Knowledge Discovery in Neuroblastoma-related Biological Data

Using Gene Locus Information

Edwin van de Koppel

Cognitive Artificial Intelligence

Outline

- Neuroblastoma: a type of cancer
- Data Sets
- Safarii Multi-Relational Data Mining Tool
- Single Table Analyses
- Aggregation & Integration: combining knowledge from different data sets
- Conclusion

European Embryonal Tumour Pipeline

- EU-funded research project
- Goal: improve treatment of certain types of cancer affecting small infants
- Of all tumour types, neuroblastoma is has the most complete data sets

The Disease

- One of the most common childhood cancers (15% of cancer deaths in children)
- Usually starts during the development of the adrenal glands
- 88% of the patients is 5 years or younger
- Ideal to study genetically based changes leading to cancer, due to the disease's manifestation in early life

Stage

- *Stage* based on physiological characteristics
- Can be diagnosed relatively easy by doctors
- Does not necessarily have a direct genetic basis

Stages

- 1: Localised tumour, confined to area of origin
- 2A: Tumour not fully removable by surgery, no lymph nodes infected
- 2B: Infected lymph nodes on one side of the body
- 3: Tumour spread across body's midline
- 4: Tumour spread to distant lymph nodes, bone (marrow) and organs (survival: $<35\%$)
- 4S: Stage 4 like symptoms that often spontaneously regress

Clinical Course - *NBstatus*

Describes the patient's health status up to 5 years after treatment

NBstatus

- 'alive without event'
- 'alive with relapse/ primary tumour': Cancer reappeared after treatment
- 'deceased'

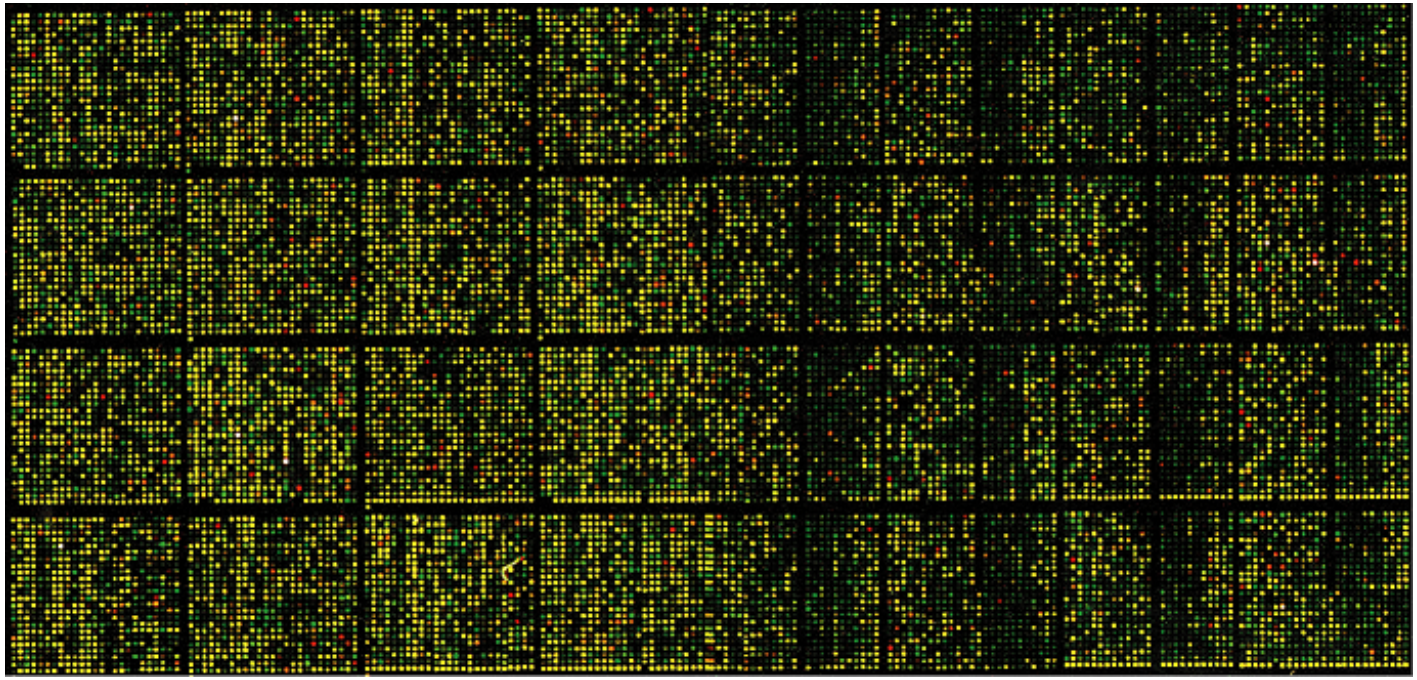
Data Sets

4 data sets:

- 3 genetic data sets:
  - Affymetrix gene expression microarray
  - microRNA expression profiling
  - arrayCGH
- Mass spectrometry data

Gene Microarray

- Measures the expression levels of 12625 genes
- For 63 neuroblastoma patients



MicroRNA

MicroRNAs are small mRNA molecules that do not code for proteins (like genes do), but each miRNA inhibits the activity of many genes.

- 384 miRNA expression levels have been measured
- For 25 patients

ArrayCGH

ArrayCGH data measures the deletion or gain of certain parts of chromosomes. These parts are called Bacterial Artificial Chromosomes (BACs).

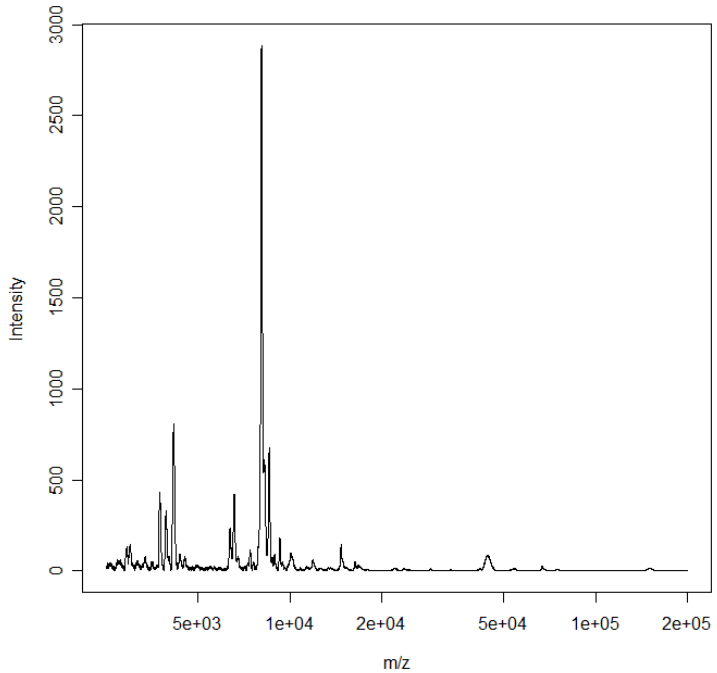
- 6228 BACs measured
- For 28 patients

Genetic Objects

The differences in size between the genetic objects.

SELDI Mass Spectrometry

- The mass spectrometer measures the amount of proteins in the blood of patients
- Each peak represents a type of protein
- It is non-trivial to find out which protein belongs to a peak
- Over 55000 measuring points per patient
- 44 neuroblastoma patients in the data set



Safarii Multi-Relational Data Mining Tool

- Developed by Arno Knobbe from the UU
- Focuses on the discovery of interesting subgroups (patterns)
- Only makes binary comparisons (e.g. 'stage 4 vs others', 'alive without event vs others')
- Using an interestingness measure (accuracy, coverage, novelty, ...)
  - *gene VDAC1* $\leq$ *2345.0* $\rightarrow$ *stage 4*
- These lists of patterns are intelligible to the user
- Multi-Relational: can search through multiple data sets at once, combining their info into patterns
  - *gene VDAC1* $\leq$ *2345.0* $\wedge$ *miRNA hsa-miR-137* $\leq$ *0.2* $\rightarrow$ *stage 4*
- Create cross-validated classifiers (like Support Vector Machines and Decision Table Majority Classifiers)

Pattern Teams

- Lists of patterns contains lots of redundancy
- Reduce this by creating a Pattern Team:
- A small subset of patterns that all contribute something unique
- And that together describe most of the data's characteristics
- According to a certain measure
  - DTM Accuracy: uses the patterns as a simple classifier that predict each patient's class label
  - Joint Entropy: describes complementary sets of individuals; subgroups that do not overlap
  - ...

Single Table Analyses

Problem: only 3 patients can be found in all 4 data sets, so no Multi-Relational Data Mining by Safarii

- Use Subgroup Discovery to create a list of interesting patterns according to the Novelty measure
- Novelty: measures the interestingness or the unusualness of a rule
- Find a Pattern Team of 2 patterns, using the DTM Accuracy
- Visualise the Pattern Team in a scatter plot

Gene Expression Profiling - Stage 4

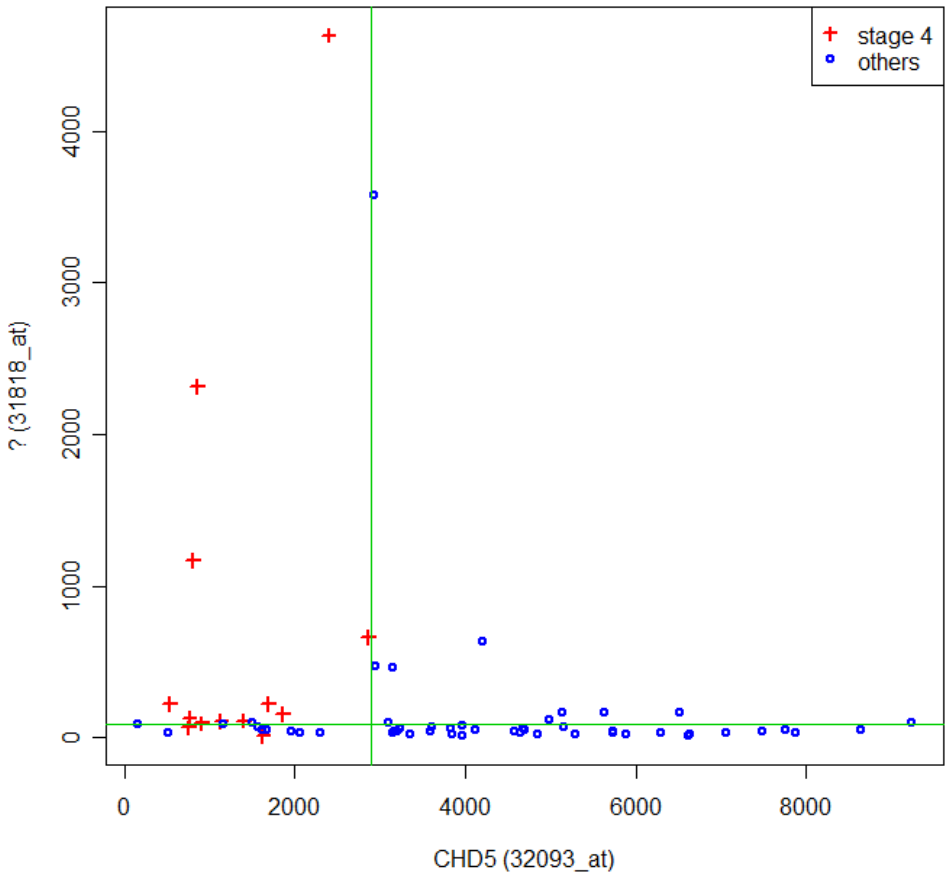
The first 10 patterns from the list:

Pattern Rank	Coverage	Novelty	Condition list
1	23 (37%)	0.13	CHD5 (32093_at) $\leq$ 2890
2	20 (32%)	0.12	PGK1 (37677_at) $\geq$ 11000
3	25 (40%)	0.12	ADCY1 (33353_at) $\leq$ 3370
4	25 (40%)	0.12	COX7B (36687_at) $\geq$ 1569
5	16 (25%)	0.12	NOTCH4 (39048_at) $\leq$ 294
6	21 (33%)	0.12	PRDM2 (33922_at) $\leq$ 1067
7	21 (33%)	0.12	? (39691_at) $\leq$ 4070
8	21 (33%)	0.12	DRD2 (40371_at) $\leq$ 460
9	26 (41%)	0.12	PRKCZ (362_at) $\leq$ 580
10	22 (35%)	0.12	LDHA (41485_at) $\geq$ 8550

Green genes: reported on in recent literature on neuroblastoma

Pattern Team

Pattern Rank	Coverage	Novelty	Condition list
1	23 (37%)	0.13	CHD5 (32093_at) $\leq$ 2890
77	22 (35%)	0.10	? (31818_at) $\geq$ 90



SVM scores

Data Set	Stage		NBstatus	
	Def (%)	SVM (%)	Def (%)	SVM (%)
Microarray	79.37	76.51 ($\pm$ 0.71)	71.43	79.68 ($\pm$ 2.61)
MicroRNA	60.00	68.80 ($\pm$ 4.38)	60.00	46.40 ($\pm$ 7.80)
ArrayCGH	75.00	80.00 ($\pm$ 1.96)	64.29	57.11 ($\pm$ 6.68)
SELDI MS	79.54	80.47 ($\pm$ 2.65)	72.72	63.26 ($\pm$ 1.95)

The poor performances are due to the small volumes of data and the skewness of the data sets. Moreover, the miRNA data misses half of the *NBstatus* targets, and the arrayCGH data has many missing values.

Conclusion

- Subgroup Discovery finds patterns that were reported on in recent neuroblastoma literature
- However, it is time consuming to compare each pattern to recent literature by hand
- Pattern Teams reduce the redundancy, and provide helpful scatter plots
- Drawback: they only consider a few patters
- SVMs are not really succesful, due to data-related problems
- So far, no data sets could be combined

Aggregation & Integration

Aggregation

Combining low-level knowledge (here: the patterns), discovered on a single data set, and turning them into meaningful, intelligible biological explanations.

Integration

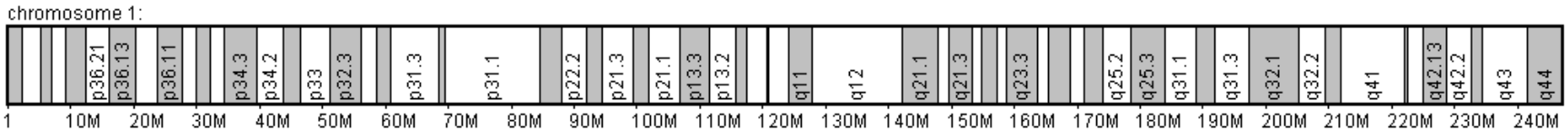
Combining multiple data sets (data-level integration), or combining knowledge discovered on multiple data sets (knowledge-level).

Combining Knowledge

- 3 data sets contain genetic data
- Their patterns can easily be linked to a location on the genome
- And a statistical value can be calculated for parts of the genome that contain patterns (potentially from different data sets)
- Crucial areas of the genome for neuroblastoma can be found
- As well as interactions between genes, miRNAs and BACs
- The mass spectrometry data cannot be considered

Cytogenetic Band

- Every chromosome can be divided into smaller parts: so-called cytogenetic bands
- For every cytoband, calculate a statistical value
- Based on the patterns' characteristics that are located on it



Statistics

p-values

From each pattern's data, calculate a *p*-value, using Fisher's exact test, then combine the *p*-values from each pattern on a certain cytoband with Fisher's method.

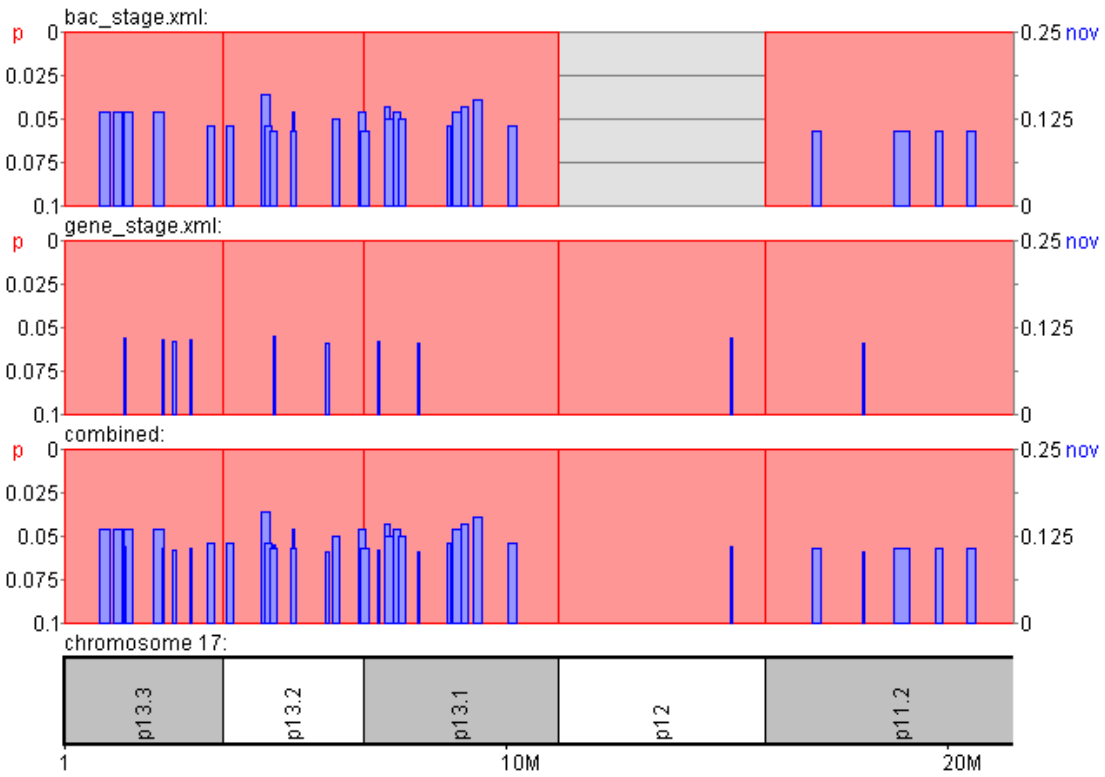
Rank-based

Rank all patterns found on different data sets according to their novelty value. Calculate the Area Under the ROC Curve for a cytoband, using the rank of its patterns.

p-values seem to favour cytobands with many patterns; rank-based statistics those with just 1, high ranked pattern.

Stage 4; *p*-values

Chromosome 17, band p13.3, p13.2, p13.1 were rated the most significant (smallest *p*-values) of all cytobands on the human genome. Probably, the genes on these areas are influenced by gains or losses of these areas.



Stage 4; AUC

The AUC statistic points to many areas with a single, high-ranked miRNA.
Most miRNAs influence many genes, and if they are mutated, they could start a chain-reaction throughout the genome.

Cytoband table for ROC AUC

chr	band	# of objects	auc	best object (p)	best p	best object (nov)	best nov
7	q36.3	1	0.99079...	hsa_miR_153	0.0017...	hsa_miR_153	0.152...
X	q27.1	1	0.98773...	hsa_miR_505	0.0017...	hsa_miR_505	0.152...
7	p22.2	1	0.98159...	bA106E3	1.2499...	bA106E3	0.151...
1	q22	1	0.96625...	hsa_miR_9...	0.0051...	hsa_miR_9_AS	0.143...
5	q14.3	1	0.96319...	hsa_miR_9...	0.0051...	hsa_miR_9_AS	0.143...
15	q26.1	1	0.96012...	hsa_miR_9...	0.0051...	hsa_miR_9_AS	0.143...
7	q32.3	1	0.95705...	hsa_miR_29b	0.0051...	hsa_miR_29b	0.143...
1	q32.2	1	0.95398...	hsa_miR_29b	0.0051...	hsa_miR_29b	0.143...
5	q33.3	1	0.95092...	hsa_miR_1...	0.0051...	hsa_miR_146a	0.143...
X	q26.2	2	0.95076...	hsa_miR_92	0.0027...	hsa_miR_92	0.144
7	q33	1	0.94785...	hsa_miR_490	0.0011...	hsa_miR_490	0.144...
14	q22.2	1	0.94171...	bA533L7	6.6889...	bA533L7	0.142...
21	q22....	1	0.93865...	dJ255P7	4.7971...	dJ255P7	0.142...
8	p22	1	0.93251...	bA161I2	4.7971...	bA161I2	0.142...
22	q11....	1	0.92638...	hsa_miR_1...	0.0118...	hsa_miR_130b	0.136
5	q33.1	1	0.92024...	hsa_miR_4...	0.0089...	hsa_miR_422b	0.136
2	p16.1	1	0.91717...	hsa_miR_216	0.0089...	hsa_miR_216	0.136
1	p13.3	1	0.91104...	hsa_miR_197	0.0089...	hsa_miR_197	0.136
20	q12	1	0.90797...	dJ600E6	0.0014...	dJ600E6	0.133...
13	q31.3	6	0.89771...	hsa_miR_1...	0.0027...	hsa_miR_17_...	0.144
20	p12.1	1	0.88957...	dJ348M17	2.1367...	dJ348M17	0.133...
17	q23.3	1	0.87730...	bA51F16	0.0013...	bA51F16	0.133...
17	q22	1	0.87423...	bA112J9	0.0013...	bA112J9	0.133...
13	q14.2	1	0.87116...	bA305D15	0.0013...	bA305D15	0.133...
7	q22.2	1	0.86809...	bA22N19	0.0013...	bA22N19	0.133...

Close

Conclusion

- The tool indicates most genomic areas that are known to have effect on neuroblastomas
- It also displays possible interactions between different genetic objects
- To combine knowledge from data sets, they do not necessarily have to contain the same patients anymore
- Finding interesting genetic areas is much faster now, than searching the lists of patterns by hand