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

DNA: 3.2 x 10^9 nucleotides

Human body: 100 x 10^13 cells

Cell: 23 chromosomes

Central dogma

DNA → genomics
(25,000 genes)

mRNA → transcriptomics

Alternative splicing

TRANSCRIPTION
post-translational modifications

Protein → proteomics
(10^6 proteins)

(epi)Genetics

GENOME → TRANSCRIPTOME

METABOLOME → PROTEOME
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

Microarray

Quackenbush et al. (2006)
Microarray

Array CGH

Tumor genomic DNA

Reference genomic DNA

Gain of DNA copies in tumor

Loss of DNA copies in tumor

Log₂(3/2) ~ 0.5 → duplication
Log₂(2/2) = 0 → neutral
Log₂(1/2) ~ -0.8 → deletion

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

Kernel methods

Least Squares SVM

\[
\min_{w, b, \xi} J(w, \xi) = \frac{1}{2} w^T w + \gamma \sum_{i=1}^{N} \xi_i^2
\]

subject to

\[
y_i \left[ w^T \phi(x_i) + b \right] \geq 1 - \xi_i, \quad k = 1 \ldots N
\]

with

\[
\xi_i = \begin{cases} 
N/2N_x & \text{if } y_i = 1 \\
N/2N_y & \text{if } y_i = -1 
\end{cases}
\]

Kernel function

\[
\begin{align*}
    k(x_i, x_j) &= \langle \phi(x_i), \phi(x_j) \rangle \\
    k(x_i, x_j) &= x_i^T x_j \\
    k(x_i, x_j) &= (x_i^T x_j + \tau)^d \\
    k(x_i, x_j) &= \exp\left(-\frac{|x_i - x_j|^2}{\sigma^2}\right)
\end{align*}
\]
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)

Model selection

True outcome
- Poor prognosis
- Good prognosis

Predicted outcome
- True Positive (TP)
- False Positive (FP)
- False Negative (FN)
- True Negative (TN)

Sensitivity = \(\frac{TP}{TP + FN}\)
Specificity = \(\frac{TN}{TN + FP}\)

Receiver Operating Characteristic Curve

AUC = Area under the ROC curve

AUC = Area under the ROC curve

AUC = Area under the ROC curve
Methodology

Normalized linear kernel function

\[ \hat{k}(x_i, x_j) = \frac{k(x_i, x_j)}{\sqrt{k(x_i, x_i)k(x_j, x_j)}} \]

with \( k(x_i, x_j) = x_i^T \mathbf{x}_j \)

Data fusion

\[ K = \sum_{j=1}^{n} \mu_j K_j \]
**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)

![Graph showing 36 rectal cancer patients at 3 timepoints during therapy with Microarray data and Proteomics data](image)

**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

**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**

Both microarray and proteomics necessary due to complementarity

Limited number of genes and proteins, of which many related to (rectal) cancer
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, TGFα, MMP-2, TNFα) or prostate cancer (e.g. CXCL14, ERG, VAV2)
- 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. ZNAC, Lyon, France, 5411-5415 (6 citations).
- Daemen et al. (2008), Improved decision support in cancer with limited number of variables. ZNAC, Lyon, France, 5411-5415 (6 citations).
- Daemen et al. (2009), Improved decision support in cancer with limited number of variables. ZNAC, Lyon, France, 5411-5415 (6 citations).
- Daemen et al. (2010), Integration of clinical and microarray data with kernel methods. ZNAC, Lyon, France, 5411-5415 (6 citations).
- Daemen et al. (2011), Improved decision support in cancer with limited number of variables. ZNAC, Lyon, France, 5411-5415 (6 citations).
- Daemen et al. (2012), A high-dimensional data integration toolbox for clinical applications. Submitted to ZNAC, Lyon, France, 5411-5415 (6 citations).
**Clinical kernel function**

- **Linear kernel function:** \( k(i, j) = x^T \cdot x' \) 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

---

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

---

**Methodology**

- **Continuous & Ordinal variables:**
  \[ k_i(i, j) = e^{-\frac{1}{\epsilon}(z_i - z_j)^2} \]

- **Nominal variables:**
  \[ k_{ij}(i, j) = \begin{cases} 1 & \text{if } z_i = z_j \\ 0 & \text{if } z_i \neq z_j \end{cases} \]

- **Final kernel for clinical data:**
  \[ k(i, j) = \frac{1}{p} \sum_{i \neq j} k_i(i, j) \]

- **Polynomial kernel:**
  \[ (x^T \cdot x')^p \rightarrow \left( \frac{1}{p} \sum_{i \neq j} k_i(i, j) \right)^p \]
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. Release: yes vs. no
- 177 patients: 65/112
- 5 variables: 2C, 3N

Limited sample size: 1=1
Three settings:
- 1 CL + 0 MA
- 0.5 CL + 0.5 MA
- μ CL + (1-μ) MA
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).

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

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

- Hidden Markov Model
  - Hidden states = underlying copy number (loss, neutral, gain)
  - Observations = observed log₂ 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 3K7, 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

<table>
<thead>
<tr>
<th>Data set</th>
<th>Nb regions</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leunen data</td>
<td>13</td>
<td>0.517/0.713</td>
<td>0.95/0.972</td>
<td>0.875/0.962</td>
<td>0.959</td>
</tr>
<tr>
<td>Snijders*</td>
<td>10</td>
<td>0.56/0.710</td>
<td>0.92/0.954</td>
<td>0.864/0.962</td>
<td>0.884</td>
</tr>
<tr>
<td>Garnis*</td>
<td>6</td>
<td>0.953/0.972</td>
<td>0.923/0.972</td>
<td>0.90/0.972</td>
<td>0.983</td>
</tr>
</tbody>
</table>

* 5-fold CV performance, *LOO performance

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 (BAC57, HOXA5, LAMA3, CUTF1, FGF-10)

Publications
- Daemen et al. (2009) Classification of sporadic and BRCA1 ovarian cancer based on a genome-wide study of copy number variations. KBS (Lecture Notes Comp Science), Sagiya, Chiesa, 162-172.
- Daemen et al. (2009) A genome-wide computational study of copy number variations: an approach to improve the classification of sporadic ovarian cancer. KBS (Lecture Notes Comp Science), Sagiya, Chiesa, 162-172.
Spectral graph theory

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

\[
A = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 \\
0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\]

Degree matrix \( D = \begin{bmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix} \)

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

<table>
<thead>
<tr>
<th>Metabolic pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = genes/proteins belonging to same pathway</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein-protein interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = interacting proteins</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain-domain interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = proteins interacting via a domain-domain interaction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein domains and families</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Transcription factors</th>
</tr>
</thead>
</table>

- **DOMINE**
- **UniDomInt**
- **OPHID**
- **BioGRID**
- **STRING**
- **proSite**
- **Pfam**

---

### Secondary data sources

= knowledge in databases on different aspects of biological systems

<table>
<thead>
<tr>
<th>Metabolic pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = genes/proteins belonging to same pathway</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein-protein interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = interacting proteins</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain-domain interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge = proteins interacting via a domain-domain interaction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein domains and families</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Transcription factors</th>
</tr>
</thead>
</table>

- **DOMINE**
- **UniDomInt**
- **OPHID**
- **BioGRID**
- **STRING**
- **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

Microarray data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Cancer type</th>
<th>Outcome</th>
<th>#samples (+/-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>breast</td>
<td>survival</td>
<td>53 (26/27)</td>
</tr>
<tr>
<td>Ivshina</td>
<td>breast</td>
<td>pathologic response</td>
<td>133 (96/34)</td>
</tr>
<tr>
<td>Pitman 1</td>
<td>breast</td>
<td>relapse</td>
<td>158 (90/68)</td>
</tr>
<tr>
<td>Pitman 2</td>
<td>breast</td>
<td>distant metastasis</td>
<td>158 (118/40)</td>
</tr>
<tr>
<td>Rosenwald</td>
<td>DLBCL</td>
<td>survival</td>
<td>220 (118/102)</td>
</tr>
<tr>
<td>Singh</td>
<td>prostate</td>
<td>tumor status</td>
<td>102 (50/52)</td>
</tr>
<tr>
<td>Sotiriou 1</td>
<td>breast</td>
<td>relapse</td>
<td>187 (136/50)</td>
</tr>
<tr>
<td>Sotiriou 2</td>
<td>breast</td>
<td>distant metastasis</td>
<td>179 (130/40)</td>
</tr>
<tr>
<td>Wang</td>
<td>breast</td>
<td>metastasis within 5 yrs</td>
<td>276 (182/92)</td>
</tr>
<tr>
<td>V</td>
<td>blood</td>
<td>survival</td>
<td>133 (88/45)</td>
</tr>
<tr>
<td>Chin</td>
<td>breast</td>
<td>distant recurrence</td>
<td>129 (102/27)</td>
</tr>
<tr>
<td>Huang 1</td>
<td>breast</td>
<td>disease recurrence</td>
<td>52 (24/18)</td>
</tr>
<tr>
<td>Huang 2</td>
<td>breast</td>
<td>relapse</td>
<td>80 (33/27)</td>
</tr>
<tr>
<td>Miller</td>
<td>breast</td>
<td>death from breast cancer</td>
<td>236 (181/55)</td>
</tr>
<tr>
<td>Pittman 2</td>
<td>breast</td>
<td>loco-regional recurrence</td>
<td>138 (122/55)</td>
</tr>
</tbody>
</table>

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

Methodology

Each G-matrix exhaustively relates the gene expression profiles of multiple samples, weighted by its entries \( p_{ij} \) 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)

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)
  - naive Bayes (-logp: 0.02 – 5.47)

Overall difference:
- individual: 0.004
- fixed: 0.0039
- trained: 0.0098
- advanced: 0.557

Overall difference:
- 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