
DLBCL = diffuse large-B-cell lymphoma
Affymetrix chips except for Rosenwald (lymphochip)

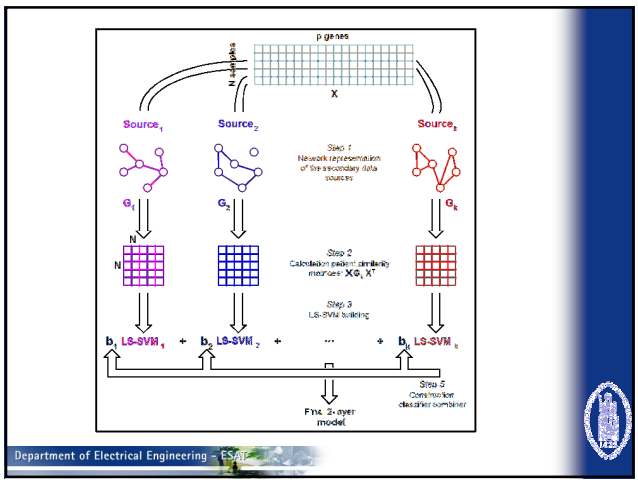
Methodology



➔ Each G-matrix exhaustively relates the gene expression profiles of multiple samples, weighted by its entries g_{kl} to obtain a more accurate patient similarity matrix

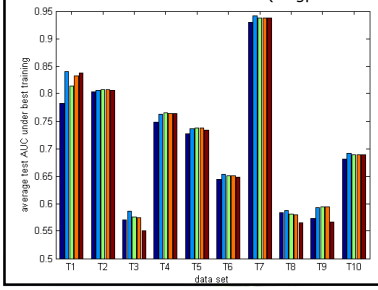






Training results

mean AUC of
- baseline model
- best individual secondary data source (-logp°: 0.7 – 24.82)
- best fixed combination rule (-logp°: 0.1 – 16.7)
- best trained combination rule (-logp°: 0.03 – 19.89)
- best advanced model (-logp°: 0 – 14.47)

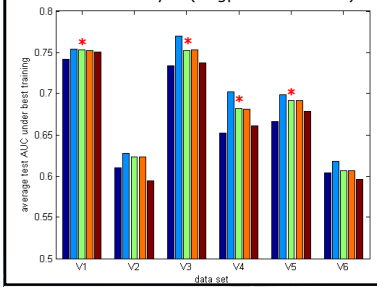


° Comparison of 200 AUC values between baseline model and all other models (one-sided paired-sampled t-test)

Overall (Wilcoxon)
individual: 0.002
fixed: 0.0039
trained: 0.0098
advanced: 0.557

Validation results

mean AUC of
- baseline model
- best individual secondary data source (-logp°: 1.03 – 16.35)
- mean rule (-logp°: 0.71 – 8.34)
- AUC weighting (-logp°: 0.73 – 8.21)
- naïve Bayes (-logp°: 0.02 – 5.47)

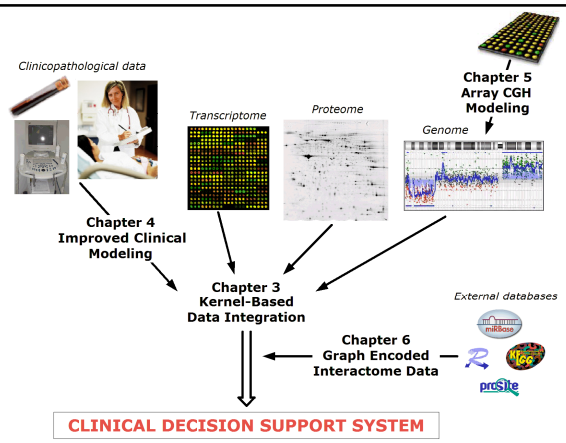


° Comparison of 200 AUC values between baseline model and all other models (one-sided paired-sampled t-test)

Overall (Wilcoxon)
individual: 0.0004
mean: 0.0005
AUC weighting: 0.001

Conclusions

- Improved decision making based on microarray data by incorporating the human interactome
- Interactome data encoded in a graph-based way
- Any type of gene-related info can be considered
- KEGG, OPHID and microRNA.org outperform other sources with regard to LS-SVM
- Mean rule for the prediction of the 3 corresponding models suffices
- Applicable to any kernel method, kernelizable method and in a general regression framework
- 2-layer approach essential
- Publications
  - Daemen *et al*. Improved microarray-based decision support with graph encoded gene-related data sources. *PLoS ONE* **5**(4): e10225 (2010) (1 citation).



Acknowledgements

Department of Electrical Engineering, KULeuven, Belgium

O. Gevaert – F. Ojeda – M. Signoretto – S. Yu – R. Van de Plas – B. Van Calster – J. Suykens – B. De Moor

University Hospital Gasthuisberg, KULeuven, Belgium

D. Timmerman – A. Pexsters – C. Holsbeke – T. Van den Bosch – K. Haustermans – A. Debucquoy – S. Tejpar – K. Leunen – E. Legius – A. Smeets – H. Wildiers – I. Vergote – E. Waelkens – I. Cadron – T. Van Gorp – J. van Pelt – C. Verslype – L. Libbrecht – H. Van Malenstein

St George's, University of London, UK

T. Bourne – C. Bottomley – A. Papageorghiou – E. Kirk – I. Sarris

Cambridge University Hospitals, NHS Trust, UK

C. Lees – O. Habayeb – Y. Abdallah – U. Hussain – A. Talmor

Erasmus University Medical Center, Rotterdam, the Netherlands

P. French – L. Gravendeel – A. Stubbs – J. de Rooi – P. van der Spek

Others

L. Valentin – D. Jurkovic – A. Testa – T. De Bie – J.-P. Machiels

Design of Clinical Decision Support Systems for Cancer based upon Clinical and Molecular Data

Anneleen Daemen
ESAT, Department of Electrical Engineering
Katholieke Universiteit Leuven, Belgium

PhD Defense
May 31, 2010
Leuven, Belgium