Knowledge Discovery in Neuroblastoma-related Biological Data

Edwin van de Koppel[1], Ivica Slavkov[2], Kathy Astrahantseff[3], Alexander Schramm[3], Johannes Schulte[3], Jo Vandesompele[4], Edwin de Jong[1], Sašo Džeroski[2], Arno Knobbe[1]

Utrecht University, P.O. box 80 089, NL-3508 TB Utrecht, The Netherlands[1], Department of Knowledge Technologies, **Jožef Stefan Institute**, Jamova 39, SI-1000 Ljubljana, Slovenia[2], Center of Pediatric Oncology, **University Children's Hospital of Essen**, Hufelandstrasse 55, 45122 Essen, Germany[3], **University of Ghent**, Centre for Medical Genetics, De Pintelaan185, 9000, Gent, Belgium[4]
contact: a.knobbe@kiminkii.com

Abstract In this paper, we provide initial Data Mining results on four sets of genetic data, collected in the context of the new European Embryonal Tumour Pipeline project. These data sets provide different views on the genetic processes involved in the genesis and development of a specific type of tumour, known as neuroblastoma. Although the project involves other types of tumours as well, with potentially similar underlying causal processes, neuroblastoma is currently the only disease for which sufficient data has been collected to analyse. We provide results on this data using systems developed at two Data Mining groups in Europe, with the aim of introducing the different Data Mining challenges involved, and outlining the approach we intend to apply throughout the project. Our descriptions focus on the analysis of individual data sets, stemming from separate analysis platforms (e.g. Affymetrix microarrays). Additionally, we provide some pointers for doing cross-platform analysis in the future.

1 Introduction

In this paper we give an overview of the many Data Mining challenges involved in a new EU-funded project called the European Embryonal Tumour Pipeline (EET Pipeline in short). The EET Pipeline attempts to improve treatment of a group of cancers affecting infants and small children through channelling information extracted from high-throughput molecular profiling of these tumours into pipelines to validate targets for novel therapy and diagnostic development. As cancer is the second cause of deaths (after accidents) among children in Europe, this is an important goal, and Data Mining will play a crucial role in the extraction of knowledge from the large quantities of data produced in this project. As the project has started only recently, we do not intend to give a complete report of the results obtained, but rather provide insights in the intended approach, and show promising initial results. The aim of this paper is to outline the types of data available, the Data Mining challenges that result

from this data, and some of the techniques we are employing to deal with these challenges. Specifically, we are considering the embryonal tumour neuroblastoma (reviewed in [3]), for which extensive molecular data is already available within the project. In Section 5 of this paper, we provide initial results for this illness. As the remaining tumour types will involve similar types of data, the results reported here will give a good indication of the activities that will be performed throughout the project. However, data for these other tumour types will only become available in later stages of the project.

One of the main characteristics of the project is its integrated approach towards embryonal tumours. This integration takes the form of a unified approach across tumour types. Furthermore, for all tumours, we are gathering data through a range of high-throughput analysis platforms, providing multiple views on the biological processes involved in the development of tumours. The analysis platforms include microarrays for gene and microRNA expression, ArrayCGH for chromosomal deletion and multiplication, and mass-spectrometry for proteomics. The diverse nature of these different data sources, in terms of data structure, is a first challenge, which we address in this paper by demonstrating how our analysis techniques can be applied to individual data sources. Further challenges lie of course in the integrated analysis of data across analytic platforms, either by combining data sources into rich unified descriptions of patients and tumour tissue, or by integrating the knowledge that is extracted using platform-specific techniques. We provide some pointers as to these activities in Section 6.

The two analysis systems we are employing are the results of years of Data Mining research at Utrecht University and the Jožef Stefan Institute (JSI) respectively (affiliations 1 and 2 in the author list). Both flavours of Data Mining can be characterised by a strong emphasis on interpretability of the models created. We are focusing on results that make sense to a domain expert and that lead to new insights about the underlying genetic processes, rather than on inducing a black box with high predictive accuracy per se. Both systems are generic Data Ming tools, with broader application than just biology. The first system, Safarii [7, 11], was developed at Utrecht University and Kiminkii, a Dutch company owned by the last author. It is based on the discovery of patterns such as rules and interesting subgroups, and combining these patterns into classifiers using a number of techniques such as Pattern Teams [8] or Support Vector Machines. A specific forte of Safarii is the support for Multi-Relational Data Mining, a technique that allows the integration of data from different sources. The second system, developed in cooperation between the Katholieke Universiteit Leuven and the JSI, implements a tree-based approach known as Predictive Clustering Trees (PCTs) [1]. Such trees combine the benefits of clustering with those of tree-based classification methods. Of specific interest in this context is the ability to induce multi-target PCTs, trees that are optimised for multiple targets (e.g. tumour subtype and developmental stage) at the same time. We describe and demonstrate both systems in Section 4.

2 Neuroblastoma

Neuroblastoma is the most common extracranial solid tumour of childhood, and 88% of neuroblastoma patients are 5 years or younger. Neuroblastoma demonstrates

many features of common interest to cancer, such as spreading of the cancer and the development of resistance to chemotherapy. However, due to its manifestation early in life, it presents an excellent model to study genetically based changes leading to cancer, relatively free from the influence of environmental factors. Additionally, the embryonal tumours, to which neuroblastoma belongs, also are unique in the high incidence of spontaneous regression and differentiation of the tumours. The understanding of how this "self-cure" mechanism works may also be applicable to develop new treatment strategies for other cancers. Treatment of neuroblastomas with polychemotherapy provokes good initial response, regardless of tumour stage. However, two major problems of the current treatment regimen exist. Disseminated (cancer spread throughout the body) stage 4 tumours frequently relapse due to minimal residual disease arising from a few resistant tumour cells, resulting in poor overall survival rates (<35%). On the other hand, overtreatment of *MYCN*-nonamplified stage 2 or 3 tumours causes most of the surviving patients to suffer from significant organ toxicity or develop secondary malignancies later in life, reducing their quality of life. Novel strategies to more precisely diagnose and treat neuroblastoma are urgently needed to improve this situation. With the recent advent of high-throughput technologies, it is now possible to assess the tumour at multiple biological levels, including the genome, transcriptome and proteome. The large amounts of molecular information resulting from these analyses holds the promise of not only a better understanding of neuroblastoma biology and progression, but also the identification of molecules that can be targeted for therapy and used to better tailor treatment for a personalised diagnosis.

3 Data Sources

For neuroblastoma tumour samples and patient serum, a total of four data sets have been collected (being ArrayCGH, Affymetrix microarray, MicroRNA and SELDI Mass Spectrometry data). In this section, we will discuss the characteristics of each of these separately, and assess their potential and problems. First, we will give a description of the target concepts that we want to investigate. Then we will address the characteristics of each data set.

Target concepts for Investigation

The data from the EET Pipeline project leads to a range of potentially interesting target concepts for Data Mining. With space limitations in mind, we have selected two targets for this paper that are of interest to the domain experts: clinical course (*NBstatus*) and neuroblastoma stage (*Stage*).

Clinical Course: Domain experts rate this as being one of the most interesting target concepts for investigation. The clinical course *NBstatus* lists the patients last recorded follow up status, being either ‘alive without event’, ‘alive with relapse/primary tumour’ or ‘deceased’. Since only deceased patients in the data who died as result of a relapse or primary tumour were chosen for analysis here, we can make a binary comparison by testing ‘alive without event’ versus the rest.

If Data Mining can succeed in showing correlations between, for example, gene expression levels in the tumour or protein levels in the blood and relapse, an ‘early

warning system' can be constructed, identifying patients with a high risk of relapse before they actually suffer from it.

Stage: The INSS staging system developed for neuroblastoma tumours is the standard in Europe, the U.S. and Japan [4]. It categorises tumours into several stages based on clinical characteristics, numbered 1 through 4 with 4 being the most severe. All tumours from children under one year of age but limited metastases to liver, bone marrow or skin (never bone) are classified into a special stage, known as 4s. These patients have a very good prognosis for recovery. The majority of tumours from this patient subset undergo spontaneous regression even with little or no chemotherapy treatment. Patients diagnosed with stage 4 tumours, however, often succumb to their disease despite aggressive multimodal therapy. We attempt to determine whether Data Mining can deliver more information about molecular characteristics specific for certain clinical subgroups of neuroblastoma. As a starting point, we will only consider the task of distinguishing less severe neuroblastoma subgroups (stages 1, 2, 3 and 4s) from stage 4.

Data Sets

Affymetrix Expression Profiling Affymetrix is one type of array platform to conduct expression profiling. The probes on the microarray recognise one or more short areas of a specific gene transcript. The signals measured give information about how many RNA transcripts of which genes are present in the sample, which is a measure of gene activity. Expression was analysed in 63 primary neuroblastomas using the Affymetrix U95Av2 oligonucleotide microarrays. These data are included in the 68 patients analysed in [12]. This array measures the expression levels for a total of 12625 probes (genes).

ArrayCGH Array-based Comparative Genomic Hybridization (ArrayCGH) analyses the status of the whole genome of a tissue sample. It is known that certain segments of the DNA in the chromosomes are often altered in neuroblastomas [14, 9]. Possible genomic alterations include amplifications or deletions in distinct areas of certain chromosomes (including several genes) and even multiple copies of the complete chromosome complement in the cell (*trisomy*). ArrayCGH utilizes DNA probes of varying sizes to represent all areas of the genome in different levels of detail. These probes are Bacterial Artificial Chromosomes (BAC's), and analysis detects the number of copies of the DNA region corresponding to a BAC that is present in the tumour sample relative to the normal DNA complement of two copies. The data is represented as negative or positive real numbers, showing deletion or amplification, respectively.

Our data set includes ArrayCGH analysis of 19 primary neuroblastomas. These 19 tumours were among the 23 analysed in [14]. Unfortunately, four patients had to be disregarded in the current analysis, since *Stage* and *NBstatus* information could not be obtained. For each tumour there are 6228 attributes (the BAC's). However, the data contain many missing values. Specifically, data for certain BAC's are missing for all tumours analysed. Removing those, we end up with only 4820 attributes that have a value for at least one patient.

MicroRNA Expression Profiling The expression of small, non-coding regulatory RNAs, or microRNAs (miRNAs), can also be analysed using a microarray platform. MicroRNAs inhibit the expression of specific groups of genes via sequence specific

binding of the mRNA molecule, inhibiting translation into the protein. The probes on these types of array measure the expression of miRNAs, which are short RNA molecules (about 21-23 nucleotides long).

The data set contains measurements from 25 primary neuroblastomas. The tumours were analysed on a 2-channel cDNA array with probes for 384 miRNAs [13]. Two records come from different tissue samples from the same tumour (so there are 24 unique patients). For all patients we have the *Stage* information. Unfortunately, there is *NBstatus* information available for only 13 patients. Each patient is characterised by 384 attributes (miRNAs) indicating the deviation in activity from the average case.

SELDI Mass Spectrometry Surface-Enhanced, Laser Desorption/Ionisation Mass Spectrometry (SELDI MS) data is a different type of data. The mass spectrometer measures the amount and size (in Daltons) of all proteins in a complex protein mixture using time-of-flight (TOF) detection. The serum from 43 neuroblastoma patients at the time of diagnosis were fractionated on anion-exchange columns and profiled on metal-binding arrays (IMAC-Cu^{++}) using SELDI-MS. Both *Stage* and *NBstatus* information were available for these patients. Data from this analysis is expressed as mass-to-charge ratios (m/z). Mapping these m/z data to a specific protein identity is a non-trivial task requiring further chemical purifications and analyses of a larger sample amount. Only data produced from serum fraction 1 were used here.

4 Methods

Predictive Clustering Trees

Predictive modelling aims at constructing models that can predict a target property of an object from a description of the object. Predictive models are learned from sets of examples, where each example has the form (D, T), with D being an object description (or set of attributes of that object) and T a target property value. While a variety of representations ranging from propositional to first order logic have been used for D, T is almost always considered to consist of a single target attribute called the class, which is either discrete (classification problem) or continuous (regression problem).

Clustering, on the other hand, is concerned with grouping objects into subsets of objects (called clusters) that are similar with respect to their description D. There is no target property defined in clustering tasks. In conventional clustering, the notion of a distance (or conversely, similarity) is crucial: examples are considered to be points in a metric space and clusters are constructed such that examples in the same cluster are close according to a particular distance metric.

Predictive clustering [1], the analysis paradigm of our interest, combines elements from both prediction and clustering. As in clustering, we seek clusters of examples that are similar to each other, but in general taking both the descriptive part and the target property into account. In addition, a predictive model must be associated to each cluster. The predictive model assigns new instances to clusters based on their description D and provides a prediction for the target property T. It should be noted

that in this predictive clustering setting, the target T is not necessarily a single value, but rather a set of target attributes.

Also a distinction is made between the target attributes T and clustering attributes C. The distance measure is calculated on $C \cup T$, i.e., we produce models that are trying to correctly predict the attributes in both T and C. The difference between the T and C attributes is purely in the semantic for the end-user. The user is interested in the accuracy of the target attributes T, while the clustering attributes are included in the model building process in order to improve it. That is why in the results section we only report the accuracy of the obtained models for the target attributes T.

A well-known type of model which is used for the predictive clustering paradigm is a decision tree [10]. A decision tree that is used for predictive clustering is called a predictive clustering tree (PCT). Each node of a PCT represents a cluster. The conjunction of conditions on the path from the root to that node gives a description of the cluster. Essentially, each cluster has a symbolic description in the form of a rule (IF conjunction of conditions THEN cluster), while the tree structure represents the hierarchy of clusters.

A generic system for constructing PCTs is available in the Clus system, which can be obtained at "http://www.cs.kuleuven.be/~dtai/clus".

Safarii

Safarii [11] is a Multi-Relational Data Mining system that has been developed over the last year at Utrecht University and Kiminkii, primarily by the last author and colleagues. It includes a range of Data Mining techniques, as well as general facilities for dealing with large (multi-relational) data stored in relational databases. The primary approach for data analysis that is relevant to the domain at hand is centred around the discovery of regularities such as rules or interesting subgroups, which we will refer to in general as patterns [5, 12]. Such patterns may capture interesting, but possibly incomplete, knowledge concerning the influence of specific genes on a selected target (e.g. neuroblastoma vs. healthy), or the interaction of two or more genes, to name but a few examples. After such patterns have been discovered, they can be combined into more ambitious models of the biological processes that involve multiple patterns. Such global models can be used as classifiers in a black-box setting, for example to aid the diagnosis of tissue from new (suspected) patients. More importantly, by focussing on fairly simple and understandable patterns and the interaction between them, our approach aims to produce useful insights into the dynamics of the domain.

For combining patterns into global models, Safarii offers a number of reasonably well-known classifiers, notably Support Vector Machines (SVM) and Decision Table Majority (DTM) classifiers [8]. It is important to note that we are applying these classifiers not directly to the original data, but rather to the set of patterns that was previously discovered. In a sense, the patterns are treated as new constructed features, which are guaranteed to be predictive because they are the result of a mining operation themselves. The benefit of this approach is that the classifiers are constructed of pieces of knowledge that are intelligible and informative, compared to, for example, the application of SVMs to the data directly, which produces classifiers that are notoriously hard to interpret.

A possible downside of the pattern discovery approach is the potentially large number of patterns reported. Especially in genetic data, where it is not uncommon for many genes to be correlated, many possible patterns may be found, involving a range of genes that essentially capture the same aspect of the biological process. Safarii offers substantial facilities for dealing with this redundancy in sets of patterns. A technique known as Pattern Teams [8] selects out of the original large set of patterns, a small but informative subset of patterns, where each pattern adds something unique to the team.

Due to the small volumes of data, we are forced to work with fairly simple patterns, typically only including a single gene or location in the mass-spectrum. With larger data sets, and therefore less risk of overfitting, there is nothing that would prevent us from discovering more complex patterns. Note that possible interactions between genes are also captured during the combination into classifiers or teams, reducing the need for finding these interactions immediately. As a further limit on the complexity and expressiveness of our models, we will build Pattern Teams involving only few patterns. Small teams have the further advantage that they can be easily visualised, aiding the understanding and communication of findings.

5 Results

We have analysed all four dataset with the two systems at our disposal. In the interest of space however, we only demonstrate the results for two arbitrarily selected datasets per analysis technique: MicroRNA and SELDI-MS in the case of Safarii, and Affymetrix microarray and ArrayCGH in the case of Predictive Clustering Trees. Predictive models were built for *NBstatus* and *Stage* attributes. Additionally, we utilised the ability of PCTs for multi-target prediction and constructed predictive models which take into account other patient information (e.g. MYCN amplification). Comparisons were made between the single and multi-target prediction models.

Affymetrix (PCTs) When analysing the Affymetrix microarray data, two target attributes were taken into account: *NBstatus* and *Stage*. As it can be seen in Table 1 and Table 2, when trying to do a single target prediction for *NBstatus* and *Stage*, the accuracy obtained from the ten-fold cross-validation was a little better (for *NBstatus*) or worse (for *Stage*) than the default distribution.

In order to improve the performance when building PCTs, we included as clustering attributes other patient information which was previously shown [9] to be connected to the outcome of the disease. Those attributes were *deletion of the 1p chromosome region* and *amplification of the MYCN gene*. Figure 1 shows a PCT which is built when considering *NBstatus* as target and *1p deletion* as a clustering attribute. As any decision tree model, a PCT can be easily interpreted. The first node of the tree, with attribute *40235_at* (TNK2, 'tyrosine kinase, non-receptor, 2'), splits the samples into two groups. In the first group there are patients with 'alive without event' and 'no deletion' of the 1p chromosome region. The remaining group is split by a node (*34480_at*, CDH16, 'cadherin 16, KSP-cadherin') of the PCT that essentially distinguishes between patients that have/do not have a 1p deletion. The last node (*g32415_at*, IFNA5, 'interferon, alpha 5') further differentiates between the patients with 1p deletion that had a relapse (i.e., 'alive with relapse/primary tumour' or 'deceased') or are 'alive without event'.

From Table 1 it can be seen that including *1p deletion* and *MYCN amplification* as clustering attributes significantly improved the predictive performance of the constructed PCTs. The results in Table 2 show that for *Stage*, it is extremely difficult to build a predictive model which will surpass the default distribution (probability of the majority class), except for the last case when as a clustering attribute *NBstatus* is included. Considering the initial distribution, which is skewed, and the few Stage 4 cases, the learning of a predictive model is a difficult task.

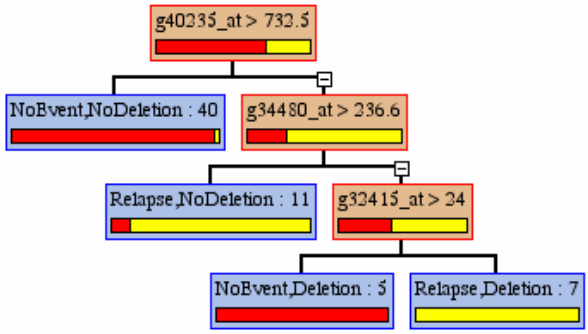


Figure 1. PCT constructed for *T = NB* status and *C = 1p*

Table 1. Results from a 10-fold cross-validation for *NBstatus*

Target and Clustering attributes	default acc. (%)	PCTs acc. (%)
T = NBstatus	71.4	74.6
T = NBstatus, C = 1p	71.4	90.5
T = NBstatus, C = MYCN	71.4	84.1
T = NBstatus, C = 1p, MYCN	71.4	74.6
T = NBstatus, C = Stage	71.4	82.5

Table 2. Results from a 10-fold cross-validation for *Stage*

Target and Clustering attributes	default acc. (%)	PCTs acc. (%)
T = Stage	79.3	77.7
T = Stage, C = 1p	79.3	73.0
T = Stage, C = MYCN	79.3	74.6
T = Stage, C = 1p, MYCN	79.3	77.7
T = Stage, C = NBstatus	79.3	80.9

ArrayCGH (PCTs) For the ArrayCGH data, a similar analysis was performed. The same target and clustering attributes were taken into account. As is evident from the results in Table 3 and Table 4, it proved to be very difficult to build PCTs with accuracy higher than the default. Including multiple attributes did not significantly improve the accuracy. The small sample size (19) and the initial class distribution

(only 3 “Stage4” samples) of this particular dataset make building accurate PCTs and predictive models difficult.

Table 3 Results from a 10 fold cross validation for *NBstatus*

Target and Clustering attributes	default acc. (%)	PCTs acc. (%)
T = NBstatus	73.6	73.6
T = NBstatus, C = 1p	73.6	78.9
T = NBstatus, C = MYCN	73.6	73.6
T = NBstatus, C = 1p, MYCN	73.6	73.6

Table 4 Results from a 10-fold cross-validation for *Stage*

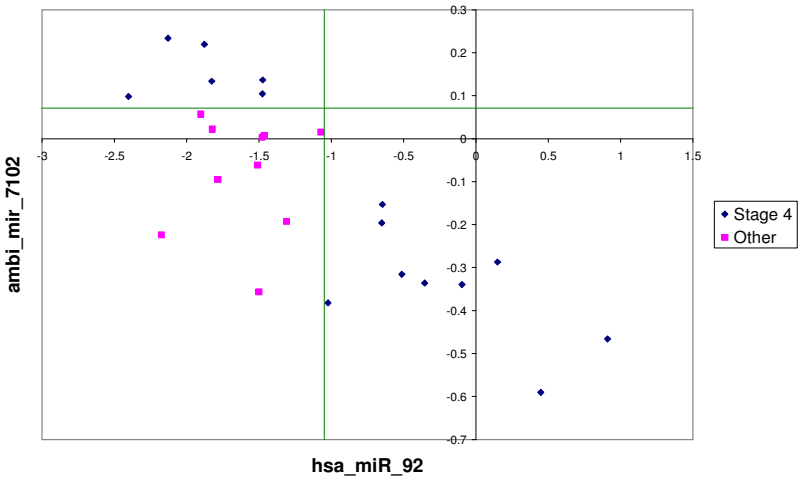
Target and Clustering attributes	default acc. (%)	PCTs acc. (%)
T = Stage	84.2	89.4
T = Stage, C = 1p	84.2	89.4
T = Stage, C = MYCN	84.2	84.2
T = Stage, C = MYCN, 1p	84.2	84.2
T = Stage, C = NBstatus	84.2	84.2

MicroRNA (Safarii) As a demonstration of the kind of knowledge that can be discovered with Safarii, we show some results for *Stage*. For the MicroRNA data, the top 100 patterns (in this case most differentially expressed probes) were identified using Safarii’s Subgroup Discovery algorithm. The resulting patterns are ranked according to the *novelty* measure (a.k.a ‘weighted relative accuracy’) [5, 7]. A minimum coverage of 6 patients was applied. For reducing the redundancy, we then applied the Pattern Team technique to the 100 patterns, producing a team of two essential probes. It reports a combination of the 2nd and 96th pattern:

Pattern Rank	Coverage	Novelty	Condition list
2	9	0.14	hsa-mir-92 ≥ -1.04
96	6	0.096	ambi-mir-7102 ≥ 0.07

A Pattern Team of size two can be easily visualised in a scatter plot, as demonstrated below. The two thresholds for the patterns involved are shown as the horizontal and vertical lines. Clearly, the lines separate the patients into three distinct clusters that appear to coincide with the target concept specified. This plot clearly demonstrates how the selected approach finds multivariate interactions that are relevant to this tumour type. Analysis of these array results using the SAM algorithm also identified hsa-mir-92 part as the most important miRNA associated with MYCN-amplified neuroblastomas (submitted). This miRNA was also identified as the first "oncomir", or miRNA which can act as an oncogene to potentially induce several tumour types [6].

Analogous results can of course be obtained for the *NBstatus*, our second target concept.

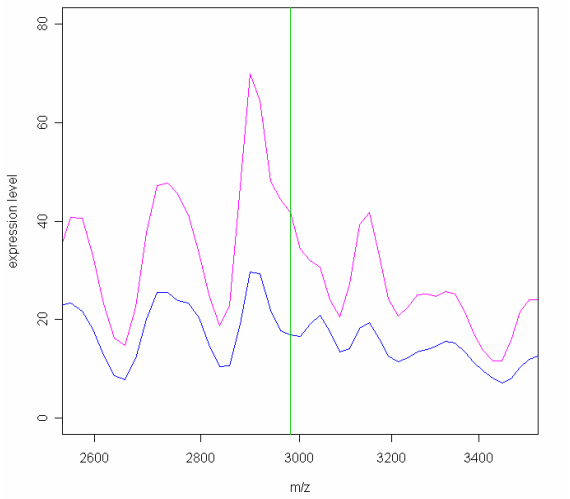


SELDI-MS (Safarii) For the SELDI-MS data some extra pre-processing was required. To reduce the effects of noise in the data as a consequence of the measurement process, a data smoothing procedure was applied, based on a Gaussian Kernel. We used the approach given by [2]. For our analyses, the kernel width was set to 101 data points (50 below the data point we want to smooth, the point itself, and 50 above). After that we reduced the resolution of the total spectrum, since it has around 56000 data points for each patient. We did this by selecting every 25th data point from the smoothed spectrum, resulting in a little over 2200 data points for each serum.

Again, the Subgroup Discovery algorithm was run with the same settings as for the MicroRNA data, creating 100 patterns for *Stage* 4 versus other stages. In the figure below, we show part of the (pre-processed) spectrum in the area of one of the patterns discovered, as an example. The two curves represent the averages of the stage 4 group (the lower line) and the remaining stages (the upper line). The vertical line corresponds to the second pattern found:

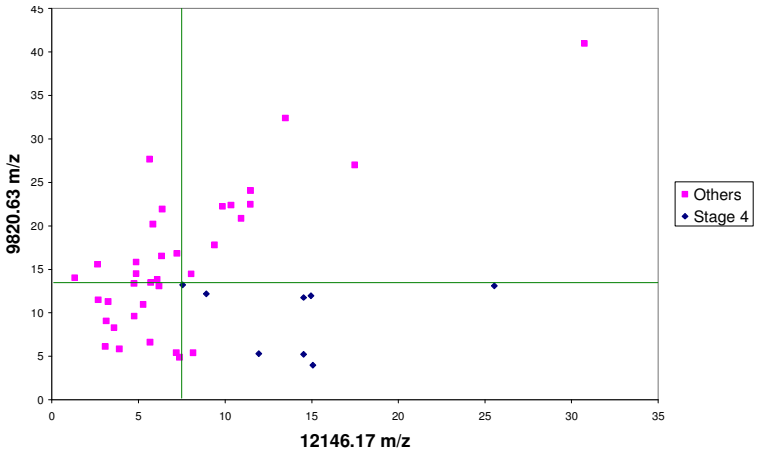
Pattern Rank	Coverage	Novelty	Condition list
2	13	0.11	2981.60 m/z ≤ 15.0

It is interesting to observe that our method does not necessarily select locations corresponding to peaks in the spectrum. Although peaks obviously correspond to specific proteins, some of which may be related to the difference between stages, apparently the exact optimum of such a peak is not guaranteed to be the most informative. As the figure demonstrates, peaks overlap to some degree, and subtle peaks may therefore not appear as actual optimums in the spectrum. The upper line in this figure shows a bump between the two adjacent peaks which is clearly missing in the stage 4 patients. Although this location does not seem promising at first hand, our method is able to identify such cases. This is in contrast to other methods (statistical and modified SVM) that have been used to analyse these data, which were incapable of analysing differences between neuroblastoma subtypes and could only be used to analyse neuroblastoma vs. healthy or related tumour patients (different targets).



We would like to add that simply taking the average over a group (as is done in the figure) does not necessarily give good insight into the distribution of individual values. As an alternative, we again show a scatter plot for a predictive pair of patterns:

Pattern Rank	Coverage	Novelty	Condition list
3	19	0.11	12146.17 m/z ≥ 7.4
24	21	0.10	9820.63 m/z ≤ 13.28



6 Conclusion and Future Developments

We have presented initial Data Mining results for a number of data sets related to neuroblastoma, in the context of the EET Pipeline project. For some of these, the methods that we used were able to construct good predictive models for the targets of

interest. For others, the small sample size and the prior distribution made the task of constructing good predictive models challenging. At this stage, we have only considered the analysis of data sets separately. The ultimate goal of the project is to combine data sets and thus obtain knowledge that spans different biological levels. As was demonstrated, there is still a considerable mismatch between the patient sets used for the different analysis platforms. This not only hinders the analysis of individual data sets, as data samples are often small, but also the integration of data sets, because the intersection of samples is even smaller. Still, with data sets becoming more complete as the project continues, integrated analysis will become important. An obvious way of integrating is to simply join data sets (over patient or tissue identifiers). Apart from scalability problems, this will be a straightforward step that will ideally lead to findings that involve cross-platform combinations of patterns. An alternative approach involves integration on the level of discovered knowledge rather than on the data level. For example, all data sets, except the SELDI-MS data, in some way map to loci on the genome. This means that if multiple data sets independently produce patterns involving the same locus, this will improve the evidence for this locus being involved in the biological process under investigation.

References

1. Blockeel, H., De Raedt, L., Ramon, J., *Top-down induction of clustering trees*. In Proceedings of ICML '98, p. 55-63, 1998
2. Brett, M., *An Introduction to Smoothing,* http://imaging.mrc-cbu.cam.ac.uk/imaging/PrinciplesSmoothing, 2006
3. Brodeur, G.M., *Neuroblastoma: Biological insights into a clinical enigma*, Nat. Rev. Cancer 3:203-216, 2003
4. Evans, A.E., D'Angio, G.J., Sather, H.N., *et al.*, *A comparison of four staging systems for localized and regional neuroblastoma: a report from the Childrens Cancer Study Group*, J. Clin. Oncol. 8:678-688, 1990
5. Fürnkranz, J., Flach, P., *ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms*, Machine Learning, 58, 39–77, Springer, 2005
6. He, L., Thomson, J., Hemann, M., Hernando-Monge, E., Mu, D., Goodson, S., *et al. A microRNA polycistron as a potential human oncogene*. Nature 435:828-833, 2005
7. Knobbe, A.J., *Multi-Relational Data Mining*, Ph.D. dissertation, 2004, http://www.kiminkii.com/thesis.pdf
8. Knobbe, A.J., Ho, E.K.Y., *Pattern Teams*, in Proceedings PKDD 2006, 2006
9. Maris, J.M., *The biologic basis for neuroblastoma heterogeneity and risk stratification*, Curr. Opin. Pediatr. 17:7-13, 2005
10. Quinlan, J.R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993
11. *Safarii Multi-Relational Data Mining Environment*, http://www.kiminkii.com/safarii.html, 2006
12. Schramm, A., Schulte, J.H., Klein-Hitpass, L., Havers, W., Sieverts, H., Berwanger, B., Christiansen, H., *et al. Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling*, Oncogene 24:7902-7912, 2005
13. Shingara, J., Keiger, K., Shelton, J., Laosinchai-Wolf, W., Powers, P., Conrad, R., Brown, D., Labourier, E., *An optimized isolation and labeling platform for accurate microRNA expression profiling*. RNA 11:1461-1470, 2005
14. Vandesompele, J., Baudis, M., De Preter, K., Van Roy, N., *et al.*, *Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma*, J. Clin. Oncol. 23:2280-2299, 2005