International Journal of Biometrics and Bioinformatics (IJBB)

VOLUME 3, ISSUE 4

ISSN : 1985-2347

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

Copyrights © 2009 Computer Science Journals. All rights reserved.

Editor in Chief Professor João Manuel R. S. Tavares

International Journal of Biometrics and Bioinformatics (IJBB)

Book: 2009 Volume 3, Issue 4
Publishing Date: 30 - 08 -2009
Proceedings
ISSN (Online): 1985 - 2347

This work is subjected to copyright. All rights are reserved whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illusions, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication of parts thereof is permitted only under the provision of the copyright law 1965, in its current version, and permission of use must always be obtained from CSC Publishers. Violations are liable to prosecution under the copyright law.

IJBB Journal is a part of CSC Publishers
http://www.cscjournals.org

©IJBB Journal
Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Table of Contents

Volume 3, Issue 4, August 2009.

Toward Integrated Clinical and Gene- Expression Profiles For Breast Cancer Prognosis: A Review Paper

Farzana Kabir Ahmad farzana58@uum.edu.my
Graduate Department of Computer Science,
College of Arts and Sciences,
Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia

Safaai Deris safaai@utm.my
School of Postgraduate Studies,
Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia

Nor Hayati Othman hayati@kb.usm.my
Clinical Research Platform & Pathologist,
Health Campus Universiti Sains Malaysia,
16150 Kubang Kerian, Kelantan,Malaysia

Abstract

Breast cancer patients with the same diagnostic and clinical prognostics profile can have markedly different clinical outcomes. This difference is possibly caused by the limitation of current breast cancer prognostic indices, which group molecularly distinct patients into similar clinical classes based mainly on the morphology of diseases. Traditional clinical-based prognosis models were discovered to contain some restrictions to address the heterogeneity of breast cancer. The invention of microarray technology and its ability to simultaneously interrogate thousands of genes has changed the paradigm of molecular classification of human cancers as well as shifting clinical prognosis models to a broader prospect. Numerous studies have revealed the potential value of gene-expression signatures in examining the risk of disease recurrence. However, most of these studies attempted to implement genetic-marker based prognostic models to replace the traditional clinical markers, yet neglecting the rich information contained in clinical information. Therefore, this research took the effort to integrate both clinical and microarray data in order to obtain accurate breast cancer prognosis, by taking into account that these data complement each other. This article presents a review of the development of breast cancer prognosis models, concentrating precisely on clinical and gene-expression profiles. The literature is reviewed in an explicit machine-learning framework, which includes the elements of feature selection and classification techniques.

Keywords: Breast cancer, Prognosis, Gene-Expression Profiles, Feature selection, Classification.

1. INTRODUCTION

Cancer is a class of disease or disorder characterized by the uncontrolled division of cells to spread, either by direct growth into adjacent tissues through invasion or by implantation into distant sites by metastasis. Breast cancer, on the other hand, has become a major cause of cancer-related morbidity and mortality among female worldwide and remains a major health burden. Although in previous years most of the researches were concerned about diagnosing breast cancer, it is only recently that cancer researchers have attempted to look at cancer prognosis. This idea actually is a part of a growing trend towards personalized and predictive medicine. Prognosis can be defined as (pro: before; gnoscere: to know) foreknowledge of an event before its possible occurrence [1]. There are three prognosis foci, which are cancer susceptibility, cancer recurrence and cancer survivability [2].

Cancer recurrence has been attracting a lot of attention from patients and physicians. The major clinical problem of breast cancer recurrence is by the time primary tumor was diagnosed, microscopic or clinically evident metastases have already occurred. Although breast cancer patients are prescribed various types of treatments such as chemotherapy, endocrine and radiation therapy or even go through surgery there is no assurance that metastases will never occur. Despite significant advances in cancer treatment, the ability to predict the metastases behavior of tumors still remains one of the greatest clinical challenges in the oncology field. Various studies have been conducted to predict breast cancer recurrence. Traditional cancer prognosis relies on a complex and inexact combination of clinical and histopathological data. Age, tumor size, estrogen and progesterone receptors, and lymph-node involvement are some of the clinical and histopathological factors used in the conversional prognosis method. These classic approaches, however, may fail when dealing with atypical tumors or morphologically-indistinguishable tumor subtypes. The cause of these incidents is breast cancer is an extensively heterogeneous disease, which is not only based on clinical information but also involves genes-cellular proliferations.

Advances in the area of microarray-based expression have led to the promise of cancer prognosis using new molecular-based approaches. It has become a standard tool in many genomic research laboratories. The reason for this popularity is that microarrays have revolutionized the approach of biology research. Instead of working on a single gene basis, scientists can now study thousand of genes at once. Unfortunately, microarray data are often overwhelmed, over-fitting and confused by the complexity of data analysis. Although many studies are trying to solve these issues, most results reported are data dependent. Moreover, it is noticed that clinical data are often underused and a lot more focus is given to microarray data. This paper attempts to review the classification techniques employed in clinical and microarray data as well as explore feature-selection techniques that have been applied in extracting significant gene signatures. In addition, related works that integrate both clinical and microarray data were reported, even though only a few studies had been conducted in this area. The aim of this review is to develop a breast cancer prognosis model that incorporates both clinical and gene-expression profiles data, which could enhance and accurately predict the outcome. Furthermore, it would assist physicians make informed decisions regarding the potential necessity of adjuvant treatment and consequently this could ultimately contribute to the decrease in overall breast cancer mortality.

The remainder of this paper is organized as follows. Section 2 briefly explains the breast cancer domain, differentiates between benign and malignant tumors and identifies current prognostic indices that are applied in classifying breast cancer patients. The discussion is followed by revealing the most dominant classification techniques in cancer prognosis using clinical data in section 3. DNA microarray technology and its complexities which are associated with microarray data analysis are discussed explicitly in section 4. Section 5, meanwhile, addresses the works that have been done in feature selection using microarray data, which can be divided into three main groups; univariate, wrapper and embedded approach. The discussions of classification techniques in microarray data analysis are compared and related works in integrating clinical and

gene-expression data are conveyed in section 6 and section 7 respectively. Subsequently, validation methods in estimating prediction errors for microarray data analysis are explained in section 8. Section 9 then reveals the trends and directions in prognosis models. Lastly, section 10 offers concluding remarks.

2. BREAST CANCER

Breast cancer is a neoplastic disease, where normal body cells can be transformed into malignant (cancerous) ones. It is the most common cancer in women worldwide and the second leading cause of morbidity after lung cancer among Malaysian women [3]. Breast cancer can be grouped into two different tumor types, benign and malignant tumors. These tumors are different from one another in such a way that benign tumors do not spread, but malignant tumors, as in breast cancer, are made up of cells that can spread to and damage other parts of the body through the lymphatic systems or invade adjacent tissues. The cancer spreading mechanism can happen in three stages; local, where the cancer is confined to the breast or certain parts, which means the lymph nodes, primarily those in the armpits are involved. It is also possible cancers are found in other parts of the body, known as distant spreading.

The guidelines for early detection of breast cancer include breast self–exams (BSE), clinical breast examination (CBE) and screening mammogram [4]. BSE is a visual and manual examination of the breast that can be easily carried out by women, while CBE is the physical examination of the breast conducted by a trained medical or health professional. On the other hand, screening mammography is the most common imaging procedure for diagnosing breast cancer usually among women who are asymptomatic (have no complaints or symptoms of breast cancer). The goal of screening mammogram is to detect cancer when it is still too small to be felt by a woman or her physician. In order to determine whether an area of concern in a breast (found by BSE, CBE, or screening mammogram) is malignant (cancerous) or benign (not cancerous), a physician may perform a biopsy test. A breast biopsy is the removal of a sample of breast tissue for laboratory examination by a pathologist and is the only definitive way to determine if an abnormality is cancerous or not. Moreover, the biopsy result also indicates the cancerous stage as well as the appropriate treatments to be prescribed.

Despite the advance in diagnosis, breast cancer prediction remains a challenging task to physicians and patients. Currently, four prognostic indices are used to predict breast cancer patients, which include TNM Classification of Malignant Tumors (TNM), the National Institute of Health (NIH), the St. Gallen criteria and the Nottingham Prognostic Index (NPI) [5]. However, these cancer classifications have been based primarily on the morphological appearance of the tumor and have serious limitations. Tumors with similar histopathological appearances can follow significantly different clinical courses and show different responses to therapy. It is estimated that 70% of patients receiving chemotherapy or hormone therapy would have survived without them [6]. Nevertheless, many patients do not respond to specific treatments such as tamoxifen, which is a standard adjuvant treatment for patients with primary estrogen receptor-positive breast cancer [7]. This phenomenon proves that physicians have difficulties in deciding the appropriate treatments, which may lead to unnecessary adjuvant treatments, associated risks and expensive medical costs, whereas patients are more aware and demand for treatment that could improve the quality of life.

3. CLASSIFICATION TECHNIQUES IN CANCER PROGNOSIS USING CLINICAL DATA

Prognosis plays an important role in patient management tasks like treatment planning as well as evaluating the quality of health and the consequences of disease progression. Approaches to develop prognosis models vary from using traditional probabilistic techniques, obtained from the field of statistics, to more qualitative and model-based techniques, originating from artificial intelligence (AI). In the last decade, most of the prognosis models were based on regression

analysis such as the proportional hazard model and the Kaplan-Meier Curve [8, 9]. The Kaplan-Meier Curve is a nonparametric analysis and usually has some problems due to confidence bounds, which is wider than those calculated via parametric analysis. As a result, predictions outside the range of observations are not possible. In 1994, Burke *et al*.[10], compared the performance of Artificial Neural Networks (ANN), Logistic Regression and Principal Component Analysis (PCA) with traditional staging system termed TNM staging system (primary tumor, regional lymph nodes and distant metastases) and the results showed that ANN was superior than other statistical methods. The results were later confirmed by Laurentiis *et al*.[11], Jerez-Aragones *et al*.[12] and Kates *et al*.[13]. The idea behind this finding was the ANN ability in adding a large number of parameters that could enhance the accuracy of the prognosis model.

ANN is an information processing paradigm that is inspired by the nervous system, such as the brain. The key element of this paradigm is the interconnected processing elements called neurons working in unison to solve specific problems. The learning mechanism in the ANN system involves adjustments to the synaptic connections that exist between the neurons. Moreover, the ANN methodology represents a useful alternative to classical modeling techniques when applied to variable data sets presenting non-linear relationships. Therefore, ANN has broadly used in implementing various cancer prognosis models [14-18] to address the problems of highly correlated prognostic factors and censored data handling. Although the ANN technique has dominated many cancer prognosis models, it suffers mainly from two problems; first the selection of architecture and the value of the parameter involved and second, understanding the underlying rules is impossible since it is a black box processing system.

In contrast to ANN, Decision Tree (DT) represents outputs as a set of symbolic rules. Formally a DT is structured in a graph or a flow chart of nodes which will be used to determine the ultimate goal. In the case of cancer prognosis, the aims of most researchers usually can be categorized into two distinct classes; i) decision support system [16, 19] (for example, the probability of survival, recurrence within 5 year interval time) or ii) identifying prognostic factors in cancers [20-22]. Although DT is easy to interpret and can handle various types of data including numeric, nominal, and categorical data, missing values for an attribute can lead to ambiguity in choosing the right branch. Moreover, it may generate too many rules, which make it hard to be understood. Concerned with the importance to provide rules clarification for determining cancer prognosis and addressing the limitation of DT, another type of technique called XCS was introduced in [23]. XCS is a type of learning classifier system that consists of a set of rules and procedures for performing and discovering patterns. Later, a new rule-driven compaction approach was employed to obtain a new piece of knowledge and the results exemplified that XCS outperformed DT.

Support Vector Machine (SVM) is another type of classification technique. The underlying concept in SVM algorithm is to create a hyperplane that separates the data into two classes within the maximum margin. Like ANN, SVM can be used to perform non-linear classification using non-linear kernel. Lee *et al*.[24] has applied SVM to extract prognostic factors and to classify breast cancer patients into 3 different classes; i) good prognosis\node-negative patients (patients with no metastasized lymph nodes), ii) intermediate prognosis\node-positive patients (patients with 1 to 4 metastasis lymph nodes) iii) poor prognosis\node-positive patients (patients with more than 4 metastasis lymph nodes). However, from the literature review, we found out that SVM is almost unfamiliar compared to ANN and DT in the field of cancer prognosis. The same conclusion was mentioned in [2]. Other technique such as k-nearest neighbor is also rarely applied in this domain.

In addition, several common clinical prognostic factors that frequently had been employed to predict breast cancer recurrence were noted. The common ones were; age, lymph-node involvement, tumor size, histological grade. The next section will describe the problems associated with DNA microarray data analysis in examining cancer prognosis.

4. DNA MICROARRAY AND COMPLEXITIES

Microarray offers an efficient method of gathering data that can be used to determine the expression pattern of thousands of genes. The mRNA expression pattern from different tissues in normal and diseases states could reveal which genes and environmental conditions can lead to disease.

The experimental steps of typical microarray began with extraction of mRNA from a tissues sample or probe. The mRNA is then labeled with fluorescent nucleotides, eventually yielding fluorescent (typically red) cDNA. The sample later is incubated with similarly processed cDNA reference (typically green). The labeled probe and reference are then mixed and applied to the surface of DNA microarrays, allowing fluorescent sequences in the probe-reference mix to attach to the cDNA adherent to the glass slide. The attraction of labeled cDNA from the probe and reference for a particular spot on microarray depends on the extent to which the sequences in the mix (probe-reference) complement the DNA affixed to the slide. A perfect compliment, in which a nucleotide sequence on a strand of cDNA exactly matches a DNA sequence affixed to the slide, is known as hybridization. Hybridization is the key element in microarray technology.

The populated microarray is then excited by a laser and the consequential fluorescent at each spot in the microarray is measured. If neither the probe nor the reference samples hybridize with the gene spotted on the slide, the spot will appear in the black color. However, if hybridization is predominantly with the probe, the spot will be in red (Cy5). Conversely, if hybridization is primarily between the reference and DNA affixed to the slide, the spot will fluoresce green (Cy3). The spot can also incandescent yellow, when cDNA from probe and reference samples hybridize equally at a given spot, indicating that they share the same number of complementary nucleotides in particular spot. The process of microarray experiment is illustrated in Figure 1.

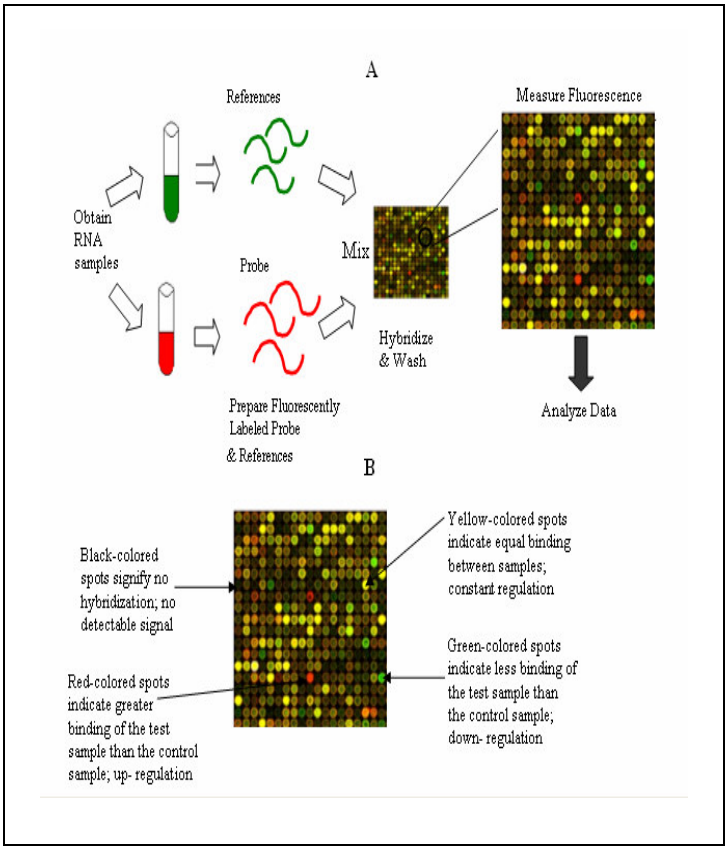

FIGURE 1: Microarray Experiment

Using image processing software, the red-to-green fluorescence will be digitized and providing the ratio values output indicating the expression of genes. Finally, the data of all samples are incorporated into one table constructing gene expression matrix G as shown in Figure 2. The rows of G correspond to single genes and the columns to single samples. Due to its high throughput nature, microarray data poses new challenges for data analysis. Computational approaches are generally necessary to divulge data structures. Although the type of analysis depends on the research questions posed, typical steps in the analysis of microarray data are; i) pre-processing and normalization, ii) detection of genes with significant fold changes, iii) classification and clustering of expression profiles. However, this paper will only focus on feature selection and classification techniques.

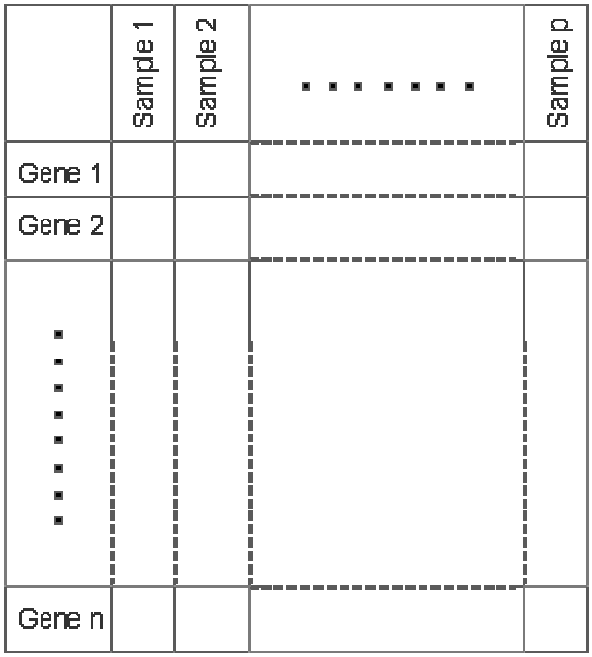

FIGURE 2: Gene Expression Matrix G

4.1 Microarray Data Analysis Problems

Although the invention of DNA microarray has opened a new opportunity to monitor thousands of genes simultaneously, there are many challenging problems in microarray data analysis that need to be addressed before new knowledge about gene expression can be revealed. Some of the problems are:

i. Bias and confounding problems, which occurred during the study-design phase of microarray which can lead to erroneous conclusion [25, 26]. Technical factors, such as differences in physical, batch of reagents used and various levels of skill in technicians could possibly cause bias. Confounding, on the other hand, takes place when another factor distorts the true relationship among the variables of interest.
ii. Cross-platform comparisons of gene-expression studies are difficult to conduct when microarrays are constructed using different standards. Thus, the results cannot be reproduced. To deal with this problem, Minimal Information About a Microarray Experiment (MIAME) [27] has been developed to improve reproducibility, sensitivity and robustness in gene-expression analysis.
iii. Microarray data is high dimensional data characterized by thousands of genes in a few sample sizes, which cause significant problems such as irrelevant and noise genes,

complexity in constructing classifiers, and multiple missing gene-expression values due to improper scanning. Moreover, most of the studies that applied microarray data suffered from data over-fitting, which required additional validation.

iv. Mislabeled data or questionable tissue results by experts are also other types of drawbacks that could decrease the accuracy of experimental results and lead to imprecise conclusions about gene-expression patterns [28].

v. Biological relevancy result is another integral criterion that should b taken into account in analyzing microarray data rather than only focusing on accuracy of cancer classification. Although there is no doubt in them gaining high accuracy, classification results are important in microarray data analysis. However, revealing biological information during the process of cancer classification is also essential. For instance, determination of genes that are under-expressed or over-expressed in cancerous cells could assist domain experts in designing and planning more appropriate treatments for cancer patients. Therefore, most of the domain experts are interested in classifiers that not only produce high classification accuracy but also reveal important biological information [29] .

5. FEATURE-SELECTION TECHNIQUES IN MICROARRAY DATA ANALYSIS

Feature-selection techniques, also known as gene-selection techniques have become a prerequisite in many large-scale gene-expression data analysis. The advance in genomic studies along with the exponential accumulation of microarray data has altered the feature-selection paradigm from being an optional to a compulsory need. It is because by cutting down the number of features to a sufficient minimum, classification performance can be improved. The taxonomy of dimensionality-reduction techniques can be divided into two categories; transformation or selection-based reduction. The key distinction made within the taxonomy is whether a dimensionality-reduction technique will transform or preserve the data set semantics in the process of reduction. Transformation-based reduction such as Principal Component Analysis (PCA) transforms the original features of a data set with a typically reduced number of uncorrelated ones, termed principal component. In contrast, selection-reduction techniques attempt to determine a minimal feature subset from a problem domain while retaining the meaning of the original feature sets. Thus, selection-based reduction techniques have become the main preference in many bioinformatics applications, especially microarray data analysis. This is due to its advantage of interpretability by a domain expert. The objectives of feature-selection techniques are various. The major ones are [30]:

i. To avoid over-fitting and improving model performance, for example, selecting highly informative genes could enhance the accuracy of the classification model.
ii. To provide faster and more cost-effective models, and
iii. To gain a deeper insight into the underlying processes that generated the data.

Although, feature-selection techniques have many benefits, it also introduces an extra complexity level, which requires a thoughtful experiment design to address the challenging tasks, yet provide fruitful results. Feature-selection methods can be structured into three factions; filter methods, wrapper methods and embedded methods. Filter methods rank each feature according to some univariate metric, and only the highest-ranking features are used while the remaining low-ranking features are eliminated. This method also relies on the general characteristics of the training data to select some features without involving any learning algorithm. Therefore, the results of the filter model will not affect any classification algorithm. Moreover, filter methods also provide very easy ways to calculate and can simplify large- scale microarray data sets since it only has a short running time.

Univariate filter methods such as Bayesian Network [31], Information Gain (IG), Signal-to-Ratio (SNR) [32-35] and Euclidean Distance [33, 34], have been extensively used in microarray data to identify informative genes. Information Gain has been reported to be the superior gene-selection technique by Cho et al. and Hu et al. [33, 36]. However, different types of univariate techniques

appear to be significant when they are trained over various data sets. Bayesian Networks, on the other hand appear to be the ideal platform for the integration of heterogeneous sources of information [37]. Besides the application of parametric techniques in determining informative genes from microarray data, Ben Dor *et al.*[38], Barash *et al.* [39] and Rogers *et al.* [40] had applied non-parametric techniques such as the threshold number of misclassification or TNoM score. This technique basically separates the informative gene by assigning a threshold value. However, it is hard to determine the most appropriate threshold. Other non-parametric techniques such as Pearson correlation coefficient [33, 34] and Significant Analysis of Microarray (SAM) [41] have been reported to be top feature-selection techniques.

Univariate filter methods have been widely utilized in microarray data analysis. This trend can be clarified by a number of reasons, for instance, the output or the result provided by univariate gene rankings are intuitive and easy to understand. These simplified versions of output could fulfill the aims and expectations of biology and molecular-domain experts who demand for validation of results using laboratory techniques. In addition, filter methods also offer less computational time to generate results which is an extra point to be preferred by domain experts. However, gene-ranking based on univariate methods has some drawbacks. The major one is the genes selected are most probably redundant. This means highly-ranked genes may carry similar discriminative information toward the defined class. Although we eliminate one high-ranked gene it may not cause any degradation of classification accuracy.

Since univariate filter methods do not count the relationship between genes, Koller and Sahami [42] developed an optimal gene-selection method called Markov Blanket Filtering, which can remove redundant genes to eliminate this problem. Based on this method, Yu and Liu [43] proposed the Redundancy Based Filter (RBF) method to deal with redundant problems and the results are quite promising. While the filter techniques handle the identification of genes independently, the wrapper method embeds a gene-selection method within a classification algorithm. In the wrapper methods [44] the search is conducted in the space of genes, evaluating the goodness of each found gene subset by the estimation of the percentage of accuracy of the specific classifier to be used, training the classifier only with the found genes. The wrapper approach, which is very popular in machine-learning applications, is not comprehensively used in DNA microarray tasks and only few works in the field make use of it [45, 46]. It is claimed by many authors [45, 47] that the wrapper approach obtains better predictive accuracy estimates than the filter approach. However, its computational cost must be taken into account. Wrapper methods can be divided into distinct groups; deterministic and randomized-search algorithm. Genetic Algorithm (GA) is a randomized-search algorithm and optimizes the mimicking of evolution and natural genetics. It has been employed for binary and multi-class cancer discrimination studies [48, 49]. A common drawback of wrapper methods, such as GA, is that they have a higher risk of over-fitting than filter techniques and are very computationally intensive. In contrast, wrapper methods incorporate the interaction between gene selection and classification model, which make them unique compared to filter techniques.

The third class of feature-selection approaches is embedded methods. The difference of embedded methods with other feature-selection methods is the search mechanism is built into the classifier model. Identical to wrapper methods, embedded methods are therefore, specific to a given learning algorithm. Embedded methods have the advantage in that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods. Choosing an appropriate feature-selection technique is essential in obtaining accurate and precise results. The next section will describe, in detail, the classification techniques that have been applied in microarray data analysis.

6. CLASSIFICATION TECHNIQUES IN MICROARRAY DATA ANALYSIS

The development of microarray, based high throughout gene profiling, has led to the promising endeavor to classify tumors with an accurate and efficient means for predicting prognosis as well

as providing effective treatments. Many researchers have been studying problems associated with cancer classification using different gene-expression profiles data and attempting to propose the optimal classification technique to solve these problems. Several machine-learning techniques were previously used in classifying gene-expression data, including Fisher Linear Discriminant Analysis [51], k-Nearest Neighbor [49],[33],[41], Decision Tree, Multi-layer Perceptron [52, 53], Support Vector Machine [28],[54], Boosting [38], and Self-Organizing Map [35]. However, there is no single classifier that is superior over the rest, since the performances of classifiers also depend on gene-selection methods and the size of the data sets employed. Moreover, some of the methods only work well on binary-class problems and are not extensible to multi-class problems, while others are more general and flexible.

K-Nearest Neighbor is a non-parametric classifier that classifies the expression values of each gene based on the majority voting. It has been extensively used in microarray data analysis due to it robust characteristic to noisy and enormous training data and has been one of the first choices for a classification study when there is little or no prior knowledge about whether the distribution of the data is available [33, 41, 49, 55, 56]. Utilized as binary categorical classifiers, kNN has been noted to be a prominent technique in order to identify a subset of predictive genes from large noisy data [33, 49], which has been tested over the same three benchmark data sets; colon, leukemia and lymphoma data. However, these studies showed some diversity in results, which is basically due to different types of feature-selection methods and the choice of distance functions used in such as Euclidean Distance, Manhattan, and Pearson. Although kNN is reported as a known technique for classification, the main drawback of this technique is due to its non–scalability restriction, which is computationally intensive for large data sets. Therefore, this technique may be inappropriate to be used in cancer classification since the availability of gene-expression data sets probably increase and it requires too much computational time, unless prior efficient gene selection is done. In addition, the choice of the number of neighbors (k) is also another problem that needs to be taken into account [57].

Unlike kNN, Support Vector Machine (SVM) is scalable. SVM was introduced by Vapnik [58, 59] and successively extended by many other researchers. The fundamental idea behind this classifier can be viewed as a process of finding a max-margin hyperplane that separates the training tuples into different groups according to their corresponding classes. SVM's remarkable robust performance with respect to sparse and noisy data makes it preferable in a number of applications, especially in microarray data analysis, whether in binary or multi-class cancer classification [28, 33, 34, 38, 48, 54, 60]. Furey *et al.* [28] has applied SVM linear kernel with a signal-to-ratio feature-selection technique on colon, leukemia and ovarian data sets. Their results demonstrated that SVM not only can accurately classify new samples, but also assist in the identification of mislabeled samples by experts. However, this classification is fragile with respect to SVM parameter settings. Softness of margin and the number of genes selected as input could affect the correctness of the classifying sample.

Ben Dor *et al.* [38] on the other hand attempted to evaluate SVM linear kernel and SVM quadratic kernel performance using the same data sets. The results gained using colon data set were found aligned with the finding of Furey *et al.* [28], which stated that SVM linear kernel work well compared to complex kernel. However, it was noticed to be contradictory in the leukemia data set. This inconsistency may be due to the amount of gene-expression values in the leukemia data set, which is enormous, compared to the colon data set. Therefore, more complex kernel was required to be applied. Linear SVM also reported to be the most successful classifier in the studies of Symons et al. and Al-Shalalfa et al. [61, 62] and has been shown to consistently outperform other classification approaches including kNN[48, 63].

It is also noted that in past years, researchers relied on a single classifier and gene-selection method to analyze gene-expression data. However, the trends then shifted to investigating the performance of several classifiers over a few selected gene-selection techniques as were being done by Cho and Won [33] and Hu *et al.* [34] but it has become apparent that no particular classifier works well over different data sets. The main drawback of the SVM classification

technique is, similar to kNN, it is computationally expensive, thus the run-time is long and slow. Moreover, it originally suited binary class problems. As a result multi-class SVM lately is being studied [48, 64] and is still an on-going research problem.

Artificial Neural Networks (ANN) is another classification technique that was used in analyzing microarray data sets. It can model and reveal complex relationships among inputs (gene-expression patterns) and outputs (class-decisions) exemplified or embedded in the training data through different structures, linear or non-linear transfer functions and adjustment of weight-connection between nodes. Although there is still considerable skepticism about ANN among statisticians and bioinformaticians due to its black box approach, ANN has been applied in a broad category of class-prediction problems especially by domain experts. Examples of ANN in gene-expression profiles classification can be seen in studies of Khan *et al.* [52] Peterson *et al.* [65], Ringner *et al.* [66], Tusch *et al.* [67], Wei *et al.* [68], Bevilacqua *et al.* [69] and Eden *et al.* [70], which discussed the parallelism that exists among different ANN and concluded that ANN does offer several advantages such as unified approaches for feature extraction and classification and flexible procedures for finding good, moderately non-linear solutions.

The Bayesian Network proposed by Pearl [71], is a graphical model that encodes probabilistic relationships among variables of interest with mathematically-grounded framework. This graphical model has been used widely in analyzing gene-expression data [72, 73]. Meanwhile, Huang *et al.* [74] and West *et al.*[75] have applied the Bayesian technique to classify gene-expression values which are associated with the lymph node and estrogen-receptor status for breast cancer patients. These studies showed that the prognosis of the lymph node and the estrogen-receptor status are important elements and significant factors in accurate prediction of disease course. The preference toward this technique relies on the structure of the model, which encodes dependencies among all variables, thus it readily handles situations where some data entries are missing. Moreover, Bayesian Network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. In addition, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. However, the main limitation of this technique is its assumption of a linear dependency of a child node on its parents, which is unrealistic since most regulatory relationships between genes are highly non-linear.

Despite the success of classification techniques reported in past years, none of them are superior to others. Hence, it is more desirable to make a decision by combining the results of various expert classifiers rather than by depending on the result of only one classifier. Ensemble approaches lately have become an on going research area [76-78]. Liu *et al.* [79] attempted to propose a combinational feature-selection method in conjunction with ensemble neural networks. Three feature-selection methods was adopted, which consisted of the Ranksum test, PCA, clustering and t test and each gene-selection result was then presented as input into three neural networks. In contrast, Kim *et al.* [80] had selected seven correlation analysis of feature selections in combination with multi-layer perceptron (MLP), k-nearest neighbor (KNN), the support vector machine (SVM) and the structure adaptive self-organizing map system (SASOM). Although these studies analyzed gene-expression data from different angles, they proved to enhance generalization capability compared to the single classifier.

7. RELATED WORKS IN INTEGRATING CLINICAL GENE-EXPRESSION DATA

In spite of enormous work being done in analyzing clinical and gene-expression data, only a few studies have focused on integrating clinical and gene-expression data, although many researchers agreed that clinical and genetic markers do complement each other and improve the prediction accuracy compared to those made by using clinical or gene expression alone [2, 74, 81-85]. To our knowledge, initial efforts to combine these two different data was done by Futschik *et al.* [81] by constructing separate classifiers for microarray and clinical data to predict the

outcome for diffuse large B-cell lymphoma (DLBCL). Evolving fuzzy neural network (EFuNN) has been used to construct a microarray predictor module. Meanwhile, the International Prediction Index (IPI) as Bayesian Network was applied to develop a clinical predictor module. The predictions of the two independent modules were merged into a single prediction, which led to higher accuracy compared to the previously most accurate prognostic model.

Furthermore, Bayesian Network was noted to be a preference technique in combining clinical and gene-expression data. It also has been used by Nevins *et al.* [82] to extend the invasion of axillary lymph nodes with meta-gene signatures, while in 2006 Gavaert *et al.* [31] employed Bayesian Network to evaluate three methods for integrating clinical and microarray data; decision integration, partial-integration and full-integration to perform breast cancer prognosis. These studies revealed that Bayesian Network can be used to combine clinical and gene-expression data and boost the performance of breast cancer prognosis. Given the complexity of breast cancer prognosis and the difficulties in extracting significant genes and clinical source, Sun *et al.* [83] developed I-RELIEF algorithm to identify hybrid factors from clinical and microarray data. This study has shown that a hybrid signature can provide significantly improved prognostic specificity over the existing gene signatures and current clinical systems.

On the other hand, Li *et al.* [84] applied SVM with linear kernel to combine clinical information and gene-expression profiles to accurately discriminate ovarian cancer patients who were likely to respond to therapy treatment. Features-selection using the T-test statistical analysis was used to extract significant gene expression and clinical data for developing prediction mode and the results showed an increase in average accuracy of the integrated model compared to the base SVM model. Even though some issues were addressed in these studies like heterogeneous factors involved in prognosis cancer and the challenging task in integrating both data, it was confirmed that gene-expression data add immense detail to traditional clinical source. Thus, the combination of clinical and gene-expression data could lead to customized health care strategies. However, a more practical and appropriate strategy needs to be developed to encounter heterogeneity in clinical and gene-expression data.

8. VALIDATION METHODS IN ESTIMATING PREDICTION ERRORS FOR MICROARRAY DATA ANALYSIS

The growing avalanche of microarray data has driven an explosion of high-throughput and discovery-based research during the past decade. Although a large number of researchers claimed to have successfully discovered the gene-expression markers for cancer prognosis, most of the researches cannot be reproduced, which consequently lead to disappointment and erroneous conclusions. These issues often arise when miniature or no validation is carried out during the research process. The sources of ambiguity in microarray studies are numerous and can occur from different stages, for example, experimental design, data quality (laboratory, platform, and batch effects), preprocessing (image analysis, normalization and filtering) and data analysis [25, 86, 87]. Each of these sources could generate uncertainty in gene-expression data, therefore requiring careful consideration and validation.

The predictive accuracy of a model can be validated using a cross-validation study, in which the analysis is repeatedly performed while removing a group of samples at reanalysis and predicting the outcome for the remaining group. Cross-validation can be used simply to estimate the generalization error of a given model, or it can be used for model selection by choosing one of several models that has the smallest estimated generalization error. Two types of cross-validation techniques have been widely used in microarray data analysis which includes leave-one-out cross-validation (LOOCV) [32, 38, 48, 88-90] and k-fold cross validation [91, 92]. LOOCV often works well for estimating the generalization error for continuous error functions such as the mean-squared error, but it may perform poorly for discontinuous error functions such as the number of misclassified cases. In the k-fold cross validation, the generalization error is preferred. However, if k gets too small, the error estimate is pessimistically biased because of the difference in

training-set size between the full-sample analysis and the cross-validation analyses. A value of 10 for k is popular for estimating generalization error[17, 19, 20, 23, 34].

Another validation method, which is extensively applied in the microarray data analysis, is receiver-operating characteristic, also known as ROC curves[31, 41, 66, 70, 83]. Most of the microarray studies are concerned about correctly classifying tumors by measuring the fraction of false positive (also known as false positive rate (FPR)) and true positive (also known as true positive rate (TPR)). A ROC curve plots the tradeoff between the sensitivity versus 1-specificity by contriving FPR and TPR in the x and y axes respectively. The best possible result would yield at coordinate (0, 1), where all positives cases are classified as positive and all negative are cases classified as negative, representing 100% sensitivity and 100% specificity. Although this is the best case, the procedure to get them can be very restricted with respect to gaining false-positive error with no false-negative price to pay.

9. TRENDS AND DIRECTIONS IN PROGNOSIS MODELS

Prognosis models have been evolved drastically during the past several decades. In ancient times, clinical data such as age, estrogen and progesterone receptors, lymph-node involvements and other prognostic factors have been extensively used to determine the recurrence of breast cancer among patients. Various approaches were applied to develop the prognosis model varying from the traditional probabilistic techniques, originating from statistical methods, for instance Kaplan Meier, and the proportional hazards-regression model to more qualitative and model-based techniques derived from the artificial intelligence domain. Although many artificial intelligence techniques have been applied, ANN was identified as the dominant technique in developing clinical prognosis models instead of other interpretable techniques due to its robust characteristics to noisy data and being capable of expressing complicated interactions. However, this technique is prone to over-fitting, which requires appropriate validation to be executed.

Numerous researches have been done to determine the recurrence of breast cancer using clinical data but this approach conveys drawbacks as it difficult to distinguish tumors with similar morphological subtypes. The invention of microarray technology, with the opportunity to examine thousands of genes simultaneously, has shifted the cancer-prognosis model to a new post-genomic era. Unlike, clinical prognosis models, gene-expression profiling offers a novel ways to understand the cancer-related cellular process, thus enhancing classification accuracy. However, overwhelming data generated from microarray technology requires proper data analysis to be done. Microarray data analysis mainly consists of two parts; feature selection and classification. Many studies have been conducted to address these problems. The classification trends have changed from using a single classifier to ensemble several classifiers into one to examining the difference in gene expression and recently to multi-class classification techniques. Moreover, it also noted heavy reliance toward univariate filter-feature selection techniques compared to wrapper and embedded methods. Currently, prognosis models show an imperative growing direction toward using integrated data such as microarray and clinical, or genomic and proteomic data instead of examining cancer recurrence in a separate manner.

10. CONCLUSIONS

This paper reviewed prognosis models for clinical and microarray data, precisely focusing on feature selection and classification techniques that have been employed in cancer prognosis. The main problems emerging from the breast cancer prognosis domain was explained in detail. Due to limitation in current practice clinical-prognostics has derived attention from researchers to develop the genetic marker-based prognosis model, particularly using microarray data. However, this approach also has its own dilemma in making sense of thousands of gene-expression values. Feature-selection techniques have become a prerequisite step in analysing gene-expression data. Currently, filter methods are more prominent techniques among the Bioinformatics

community compared to wrapper and embedded methods. On the other hand, various classification techniques have been used but none of them is superior than the others. Moreover, most classification techniques are found to be data dependent. In general, it was observed that most of the researchers have underrated the power of clinical factors, although it could add complementary information to gene-expression data. The proposal of integrating clinical and gene-expression profiles can be considered as one of the most promising future lines of the work, although but a lot of work needs to be addressed to minimize heterogeneity in clinical and gene-expression profiles for breast cancer prognosis.

11. REFERENCES

[1] A. Hanna and P. Lucas. *"Prognostic models in medicine; AI and statistical approaches"*. Methods of Information in Medicine, 40:1-5, 2001.

[2] A. Joseph Cruz and D. S. Wishart. "*Applications of machine learning in cancer prediction and prognosis"*. Cancer Informatics, 2: 59-78, 2006.

[3] G. C Lim., Y. Halimah, and T.O Lim. "*The First Report of The National Cancer Registry Cancer Incidence In Malaysia* 2000", National Cancer Registry 2002.

[4] A. Mann, *"Women's health issues and nuclear medicine, Part II: Women and breast cancer"*. Journal of Nuclear Medicine Technology, 27: 184-187,1999.

[5] J. Lundin. *"The Nottingham Prognostic Index - from relative to absolute risk prediction"*. European Journal of Cancer, 43: 1498-1500, 2007.

[6] M. J. Duffy. *"Predictive markers in breast and other cancer: A review"*. Clinical Chemistry, 51: 494-503, 2005.

[7] R. S. Uma and T. Rajkumar. *"DNA microarray and breast cancer - A review"*. International Journal of Human Genetics, 7: 49-56, 2007.

[8] B. Efron. *"Logistic regression, survival analysis, and the Kaplan-Meier Curve".* Journal of the American Statistical Association, 83: 414-425, 1988.

[9] R. L. Prentice and L. A. Gloeckler. *"Regression analysis of grouped survival data with application to breast cancer data"*. Biometrics, 34: 57-67, 1978.

[10] H.B. Burke, P.H. Goodman, D.B. Rosen, et al. *"Artificial neural networks improve the accuracy of cancer survival prediction"*. Cancer, 79: 857-862, 1997.

[11] M. D. Laurentiis, S. D. Placido, A. R. Bianco, et al. *"A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients"*. Clinical Cancer Research, 5: 4133-4139, 1999.

[12] J.M. Jerez-Aragones, J.A. Gomez-Ruiz, G. Ramos-Jimenez, et al. *"A combined neural network and decision trees model for prognosis of breast cancer relapse"*. Artif Intell Med, 27: 45-63, 2003.

[13] R. Kates, N. Harbeck, and M. Schmitt. "*Prospects for clinical decision support in breast cancer based on neural network analysis of clinical survival data"*. In Proceeding of the Fourth International Conference on Knowledge-Based Intelligent Enginerring Systems & Allied Technologies, Brighton,UK, 2000.

[14] R. N. G. Naguib, A. E. Adams, C. H. W. Horne, et al. *"The detection of nodal metastasis in breast cancer using neural network techniques"*. Physiological Measurement, 17: 297-303,1996.

[15] L. Bottaci, P.J. Drew, and J.E. Hartley. *"Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions"*. Lancet, 1997; 350: 469-72.

[16] Delen D., Walker G., and Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 34: 113-127, 2004.

[17] J. A. Gomez Ruiz, J. M. Jerez Aragones, and J. A Munoz Perez. *"Neural network based model for prognosis of early breast cancer"*. Applied Intelligence, 20: 231-238, 2004.

[18] J. M. Jerez, I. Molina, J. L. Subirats, et al. "*Missing data imputation in breast cancer prognosis"* . In the Proceeding of the 24th IASTED International Multi-Conference, Biomedical Engineering, Innsbruck, Austria, 2006.

[19] L. Franco, J. L.Subirats, E. A. I. Molina, et al. *"Early breast cancer prognosis prediction and rule extraction using a new constructive neural network algorithm" Computational and Ambient Intelligence*: Springer Berlin / Heidelberg, pp. 1004-1011, (2007).

[20] X. Xiong, Y. Kim, Y. Baek, et al. *"Analysis of breast cancer using data mining & statistical techniques".* In the Proceeding of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05), 2005.

[21] H.Seker, M.O.Odetayo, D.Petrovic, et al. *"An artificial neural network based features evaluation index for the assessment of clinical factors in breast cancer survival analysis".* In the Proceeding of the IEEE Canadian Conference on Electrical & Computer Engineering, 2002.

[22] W. L. McGuire. *"Breast cancer prognostic factors; Evaluation guidelines"*. Journal of the National Cancer Institute, 3: 154 -155, 1990.

[23] F. Kharbat, L. Bull, and M.Odeh. *"Mining breast cancer data with XCS*". In the Proceeding of the 9th Annual Conference on Genetic and Evolutionary Computation, London, England, 2007.

[24] Y.J. Lee, O. L. Mangasarian, and W. H Wolberg. *"Breast cancer survival and chemotherapy: A support vector Machine analysis"*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 55, 2000.

[25] A. V. Tinker, A. Boussioutas, and D. D. L Bowtell. *"The challenges of gene expression microarrays for the study of human cancer"*. Cancer Cell, Elsevier, 9: 333-339, 2006.

[26] J. D. Potter. *"Epidemiology, cancer genetics and microarrays: Making correct inferences using appropriate designs"*. Trends in Genetics, 19: 690–695, 2003.

[27] A. Brazma, P.Hingamp, J.Quackenbush, et al. *"Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data"*. Nature, 29: 365-371, 2001.

[28] T. S.Furey, N.Cristianini, N.Duffy, et al. *"Support vector machine classification and validation of cancer tissue samples using microarray expression data"*. Bioinformatics, 16: 906-914, 2000.

[29] Y. Lu and J.Han. *"Cancer classification using gene expression data"*. Information Systems; Data Management in Bioinformatics, 28: 243 - 268, 2003.

[30] Y.Saeys, I.Inza, and P.Larranaga. *"A review of feature selection techniques in bioinformatics"*. *Bioinformatics*, 1-10, 2007.

[31] O. Gevaert, F. D Smet., D.Timmerman, et al. *"Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks"*. Bioinformatics, 22: e184 - e190, 2006.

[32] Z.Wang. "Neuro-fuzzy modeling for microarray cancer gene expression data", Oxford University Computing Laboratory 2005.

[33] S. B. Cho and H. H. Won. *"Machine learning in DNA microarray analysis for cancer classification"*. In the Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, 2003.

[34] H. Hu, J. Li, H. Wang, et al. *"Combined gene selection methods for microarray data analysis",* Knowledge-Based Intelligent Information and Engineering Systems; 4251: Springer-Verlag Berlin Heidelberg, pp. 976–983, 2006.

[35] T. R. Golub, D. K. Slonim, P.Tamayo, et al. "*Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring"*. Science, 286: 531-537, 1999.

[36] E. P. Xing, M. I.Jordan, and R. M.Karp. *"Feature selection for high-dimensional genomic microarray data".* In the Proceeding of the 18th International Conf. on Machine Learning, 2001.

[37] C. Giallourakis, C.Henson, M.Reich, et al. *"Disease gene discovery through integrative genomics".* Annual Review of Genomics and Human Genetics, 6: 381-406, 2005.

[38] A. Ben-Dor, L. Bruhn, N.Friedman, et al. *"Tissue classification with gene expression profiles"*. Journal of Computational Biology, 7: 559-583, 2000.

[39] Y.Barash, E.Dehan, M.Krupsky, et al. *"Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays"*. Bioinformatics, 20: 839–846, 2004.

[40] S.Rogers, R. D.Williams, and C.Campbell. *"Class prediction with microarray datasets"*. Bioinformatics Using Computational Intelligence Paradigms; 176: Springer, pp. 119-141, 2005.

[41] B. Y. M.Fung and V. T. Y. Ng. *"Classification of heterogeneous gene expression data"*. ACM Special Interest Group on Knowledge Discovery and Data Mining, SIGKDD Explorations, 5: 69 - 78, 2003.

[42] D.Koller and M.Sahami. *"Toward optimal feature selection"*. In the Proceeding of International Conference of Machine Learning, p. 284-292, 1996.

[43] L. Yu and H. Liu. "*Redundancy based feature selection for microarray data*". In the proceeding of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 2004.

[44] R. Kohavi and G. H. John. *"Wrappers for feature subset selection"*. Artificial Intelligence, 97: 273-324, 1997.

[45] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz. *"Incremental wrapper-based gene selection from microarray data for cancer classification"*. Elsevier, 39: 2383 – 2392, 2006.

[46] H. Zhang, T. B. Ho, and S. Kawasaki. *"Wrapper feature extraction for time series classification using singular value decomposition"*. International Journal of Knowledge and Systems Science, 3: 53-60, 2006.

[47] I. Inza, P. Larranaga, R.Blanco, et al. *"Filter versus wrapper gene selection approaches in DNA microarray domains"*. Artificial Intelligence in Medicine, 31: 91—103, 2004.

[48] J. J. Liu, G. Cutler, W. Li, et al. *"Multiclass cancer classification and biomarker discovery using GA-based algorithms"*. Bioinformatics, 21: 2691-2697, 2005.

[49] L. Li, T.A. Darden, C.R. Weingberg, et al. *"Gene assessment and sample classification for gene expression data using a genetic algorithm / k-nearest neighbor method"*. Combinatorial Chemistry & High Throughput Screening, 4: 727-739, 2001.

[50] I. Guyon, J. Weston, M. D. Stephen Barnhill, et al. *"Gene selection for cancer classification using support vector machines"*. Machine Learning, 46: 389-422, 2002.

[51] S. Dudoit, J. Fridlyand, and T.P. Speed. *"Comparison of discrimination methods for the classification of tumors using gene expression data"*. Journal of the American Statistical Association, 97, 2002.

[52] J. Khan, J. S. Wei, M. Ringnér, et al. *"Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks"*. Nature Medicine, 7, 2001.

[53] Y. Xu, M. Selaru, J. Yin, et al. *"Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer"*. Cancer Research, 62: 3493-3497, 2002.

[54] M. P. S. Brown, W. N. Grundy, D. Lin, et al. *"Knowledge-based analysis of microarray gene expression data by using support vector machines"*. In the Proceeding of the National Academy of Sciences, USA, 2000.

[55] D. Liu, T. Shi, J. A. DiDonato, et al. *"Application of genetic algorithm/k-nearest neighbor method to the classification of renal cell carcinoma"*. In the Proceeding of the Computational Systems Bioinformatics Conference (CSB 2004), IEEE, 2004.

[56] M. L. Zhang and Z. H. Zhou. "*A k-nearest neighbor based algorithm for multi-label classification*". In the Proceeding of the International Conference on Granular Computing, IEEE, 2005.

[57] D. Berrar, C.S. Downes, and W. Dubitzky. *"Multiclass cancer classification using gene expression profiling and probabilistic neural networks"*. In the Proceeding of the Pacific Symposium on Biocomputing, New Jersey, 2003.

[58] V. N. Vapnik. *"Statistical Learning Theory"*. New York, NY: John Wiley & Sons, 1998.

[59] V. N. Vapnik. "*The Nature of Statistical Learning Theory*", 2nd ed. New York, NY: Springer, 2000.

[60] Y. Mao, X. Zhou, D. Pi, et al. *"Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection"*. Journal of Biomedicine and Biotechnology, 2: 160-171, 2005.

[61] S. Symons and K. Nieselt. *"Data mining microarray data - Comprehensive benchmarking of feature selection and classification methods"*. Center for Bioinformatics T¨ubingen, Wilhelm-Schickard Institute for Computer Science, University of T¨ubingen, Sand 14, 72076 T¨ubingen, 2006.

[62] M. Al-Shalalfa and R. Alhajj. *"Application of double clustering to gene expression data for class prediction"*. In the Proceeding of the International Conference on Advanced Information Networking and Applications Workshops, 2007.

[63] S. Ramaswamy, P. Tamayo, R. Rifkin, et al. *"Multiclass cancer diagnosis using tumor gene expression signatures"*. National Academy of Sciences, 98: 15149-15154, 2001.

[64] T. Li, C. Zhang, and M. Ogihara. "*A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression"*. Bioinformatics, 20: 2429–2437, 2004.

[65] L. E. Peterson, M. Ozen, H. Erdem, et al. *"Artificial neural network analysis of DNA microarray-based prostate cancer recurrence"* In the Proceeding of the Computational Intelligence in Bioinformatics and Computational Biology, IEEE, 2005.

[66] M. Ringnér and C.Peterson. *"Microarray-based cancer diagnosis with artificial neural networks".* Complex Systems Division, Department of Theoretical Physics, Lund University, Sweden 2003.

[67] G. Tusch. *"Sequential classification for microarray and clinical data"* In the Proceeding of the Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts. IEEE, p. 5-6, 2005.

[68] J. S. Wei, B. T Greer, F. Westermann, et al. *"Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma"*. Cancer Research, 64: 6883–6891, 2004.

[69] V. Bevilacqua, G. Mastronardi, F. Menolascina, et al. *"Genetic algorithms and artificial neural networks in microarray data analysis: A distributed approach".* Engineering Letters, 13: 1-9, 2006.

[70] P. Eden, C. Ritz, C. Rose, et al. *""Good Old'' clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers"*. European Journal of Cancer, 40: 1837 - 1841, 2004.

[71] J. Pearl. *"Fusion, propagation and structuring in belief networks*". Artificial Intelligence, 29: 241-288, 1986.

[72] N. Friedman, M. Linial, I. Nachman, et al. *"Using bayesian networks to analyze expression data"*. Journal of Computational Biology, 7: 601–620, 2000.

[73] P. Helman, R. Veroff, S. R. Atlas, et al. *"A bayesian network classification methodology for gene expression data"*. Journal of Computational Biology, 11: 581-615, 2004.

[74] E. Huang, S. H. Cheng, H. Dressman, et al. *"Gene expression predictors of breast cancer outcomes"*. Lancet, 361: 1590-1596, 2003.

[75] M. West, C. Blanchette, H. Dressman, et al. *"Predicting the clinical status of human breast cancer by using gene expression profiles"*. National Academy of Science of United States of America (PNAS), 98: 11462-11467, 2001.

[76] S. B. Cho and C. Park. *"Speciated GA for optimal ensemble classifiers in DNA microarray classification"*. In the proceeding of the Congress on Evolutionary Computation, 2004.

[77] J. H. Hong and S. B. Cho. *"The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming"*. Artificial Intelligence in Medicine, 36: 43-58, 2005.

[78] A. C. Tan and D. Gilbert. *"Ensemble machine learning on gene expression data for cancer classification"*. Applied Bioinformatics, 2: S77-83, 2003.

[79] B. Liu, Q. Cui, T. Jiang, et al. *"A combinational feature selection and ensemble neural network method for classification of gene expression data"*. Bioinformatics, 5, 2004.

[80] K. J. Kim and S. B. Cho. *"Ensemble classifiers based on correlation analysis for DNA microarray classification"*. Elsevier, 70: 187–199, 2006.

[81] M. E.Futschik, M. Sullivan, A. Reeve, et al. *"Prediction of clinical behaviour and treatment for cancers"*. Applied Bioinformatics, 2: S53-S58, 2003.

[82] J. R. Nevins, E. S. Huang, H. Dressman, et al. *"Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction"*. Human Molecular Genetics, 12: R153-R157, 2003.

[83] Y. Sun, S. Goodison, J. Li, et al. *"Improved breast cancer prognosis through the combination of clinical and genetic markers"*. Bioinformatics, 23: 30-37, 2007.

[84] L. Li, L. Chen, D. Goldgof, et al. *"Integration of clinical information and gene expression profiles for prediction of chemo-response for ovarian cancer."* In the Proceeding of the Annual Conference of Engineering in Medicine and Biology, IEEE, Shanghai, China, 2005.

[85] L. Li. *"Survival prediction of diffuse large -B-cell lymphoma based on both clinical and gene expression information"*. Bioinformatics, 22: 466-471, 2006.

[86] Y. F. Leung and D. Cavalieri. *"Fundamentals of cDNA microarray data analysis*". Elsevier, 19: 649-659, 2003.

[87] J. Pittman. *"The importance of validation in genomic studies of breast cancer"*. Breast Diseases: A Year Book Quarterly, 16: 16-19, 2005.

[88] C. Tago and T. Hanai. *"Prognosis prediction by microarray gene expression using support vector machine"* . Genome Informatics, 14: 324-325, 2003.

[89] R. Xu, X. Cai, and D. C. W. II. "*Gene expression data for DLBCL cancer survival prediction with a combination of machine learning technologies".* In the Proceeding of the Engineering in Medicine and Biology 27th Annual Conference, IEEE, Shanghai, China, 2005.

Towards a Query Rewriting Algorithm over Proteomics XML Resources

Kunalè KUDAGBA kknale@fsdmfes.ac.ma
Faculty of Sciences/Department of Computer Science
USMBA University
Fes, 30 000, Morocco

Hassan BADIR hbadir@ensat.ac.ma
Department of Computer Science/ SIC Division
National School of Applied Sciences, ENSA
Tanger, 45 000, Morocco

Omar El BEQQALI oelbeqqali@fsdmfes.ac.ma
Faculty of Sciences/Department of Computer Science
USMBA University
Fes, 30 000, Morocco

Abstract

Querying and sharing Web proteomics data is not an easy task. Given that, several data sources can be used to answer the same sub-goals in the Global query, it is obvious that we can have many candidates' rewritings. The user-query is formulated using Concepts and Properties related to Proteomics research (Domain Ontology). Semantic mappings describe the contents of underlying sources. In this paper, we propose a characterization of query rewriting problem using semantic mappings as an associated hypergraph. Hence, the generation of candidates' rewritings can be formulated as the discovery of minimal Transversals of an hypergraph. We exploit and adapt algorithms available in Hypergraph Theory to find all candidates rewritings from a query answering problem. Then, in future work, some relevant criteria could help to determine optimal and qualitative rewritings, according to user needs, and sources performances.

Keywords: Proteomics, Ontology, XML, Trees, Semantic Web, Query Rewriting, minimal Transversals.

1. INTRODUCTION

The rapid progress of biotechnologies and the multiple genome project [7], [5] about organisms and diverse species have generated an increasing amount of proteomic data stored in many sources [17] available and publicly accessible on the Web. They contain data about metabolic pathways, protein 3D structures, DNA Sequences, organisms, diseases, and so on.
Many biological questions require that data from several data sources are queried, searched, and integrated. The first step in this process of Data Integration is the Query rewriting which consists of reformulating a global query in several specific local queries. The query rewriting problem has recently received significant attention because of its relevance to a wide variety of data

management problems [14]: query optimization, maintenance of physical data independence, data integration, and data warehouse design.
Semantic mappings can be used to adapt a global query expressed in terms of a Domain Ontology in terms of specific local sources. In fact, as several sources could provide expected resources, more than one could be relevant to rewrite the global query.
The motivation of this work is to investigate how to generate candidate rewritings when user-query is posed over XML Sources, and mappings are expressed in LAV Approach. To achieve this goal, we characterize the query rewriting problem as minimal Transversals Discovery from an associated Hypergraph. From this set of generated rewritings, we would compute an optimal and best quality rewriting based on some defined and relevant criteria. We illustrate an intuitive execution of the rewriting algorithm proposed, using a scenario of Proteomics data sources.
The following two XML sources contain data that are semantically similar, but are described with autonomous and heterogeneous schemas. They both represent the same proteomics data, but not identically. They give an idea of differences in terms of terminologies, structures and contents.

```
<PROTEIN_SET>
<PROTEIN>
<ACCESSION>P26954</ACCESSION>
<ENTRY_NAME>IL3B_MOUSE</ENTRY_NAME>
<PROTEIN_NAME>Interleukin-3 receptor
class II beta chain [Precursor]
</PROTEIN_NAME>
<GENE_NAME>CSF2RB2</GENE_NAME>
<ORGANISM taxonomy_id="10090">Mus
musculus</ORGANISM>
</PROTEIN>
…
</PROTEIN_SET>
```

FIGURE 1: XML Data Source 1

```
<PROTEIN_BASE>
<PROTEIN  ACC_NUMBER="P26954 >
<ENTRY>IL3B_MOUSE</ENTRY>
<PROTEIN_NAME>Interleukin-3 receptor
class II beta chain [Precursor]
</PROTEIN_NAME>
<GENE>CSF2RB2</GENE>
<ORGANISM >
    <TAX_ID >10090</TAX_ID >
<NAME> Mus musculus</NAME>
</ORGANISM>
</PROTEIN>
   …
</PROTEIN_BASE>
```

FIGURE 2: XML Data Source 2

We can remark also some semantics heterogeneities, like ACC_NUMBER attribute in Source 2 which is equivalent with ACCESSION Element in Source 1.
Although both data sources contain semantically similar proteomic data, the simple user query *"Which are proteins that are encoded by the gene named by CSF2RB2 in Mus Musculus Organism? "* need to be formulated quite differently with existing XML query languages, like Xquery or XPath for both sources.
So, flexible, user-centric and semantic strategies of discovering relevant sources are needed to compute optimal and best quality rewriting, according to suitable criteria.
The rest of the paper is structured as follows. Section 2 gives a brief survey on our related work regarding existing query rewriting approach in Mediators. Section 3 discusses the basic concept of Knowledge representation. The hypergraph-based semantic query rewriting algorithm is presented in Section 4. The last section 5 draws the conclusions and future work.

2. RELATED WORK

One of the first LAV systems that allow the integration of XML is AGORA [21]. But AGORA still makes an extended use of the relational model: Although it offers an XML view for relational and XML data, this view is translated into a generic relational schema, XML resources are described as relational views over this schema and XQuery expressions are translated to standard SQL queries, which are then decomposed and evaluated.
Information Manifold [19] also follows a local-as views approach. In this system, the global schema is a flat relational schema, and Description Logics are used to represent hierarchies of classes. The sources are expressed as relational views over this schema. Query rewriting is done

by the Bucket algorithm which rewrites a conjunctive query expressed of the global schema using the source views. It examines independently each of the query sub-goals and tries to find rewritings but loses some by considering the sub-goals in isolation.
STyX system [12] already uses a domain Ontology as global schema language and translates an OQL-like global query language to XQuery expressions on the heterogeneous XML sources. STyX maps XPath expressions to ontology concepts. In [1], a data integration system whereby XML sources are mapped into a simple ontology (supporting inheritance and roles), is discussed.
In [18], authors have also used to integrate XML heterogeneous data sources. Their work consists to map XML schema constructs to concepts. The main difference to STyX is the approach to semantic mapping. Although Lehti's approach is not as flexible and powerful as using XPath mappings, it is in principle able to detect inconsistencies in the mapping with the help of a description logic reasoner [2].
Other data integration approaches that use an Ontology as Global Schema are either based on an extended data warehouse. Semantic mediation in C-Web [2] is based on thesauri. In Xyleme mediator [8], the global schema is a set of abstracted DTDs which are terms Trees according to domain vocabularies such as Culture or Tourism. Both follow GLAV (GAV and LAV together) because correspondences between Mediator vocabulary and Sources vocabularies are expressed by simpler path mappings.
Recently, with the development of Semantic Web, mediation systems have been developed. Project Piazza [15] proposes an infrastructure based on Peer-to-Peer (like a decentralized mediator) for RDF and OWL data integration.
According to the query rewriting algorithm, you can refer to [14] for a large survey on the Query rewriting problem. Our approach is inspired by WS-CatalogNet's semantic-driven Algorithm. In this work, [3] have developed a novel and more advanced query rewriting techniques for flexible and effective E-Catalogs selection.

3. REPRESENTING KNOWLEDGE

In order to rewrite semantically a global query, it is essential to make a choice of an adequate abstraction model for local sources and to express in a common formalization language all available knowledge. This last case concerns the domain ontology, the semantic mappings, and the user query.
The proteomics sources which we are working with are stored and available as XML Documents according to their XML Schemas. XML [6] is presently becoming the standard for the exchange of biological data sources. So, the reason for the use of XML Sources for the data Integration is obvious. XML Schemas [24] are more suitable than DTDs for expressing the syntax, structural, cardinality and typing constraints required by proteomics data. We propose to abstract sources XML Schemas as unordered Trees, and we try to propose a specification language based on description logic (DL) [2] formalization and reasoning (Trees Logics).

3.1 Trees abstract model

We know that XML Schemas are special XML Documents. Various models have been proposed to represent XML Documents. The W3C proposed a generic model named Document Object Model (DOM). In this model, presented in the current section, XML Documents can be abstracted as Trees. Our motivation using this way of abstraction is to further exploit some achieved and well known results on Trees Embedding Problem [24] as knowledge semantic retrieval in an integration framework.

3.2 Logic-based Trees descriptions

To provide a semantic formalization, necessary for rigorous characterization of proteomic queries and knowledges, we propose to use a description language of hierarchic structures such as Trees, based on Logics and called *Trees-Logics*.
Many researchers have addressed the question of using logics over Trees. In [9] and [10] authors have translated XQuery global queries into local conjunctive queries over Trees in a Data integration processes. In [24] a language called ApproXQL, which exploits among others logical

operators, to formalize more richer and expressive requests, has been developed. These requests expressed as conjunctive queries could be illustrated and interpreted as Trees. Then, in the paper [13] authors have studied the complexity and the expressive power of conjunctive queries over Trees.
So, we believe strongly that it is possible to describe data Trees with a suitable subset of logical formalisms [2]. We want to exploit all these works in order to provide a logical description of hierarchical structures, such as Trees and consequently Paths in particular. In our final integration framework, both the phases of Trees generation and their specification in *Trees-Logics* are totally transparent for the user. We precise that Description Logics [2] are a family of logics which were developed for modeling complex hierarchical structures and to provide a specialized reasoning engine to do inferences on these structures.
Due to the space limitation, we could not give more details on *Trees-Logics* and so this paper will focus only on the query rewriting problem.

4. QUERY REWRITING

In this section, we begin by presenting an abstraction of our approach for query rewriting. Then, we show that using some hypergraph Theory results can help generate candidate rewritings. Therefore we present the Classical algorithm to compute minimal Transversals of a hypergraph. Finally, we illustrate an execution of this algorithm to find candidate rewritings, given concrete case of bio-query reformulation.
We recall that the main goal of query rewriting phase is to reformulate a Global query Q expressed as *Trees-Logics* over Domain ontology, into Local queries Q_j that are expressed in terms of Local schemas. This operation is realized using semantic mappings pre-calculated and stored on the mediator. Semi-automatic detection of semantic mappings has no impact on the processing time of the user query.

The concrete algorithm showing how these semantic mappings are calculated is out of scope of this paper.

The domain Ontology is abstracted as a Tree and expressed using the defined language, Trees-Logics. The knowledge domain concerns proteomics research including concepts such Protein Family, 3D Structures, Coding Genes, Motifs, Domains, amino-acids sequences, Active Sites, Binding Sites, Enzymes, Chains, Chemical Bonds, … and their relative properties, so we call the ontology by *O'proteomics*. Due to space constraints, we could not give more details on *O'proteomics*.

The Ontology constitutes a support for user query formulation and gives an idea of which concepts, it is possible (but not obliged) to find or retrieve in the underlying Proteomics sources. Therefore, the first initiative consists of determining the part of the query that cannot be answered by available proteomics sources.

4.1 Logic-based Trees descriptions

We represent by the following couple $Sch'O = (O'\ proteomics, M\ mappings)$, the set of semantic knowledges about our domain of interests, which is proteomics.

The concepts annotations, defined in *O'proteomics* will serve to enrich Global query before rewriting process. The semantic mappings will show query answering capabilities of the underlying sources.

Given a Global query Q and the knowledges couple $Sch'O$, our rewriting approach consist to determine two sub-queries Q_{valide} and $Q_{invalide}$. Explicitly, we shall calculate:

- $Q'' = Q_{invalide}$ having a size as minimal as possible. The Size of a query is the number of atomic goals that it contains. Sub-Query Q'' cannot be answered by underlying sources, at the moment of the sending of the Global query Q. This initial operation has the role of cleaning up Q of domain concepts/properties which are

not yet available, as Web proteomics registered resources. So, no processing will be realized on Q'', in the future.

- $Q' = Q_{valide}$ is the part of Q that will be rewrite using semantic mappings M of $Sch'O$. Sub-query Q' can be answered by registered sources. Our final goal is to propose an intelligent subdivision of $Q' = Q_{valide}$ into sub-queries $Q'_1, Q'_2, \ldots, Q'_m$ with $1 \leq m \leq n$, n is the number of sources available in the integration while m denotes the number sources which are necessary to provide an answer to the query Q. So, we might find the set $Q' = \{(Q'_j, m_j)\}$ of couples (Q'_j, m_j) such as Q'_j be an atomic subdivision of Q' that will be answered by mapping m_j.

We can easily remark that several rewritings can be proposed, we will call them Candidates rewritings. In fact, more than one source, and so mapping could provide the same resources researched.
The algorithm receives as input a global query Q, a schema $Sch'O$ and generate as output all candidates' rewritings $r_i(Q)$.

4.2 Hypergraph-based Algorithm

In practice, Global queries are expressed like conjunctive queries using *Trees-Logics*. So, a rewriting Q' is a suitable conjunction of constraints. These constraints might be checked by all final answers of the global query Q, because they constitute an indication of resources that may be retrieved from adequate sources.
In order to provide a characterization of our query rewriting problem, we give an alternative formulation of the rewriting formalization.
Given a Global Query Q and the semantic knowledges couple $Sch'O$, query rewriting consists to compute two sub queries Q_{valide} and $Q_{invalide}$ on the basis of mappings set M, such as:

$$Q = Q_{valide} \wedge Q_{invalide} \qquad (1)$$

We are searching for all candidates rewritings, formulated as the conjunction of constraints:

$$Q_{valide} = Q' = \bigwedge_{i=1}^{m} C_i \qquad (2)$$

All constraints C_i are logical representation of the user specific needs. Finally, our motivation is to answer the fundamental question which is to find, given Q a new query called rewriting expressed by $Q_{valide} = Q' = \bigwedge_{i=1}^{m} C_i$ such as Q' denotes as much as possible the resources expected by query Q ?
We have said that several and alternatives rewritings are possible, due to the fact that more than one mapping could be used to reformulate an atomic goal of the Global query. From this point of view, rewriting problem which requires generating all candidates' rewritings can be characterized as a current Hypergraph Theory problem of computing all minimals Transversals of an Hypergraph. Generate a Transversal Hypergraph consists of generate all minimals Transversals.

4.2.1 Definition of Hypergraph [16].

An Hypergraph H is an ordered pair $H = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is a finite set of elements and $E = \{E_1, E_2, \ldots, E_m\}$ is a family of subsets of V such that:

- $E_i \neq \varnothing, (i = 1, \ldots, m)$

- $\bigcup_{i=1}^{m} E_i = V$

The elements of V are called nodes while the elements of E are called hyperedges of the hypergraph H . A hypergraph can be seen as a generalization of a graph where the restriction of an edge having only two nodes does not hold.

4.2.2 Definition of Transversals [1].

Let $H = (V, E)$ be an hypergraph. A set $T \subseteq V$ is called a Transversal of H if it intersects all its hyperedges, i.e, $T \cap E_i \neq \varnothing$, $\forall E_i \in E$. A **Transversal** T is called minimal if no proper subset T' of T is a transversal of H . The **Transversal Hypergraph** $Tr(H)$ of an hypergraph H is the family of all minimal transversals of H .

From a rewriting query problem, we need to give a mathematical characterization, by defining an associated Hypergraph $H_{Q,M}(V, E)$, built as follows:

- For every mapping m_i , describing a local concept from M , as a logical function of O'proteomics global concepts, we associate a vertice V_{m_i} in the hypergraph $H_{Q,M}(V, E)$ and $V = \{V_{m_i}, i \in [1, n]\}$.
- For every constraint C_i of the Global query Q , we associate an hyperedge E_{C_i} in the hypergraph $H_{Q,M}(V, E)$. To simplify, we suppose that all these constraints are describing atomics goals. So, each hyperedge E_{C_i} is a set of mappings, calculated by considering those mappings which are relevant to answer these goals.

A classical algorithm to compute minimal transversals of an hypergraph is proposed and available in [20]. Many papers [11] have discussed about algorithm of generation of Hypergraph Transversal, which is a set of minimal transversals. One of the first results remains Berge's Algorithm [4], but several variants have been proposed in order to deal with the algorithm complexity [22].

Now, we present our Query rewriting algorithm called *Q-Candidates'Finder*, which integrates the better and efficient complexity of the classical Algorithm:

Input: A Query Q and Sch'O=(O'proteomics, M mappings)

Output: The set of candidates rewriting such as $Q_{candidates} = \{(Q_{valide}, Q_{invalide})\}_{H_{Q,M}(V,E)}$

1: **Build the associated Hypergraph**

2: **Compute** $Q_{invalide} = \bigwedge_{i=1}^{k} C_i$ such as C_i is not provided by any mappings in M .

3: **Build the associated Hypergraph** $H^*_{Q,M}(V, E^*)$

4: $Q_{candidates} = \varnothing$

5: **Generate the Hypergraph Transversal of** $H^*_{Q,M}(V, E^*)$

- Let be HypTransv - Using the Classical Algorithm [Mannila, 1994]

6: **For all** edge $X = \{V_{m_1}, V_{m_2}, ..., V_{m_p}\} \in HypTransv$ **do**

7: $Q_{valide} = r(Q) = Q' = \{(Q'_j, m_j), j \in [1, p]\}$

where Q'_j is a subdivision of Q_{valide} that will be answered by the mapping m_j .

8: $Q_{candidates} = Q_{candidates} \cup Q_{valide}$

9: **End For**

10: **Return** $Q_{candidates}$

FIGURE 3: Q-Candidates' Finder Algorithm

4.3 Q-Candidates' Finder Illustration

To illustrate the proposed rewriting approach, let us consider the following mappings (L.A.V approach). As the domain Ontology is characterized as a Tree, semantic mappings might express subsumption (**Sub**) and equivalence (**Eq**) relations that exist between Local XML Schemas, also abstracted as Trees, and the Ontology's Tree model. We suppose in order to simplify this illustration that we have only simple paths mappings (and so, no sub-trees) between Concepts, according to 1:1 cardinality. For every registered proteomics' sources, we provide the **LHS** (Left Hand Side expressing Local Concepts/Properties), the **TYPE** (the type of mappings), and the **RHS** (Right Hand Side, expressing Ontology Concepts/Properties) of the current mapping.

	LHS	**TYPE**	**RHS**
O'proteomics Ontology	**Gene** (Genes, Proteins, Species, Organisms) **Proteine** (IdProteins, Peptides, DevStadium) …		
Mapping *m1*: Description Of Source I	**S1_Gene** (S1_GeneName, S1_ProteinName, S1_species, S1_organisms)	**Eq** …	**Gene** (Genes, Proteins, Species, Organism)
Mapping *m2*: Description Of Source II	**S2_Gene** (S2_NomGenes, S2_NomProteine, S2_Especes, S2_Organismes)	**Eq** ….	**Gene** (Genes, Proteins, Species, Organisms)
Mapping *m3*: Description Of Source III	**S3_TreeLife** (S3_Species, S3_Genus…)	**Sub** …	**Gene** (Species, Organisms)

TABLE 1: Mapping Table

We have just shown mappings which are relevant for the Query we shall process. Note once again, that we are considering corresponding paths of these Concepts/Properties in their abstract trees.

According to this mapping table, we can say that the ontology includes Concept Gene and Proteine with their relative Properties such as Genes, Proteins, and Species … Mappings $m1$ and $m2$ show that Source I and Source II provide the properties such as Gene, Proteins, Species, and Organisms, while the mapping $m3$ illustrate that Source III only provides properties such as Species and Organisms.

Let us consider now the following query which is expressed over the domain ontology, $O'proteomics$:

What are the genes which proteins could have a peptide Signal and for which, it is assumed that they are expressed at Tardive Shizont stadium for the Plasmodium falciparum?

4.3.1 Hypergraph Construction

Intuitively, Q can be expressed like a conjunction of the following constraints:

$$Q = C_{Genes} \wedge C_{Proteins} \wedge C_{Peptides} \wedge C_{DevStadiums} \wedge C_{Organisms} \wedge C_{Species} \quad (3)$$

In practice, the user will formulate this request by using an user-friendly graphic interface, and the generation of its Trees-Logics version is done automatically. He will choose the Concept Gene and indicate for each property Genes, Proteins, Species, Organisms, Peptides, DevStadiums the expected values.
The associated hypergraph $H_{Q,M}(V,E)$ consists of the following sets of vertices and edges:

$$V = \{V_{S2\_Gene}, V_{S1\_Gene}, V_{S3-TreeLife}\} \quad (4)$$

and

$$E = \begin{Bmatrix} E_{C\_Genes}, E_{C\_Proteins}, E_{C\_Peptides}, \\ E_{C\_DevStadiums}, E_{C\_organisms}, E_{C\_Species} \end{Bmatrix} \quad (5)$$

We could see this illustration using a Sets' Theory point of view. We materialize all query constraints as Sets that contain the providers' mappings.
We show graphically these sets of mappings:

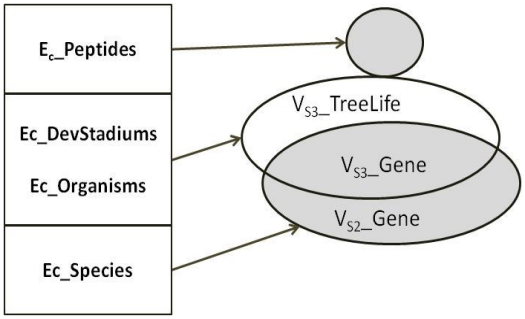

FIGURE 4: Associated Hypergraph of the Illustration

4.3.2 Determination of $Q_{invalide}$

$Q_{invalide}$ is the part of Q that cannot be answered by available sources. That means, we might constitute $Q_{invalide}$ with all Q's constraints, which are characterized by empty hyperedges of hypergraph $H_{Q,M}(V,E)$.
We can easily see that no mapping provides the hyperedges DevStadiums and Peptides. These hyperedges are empty of associated mappings (Q-Candidates'Finder: Line 2).
Hence:

$$Q_{invalide} = C_{Peptides} \wedge C_{DevStadiums} \quad (6)$$

4.3.3 Determination of Q_{valide}

Q_{valide} is the part of Q that can be answered by available sources. That means, we might constitute Q_{valide} with all Q's constraints, which are characterized by non-empty hyperedges of hypergraph $H_{Q,M}(V,E)$.
So, we have:

$$Q_{valide} = C_{Genes} \wedge \mathrm{C}_{\mathrm{Proteins}} \wedge C_{Organisms} \wedge C_{Species} \qquad (7)$$

In fact, Q_{valide} means that we could only answer the following request, given semantic mappings:

Give the genes which code all proteins, for the Plasmodium falciparum?

4.3.4 Determination of $Q_{candidates}$

From calculated above, we generate associated hypergraph $H^*_{Q,M}(V, E^*)$, (see Line 3).

Intuitively, according to Sets'Theory vision, finding all candidates rewriting suppose firstly to construct the Cartesian product of all sets of mappings. It is obvious that Elements of the Cartesian product are 4-uplets in our illustrative example. So, we must generate for each 4-uplet, an associated Set (These sets correspond to Transversals). In fact, it will be useful to use a minimal number of Sources that would be requested. This condition is guaranteed if we consider only associated sets that not contain another associated Set (These sets are minimals Transversals). That is why we call the Classical Algorithm (see Line 5). The maximal cardinality of our example Transversals is 4, it is the size of Q_{valide}.

From Sets' Theory point of view we can say that any set that contain V_{S1_Gene} and V_{S2_Gene} is a Transversal and constitutes possible rewritings of Q. The minimal Transversal $\{ V_{S1\_Gene} \}$ and $\{ V_{S1\_Gene} \}$ constitute two candidates rewritings. In our case, we could find 36 quadruplets, associated with 6 Transversals but only 2 are minimal Transversals.

5. CONCLUSION & FUTURE WORK

This paper deals with Global query rewriting, which consists in data integration context, to rewrite the global query expressed in terms of concepts and their properties defined in global schema domain ontology) into suitable terms of local data sources.

The Query rewriting process is based on semantic mappings. Our knowledge domain concern Proteomics research, and so we have proposed ontology according to interviews and talks with biologists and bio-informaticians, called O'proteomics. We provide a characterization of the Query rewriting problem based on Hypergraph Theory. We have presented the classiscal algorithm that computes all minimals Transversals, given an Hypergraph. We observe that those minimals Transversals correspond to Candidates rewritings of the Global Query.

Therefore, we need to better defined a logical formalism or language to specify the syntax and the semantics of data trees. It could be seen as a subset of Description Logics or based on Psi-terms formalism. After this essential choice, we will try to provide a prototype.

This paper shows briefly our current research that aims to provide a semantic framework to realize a Data Integration over XML bio-informatics sources on the Web. We will define some relevant criteria to rank candidates rewritings, necessary to select an optimal and qualitative rewriting $Q_{optimal}$. An efficient way for selecting best rewritings, iteration by iteration, will permit us to investigate the properties and the optimization of our algorithm. These relevant criteria could concern user preferences, quality of underlying sources, cost communication, etc…

6. REFERENCES

1. B. Amann, C. Beeri, I. Fundulaki, and M. Scholl, *"Ontology-based integration of XML web resources"*. In Proceedings of International Semantic Web Conference '02, pp. 117-131, 2002.

2. F. Baader, D. Calvanese, D. McGuinness, E.D. Nardi, and P. Patel-Schneider, "*The Description Logic Handbook, Theory, Implementation and Applications"*, Cambridge University Press, Cambridge, (2003).

3. B. Benatallah, M-S. Hacid, H-Y. Paik, C. Rey, and F. Toumani, "*Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities"*, In Elsevier's Journal of Information Systems, pp. 266-294, 2006.

4. C. Berge, "*Hypergraphs"*. North Holland, Amsterdam, ISBN 0 444 874895; QA166.23.B4813 (1989).

5. M. Bishop, "*Genetics Databases"*, Academic Press (1999).

6. T. Bray, J. Paoli, and Sperberg-McQueen, *"Extensible Markup Language (XML) 1.0"* W3C February Recommendation, 1998.

7. S.B. Davidson, C. Overton, P. Buneman, 1995. "*Challenges in integrating biological data sources"*. Journal of Computational Biology, 2(4):557–572.

8. C. Delobel, C. Reynaud, M-C. Rousset, J-P. Sirot, and D. Vodislav "*Semantic integration in Xyleme: a uniform tree-based approach"*, In Elsevier's Journal of Data & Knowledge Engineering 44, pp: 267–298, 2003.

9. A. Deutsch, V. Tannen, "*Reformulation of XML Queries and Constraints"*, In Proceedings of the 9th International Conference on Database Theory (ICDT), pp. 225-241, 2003.

10. A. Deutsch, V. Tannen, "*XML queries and constraints, containment and reformulation"*, Elsevier's Journal of Theoretical Computer Science, 33(6): 57-87, 2005.

11. T. Eiter, G. Gottlob, "*Identifying the minimal transversals of a hypergraph and related problems"*. SIAM Journal on Computing, 24(6):1278-1304, 1995.

12. I. Fundulaki, B. Amann, C. Beeri, and M. Scholl, "*STYX: Connecting the XML World to the World of Semantics Web resources"*. In Proceedings of EDBT Conference, Prague, Czech Republic, 2002.

13. G. Gottlob, C., Koch, and K.U. Schulz, "*Conjunctive Queries over Trees"*, In Proceedings 23rd ACM SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2004), Paris, France. ACM Press, New York, USA, pp. 189-200, 2004.

14. A. Halevy, "*Answering queries using views: a survey"*, In Proceedings of Very Large Data Bases. 10 (4), pp. 270–294, 2001.

15. A. Halevy, Z. Ives, I. Tatarinov, and P. Mork "*Piazza: Data management infrastructure for semantic web applications"*. In Proceedings of the International World Wide Web Conference, 2003.

16. D. Kavvadias, E., Stavropoulos, "*An Efficient Algorithm for the Transversal Hypergraph Generation"*, In Journal of Graph Algorithms and Applications, 9 (2), pp. 239-264, 2005.

17. J. Kohler, "*Integration of Life Science databases"*, In Elsevier's Drug Discovery Today Journal, BIOSILICO, 2(2), 2004.

18. P. Lehti, P. Fankhauser, "*XML Data Integration with OWL: Experiences and Challenges"*, In Proceedings of Symposium on Applications and the Internet (SAINT'04), pp. 160 -170, 2004.

19. A. Levy, A. Rajaraman., and J. Ordille "*Querying Heterogeneous Information Sources Using Source Descriptions"*, In Proceedings of Very Large Data Bases Conference, Mumbai, India, pp. 251-262, 1996.

20. H. Mannila, K-J. Raiha, "*The Design of Relational Databases"*. Addison-Wesley, Wokingham, England, (1994).

21. I. Manolescu, D. Florescu, and D.K. Kossmann, "*Answering XML queries over heterogeneous data sources"*. In Proceedings of the 27th International Conference on Very Large Data Bases (VLDB '01), Orlando, pp. 241–250, 2001.

22. C. Rey, F. Toumani, M.-S. Hacid, and A. Leger, "*An algorithm and a prototype for the dynamic discovery of e-services"*, Technical Report, LIMOS, Clemont-Ferrand, France, 2003.

23. T. Schlieder, "*ApproXQL: Design and Implementation of an Approximate Pattern Matching Language for XML"*, Technical Report, Freie Universitat Berlin, 2001.

24. H.S. Thompson, D. Beech, M. Maloney, N. Mendelson, *"XML Schema Part 1: Structures"* Second Edition, W3C Recommendation.Technical report, available on http://www.w3.org/TR/xmlschema-1/(2004), 28 October 2004.

JEVBase: An Interactive resource for protein annotation of JE Virus

Manas Ranjan Dikhit manasranjandikhit@gmail.com
Scientist, BioMedical Informatics Division
Rajendra Memorial Research Institute
of Medical Sciences
Agam Kuan, Patna, India-800007

Ganesh Chandra Sahoo ganeshiitkgp@gmail.com
Scientist & Head, BioMedical Informatics Division
Rajendra Memorial Research Institute
of Medical Sciences
Agam Kuan, Patna, India-800007

Pradeep Das drpradeep.das@gmail.com
Director
Rajendra Memorial Research Institute
of Medical Sciences
Agam Kuan, Patna, India-800007

Abstract

Databases containing proteomic information have become indispensable for virology related studies. Rajendra Memorial Research Institute of Medical Sciences (RMRIMS) has compiled and maintained a functional and molecular annotation database (http://www.jevbase.biomedinformri.org) commonly referred to as JEVBase. This database facilitates significant relationship between molecular analysis, cleavage sites and possible protein functional families assigned to different proteins of Japanese encephalitis virus (JEV). Identification of different protein functions and molecular analysis facilitates a mechanistic understanding of (JEV) infection and opens novel means for drug development. JEVBase database aims to be a resource for scientists working on JE virus.

Keywords: JEVBase, *Japanese encephalitis,* SVMProt, Protein annotation database, Functional database

1. INTRODUCTION

Japanese encephalitis virus (JEV) is the most common agent of viral encephalitis in the world, causing an estimated 45,000 cases and 10,000 deaths annually [1]. Epidemic form of Japanese encephalitis has been known since 1924 when 4,000 human deaths were recorded in Japan [2]. JE virus infection is also wide spread in southern states of India.

JEV contains a single positive sense RNA strand with about 11Kb nucleotides [3]. A single precursor polyprotein derived from JEV genome is subsequently processed by the host and viral protease to produce three structural proteins(Capsid (C), membrane(prM/M) and envelope (E)) and seven nonstructural proteins(NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) [4].

These three structural proteins are synthesized in the order of C, M and E from the 5' half of a single long open reading frames of the flavivirus genome. The glycosylated preM (precursor of M protein) and E proteins appear to be released from the nascent polyprotein following co-translational cleavage by signal peptidases. Late in virion maturation, preM is cleaved to M [5].

Different protein functions and molecular analysis facilitates for finding potential anti-viral inhibitors. Knowledge about protein function is essential in the understanding of biological processes [6]. The presence of a shared domain within a group of proteins does not necessarily imply that these proteins perform the same function [7]. One approach for function prediction is to classify a protein into functional family. Support vector machine (SVM) is a useful method for such classification, which may involve proteins with diverse sequence distribution [8]. Cloning and expression of different proteins practiced by molecular biologist will be helped by in silico restriction site analysis in the database.

In virology research, virus-related databases and bioinformatics analysis tools are essential for discerning relationships within complex datasets about viruses [9]. Chemoinformatics is the use of informatics methods to solve chemical problem[10]. Computational analysis and chemical information on *Japanese encephalitis* viruses involves the general tasks related to the analysis of any novel sequences, such as molecular analysis, functional annotation, and analysis of cleavage sites of the sequences. Support vector machines (SVM), useful for predicting the functional class of distantly related proteins, is employed to ascribe a possible functional class to *Japanese encephalitis* virus protein . Novel JEV protein functions have been analyzed through SVMProt have been earlier reported [11].

As the gap between the amount of sequence information and functional characterization widens, increasing efforts are being directed to the development of databases. For virologist, it is therefore desirable to have a single data collection point which integrates research related data from different domains. JEVBase is our effort to provide virologist such a one-step information center. We describe herein the creation of JEVBase, a new database that integrates information of different proteins in to a single resource. For basic curation of protein information, the database relies on features from other selected databases, servers and related papers.

2. MOTIVATION FOR JEVBase

Virology was slower to embrace bioinformatics [12]. No computational functional analysis of different proteins of JE virus is available till date. Protein identification and analysis software performs a central role in the investigation of proteins from two-dimensional (2-D) gels and mass spectrometry. For protein annotation, the user matches certain empirically acquired information against a protein database to define a protein function as already known or as novel. For protein analysis, information in protein databases can be used to predict certain properties about a protein, which can be useful for its empirical investigation.

All these in silico analysis give us an idea concerning the role of different proteins of JEV in replication, survival and spread of JEV in the host. Considering the biological significance of JEV protein and with the aim of providing easy access to the large and growing volume of data, we have developed JEVBase, a repository of all known JEV protein. JEVBase is the first known web resources, which provide the sequences as well as annotation information. The JEV protein have been analyzed, organized and integrated to develop a user friendly database and analysis system. The web interface enables the user to execute a quick and efficient search on JEVBase data. The database can be queried comprehensively through arguments such as National Center for Biotechnology Information **(**NCBI**)** Locus number, different protein name, different predicted functional family, stability etc. JEVBase will be an extremely useful resource for computational and experimental biologist working in this and related areas.

3. DATA PROCUREMENT AND REFINEMENT

The large scale of protein sequences have been reported in the NCBI protein database and supplementary data in the published literature. The sequences of *Japanese encephalitis* have been downloaded from the National Center for Biotechnology Information (NCBI) Protein database. Sequence redundancy is another problem of JE virus sequences in public protein databases. Different strains of the same species from samples collected in different location or at different times may possess completely identical sequences. Redundancy and repetition in protein sequences has been carefully removed by using ALIGN software to obtain a unique dataset [13]. Exactly matching sequences taken from multiple sources were eliminated while constructing the dataset. The raw dataset was preprocessed to remove the sequence smaller than 50bp while analyzing with different software.

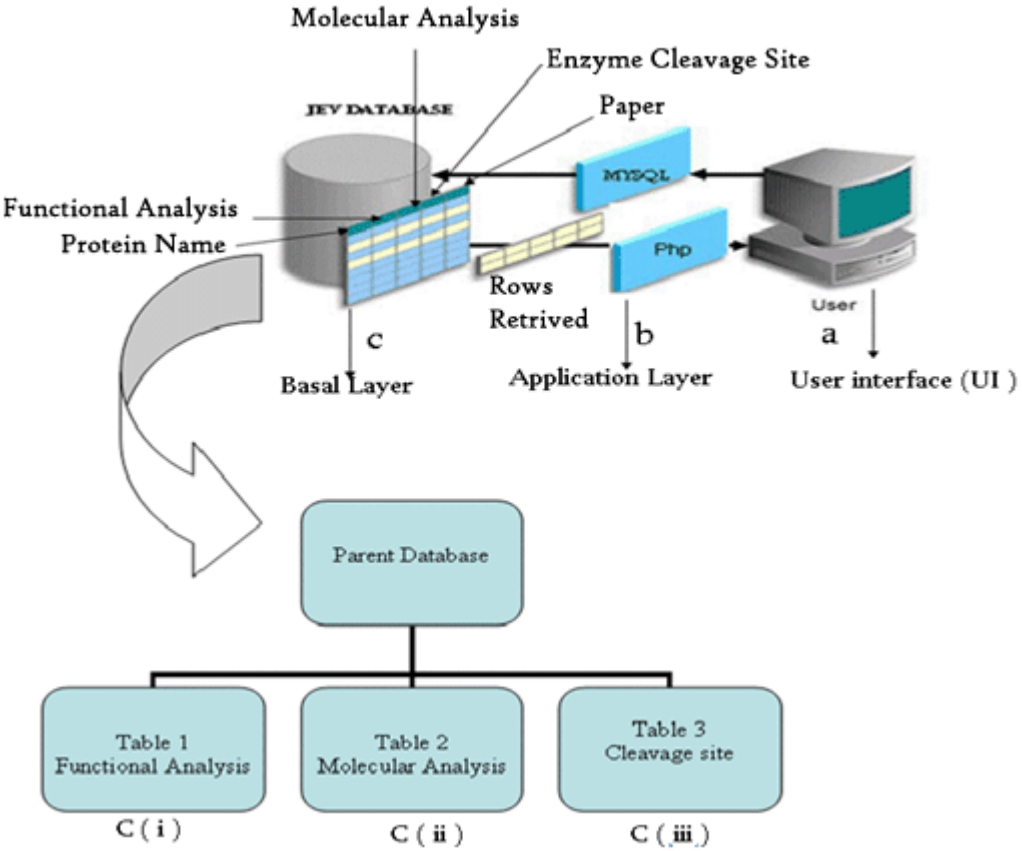

FIGURE 1. System architecture of the JEV database

4. SYSTEM ARCHITECTURE AND DESIGN

A relational database was constructed in MySQL for storage and query of data. It includes three key entities: 'functional analysis', 'molecular analysis' and 'cleavage sites', which simply analyze the protein. The JEVBase consists of three layers: (Fig1), 'The basal layer', 'Application layer' and 'UI'. The user interface (UI) layer has been developed using Php, CSS and JavaScript. Hypertext Preprocessor (php) is a widely used, general-purpose scripting language that was originally designed for web development, to produce dynamic web pages and Cascading Style Sheets (CSS) is a style sheet language used to describe the presentation semantics [14]. The basal layer that is parent database has been divided in to 3 tables. The JEVBase data and

information have been stored in MySQL relational database tables. Meta-information for different types of biological data is placed as individual table in this layer. The application layer between the web interface and the backend relational tables has been implemented using Php. The three layers of JEVBase can be manipulated and developed independently, which provides an optimal environment for maintenance and expansion of the JEVBase. Most of the interface component and application layer were standardized. This was made possible by employing a standardized scheme in building each layer.

5. DATABASE FEATURES

5.1. Data access

JEVBase can be queried to obtain the information about the protein sequences in many ways. Data stored in JEVBase can be accessed in the following ways:

(i) Search by protein name: The user can enter the desired protein name to access the Meta information about the protein sequences.

(ii) Search by protein functional family: The user can select the different protein functional family to find out the protein functional group of different structural and non-structural proteins.

(iii) Search by NCBI locus ID: The user can enter the NCBI locus ID to obtain JEV protein sequence information.

(iv) Search by Instability Index: To find out the stable and unstable protein, user can search by instability index.

JEVBase can be queried to obtain the information about protein-protein comparison. The user can enter the corresponding NCBI locus ID to compare two proteins.

5.2. Visualization

Database visualization helps users process, interpret and act upon large stored data sets. JEVBase provides a number of web-based forms for querying the dataset and selecting one or more protein for either a more detailed view of molecular annotation, Cleavage site and functional family or for viewing the comparison between two selected proteins.

In an effort to improve access to diverse JEV data, The JEVBase has been modified to include an abundance of linkage to other database including pubmed [www.ncbi.nlm.nih.gov/sites/entrez] for related papers and NCBI [www.ncbi.nlm.nih.gov] for corresponding sequences.

After performing a typical search the user is first presented with a summery page detailing the number of proteins matching the search (Fig 2). The following result page then provides the user with a list of proteins and brief descriptions from which individual proteins may be selected for either a detailed view (functional family, molecular annotation & cleavage sites) or a view of the related paper.

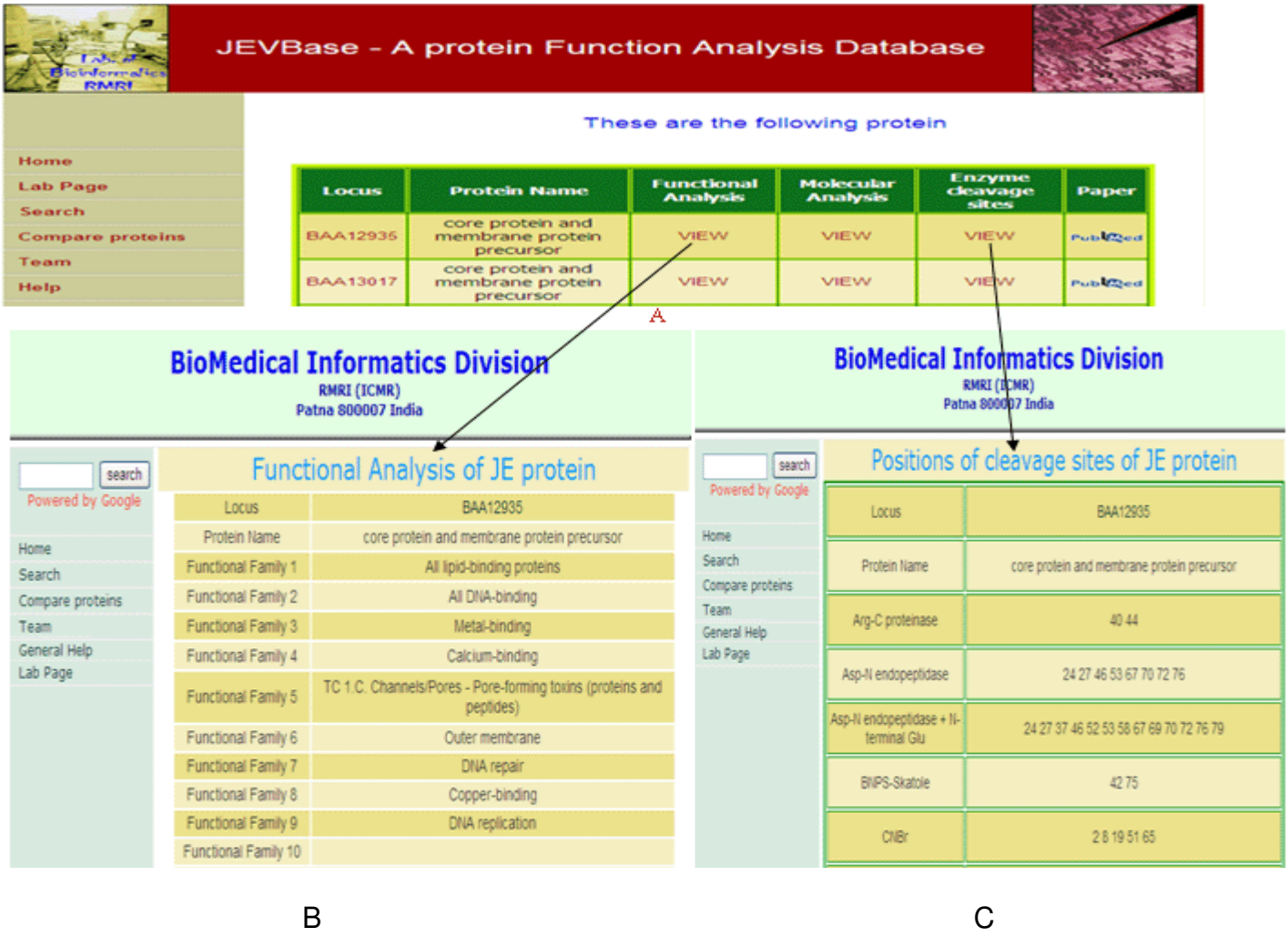

FIGURE 2. Typical screenshots from JEVBase. (A) Search result pages. These pages provide detailed view of each protein identified by a search. (B) Functional family view. This option gives the detailed view of protein functional family. (C) Cleavage sites view. These options give the detailed view of cleavage sites of the protein.

5.3. Data analysis

The protein function family predicted by SVMProt and other research papers are different for each structural and non-structural protein of JE virus strain, some of which may be responsible for virulence or pathogenicity of the virus and others for replication of the virus in the host. Prediction of the functional roles of lipid binding proteins is important for facilitating the study of various biological processes and the search for new therapeutic targets. Comparison of two amino acid sequences of any JE protein will reveal the user the distinguished functional properties of the corresponding protein if there is any amino acid change at any position as SVM works on the basis of physico-chemical properties of the amino acids of the protein e.g. when comparing function assignment of protein of two different NCBI locus numbers ABD84344 and ABD84370 , functions like metal binding, copper binding and DNA repair are common to both the strains whereas lipid degradation, calcium binding , iron binding and DNA binding functions are specific to ABD84344 and outer membrane and magnesium binding functions are specific to ABD84370(fig 3). Like this in molecular analysis of PreM protein, molecular

properties of this protein are also found to be different for each strain. From peptide-cutter and protparam analysis, pattern of restriction sites for all types of restriction enzymes for JE virus protein are visualized from the web server.

Hence the protein function family predicted is different for each structural and non-structural protein of JE virus strain, some of which may be responsible for virulence or pathogenicity of the virus and others for replication of the virus in the host. Prediction of the functional roles of lipid binding proteins is important for facilitating the study of various biological processes and the search for new therapeutic targets.

From this analysis, it is predicted that there is presence of network of functions performed by JEV proteins which brings about severe complicated clinical manifestations e.g. toxin-like pore forming property of core, matrix and NS4B proteins of JEV is responsible for causing acute flaccid paralysis as pore formation in the host causes release of water, micronutrients and macronutrients which can also occur in nerve cells causing severe inflammation of nerve cells and hence the patient suffers from acute meningitis.

This is comparision between two given proteins

FUNCTIONAL ANALYSIS

	PrM/PreM	PrM/PreM
	ABD84344	ABD84370
Functional Family 1		
Functional Family 2		
Functional Family 3	Lipid degradation	
Functional Family 4		
Functional Family 5	Metal-binding	Metal-binding
Functional Family 6	Calcium-binding	
Functional Family 7	Copper-binding	Copper-binding
Functional Family 8	DNA repair	DNA repair
Functional Family 9	Iron-binding	
Functional Family 10		Outer membrane
Functional Family 11		
Functional Family 12		Magnesium-binding
Functional Family 13	All DNA-binding	
Functional Family 14		
Functional Family 15		
Functional Family 16		
Functional Family 17		
Functional Family 18		
Functional Family 19		
Functional Family 20		

MOLECULAR ANALYSYS

	PrM/PreM	PrM/PreM
	ABD84344	ABD84370
Number of amino acids	80	80
Molecular weight	8891.2	8934.2
Negatively charged residues (Asp + Glu)	12	13
Positively charged residues (Arg + Lys)	7	7
Total number of Carbon atom	381	384
Total number of Hydrogen Atom	607	612
Total number of Nitrogen Atom	101	100
Total number of Oxygen Atom	123	124
Total number of Sulfur Atom	10	10
Total number of Atom	1222	1230
Instability index	stable	stable

FIGURE 3 . showing comparative functional analysis of same protein but with different NCBI locus numbers

6. CONCLUSION

JEVBase has been designed to manage and to explore the vast amount of protein data analysis. The current version of JEVBase has provided the basic molecular and functional analysis data of different proteins of *Japanese encephalitis* virus. JEVBase has been developed by keeping pace with the progress of the availability *Japanese encephalitis* proteins. User can search either by protein Functional family or protein name to access the Meta information about the protein sequences. This database facilitates significant relationship between molecular analysis, cleavage sites in the sequence, related paper and possible protein functional family of different proteins. Understanding of the structure-function correlation in viruses is important for finding potential anti-viral inhibitors and vaccine targets.

In near future we aim to include the modeled structures of different JE proteins and analyze quantitative structure–activity relationship of novel ligands targeting different proteins of JE virus. The database will be updated weekly on the basis of availability and analysis of the JE virus information and the amino acid sequences from NCBI and other reliable resources.

7. AVAILABILITY

The JEVBase database is freely available at http://www.jevbase.biomedinformri.org. All questions, comments and requests should be sent by e-mail to ganeshiitkgp@gmail.com .

8. ACKNOWLEDGEMENTS

The authors would like to thank Priya Darsan Sahu, Lingaraj Jena, Mukta Rani and Dr. Sindhuprava Rana for helpful discussion and valuable suggestions. We are thankful to Dr. Meera Singh for helping us during establishment of our division. This work is supported by Indian Council of Medical Research (ICMR), Govement of India.

9. REFERENCE:

[1] T. Solomon, L. T. T. Thao, N. M. Dung, R. Kneen, N. T. Hung, A. Nisalak, D. W. Vaughn, J. Farrar, T. T. Hien, N. J. White and M. J. Cardosa. " Rapid Diagnosis of Japanese Encephalitis by Using an Immunoglobulin M Dot Enzyme Immunoassay". Journal of Clinical Microbiology, 36**,** 2030-2034, 1998

[2] E.L. Buescher and W. F. Scherer. "Ecologic studies of Japanese encephalitis virus in Japan. IX. Epidemiologic correlations and conclusions". Am. J. Trop. Med. Hyg., 8,719-722, 1959

[3] H. Sumiyoshi, C. Mori, I. Fuke, K. Morita, S. Kuhara, J. Kondou, Y. Kikuchi, H. Nagamatu and A. Igarashi. "Complete nucleotide sequence of the Japanese encephalitis virus genome RNA". Virology, 161, 497-510, 1987

[4] F. Zhang, W. Ma, L. Zhang, M. Aasa-Chapman and H. Zhang. "Expression of particulate-form of Japanese encephalitis virus envelope protein in a stably transfected Drosophila cell line". Virology, 4, 17-24, 2007

[5] X. Liu, S. Cao, R. Zhou, G. Xu, S. Xiao, Y. Yang, M. Sun, Y. Li and H. Chen. "Inhibition of Japanese encephalitis virus NS1 protein expression in cell by small interfering RNAs", Virus Genes, 33, 69-75, 2006

[6] D. Eisenberg, C.A. Marcotte, I. Xenarios and T.O. Yeates. "Protein function in the post-genomic era". Nature , 405, 823–826, 2000

[7] S. Henikoff, E. A. Greene, S. Pietrokovski, P. Bork,T.K. Attwood and L. Hood. "Gene families: the taxonomy of protein paralogs and chimeras". Science, 278**,** 609–614, 1997

[8] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen and Y.Z. Chen. "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence". Nucleic Acids Research, 31, 3692–3697, 2003

[9] M. R. Dikhit , S.P. Rana, P. Das, and Ganesh Chandra Sahoo. " CHPVDB - a sequence annotation database for Chandipura Virus". Bioinformation. 3(7): 299–302, 2009

[10] A. Abdo and N. Salim. "Inference Networks for Molecular Database Similarity Searching". International Journal of Biometric and Bioinformatics, 2 (1) 1-16, 2008

[11] G.C. Sahoo, M.R. Dikhit and P. Das. " Functional assignment to JEV proteins using SVM". Bioinformation, 3**,** 1-7, 2008

[12] Y. Qing. "Bioinformatics databases and tools in virology research: An overview". In Silico Biology, 8, 0008 ,2008

[13] http://www.ebi.ac.uk/Tools/emboss/align/index.html .

[14] www.phpro.org

COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA