Integrative Visual Data Mining Approach to Biomedical Data: Investigating cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia.

Paul Kennedy[2], Simeon J. Simoff[1,2], Daniel R. Catchpoole[3,2], David B. Skillicorn[4], Franco Ubaudi[2] and Ahmad Al-Oqaily[2]

[1]School of Computing and Mathematics, University of Western Sydney
NSW 2007 Australia
s.simoff@uws.edu.au
[2]Faculty of Information Technology, University of Technology, Sydney
PO Box 123 Broadway NSW 2007 Australia
{paulk, simeon, faubaudi, aaoqaily}@it.uts.edu.au
http://research.it.uts.edu.au/emarkets
[3]The Oncology Research Unit, The Children's Hospital at Westmead,
Locked Bag 4001, Westmead NSW 2145, AUSTRALIA
DanielC@chw.edu.au
[4] School of Computing, Queen's University, Kingston, CANADA
skill@cs.queensu.ca

Abstract. This chapter presents an integrative visual data mining approach towards biomedical data. This approach and supporting methodology are presented at a high level. They combine in a consistent manner a set of visualisation and data mining techniques that operate over an integrated data set of several diverse components, including medical (clinical) data, patient outcome and interview data, corresponding gene expression and SNP data, domain ontologies and health management data. The practical application of the methodology and the specific data mining techniques engaged are demonstrated on two case studies focused on the the biological mechanisms of two different types of diseases: Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia, respectively. The common between the cases is the structure of the data sets.

Introduction

Molecular and genomic information are becoming an important part of methods for diagnosing diseases, based on biological indicators. There is a very large and increasing level of effort towards improving the overall methodology for utilising the data gathered through gene expression profiling. The efforts are focused on the measurement procedures and data collection technology, experiment designs, and diverse data analysis and mining methods (Parmigiani *et al*., 2003). Some of the best practices have been discussed in (Piatetsky-Shapiro *et al*., 2003; Hoffman *et al*., 2004).

Mining microarray data on it's own is a challenging task (Piatetsky-Shapiro and Tamayo, 2003), due, on the one hand, to the superposition of a number of physical processes in the data collection, on the other, to the need to convert extracted patterns to biological knowledge. Consequently, there has been an increasing interest towards complementary techniques for analysing simultaneously gene expression data and other data sources, for example, literature-based information (Glenisson *et al*., 2003), DNA sequence database (Curran *et al*., 2003) or from several sources (Seifert *et al*., 2005). This increasing tendency in extending data mining techniques, for example, association rule mining (Georgii *et al*., 2005; Carmona-Saez *et al*., 2006), is reflected in some of the tools developed recently (Shamir *et al*., 2005; Dietzsch *et al*., 2006). These "joint" methods, however, have emerged somewhat on an ad-hoc basis. Though biologists often focus on data, collected from microarray-based expression profiles, other molecular data, including the organisation and function of genes in the context of the cell, the physical genome and sequence, the relationships between species in terms of this organisation, can provide important insights into the phenomenon. Overall, in the biomedical and health sciences, various databases collect these diverse data sets, each providing a basis for knowledge discovery within a specific area of understanding. This is illustrated in Fig. 1. Biomedical and health data and patterns discovered from it often consist of many small interactions contributing to the explanation of the phenomenon. Developing a consistent methodology and the corresponding combinations of supporting algorithm is the aim of the work, presented in this chapter.

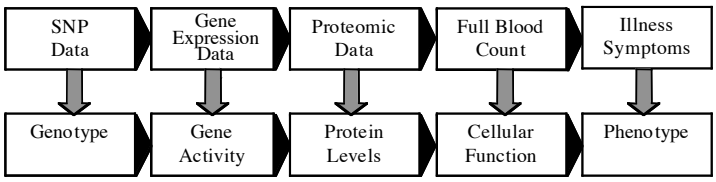

Fig. 1. Relationship of data source to the biological genotype - phenotype spectrum.

Fig. 2 shows the broader picture of the data sources that are involved on the biomedical side in modern healthcare. There is a growing opinion that the analysis of biomedical data requires the integration of various data sources to build up a more complete picture of the various levels of biology, clinical understanding and optimal patient management. Consequently, there is a need for a consistent methodology that enables combined analysis of clinical traits, marker genotypes, comprehensive gene expression, SNP data, in order to dissect the biological mechanisms of complex disease. Recent research recognises also the necessity in automatic utilisation of existing knowledge compiled in various "omic" electronic libraries in order to understand and interpret the outcomes of microarray and SNP data in the context of existing biological knowledge (Hasegawa *et al*., 2006).

A brief overview of the different types of data (data sources), their characteristics and issues of integration with the other data are presented in Table 1 - Table 4. We consider seven types of data sources, grouped in four categories:

- Medical and Clinical data sources (including Medical Data, Patient Outcome Data and Patient Questionnaire Data) presented in Table 1;

- Biological data sources (including Gene Expression Profiles and Single Nucleotide Polymorphisms (SNPs)) presented in Table 2;
- Biological knowledge bases (including Domain Ontologies and other Databases) presented in Table 3;
- Healthcare data sources (including Health Management Data) presented in Table 4.

Ideally, each of these types of data should be present in the integrated data set, however, the final selection depends on the available data and the study scenario.

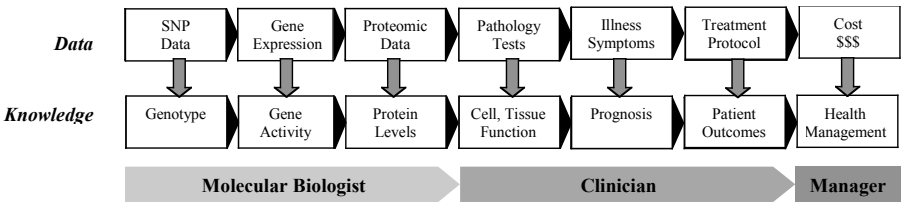

Fig. 2. The diverse biomedical and healthcare data sets are the source for knowledge discovery within a specific area of understanding associated with the management of patients.

Further in the chapter we present an overview of the general methodology and demonstrate its application in two case studies: the identification of biological markers underlying Chronic Fatigue Syndrome and to the sadly common childhood malignancy Acute Lymphoblastic Leukaemia.

Table 1. Medical and Clinical data sources

Data Sources	Characteristics of the data	Integration Issues
Medical Data: Prognostic indicators are used for the empirical diagnosis of disease. In the case of ALL patients, this is a risk-based directed therapy (Felix *et al.*, 2000).	Patient age, sex, ethnicity, white blood cell count, cytogenetic analysis, cell surface antigens and response to initial chemotherapy.	Data available is often derived from patients presenting at hospitals and treated on specific drug trials. Data types are mixed but may be available only as unstructured text.
Patient Outcome Data: Studies and trials are generally designed to compare potentially better therapy with therapy that is currently accepted as standard.	Treatment protocols list drug schedules for patients in different risk categories and modifications for patients with abnormal response to drugs.	Therapies and outcome data are in unstructured text and must be encoded into a computer representation, bearing in mind the heterogeneity of response.
Patient Questionnaire Data: Specifically designed questionnaires are used in studies of diseases with psychosocial basis. Analysis of such data usually provides a starting point for classification of cases and then for further investigation of the existence of a possible biological background.	Questionnaires usually include both close- and open-ended questions. The close-ended questions generate numerical attributes. The open-ended questions result in an unstructured data	Open ended questions generate data which may be further mined using computation linguistic approaches. This may require specialist domain ontologies such as those associated with the UMLS (Nelson *et al.*, 2002).

Table 2. Biological data sources

Data Sources	Characteristics of the data	Integration Issues
Gene Expression Profiles: The mRNA expression profile of diseased cells may reflect the unique genetic alterations present and has been shown to be predictive of clinical and biological characteristics of illness for many diseases. A major issue in these data is the unreliable variance estimation, complicated by the intensity-dependent technology-specific variance (Weng *et al.*, 2006).	cDNA microarray is the high throughput analysis of global gene expression within a biological specimen. Gene expression measurements (e.g. relative levels of expression between tumour and normal cells) are made simultaneously for many thousands of genes.	Comparing gene expression measurements between different technologies and between measurements on the same technology at different times is a challenge handled by normalisation techniques. A specialised markup language for microarray data is in (Spellman *et al.*, 2002). Furthermore, the number of replicated microarrays is usually small because of cost and sample availability, resulting in unreliable variance estimation and thus unreliable statistical hypothesis tests.
Single Nucleotide Polymorphisms (SNPs): The analysis of SNPs within the human genome will enhance our understanding of underlying genetic variations that exist in the human population. Individual SNPs are being associated with specific diseases and have been correlated to altered drug response in pharmacogenomic analyses (Aplenc and Lange, 2004).	Increasing numbers of examples of single base pair variations within the coding region of genes which, whilst not being a mutation which leads to a defective protein, are associated with altered activity of the protein (Goto *et al.*, 2001). Larger blocks of genetic variation, called haplotypes, are also being assessed in so called *Haploblock* studies.	There is a need to establish statistically significant correlations between SNPs and disease or outcome of treatment through association studies. High throughput analysis of SNPs, with up to 100000 different variations are now achievable.

Table 3. Biological knowledge bases

Data Sources	Characteristics of the data	Integration Issues
Domain Ontologies and other Databases: GO (The Gene Ontology Consortium, 2000) and other biological and medical ontologies and databases (e.g PubMed, TRASER, Swiss-Prot, etc.) are publicly available over the Internet.	GO (The Gene Ontology Consortium, 2000) is a main public curated vocabulary of over 17,000 terms and allows association of biological 'functionality' with gene products.	Integration adds context and knowledge about genes. Issues arise when matching records between databases as the primary key used to index entities often differs depending on the owner of the database.

Table 4. Healthcare data sources

Data Sources	Characteristics of the data	Integration Issues
Health Management Data: Retrospective cost-benefit assessment of clinical trials are often conducted by health managers so as to improve the broader management strategies and financial resource allocation for Departments.	Patient visits to inpatient and outpatient wards/clinics, total cost of medication, efficiency of service delivery, consultation time and study comparison analysis. Quality of life measurements. palliation vs effective cure. Effect of new screening test with regards to benefit etc.	Privacy issues, updating costs of drugs over times. Comparison of different drugs between countries. This is performed retrospectively with findings difficult to implement in future trials.

The "Extract-Explain-Generate" Methodology

We present the general outline of the "Extract-Explain-Generate" methodology, motivated by the multifactorial and multilevel nature of biomedical data. A schematic of the methodology is shown in Fig. 3.

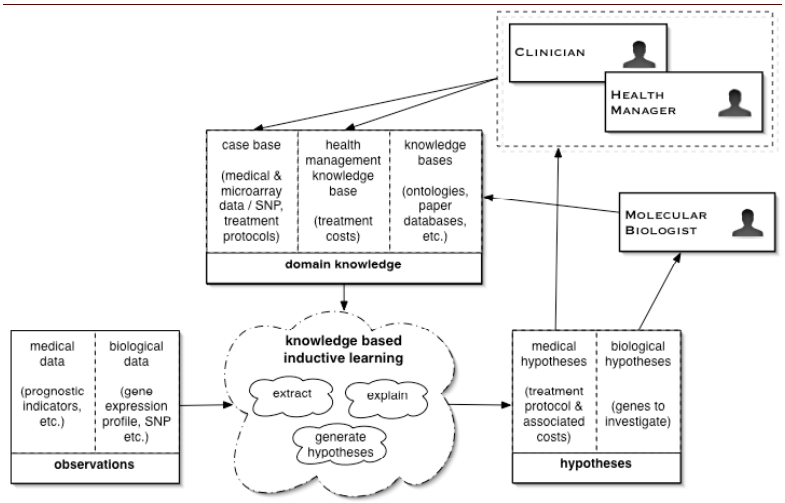

Fig. 3. The "Extract-Explain-Generate" methodology.

The methodology is centred on our technology-mediated knowledge based inductive learning process. It analyses new observations in the context of the available domain knowledge. As illustrated in Fig. 3, the observations of a new patient and domain knowledge related to the case, possibly including existing biological hypotheses, are the inputs to the knowledge based inductive learning process. The output of the process includes one or more medical hypotheses. These hypotheses assist

(i) the clinician to formulate a treatment protocol and understand how a patient differs from other previous patients;
(ii) the biological researcher to identify biological markers (possibly genes or other indicators) for future investigation, and;
(iii) the health manager to understand the costs associated with treatment.

Once validated by these end-users, the hypotheses are used to update the domain knowledge. Domain knowledge can be categorised into three classes:
(i) a case base of previous patients together with outcomes of the treatment protocol applied;
(ii) public knowledge bases of biomedical information (including domain ontologies and databases); and
(iii) health management information.

The knowledge-based inductive learning step facilitates reuse of acquired knowledge in the context of prior domain knowledge. Although the processes in the knowledge based inductive learning step are different for each of the types of hypotheses and for different problems, these processes may be grouped under three categories: "extract", "explain" or "generate hypotheses". Medical hypotheses (together with treatment outcomes) are used to update the database of cases. Biological hypotheses eventually lead to updates of the knowledge bases involved in the process in Fig. 3.

In order to position and interpret the results of the analysis of microarray data in the context of other existing biological knowledge, we utilize available domain ontologies, electronic libraries and other databases (referred in the literature also as "omic knowledge" libraries). Currently, the bioinformatics tools that process such sources are restricted to deal with one type of "omic knowledge", e. g. particular gene ontology, or interactions. Promising for our approach are the efforts in the development of mechanisms and protocols to deal with any type of omic knowledge for example the work on Omic Space Markup Language (OSML) (Hasegawa *et al*., 2006).

Overall, the proposed integrative methodology takes into account and incorporates biological, clinical and economic aspects of the medical treatment. This is also indicated by the explicit presence of the roles of Clinician, Healthcare Manager and Molecular Biologists in Fig. 3. As "Extract-Explain-Generate" methodology involves in each step one or more data mining and analytics experts, these roles are not explicitly shown in the diagram in Fig. 3. It provides a broad framework for constructing consistent instances of case study designs, including required data mining support for specific cases. We illustrate how these instances are formed on the cases of Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. In the first case study the focus is on the anatomy of the "Extract" step, when in the second study the focus is on the anatomy of the "Explain" step.

Case Study 1: Chronic Fatigue Syndrome

In this section we demonstrate an instance of the application of the "Extract-Explain-Generate" methodology to a biomedical problem: identification of biological markers underlying Chronic Fatigue Syndrome. In the following subsections we describe the problem and the goals of our study. Then we construct and apply an instance of our methodology to this particular problem and describe the outcomes of the investigation.

Problem

Chronic Fatigue Syndrome (CFS) (Afari and Buchwald, 2003) is an illness with a primary symptom of debilitating fatigue over a six month period. Currently diagnosis of CFS is generally made by clinical assessment of symptoms using a number of surveys measuring functional impairment, quantifiable measurements of fatigue and occurrence, duration and severity of the symptoms (Reeves *et al*., 2005). A primary goal of current research is to derive a definition of the syndrome, which goes beyond a clinical assessment of symptoms to an empirical diagnosis founded on an established biological lesion. The motivation for this kind of research is to gain a clearer understanding of the illness and to find empirical guidelines for its diagnosis.

The Goals of the Study

The goal of our study is to investigate whether there is a biological basis to CFS. To this end we interrogate an integrated dataset of clinical, blood evaluation and gene expression data to identify patterns differentiating patients suffering from fatigued (CFS and other fatigued individuals (ISF) with insufficient severity of symptoms to be classified as suffering from CFS) to non-fatigued (NF) individuals.

We use publicly available data for CFS and NF individuals from the Critical Assessment of Microarray Data Analysis (CAMDA 2006) competition datasets (CDC Chronic Fatigue Syndrome Research Group, 2006). The integrated data set that we composed, comprises two clinical data sets, one giving survey results for patients for the above mentioned fatigue and symptom questionnaires, the other giving complete blood evaluation results for patients. This involved 139 CFS/ISF patients and 73 NF individuals. Gene expression data for a subset of the CFS/ISF and NF patients (118 CFS/ISF and 53 NF) was also available (consisting of around ten thousand genes for each sample), together with SNP data and proteomics data. We did not use the SNP or proteomics data in our investigation although it can easily be incorporated within our methodology.

These data cover the full biological spectrum from genotype to phenotype (see Fig. 1). Investigators have developed a stratification of CFS which characterises its clinical significance (National Center for Infectious Diseases, 2006). Their initial hypothesis stated that gene expression profiling would allow them to establish prognostic indicators of the syndrome. We have queried this assumption and asked whether there is a biological basis to CFS or does it have a purely psychosocial

aetiology? In particular we have focussed on whether we can identify a pathological lesion for CFS in peripheral blood. Put more simply, in what way, and to what extent, are the SNP, gene expression, proteomic and blood chemistry profiles different between non-fatigued (NF) subjects and those with CFS or a fatigue syndrome? These questions influenced the outline of the study scenario and the components included in the integrated data set.

The Study Scenario

A specific instance of the "Extract-Explain-Generate" methodology was applied to the problem of identifying a biological basis to CFS. It is thought that CFS is unlikely to be caused by a single agent (Afari and Buchwald, 2003). This multifactorial nature of the problem domain motivates us to extend the "Extract-Explain-Generate" methodology to take a complex systems approach towards the analysis, illustrated in the schematic in Fig. 4. We use the labels "global" and "local" patterns to distinguish between patterns derived from and valid over the integrated data set and patterns that are generated from a subset of attributes. For example, clusters of patients can be global patterns, if they are generated out of clinical and gene expression data in the integrated data set. Global patterns are aimed at establishing deep linkages between the attributes and within and between the components of the integrated data set that explain or question assumptions about the phenomenon. Therefore, we label approaches and algorithms seeking them as "constructionist". If we use just a subset of attributes, like a list of individual genes, but not all genes, then the patterns derived are local. Such reductionist approaches are common in microarray data analysis and are used to discover biomarkers for diseases. Local patterns can be viewed as the output of a reductionist approach in predictive modeling, where one looks at the attributes that allow generating accurate predictive models, without necessary providing an explanation about the underlying phenomenon.

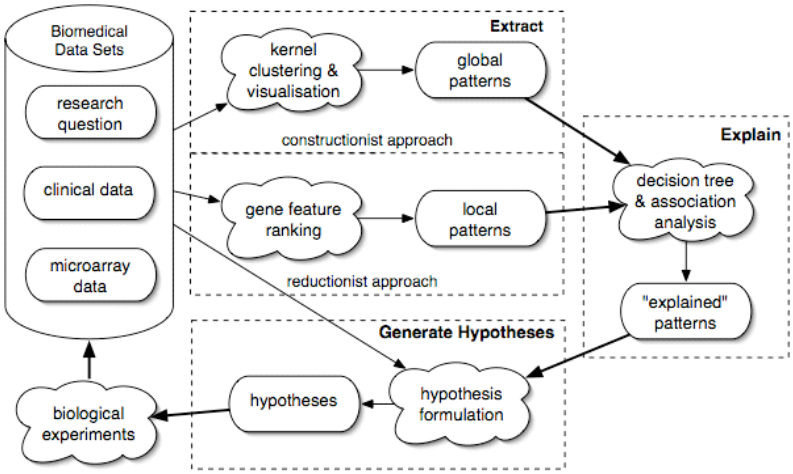

Fig. 4. An instance of "Extract-Explain-Generate" applied in the Chronic Fatigue Syndrome case study.

PK: the caption has separated from the figure when I view it.

We posit that such reductionist approaches are less useful for highly dimensional datasets and for multifactorial diseases for two reasons. Classification between classes of patient in high dimensional datasets is susceptible to the "curse of dimensionality" where the biological markers (genes) chosen differentiate training examples but do not generalise well to unseen data. This is a result of insufficient patient samples compared to data items collected per sample. It is infeasible to collect sufficient data items for the extremely high dimensional gene expression data (consisting of thousands of attribute values i.e. genes per patient). Secondly, with the predicted 'multifactorial' nature of CFS, it is likely to be multigenic in nature, governed by small changes in many genes rather than a simple genetic defect involving a single gene.

Consequently, we take a data driven approach towards the interrogation with the aim of getting a better understanding of the phenomena before phrasing a specific biological hypothesis. This approach aims to avoid the introduction of unnecessary bias.

"Extract" step

The data is pre-processed before the "Extract" stage. In particular, some attributes of the "illness" dataset, the clinical dataset containing the patient's answers to the illness questionnaires, are omitted because they are

(i) skewed with almost all individuals having the same attribute value;
(ii) not deemed useful for the data mining effort; or
(iii) calculated by the original researchers and would bias our efforts.

The attributes concerned are "DOB", "intake classific", "cluster", "onset", "yrs ill", "race" and "ethnic". The dependent variable "Empiric" is used as the patient class and patient subtypes are combined to make three classes CFS, ISF and NF. In the other clinical dataset, concerning the complete blood evaluation for patients, we add a copy of the "Empiric" attribute. The datasets are linked by the patient identifier attribute "ABTID". The gene expression datasets are combined into a single dataset for all individuals and the "Emipric" patient class linked as described above. Each gene for each patient in the gene expression dataset consists of four attribute values: "Spot Label" with the gene name, and three statistical measures of the gene expression value including standard deviation and mean of values within the spot. The statistical measures of the gene expression are normalised over all arrays and patients by multiplying values with the average value of every gene over all arrays divided by the average value of every gene over the individual array. We create integrated datasets of pairs and the triplet of the individual pre-processed datasets.

As discussed above, the "Extract" step in this case takes a complementary constructionist and reductionist approach. The constructionist schema combines a kernel-based clustering and visualisation method to the integrated data set. This method (Shawe-Taylor and Cristianini, 2004) finds a low-dimensional projection of the integrated dataset such that the distance between points in the projection is similar to the distance in the kernel induced feature space. The linear kernel is used and

additional pre-processing is applied to data in the clinical datasets where all attribute values are recoded to numeric values, patient class information is omitted from calculation of the kernel (to get an unbiased visualisation) and the data is centred and normalised. The global patterns identified are clusters of patients in the space of low-dimensional projection of the original data.

Efficient calculation of the kernel matrix for the gene expression data requires special treatment. Each row of the gene expression dataset represents an individual gene measurement for a particular microarray (for each patient). The straightforward approach of calculating the linear kernel matrix is to concatenate the rows of the gene expression dataset into a matrix consisting of one row for each array with a set of attribute values for each spot label ("ARM Dens - Levels", "MAD - Levels" and "SD - Levels") then to calculate the linear kernel by multiplying the matrix with its transpose. Clearly this approach and the corresponding algorithms are impractical in our situation because of the large number of genes on each the array. We developed a more efficient approach, motivated by computational linguistics, for direct computation of the linear kernel matrix from gene expression data. The kernel value for two samples (i.e. microarrays) is calculated from sorted lists of genes (spot labels) associated with each array. The kernel value is calculated as the sum of the product of the attribute values for genes matching in both lists. Computation of the linear kernel matrices for the integrated datasets is simply a matter of adding the linear kernel matrices for the individual datasets.

The reductionist schema in this case is based on the Gene Feature Ranking (GFR) method, developed by the team. It calculates a rank that measures the separation between fatigued (CFS/IFS) and non-fatigued (NF) data points for genes. Each gene is assigned a rank corresponding to the Euclidean distance in terms of the normalised averaged "ARM Dens – Levels" and "MAD – Levels" values (in the gene expression dataset) for the 119 patients classified as fatigued and the corresponding averaged values for the 53 non-fatigued patients. Larger ranks correspond to spot labels that better discriminate the classes of patient. Similarly, distances are calculated for the other gene expression measures ("MAD – Levels" and "SD – Levels"). The ranked genes are evaluated through an SMO (Sequential Minimal Optimisation) Support Vector Machine (SVM) classifier (Platt, 1998; Keerthi *et al.*, 2001) with test error estimated with 10-fold cross-validation again using the linear kernel function. By analogy with the Newton family of numerical methods for finding the roots of polynomial equations, we developed a strategy with variable step size in order to identify the optimum number of genes that result in the best classification.

"Explain" step

The global patterns found with the kernel clustering and visualisation and the local patterns discovered by the gene feature ranking algorithm are explored in the "Explain" stage with decision tree and association analysis (Hastie *et al.*, 2001). In this study we focus on the global patterns by applying decision tree analysis separately to three subsets of the complete blood evaluation clinical dataset with respect to the patient class as defined by the "Empiric" attribute. Association analysis examined rules where the patient class is the "consequent" of the rule. The details are presented in the Outcomes subsection of this case study.

The Outcomes

We present the outcomes of the case study following the same structure as in the Study Scenario section, distinguishing the outcomes related to the "Extract" and "Explain" steps respectively.

"Extract" step

The "Extract" step of the methodology identifies patterns in the data for further explanation. The constructionist approach identifies global patterns in the integrated dataset, whilst the reductionist approach, in this problem, looks for patterns in the gene expression data set only.

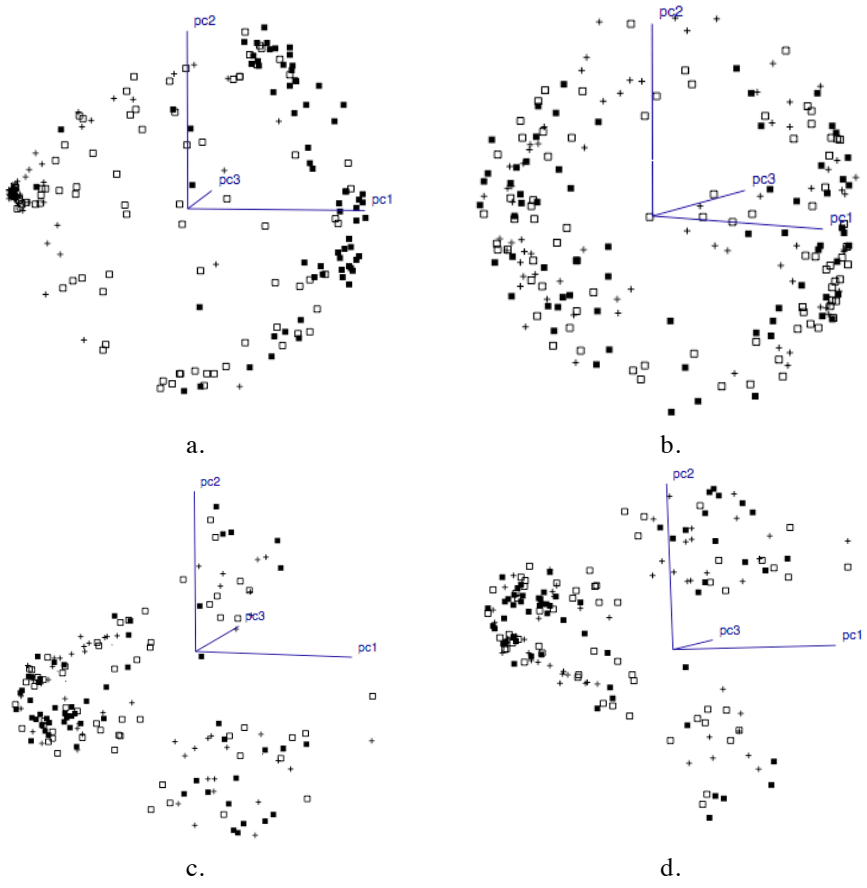

Fig. 5. Kernel-based clustering and visualisation in 3 dimensions of (a) illness dataset, (b) blood dataset, (c) gene expression dataset and (d) the integrated blood, illness and gene expression dataset. Legend: + = NF patient, □ = ISF patient, ■ = CFS patient.

Fig. 5 shows some of the global patterns found as a result of the constructionist part of the "Extract" step. These results are the input to the interactive 3D visual data

mining system, which offers various functions for exploring the visual space. The global patterns are evident as points in the 3-dimensional space comprising the first three principal components of the projection of points into the kernel feature space. Results are shown for the individual datasets and the triplet, but not for the pairs of datasets.

The kernel based visualisation of the illness dataset (i.e. the survey information) (Fig. 5a) clearly shows that the NF patients cluster together. This is expected because medical professionals make the classification of patients into CFS, ISF and NF on the basis of information in the survey data. One CFS patient near this region appears to be clustered incorrectly. However, medical professionals use two different schemes to classify patients and, using the other classification scheme, this patient is categorized as ISF. In other words, this patient is a border line case.

Less structure is evident in the visualization of the blood dataset in Fig. 5b. This suggests that there may not be strong biological markers evident in the complete blood evaluation of patients. The clustering of the gene expression dataset in Fig. 5c shows three clear clusters which do not strongly correspond to the patient classes. This also suggests to us that there may not be a clear biological basis in the gene expression values. These results reinforce the multifactorial notions of this disease.

Results from the reductionist (GFR) approach within the "Extract" step are illustrated in Fig. 6. This diagram shows the accuracy of SVM classification with different numbers of the top gene feature ranking ranked genes/spots.

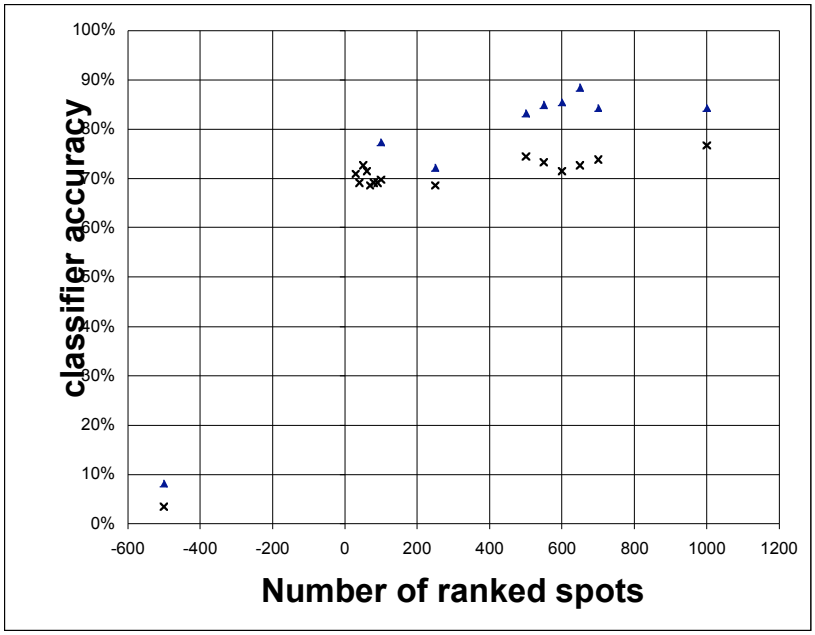

Fig. 6. Accuracy of ranked spots classifiers. Legend: × = "ARM Dens – Levels" and "SD – Levels"; ▲ = "MAD – Levels" and "SD – Levels".

The leftmost points on the graph use the 500 *lowest* ranked genes for classification to show the magnitude of the difference between classification accuracy at both ends of the ranking scale. The graph shows that many spots are required to reach acceptable classification accuracy. Reductionist approaches like this assume (most likely incorrectly) that factors affecting the outcome of the classification act independently. If there is an attribute that is strongly correlated with another attribute, the advice (Occam's razor) is to remove it. Hence, the fewer attributes - the better. However, genes may not necessarily fit well in this modeling scheme, due to variety of possible interactions. Hence they are most likely not independent. The large number of genes necessary to achieve reasonable classification accuracy in Fig. 6 suggests, again, that there is not a clear biological marker consisting of a small number of genes to discriminate between NF and CFS/ISF patients.

"Explain" step

The goal of the "Explain" step of the methodology is to provide an explanation or background to the patterns found in the "Extract" step. As discussed above, decision tree and association analysis was applied to the clinical datasets.

Interrogation of the complete blood evaluation data with decision tree and association analysis indicated few differences at the cellular level between the blood samples obtained from CFS vs ISF vs NF patients. There were slight imbalances in a range of attributes associated with red blood cells (RBC) and this may be characteristic of a fatigued patient. The 'imbalances' however, were mostly in the normal range for these attributes within the general population and could not independently be used to diagnose a fatigue syndrome. For example, the trained decision tree found that all CSF patients had a Mean Corpuscle Volume (MCV) ≥ 81.15 fl (normal range 86±10fl). Similarly, the ISF patients were found to have a Mean Corpuscle Haemaglobin (MCH) ≥ 26.45pg, however, the normal range is 29.5±2.5pg. The biological interpretation of this is that, whilst the RBC attributes are not sufficient to characterise a 'fatigued' patient as having a form of anaemia, the imbalances may point to slight inefficiencies in O2 distribution of CFS and ISF patients. The attributes, however, are not sufficient of themselves to be used as a diagnostic marker for a fatigue syndrome nor may they reflect the underlying biological basis for the syndrome. That said, decision tree analysis identified that the NF samples were identified by CO2 ≥ 21.4 units (58 of the 63 patients matching also MCH > 33.45 and anion gap ≥ 21.4). However, the CFS patients were identified by CO2 ≤ 28.9 units. Given that the normal range is between 20-30 units it appears that, for this attribute at least, the difference identified by the decision trees may represent different distributions between the test and control patient cohorts, with both cohorts having values within a range found to be normal within the wider general population. Clearly however, the biological differences between the blood count and chemistry of the fatigued and NF patients is minimal and not useful as an independent classifier of CFS. The imbalances detected may however, in combination with the other data available allow for the construction of a multifactorial or multigenic classifier for fatigue syndromes. Indeed, F Test analysis for the MCV and MCH variables indicate that sample variances between the CSF and NF populations (excluding the ISF samples) were significantly different ($p<0.05$ or 0.028 and 0.048 respectively).

Decision tree and association analysis of the illness (survey) dataset identified rules that agreed with those used to empirically classify patients (National Center for Infectious Diseases, 2006). For example, NF patients were characterised with General Fatigue < 12.5, Reduced Activity < 9.5, no exclusion and no current MDDM. Similar rules used by (National Center for Infectious Diseases, 2006) were identified for the CFS and ISF patients.

Case Study 2: Acute Lymphoblastic Leukaemia

In this section we demonstrate the application of the "Extract-Explain-Generate" methodology to another biomedical domain: the treatment of cancer, in particular, the sadly common childhood cancer acute lymphoblastic leukaemia (ALL). We follow similar presentation structure and in the following subsections we describe the problem and the goals of our study. Then we show how to apply an instance of our methodology for this particular problem together with the outcomes of the investigation.

Problem

ALL is the, sadly, most common childhood malignancy. It represents 24% of all new cancers occurring in children between 1995 and 1999 (240 ALL/985 Cancer patients) in NSW, Australia (Australian Institute of Health and Welfare (AIHW) & Australasian Association of Cancer Registries (AACR), 2004). Today, 75-80% of children with ALL survive. Current treatments incorporate systemic therapy (eg combination chemotherapy) and specific central nervous system (CNS) preventative therapy. Successful treatment of childhood ALL requires the control of systemic disease in many body systems (bone marrow, spleen, CNS, etc) as well as treatment of extramedullary disease, specifically in the CNS to bring about 'clinical remission' (ie. the disappearance of traces of the disease). Since nearly all children with ALL achieve an initial clinical remission, the major obstacle to cure is patient relapse, i.e. the recurrence of evident disease. Relapse from remission can occur during therapy or after completion of treatment and can occur in various sites. The prognosis for an ALL patient with recurrent disease depends on the site of relapse and the duration of remission prior to recurrent disease (Henze *et al*., 1991).

The Goals of the Study

This investigation focuses on finding genes in the massive gene expression datasets most associated with the high risk ALL patients and then trying to understand how these genes are alike. For this study we interrogate clinical and gene expression data from the Children's Hospital at Westmead (CHW). The focus of this study is more on the gene expression data rather than integrated data sets. High risk (compared to normal risk) ALL patients are indicated by an attribute in the clinical dataset. The patient data includes cDNA microarray and clinical data for 9 patients. Usually

between 2 and 10 repeat experiments of the same data (ie. patient) are made and for each patient, there are around 9000 genes with between 2 and 10 log ratios (ie. experiment repeats) for each gene. Clinical data describes a patient in detail, as well as the effect of different treatment protocols. Of the nine patients, 4 are labelled as high risk. In this study we are interested only in the physician-based indicator of risk stratification and cDNA microarray values.

The Study Scenario

There is a shift in paradigm from classification and prediction of cancer and treatment outcome, respectively, to utilizing data mining technologies for getting a deeper understanding of the mechanisms that govern the cancer, in particular, understanding the molecular basis of histologic grade to improve prognosis and treatment (Sotiriou *et al*., 2006). The instance of the "Extract-Explain-Generate" methodology specifically created for this case study is shown Fig. 7.

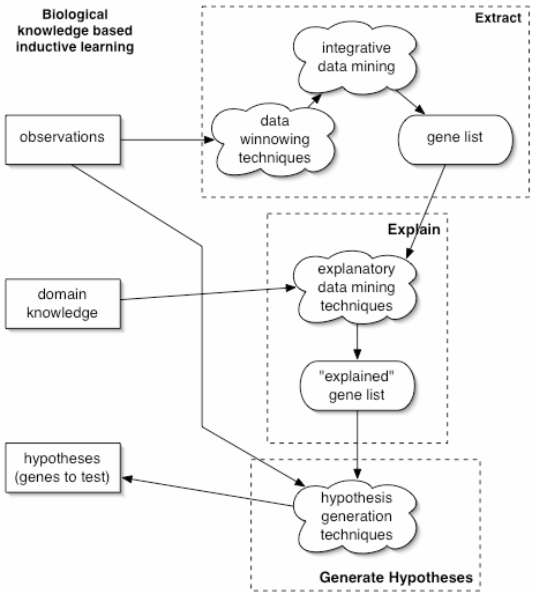

Fig. 7. An instance of "Extract-Explain-Generate" methodology applied in the ALL case study.

"Extract" step

For this problem we apply data winnowing techniques to extract local patterns in the data. We use matrix decompositions to identify genes indicative of high risk ALL patients. In particular, a combination of techniques based on singular value decompositions (SVD) and semi-discrete decompositions (SDD) are used to winnow

the thousands of genes tested for each patient to dozens of genes most indicative of ALL. See (Skillicorn *et al*., 2004) for the details of the winnowing techniques.

"Explain" step

Whilst the list of genes that is the outcome of the data winnowing techniques is interesting from a statistical point of view, it is difficult for biological interpretation. The list of genes can be characterised as "data rich" but "knowledge poor" in that the raw list of genes does not contain much domain specific knowledge. In the "Explain" step we enrich the list of genes with domain specific knowledge, specifically the associated terms from the Gene Ontology (The Gene Ontology Consortium, 2000). This large publicly available collaborative domain ontology contains over 16,000 terms from three hierarchies (biological processes, cellular components and molecular functions) and associations to genes. Gene products are described in terms of their effect and known place in the cell. Terms in the hierarchies are related by "is-a" and "part-of" parent-child relationships and we use these relationships to define a similarity function for use in cluster analysis (Kennedy and Simoff, 2003; Kennedy *et al*., 2004). Using a similarity function defined over the ontology allows us to cluster genes in groups with similar functionality defined in terms of the interrelationships between terms in the ontology. Finally, descriptions of each cluster are found by examining Gene Ontology terms that are representative of the cluster.

The "Generate" step searches for other genes that would have fallen into the same clusters the biologist is interested in. This is done through the Gene Ontology associations.

The Outcomes

"Extract" step

Results from the winnowing techniques applied to the gene expression data in the "Extract" step are shown in Fig. 8. Each point in the diagram represents a gene with distances between genes dependent on their correlation. The symbols reflect results from the semi-discrete decomposition. The details of the algorithms are presented in (Skillicorn *et al*., 2004). The genes of interest are those grouped together far from the origin. The topmost genes of interest were selected for analysis in the "Explain" step with the cluster analysis.

"Explain" step

Terms from the Gene Ontology are associated with the genes selected from the "Extract" step and cluster analysis is carried out with the MBSAS algorithm (Theodoridis and Koutroumbas, 1999) using the similarity function described in (Kennedy and Simoff, 2003). Gene clusters found are listed in Table 5.

The subset of terms associated with the clustered genes on a graph are shown in Fig. 9. The clusters are represented by the five large boxes. Nodes inside the clusters are the GO terms associated with genes in that cluster. More general terms are on the right hand side of the diagram. Edges between nodes represent the links in the

ontology. Each node is shown in only one box, but links between the boxes show where GO terms are shared by genes in the different clusters. The colour (level of grey) of the link represents the cluster that link is in. The darker colours (shades) represent GO terms and links that were in the original dataset whilst the lighter colors (shades) show relationships that are inferred from traversing the ontology. The details of the clusters are presented in Table 6.

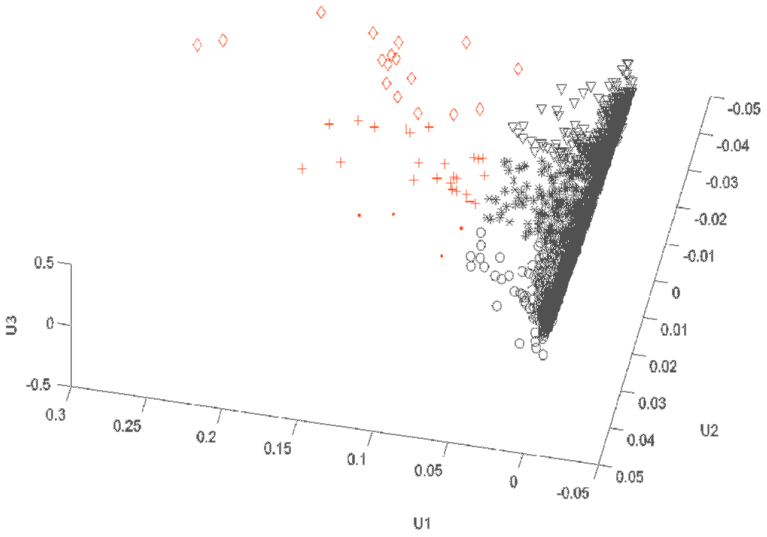

Fig. 8. Genes clustered according to the matrix decomposition methods. Axes are the first three principal components.

Table 5. Discovered clusters with Gene Ontology clustering in the "Explain" step.

Cluster	Genes (GenBank accession codes)
0	AA040427 AA406485 AA434408 AA487466 AA609609 AA609759
1	AA046690 AA644679
2	AA055946 AA398011 AA458965 AA487426 AA490846 AA504272
3	AA112660 AA397823 AA443547 AA447618 AA455300 AA478436 AA608514 AA669758 AA683085
4	AA126911 AA133577 AA400973 AA464034 AA464743 AA486531 AA488346 AA488626 AA497029 AA629641 AA629719 AA629808 AA664241 AA664284 AA668301 AA669359 AA683050 AA700005 AA700688 AA775874

Though clustering methods have been adapted and applied to microarray data, their mathematical techniques do not show biologically relevant information on the clustering results. Cluster analysis with a domain ontology permits the automatic

induction of cluster description from the most general terms associated only with each cluster (Kennedy and Simoff, 2003; Kennedy *et al*., 2004). This novel methodology for biological interpretation of gene clusters utilizes the hierarchical nature of GO terms to select possible biological interpretation of the gene clusters. Similar to our approach has been proposed in (Lee *et al*., 2004). The BayGO developers (Vêncio *et al*., 2006) also went in a similar direction. They used Bayesian models to separate the the genes from a given category that are not observable in the are observable in the microarray data due to low intensity signal, quality filters, genes that were not spotted and so on.

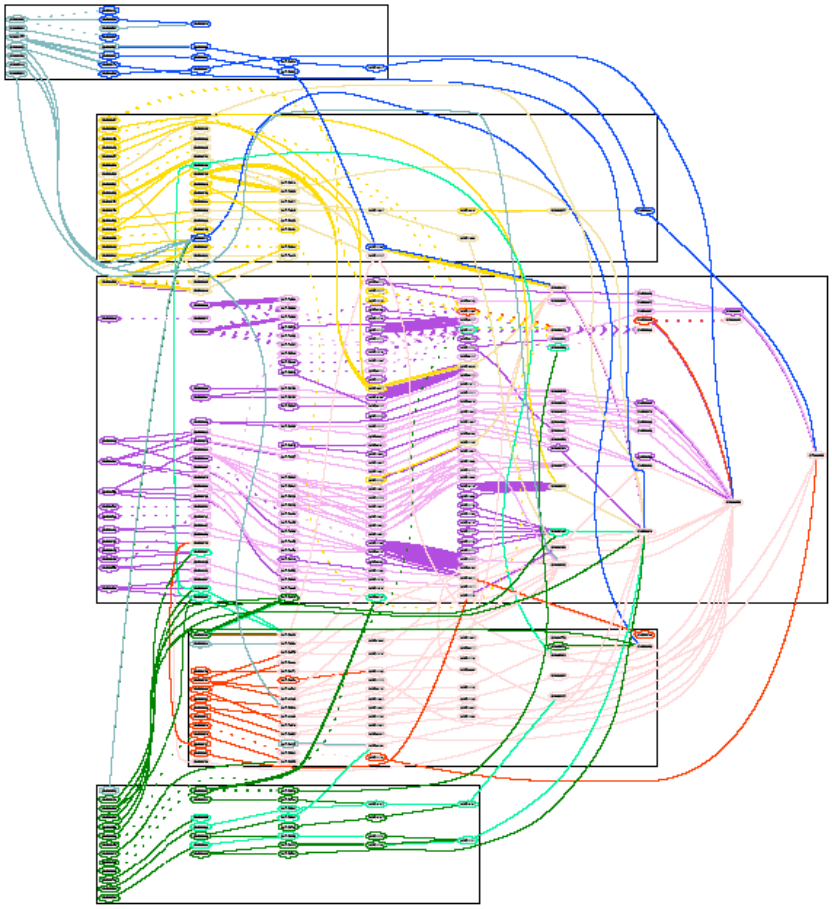

Fig. 9. Diagram showing parts of the GO hierarchy associated with genes being clustered. More general terms are at the right of the diagram. See text for description of graph.

Table 6. Principal cluster descriptions for genes.

GO ID	GO Term and number of associated genes	
Cluster 0 - 6 genes		
20 GO terms but each associated with only one gene		
Cluster 1 - 2 genes		
GO:0008092	cytoskeletal protein binding activity	2
GO:0007028	cytoplasm organization and biogenesis	2
GO:0003774	motor activity	2
GO:0005875	microtubule associated complex	2
5 GO terms but each associated with only one gene		
Cluster 2 - 6 genes		
GO:0004871	signal transducer activity	4
GO:0007154	cell communication	4
GO:0005887	integral to plasma membrane	3
GO:0005886	plasma membrane	3
GO:0005194	cell adhesion molecule activity	2
11 GO terms but each associated with only one gene		
Cluster 3 --- 9 genes		
GO:0030528	transcription regulator activity	4
GO:0008134	transcription factor binding activity	3
GO:0006366	transcription from Pol II promoter	3
GO:0003700	transcription factor activity	3
GO:0006357	regulation of transcription from Pol II promoter	3
5 GO terms but each associated with only two genes each		
13 GO terms but each associated with only one gene		
Cluster 4 --- 20 genes		
GO:0003723	RNA binding activity	10
GO:0030529	ribonucleoprotein complex	9
GO:0009059	macromolecule biosynthesis	9
GO:0006412	protein biosynthesis	9
GO:0005829	Cytosol	9
GO:0003735	structural constituent of ribosome	8
2 GO terms but each associated with only four genes each		
5 GO terms but each associated with only three genes each		
1 GO term associated with only two genes		
33 GO terms but each associated with only one gene		

Discussion and Conclusions

We have presented a methodology for the analysis of biomedical data with emphases on its use by clinicians for diagnosis of patients and by biological researchers for facilitating biological understanding of diseases. The methodology places emphasis on the human-centered "Extract-Explain-Generate" cycle which extracts patterns from the associated data sets, explains the data by supplementing it with additional domain knowledge or with other techniques and then generates hypotheses for future testing by clinicians and biologists.

The methodology supports the multilevel multifactorial nature of biomedical data and data sources and we described some of the kinds of data sources used in biomedical data mining and the sorts of issues inherent in the integration of this data.

We demonstrated the use of instances of the methodology in two complex biomedical case studies: investigation of the biological basis underlying chronic fatigue syndrome and investigation of genes indicative of paediatric acute lymphoblastic leukaemia. The first of these case studies, in particular, illustrated the use of the methodology in dealing with the multilevel, multifactorial nature of the biomedical domain by adaptation of the "Extract" step of the methodology to look for both global patterns spanning the integrated data and local patterns examining subgroups of attributes in the data.

The methodology has been developed by researchers from the the Children's Hospital at Westmead (Sydney, Australia), the University of Technology, Sydney, Queen's University, Kingston, and Silicon Graphics. Currently, the team has focused on the refinement of the "Generate" step of the methodology and completion of the "clinical assistant" – a computer based system that utilizes the described approach in order to provide timely expertise to clinicians and biologists working in the area of childhood cancer.

Acknowledgments

The authors would like to thank Seyhan Yildiz and Ying Du for the decision tree and association analysis work on the CAMDA datasets. The research work presented in this chapter has been supported by the Children's Hospital at Westmead (Sydney, Australia), the University of Technology, Sydney and Silicon Graphics.

References

Afari, N. and Buchwald, D. (2003). Chronic Fatigue Syndrome: A review. *American Journal of Psychiatry*, **160**: 221-236.

Aplenc, R. and Lange, B. (2004). Pharmacogenetic determinants of outcome in acute lymphoblastic leukaemia. *British Journal of Haematology*, **125**(4): 421-434.

Australian Institute of Health and Welfare (AIHW) & Australasian Association of Cancer Registries (AACR) (2004). Cancer in Australia 2001. *AIHW cat. no. CAN 23*. Canberra: AIHW (Cancer Series no. 28).

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. M. and Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, **7**: 54-70.

CDC Chronic Fatigue Syndrome Research Group (2006). CAMDA 2006 Conference Contest Datasets. *[cited; Available from: http://www.camda.duke.edu/camda06/datasets/, viewed at 12 January 2008.*

Curran, M. D., Liu, H., Long, F. and Ge, N. (2003). Statistical methods for joint data mining of gene expression and DNA sequence database. *SIGKDD Explorations*, **5**(2): 122-129.

Dietzsch, J., Gehlenborg, N. and Nieselt, K. (2006). Mayday-a microarray data analysis workbench. *Bioinformatics*, **22**(8): 1010-1012.

Felix, C. A., Lange, B. J. and Chessells, J. M. (2000). Pediatric acute lymphoblastic leukemia: Challenges and controversies in 2000. *Hematology*, **Jan 2000**: 285 - 302.

Georgii, E., Richter, L., Ruckert, U. and Kramer, S. (2005). Analyzing microarray data using quantitative association rules. *Bioinformatics*, **21**((Suppl. 2)): ii123-ii129.

Glenisson, P., Mathys, J. and Moor, B. D. (2003). Meta-clustering of gene expression data and literature-based information. *SIGKDD Explorations*, **5**(2): 101-112.

Goto, Y., Yue, L., Yokoi, A., Nishimura, R., Uehara, T., Koizumi, S. and Saikawa, Y. (2001). A novel single-nucleotide polymorphism in the 3'-untranslated region of the human dihydrofolate reductase gene with enhanced expression. *Clinical Cancer Research*, **7**: 1952-1956.

Hasegawa, Y., Seki, M., Mochizuki, Y., Heida, N., Hirosawa, K., Okamoto, N., Sakurai, T., Satou, M., Akiyama, K., Iida, K., Lee, K., Kanaya, S., Demura, T., Shinozaki, K., Konagaya, A. and Toyoda, T. (2006). A flexible representation of omic knowledge for thorough analysis of microarray data. *Plant Methods*, **2**(1): 5-46.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Heidelberg, Springer-Verlag.

Henze, G., Fengler, R., Hartmann, R., Kornhuber, B., Janka-Schaub, G., Niethammer, D. and Riehm, H. (1991). Six-year experience with a comprehensive approach to the treatment of recurrent childhood acute lymphoblastic leukemia (ALL-REZ BFM 85). A relapse study of the BFM group. *Blood*, **78**(5): 1166-1172.

Hoffman, E. P., Awad, T., Spira, A., Palma, J., Webster, T., Wright, G., Buckley, J., Davis, R., Hubbell, E., Jones, W., Tibshirani, R., Tompkins, R., Triche, T., Xiao, W., West, M. and Warrington, J. A. (2004). Expression profiling - best practices for data generation and interpretation in clinical trials. *Nature Reviews: Genetics*, **4**: 229-237.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, **13**: 637-649.

Kennedy, P. and Simoff, S. J. (2003). CONGO: Clustering on the Gene Ontology. *Proceedings 2nd Australasian Data Mining Workshop, ADM03*, Canberra, UTS Press: 181-198.

Kennedy, P. J., Simoff, S. J., Skillicorn, D. and Catchpoole, D. (2004). Extracting and explaining biological knowledge in microarray data. *Proceedings 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD 2004*, Sydney, Australia, Springer Berlin/Heidelberg: 699-703.

Lee, S. G., Hur, J. U. and Kim, Y., S (2004). A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**(3): 381-388.

National Center for Infectious Diseases (2006). Proposal: clinical assessment of subjects with Chronic Fatigue Syndrome and other fatiguing illnesses in Wichita. [cited; Available from: ftp://ftp.camda.duke.edu/CAMDA06_DATASETS/wichita_clinical_irb_protocol.doc.].

Nelson, S. J., Powell, T. and Humphreys, B. L. (2002). The Unified Medical Language System (UMLS) project. *Encyclopedia of Library and Information Science*. A. Kent and C. M. Hall, Eds. New York, Marcel Dekker, Inc**:** 369-378.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L., Eds. (2003). *The Analysis of Gene Expression Data: Methods and Software*. New York, Springer.

Piatetsky-Shapiro, G., Khabaza, T and Ramaswamy, S. (2003). Capturing best practice for microarray gene expression data analysis. *Proceedings of the 9-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD-2003*, Washington, D.C., ACM Press.

Piatetsky-Shapiro, G. and Tamayo, P. (2003). Microarray data mining: Facing the challenges. *SIGKDD Explorations*, **5**(2): 1-5.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning*. B. Schölkopf, C. Burges and A. Smola, Eds. Boston, MIT Press**:** 185-208.

Reeves, W. C., Wagner, D., Nisenbaum, R., Jones, J. F., Gurbaxani, B., Solomon, L., Papanicolaou, D. A., Unger, E. R., Vernon, S. D. and Heim, C. (2005). Chronic Fatigue Syndrome - A clinically empirical approach to its definition and study. *BMC Medicine*, **3**(19).

Seifert, M., Scherf, M., Epple, A. and Werner, T. (2005). Multievidence microarray mining. *Trends in Genetics*, **21**(10): 553-558.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y. and Elkon, R. (2005). EXPANDER – an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**: 232 - 244.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge, Cambridge University Press, pp 282-285.

Skillicorn, D. B., Simoff, S., Kennedy, P. and Catchpoole, D. (2004). Strategies for winnowing microarray data. *Bioinformatics Workshop, SIAM International Conference on Data Mining 2004*: 42-51.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M. and Delorenzi, M. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**(4): 262-272.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert Jr, C. J. and Brazma, A. (2002). Design and implementation of

microarray gene expression markup language (MAGE-ML). *Genome Biology*, **3**(9): 1-9.

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature - Genetics*, **25**: 25-29.

Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. San Diego, USA, Academic Press.

Vêncio, R. Z. N., Koide, T., Gomes, S. L. and Pereira, C. A. d. B. (2006). BayGO: Bayesian analysis of ontology term enrichment in microarray data. *BMC Bioinformatics*, **7**(1): 86-116.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. and Bassett, D. E. (2006). Rosetta error model for gene expression analysis. *Bioinformatics*, **22**(9): 1111-1121.