International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:2, No:4, 2008

Mining Genes Relations in Microarray Data Combined with Ontology in Colon Cancer Automated Diagnosis System

A. Gruzdz, A. Ihnatowicz, J. Siddiqi, and B. Akhgar

***Abstract*—**MATCH project [1] entitle the development of an automatic diagnosis system that aims to support treatment of colon cancer diseases by discovering mutations that occurs to tumour suppressor genes (TSGs) and contributes to the development of cancerous tumours. The constitution of the system is based on a) colon cancer clinical data and b) biological information that will be derived by data mining techniques from genomic and proteomic sources The core mining module will consist of the popular, well tested hybrid feature extraction methods, and new combined algorithms, designed especially for the project. Elements of rough sets, evolutionary computing, cluster analysis, self-organization maps and association rules will be used to discover the annotations between genes, and their influence on tumours [2]-[11].

The methods used to process the data have to address their high complexity, potential inconsistency and problems of dealing with the missing values. They must integrate all the useful information necessary to solve the expert's question. For this purpose, the system has to learn from data, or be able to interactively specify by a domain specialist, the part of the knowledge structure it needs to answer a given query. The program should also take into account the importance/rank of the particular parts of data it analyses, and adjusts the used algorithms accordingly.

***Keywords*—**Bioinformatics, gene expression, ontology, self-organizing maps.

I. Introduction

MATCH is an innovative project in the e-Health area, which involves the development of an automatic diagnosis system that aims to support treatment of colon cancer disease by discovering mutations that occurs to tumour suppressor genes (TSGs) and contributes to development of cancerous tumours.

MATCH is a web based multi functional platform that integrates medicine and molecular biology to provide more effective treatment and enhance pharmaceutical research and drug discovery. Clinical data derived from the electronic healthcare records and biological information derived from data mining techniques from genomic and proteomic sources could lead to more meaningful conclusions.

New techniques in distributed computing GRID environment together with the new applied scientific methods for genomic and proteomic discovery and data mining techniques combined with structured and controlled vocabularies (ontologies) make MATCH platform an advanced and innovative solution for information fusion of heterogeneous data.

The important part of the project is a bio-computing module for data integration and mining. The integration of GO, clinical and cDNA microarray data will give a unified view on tumour symptoms, independent from case-specific and source specific distortions. Especially the presence clinical data could provide a very useful insight into relationship between genetic and the cause of disease.

In order to facilitate the linking of information from different data source the basic biological knowledge models will be incorporate into the system, and, possibly, gradually expanded with the development of the software. The ontology module will take the input date from medical staff, enrich it with the researched biomolecular data and update all the information in the informational genomic – proteomic database.

The sophisticated computational problem of data integration in bioinformatics has recently become popular [12]-[15]. The main reason behind it is that, with such complexity of the model, there is no other way to completely understand gene function, protein regulation, and biological networks. Data integration also involves new, high-level view on the biological processes' connections, and thus is a key to comprehend all living organisms.

II. Data Mining Methods For Match

For the past few years of research and development, there have been many data mining machine learning and statistical analysis systems and tools available for general data analysis. They can be used in biodata exploration and analysis. General data mining and data analysis systems that can be used for biodata analysis include SAS Enterprise Miner, SPSS SPlus, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet and Inxight VizServer. There are also many bio

Manuscript received August 31, 2006.

Alicja Gruzdz Sheffield Hallam University, UK. University of Regina, Canada (e-mail: a.gruzdz@ shu.ac.uk).

Aleksandra Ihnatowicz. Sheffield Hallam University, UK (phone: +44 (0) 114 225 5329; fax: +44 (0) 114 225 5178; e-mail: a.ihnatowicz@shu.ac.uk).

Jawed Siddiqi Sheffield Hallam University, UK (e-mail: j.i.siddiqi@ shu.ac.uk).

Babak Akhgar Sheffield Hallam University, UK (e-mail: b.akhgar@ shu.ac.uk).

specific data analysis software systems, such as GeneSpring, Spot Fire and VectorNTI. These tools are rapidly evolving as well. A lot of routine data analysis work can be done using such tools. For biodata analysis, it is important to train researchers to master and explore the power of these well-tested and popular data mining tools and packages.

With sophisticated biodata analysis tasks, there is much room for research and development of advance, effective and scalable data mining methods in biodata analysis applied in MATCH decision support system . Some interesting topics are:

- Protein sequence analysis

The problem of sequence analysis can be reduced to sequence comparison similarities search and pattern finding. Most tools available today use dynamic programming algorithms for retrieving frequent pattern associated with proteins' biological functions. Another approach is to construct phylogenic trees from different probability models and sequence alignment methods. As in case of Hidden Markov Model, which are often used for identification of protein families. Among search problems, promoter search and protein motif search appear. Those problems are usually solved using probabilistic and stochastic methods, like Gibbs sampling method for example.

- RNA and Protein structure analysis

Pattern mining algorithms have been employed in order to analyze primary structures of proteins and extract useful knowledge. The analysis is usually done with the help of protein structure data basis i.e. Protein Data Bank (PDB), the structural classification of protein (SCOP) databases and Swiss- Model resource. The prediction of protein structure is still one of the most difficult and fundamental problems in the bioinformatics. New algorithms and optimization are needed to mine newly appeared proteomics data.

- Microarray analysis

DNA microarays contains information about gene expression differences in tissues and cell samples. The knowledge about genetic variation that appears in the samples allows for disease diagnosis or building new hypothesis about gene to gene relations in organisms. In here data mining techniques are also used for the identification of the most informative genes. One of the most challenging problems with gene expression mining is dimensionality reduction and specific variation in the data. The methods frequently used for microarray analysis are SVM, hierarchical clustering, neural networks and statistical measures, and algorithms for association rules discovery for example Apriori algorithm and its optimizations.

In case of MATCH Project data analysis we need to focus on:

- Dimension reduction techniques to handle multi-dimensional data
- Scalable algorithms for classification and clustering
- Parallel implementations for interactive exploration of data
- Applied statistics to ensure that the conclusions derived from the data are statistically sound.

III. Clustering

The fundamental part of MATH data analysis is gene clustering. There have already been many algorithms applied and optimized for this purpose, but none of them gave satisfying results. After the failures some even quarreled that gene clustering does not have a sense at all, while others created models so complex and unclear that occurred to be useless for biological analysis. From our experience the gene profiles clustering can be very beneficial, under the condition that data are well pre-processed and there is enough independent genetic knowledge to incorporate into the analysis. One of the methods that we have had best results on was Self-Organizing Maps (SOMs), adopted and tuned up especially to deal with a unique and complex gene expression data. SOMs will form an integrated module in MATCH system helping to retrieve the significant groups of genes for further processing.

SOM methods as it express (dis-)similarities in a very direct way, but flexibly mapping relationship between genes attributes. SOM module is equipped in a graphical interface that enables an expert work interactively to influent the SOM's learning process.

The similarities/distance functions to be mapped to the SOM grid can be calculated in various ways, beginning from the Euclidean-style comparison, through statistical correlations, up to the information theoretic measures. Following our previous research, we focus on functions interpreting the gene expression data characteristics as ranked attributes. We adapt the Spearman correlation and entropy-based distance to handle missing values in a non-invasive way, which is extremely important for this type of data.

Given functions studied in previous section with regards to capturing the gene similarities, we develop the self-organizing map (SOM) [16], [17] framework for calculating and visualizing the gene expression clusters. Every SOM forms a nonlinear mapping of a high dimensional data manifold into a regular, low- dimensional (usually 2D) grid. Both researchers and practitioners find that kind of display as useful for understanding the compound dependencies in data.

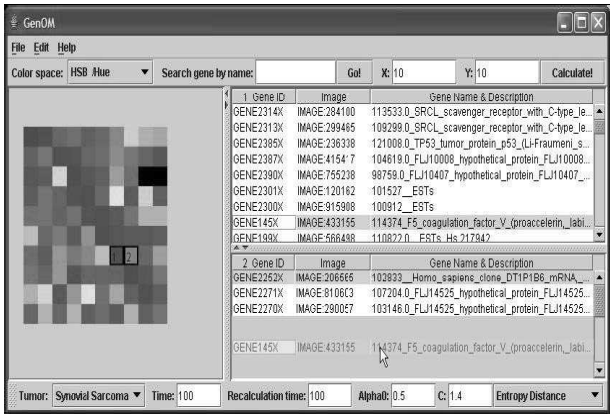

Fig. 1 Graphical GenOM interface. The left part provides the grid of clusters grouping together the most similar genes. Currently highlighted clusters are displayed in detail in the right part. There, the user can drag&drop genes between clusters. Quick recalculation of the map settings after such an operation is automatically supported

GenOM [18] is regarded as intuitive by the experts. Drag&drop function enables them to interact with the system and easily express assumptions about the gene dependencies. In this way, the following problems are solved, at least partially:

- The space of possible solutions-the grid locations of genes-is enormous. There is no guarantee that the heuristic procedure described in the previous section leads to optimum. A kind of correcting guidance would be helpful.
- Microarrays do not provide complete information about genes. They refer to genes indirectly, via their expressions. Moreover, those expressions can be imprecise depending on the applied microarray technology. An additional tool for interactive involvement of the expert knowledge seems to be required.

An important GenOM's computational feature is that drag& drop leads to the fast grid recalculation with compliance of the manually introduced changes. The program modifies the grid settings by increasing the mathematical influence of the given gene's expression characteristics on its new (after drag&dropping) neighbourhood in the grid. In the same way, the gene's influence on its previous (before drag&dropping) neighbourhood is appropriately decreased. The obtained new map's settings are automatically considered further in the learning process.

IV. Post-Clustering

In the new approach we treat the self organizing map clustering as a first attribute reduction. Then the most significant genes would be extracted and processed in next post-clustering stage to obtain better highlight on the correlations between genes. Such approach enables to us to use the methods with higher complexity, which could not be applied to high dimension data otherwise. Relevant groups of genes are already found during the first stage of processing while the interesting inner-dependencies between them are mined through post-clustering connected with ontology and clinically derived information in the further stages.

The post-cluster is done as an option and there is a possibility of selecting the technique used: SOMs, association rules mining, graph-driven feature extraction or hierarchical clustering. In the case of SOMs, the systems allows for zooming of desired cluster and performing the second computation loop with the chosen parameters and distance metric. Other methods have also been modified to enable such flexibility. Moreover, there exist interactive functions to specify the information about the known dependencies in data before processing or to correct them after the first results are calculated. An expert can simple drag&drop genes to the right clusters/positions according to the current genetic knowledge. The knowledge base built in such way can be manually saved and updated with the few findings. Once the project is already integrated this process will be automatic, retrieving and merging new information from MATCH database, that it also connected to Gene Ontology (GO) and Entrez integrated system.

V. Medcial Integrated Ontology Module

Generally, ontology is the description of an essential reality, i.e. what actually is; as opposed to what one can now. Strictly, in medical informatics ontology has come to mean a structure list of concepts which include medical knowledge. There are two main desirable properties to have in mind when designing the medical ontology: It should be an appropriate knowledge representation of the world, and it must point to a good set of terms in the search environment. Another aspect of interest is the role the ontology can play as a user guide to specify useful queries. The ontology can encapsulate key insights about inherited characteristics of medical objects that must drive the logic of software development and provide a common vocabulary is the essence of communication between software components. The ontology will be an open platform development tool that will not only interact with the other modules and parts of the system, but it will be developed in such a way that it will be able to link other medical, cellular and biomedical ontologies.

In the benefits of the usage of ontologies, we could enlist the following:

- Ontologies can help build more powerful and more interoperable information systems in healthcare.
- Ontologies can support the need of the healthcare process to transmit, re-use and share patient data.
- Ontologies can also provide semantic-based criteria to support different statistical aggregations for different purposes.

Possibly the most significant benefit that ontologies may bring to healthcare systems is their ability to support the indispensable integration of knowledge and data.

The ontology module will be used as an additional reasoning and validation component. Having medical and biological ontologies on the input, together with the results of data analysis, the tool will generate a model for validation and the enhancement of the last one. The model will reflect complex relationships among a patient's condition and current genetic and clinical knowledge about colon cancer. A patient's profile and its classification will be transformed by the ontology module to give a new view on patient's tumour development and possible ways of treatment.

We decided to use four existing ontologies listed below:

- The Gene Ontology (GO) http://www.geneontology.org/
- The Sequence Ontology Project (SO) http://song.sourceforge.net/
- Clinical Bioinformatics Ontology (CBO) https://www.clinbioinformatics.org/
- The National Cancer Institute Thesaurus (NCIT) http://ncicb.nci.nih.gov/

The Gene Ontology (GO) is especially important for our purpose and widely known in genetic society. The GO project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.

GO describes how gene products behave in a cellular context. The three organizing principles of GO are molecular function, biological process and cellular component.

GO terms are organized in structures called directed acyclic graphs (DAGs), which differ from hierarchies in that a child, or more specialized, term can have many parent, or less specialized, terms.

Medical vocabularies are important for ontologies integration. The National Cancer Institute Thesaurus (NCIT) does provide a rich terminology for carcinomas, which makes it a good starting point for ontology work in the cancer domain. When we're dealing with different ontologies the necessary integration is to be accomplished- the structure found within those terminologies and ontologies must be aligned with each other and with those found within Electronic Health Records on the basis of robust formal principles.

Medical vocabularies and relations between them make data-structure which could be mapped to data-exchange standards such as RDF and OWL. RDF, the Resource description Framework, is a W3C standard. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge. OWL stands for Ontology Web Language and is the new ontology data exchange standard accepted by the W3C (World Wide Web Consortium). It is based on RDF but has a larger vocabulary and a stronger syntax than RDF.

VI. Architectural Principles of the System

MATCH integrated systems strategy centers on Services Oriented Architecture development technologies as the common development framework spanning clients and servers, with a focus on Web Services and industry standards such as Simple Object Access Protocol (SOAP) and Extensible Markup Language (XML) for interoperability with components and services provided by other vendors.

In the MATCH project the inherent challenge in connecting diverse systems and components such as Data Mining and Colon Cancer Ontology to platform-specific information and procedural programming models. In essence a standard syntax, in which information from all systems can be unambiguously expressed, standard semantic models so that users can express their business practices in a consistent language, standard protocols so that information can be passed across boundaries between operating environments and between cancer specialist organizations,

Web service standards such as SOAP, XML, XML Schemas (XSD), Web Service Definition Language (WSDL), Universal Description, Discovery and Integration (UDDI) and WS-* specifications, such as WS-Security and WS-Policy, are used in the project in order to address the above concerns and requirements.

VII. Conclusion

The problem of genes clustering, examined excessively by the specialists, is not only a computational, but also a biological and medical challenge. There is still very little known about the genes relations, which makes it impossible to incorporate their structure logically into any sound model. These complex multi-attribute relationships haven't yet been reflected by any of the currently proposed techniques based on processing one kid of data. We believe that the key to the understanding of genes' mechanism, whatever is the studied biological process, is the integration of the genetic, proteomics, clinical and ontological data sources. The last two of those sources can especially help to solve the problem of dimensionality reduction. Maybe not completely, but they can certainly influence the groups of the analyzed genes by retrieving the ones that are known to be associated with some processes or ignore them if they do not appear in the interesting ontology terms or laboratory tests. The other method to deal with a high-dimensional microarray data is a pre-clustering, that allows others, possibly more multi-relation oriented techniques to be applied in the next step. The advantage of this approach is a focus on the interested genes according to their level of significance. Targeting such expressively suspected groups can be achieved through statistical measures, while the phase of dependencies' mining can be done with e.g. algorithms for extracting association rules. In case of MATH system such combination will give a unified view on tumor symptoms, independent from case-specific and source specific distortions, which will allow not only for patient's diagnosis, but also for inferring new genetic hypothesis.

REFERENCES

[1] http://www.match-project.com/
[2] Pawlak Z. (1982) Rough sets. International Journal of Information and Computer Sciences, 11(5):341-356.
[3] Pawlak Z. and Slowinski. R. (1994) Rough set approach to multi-attribute decision analysis. European Journal of Operational Research, 72(3):443-459.
[4] Slezak D. (2005) Association Reducts: A Framework for Mining Multi-attribute Dependencies. ISMIS 2005: 354-363.
[5] Wroblewski J. (1996) Theoretical Foundations of Order-Based Genetic Algorithms. Fundam. Inform. 28(3-4): 423-430.
[6] Wroblewski:J., Slezak D. (2003) Order Based Genetic Algorithms for the Search of Approximate Entropy Reducts. RSFDGrC 2003: 308-311.
[7] Yao H., Hamilton H.J., Butz C.J. (2004) A Foundational Approach to Mining Itemset Utilities from Databases. SDM 2004.
[8] Yao J.T., Yao Y.Y., and Zhao, Y. (2005) Foundations of classification, in: Lin, T.Y., Ohsuga, S., Liau, C.J. and Hu, X. (Eds), Foundations and Novel Approaches in Data Mining, Springer, Berlin, pp. 75-97.
[9] Yao Y.Y., Zhong, N. and Zhao, Y.(2004) A three-layered conceptual framework of data mining, Proceedings of ICDM'04 Workshop of Foundation of Data Mining, 215-221.
[10] Ziarko, W. (1989) A technique for discovering and analysis of cause-effect relationships in empirical data. International Joint Conference on Artificial Intelligence, Proceedings of the Workshop on Knowledge Discovery in Databases, Detroit, p.390-396.
[11] Ziarko, W. (1989) Determination of locally optimal set of features for representation of implicit knowledge. Proceedings of International Conference on Computing and Information, Toronto, North Holland, p.433-438.
[12] Baskin C., García-Sastre A., Tumpey T. (2004) Integration of Clinical Data, Pathology, and cDNA Microarrays in Influenza Virus-Infected Pigtailed Macaques Journal of Virology, October 2004, p. 10420-10432, Vol. 78, No. 19
[13] Casey R. M. (2005) Bioinformatics Data Integration. Business Intelligence Network
[14] Pasquier, C. et al. THEA: ontology-driven analysis of microarray data. Pasquier, C. et al. Bioinformatics 20(16), 2636-2643, 2004.
[15] Radetzki, U., Bode, T., Witterstein, G., Gnasa et al. (2003) A Service-Centric Computing Environment for Heterogeneous Biological Databases and Methods." In R. Spang, P. Beziat, and M. Vingron (eds.): Currents in Computational Molecular Biology (RECOMB 2003), pp. 25-26, April 2003, Berlin, Germany.
[16] Burger, M., Graepel, T., Obermayer, K.: Self-organizing maps: Generalizations and new optimization techniques. Neurocomputing 20 (1998) pp. 173-190.
[17] Kohonen, T.: Self-organized formation of topologically correct feature maps. Bio-logical Cybernetics 43 (1982) pp. 59-69.
[18] Gruzdz, A.,Ihnatowicz, A., Slezak, D.: Interactive gene clustering-A case study of breast cancer microarray data. Information Systems Frontiers (2006) 8:21-27.