INTEGRATING GENETIC AND CLINICAL DATA INTO A DECISION SUPPORT SYSTEM FOR COLON CANCER DIAGNOSIS

Nikolaos Giannakeas[1], Themis P. Exarchos[1], Yorgos Goletsis[2], Costas Papaloukas[3], Dimitrios I. Fotiadis[1], Babak Akhgar[4], Massimo Gulissano[5]

[1]Unit of Medical Technology and Intelligent Information Systems,
Dept. of Computer Science, University of Ioannina, PO Box 1186, GR 451 10 Ioannina, Greece
[2]Dept. of Economics, University of Ioannina, GR 45110, Ioannina, Greece
[3]Dept. of Biological Applications and Technology, University of Ioannina, GR 45110, Ioannina, Greece
[4]Sheffield Hallam University, City Campus, Howard Street, Sheffield S1 2BW, United Kingdom
[5]Istituto Oncologico del Mediteranneo, Viagrande, 95030, Italy
Email: fotiadis@cs.uoi.gr

ABSTRACT

In the cancer treatment domain, accuracy in patient staging and in the selection of personalised treatment plan can be of critical importance for patient's health or even survival. A Decision Support Platform that can correlate the patient clinical situation with the patient DNA Single Nucleotide Polymorphisms (SNPs) / mutations and so with his/her real condition can provide the oncologist with a better understanding of the personalized conditions of every single patient. In this paper we present the MATCH system which performs data integration between medicine and molecular biology, by developing a framework where, clinical and genomic features are appropriately combined in order to handle colon cancer diseases. Clustering (profiling) algorithms are designed and developed in order to analyze these heterogeneous data and extract useful information and knowledge. The constitution of such a decision support system is based on a) colon cancer clinical data and b) biological information that is derived from genomic sources. Through this integration, real time conclusions can be drawn for early diagnosis, staging and more effective colon cancer treatment. MATCH uses information directly from the DNA sequence of a patient in different stages of the cancer. Intelligent components are designed and developed which can identify single nucleotide polymorphisms (SNPs) from the gene sequences of the patients and create patient clinico-genomic profiles. These clinico-genomic profiles are used as a decision support tool for newly sequenced patients.

Introduction

Computer aided medical diagnosis is one of the most important research fields in biomedical engineering. Most of the efforts made, focus on diagnosis based on clinical features. The latest breakthroughs made by technology in biomolecular sciences are a direct cause of the explosive growth of biological data available to the scientific community. New technologies allow for high volume affordable production and collection of information on biological sequences, gene expression levels and proteins structure, on almost every aspect of the molecular architecture of living organisms. Additionally, bioinformatics provides tools for biological information processing, representing today's key in understanding the molecular basis of physiological and pathological genotypes (Campbell et. al. 2007). The exploitation of bioinformatics for medical diagnosis appears as an emerging field as the integration of clinical and genomic features maximizes the information regarding the patient's health status and the quality of the produced computer aided diagnosis.

Cancer is one of the prominent domains, where this integration is expected to bring about significant achievements. As genetic features play a significant role in the metabolism and the function of cells, the integration of genetic information (proteomics-genomics) to cancer related decision support is now perceived by many not as a future trend but rather as a demanding need. The usual patient management in cancer treatment involves several, usually iterative steps consisting of diagnosis, staging, treatment selection and prognosis. As the patient is usually asked to perform new examinations, diagnosis and staging status can change over time while treatment selection and prognosis depends on available findings, response to previous treatment plan and of course clinical guidelines. The integration of these evolving and changing data into clinical decision is a hard task which makes fully personalised treatment plan almost impossible. The use of decision support systems (DSSs) can assist in the processing of the available information and provide accurate staging, personalised treatment selection and prognosis. Moreover, the integration of the processed (quantitative) data that are hard to be processed by a human decision maker (the clinician), imposes the use of DSSs. The future vision - but current need - will not include genetic treatment plans according to some naïve reasoning but totally personalised treatment based on the clinico-genomic profile of the patient (Goletsis et. al. 2007).

The development of electronic patient records and technologies that produce and collect biological information have led to a plethora of data characterizing a specific patient. Although, this might seem beneficial, it can lead to confusion and weakness concerning data management. Thus, the need for DSSs that perform personalized medical care by integrating a large amount of data is evident (Louie et. al. 2007). In the following, we present a decision support system for colon cancer, namely the MATCH system, that uses clinical data and identifies mutations on tumour suppressor genes. Through this integration real time conclusions can be drawn for early diagnosis, staging and more effective colon cancer treatment.

Materials and Methods

Clinical and Genomic Data

Colon cancer includes cancerous growths in the colon, rectum and appendix. It is the third most common type of cancer and the second leading cause of death among cancers in the developed countries. There are many different factors involved in colon carcinogenesis. The association of these factors represents the base of the diagnostic process performed by medics which can obtain a general clinical profile integrating patient information using scientific knowledge. Available clinical parameters are stored together with genomic information for each patient to create a (as much as possible) complete electronic health record.

Several clinical data, that are contained in the electronic health records, are related with colon cancer (Read et. al. 1999): age, diet, obesity, diabetes, physical inactivity, smoking, heavy alcohol consumption, previous colon cancer or other cancers, adenomatous polyps, which are the small growths on the inner wall of the colon and rectum; in most cases, the colon polyp is benign (harmless). Also, other diseases or syndromes such as inflammatory bowel disease, Zollinger-Ellison syndrome and Gardner's syndrome are related to colon cancer.

In the context of genomic data related with colon cancer, malignant changes of the large bowel epithelium are caused by mutations of specific genes among which we can differentiate (Houlston et. al. 1997):

- Protooncogenes. The most popular mutated protooncogenes in colon cancer are: K-RAS, HER-2, EGFR and c-MYC
- Suppressor genes-anticogenes. In colorectal cancer the most important are DCC, TP53 and APC.
- Mutator genes. So far 6 repair genes of incorrectly paired up bases were cloned from humans, where four are related to Hereditary Nonpolyposis Colon Cancer (HNPCC) - hMSH2- homolog of yeast gene MutS, hMLH1 - homolog of bacterial MutL, hPMS1 and hPMS2 - from yeast equivalent - pair mismatch sensitive.

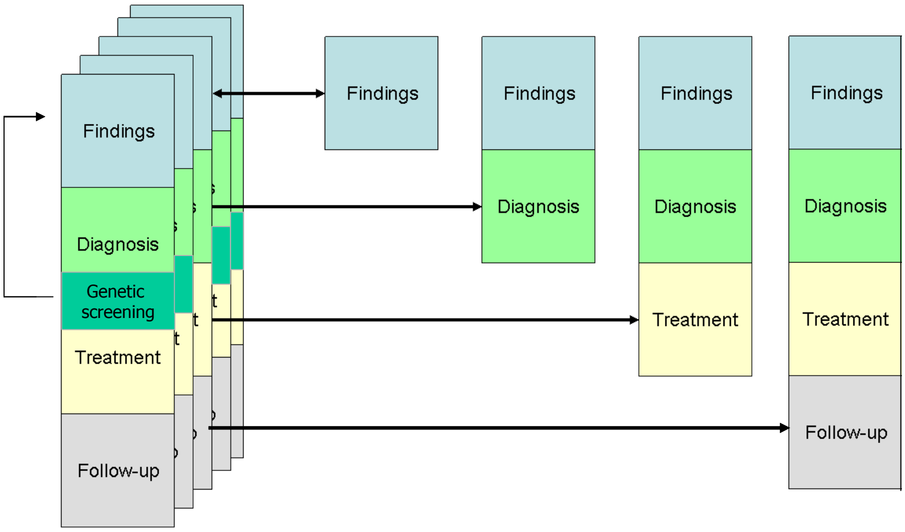
An efficient way to process the above gene sequences is to detect Single Nucleotide Polymorphisms (SNPs) (Sielinski 2005). SNP data are qualitative data providing genomic information at a specific locus of a gene. A SNP is a point mutation present in at least 1% of the genome population. A point mutation is a substitution or an addition of one base pair or a deletion, which means that the respective base pair is missing. Though several different sequence variants may occur at each considered locus, usually one specific variant of the most common sequence is found, an exchange from adenine (A) to guanine (G), for instance. Thus, information is basically given in the form of categories denoting the combinations of base pairs e.g. A/G if the most frequent variant is adenine and the SNP is an exchange from adenine to guanine.

MATCH Methodology

Conventional approaches for DSS focus on a single outcome concerning their domain of application. A different approach is to generate profiles associating the input data (e.g. findings) with several different types of outcomes. These profiles include clinical and genomic data along with specific diagnosis, treatment and follow-up recommendations. The idea of profile-based DSS is based on the fact that patients sharing similar findings are most likely to share the same diagnosis and should have the same treatment and follow-up; the higher this similarity is, the more probable this hypothesis holds.

The profiles are created from an initial dataset including several patient cases using a clustering method. Health records of diagnosed and treated patients, with clear follow-up description, are used to create the profiles. These profiles constitute the core of the DSSs; each new case that is inserted, is related with one (or more) of these profiles. More specifically, an individual health record containing only findings (and maybe the diagnosis) is matched to the centroids. The matching centroids are examined in order to indicate potential diagnosis. If the diagnosis is confirmed, genetic screening may be proposed to the subject and then, the clusters are further examined, in order to make a decision regarding the preferred treatment and follow-up. The above decision support idea is shown schematically in Figure 1.

Figure 1: Decision support based on profiles. Unknown features (diagnosis, treatment, follow-up) of a new case, described only by findings and maybe the diagnosis, are derived by known features of similar cases.



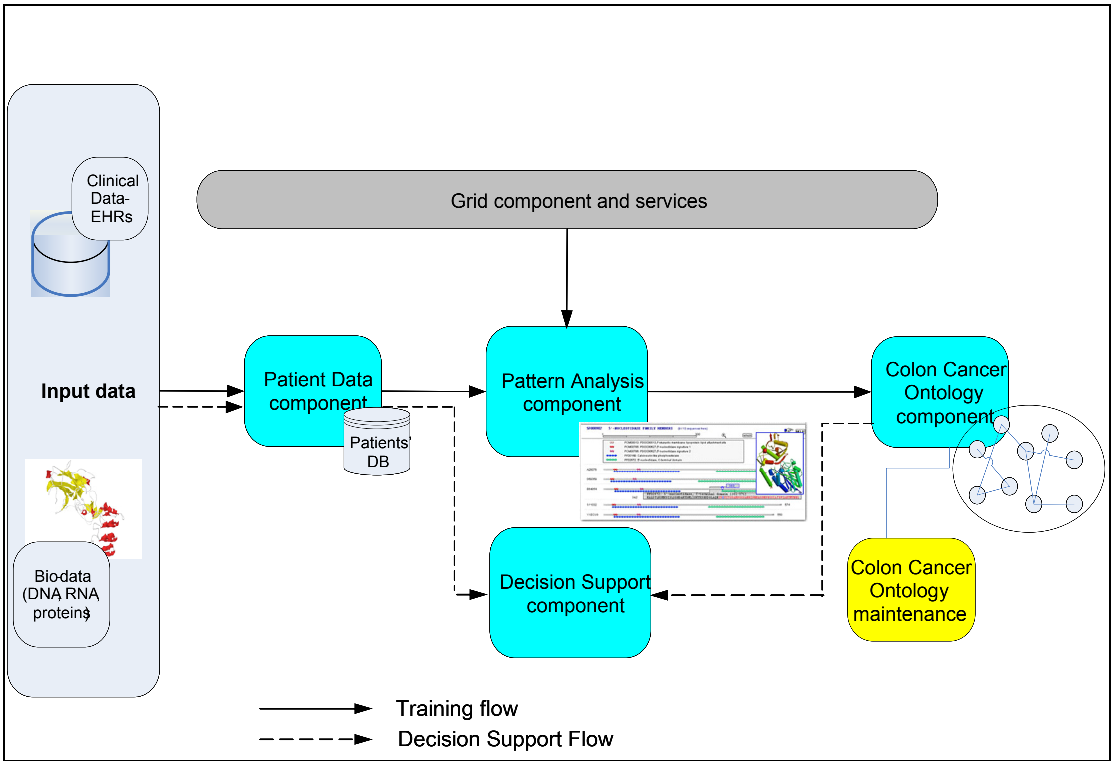
According to the above profiling methodology, sequences from genes are acquired from the subjects and based on the SNP information from every acquired gene, new features are derived, each one containing information regarding the existence or not of these SNPs in the patient's gene sequence. The derived features along with the aforementioned clinical data are the main input to the MATCH system. Then, clinico-genomic profiles are generated so as to provide advanced cancer decision support for new cases.

System Architecture

The architecture of the MATCH system is composed from a set of components which are presented in fig. 2. Two data flows are considered in MATCH functionality: the first one is a training flow dealing with the creation of the Colon Cancer Ontology from the discovery phase of the patient profiles while the second one is a decision support flow providing patient profile matching. The main components participating in system's functionality are as follows:

- Patient Data component. The Patient Data component interoperates with legacy applications, anonymizes the data coming from the MATCH patients repository (Patients DB), transforms them into an HL7 compatible form and integrates them. MATCH repository contains information coming from clinical sources (patient's electronic health record, legacy systems) and from biomolecular sources (patient's DNA mutations and other biomarkers). The component also exposes services that transform integrated data to XML structures that are the input to pattern matching and decision support components. The patient data component is responsible for the collection, integration, storage and presentation/output of the data.
- Pattern Analysis component. The Pattern Analysis component combines data output from the patient data component in order to identify metrics to be entered in the colon cancer ontology. These metrics are associated with information contained in the patients' health records (clinical and genomic information/mutations). Pattern analysis is based on k-means algorithm, which classifies patients into clusters, representing categorization into profiles. k-means can handle both continuous and discrete data and has low time and space complexity. Also, it provides straightforward distance computation, using the Euclidean distance for continuous data and city-block distance for the discrete data. A deficiency of the k-means algorithm is that the number of clusters (profiles) must be predefined, which is not always feasible. Thus, in order to fully automate the profile extraction process, a meta-analysis technique is employed, which automatically calculates the optimal number of profiles (Witten et al. 2005).
- Colon Cancer Ontology component. The Colon Cancer Ontology component is the domain ontology structure that captures information for patterns to be matched against mutations and patient profiles. The MATCH Ontology provides a conceptual scheme of complex system dealing with heterogeneous medical data. Knowledge is modelled in such a way that enables proper understanding of particular attributes from health records, show relations between them as well as provide methods for data comparison and matching (i.e. algorithms, parameters and metrics). Specifically, the ontology provides information concerning which features participate in pattern analysis and in decision support. Also, the ontology provides a model for distance computation applied in the clustering, as well as in the matching of individual health records to the profiles.
- Decision Support component. The Decision Support component (DSS) is a service oriented program whose main purpose is to provide diagnostic and therapeutic information about a specific new patient contained in the MATCH system. This is done by performing comparisons to the cluster centroids obtained by the Pattern Analysis Component. The Decision Support component, uses the ontology engine to support the doctor's diagnosis. When necessary, the Decision Support component retrieves data from the Patient Data component to continue its execution. The output of the DSS is driven to the user interfaces, part of which uses visualization tools such as the European Molecular Biology Open Software Suite, for the representation of molecular structures.

Figure 2: MATCH Architecture


Clinical Data-EHRs
Grid component and services
Input data
Patient Data component
Patients' DB
Pattern Analysis component
Colon Cancer Ontology component
Bio-data (DNA, RNA, proteins)
Decision Support component
Colon Cancer Ontology maintenance
Training flow
Decision Support Flow

DISCUSSION

The MATCH system provides health professionals with a multi-functional platform for colon cancer decision support. MATCH contributes directly to the health care sector by grouping previously unrelated data and reducing the cost of expensive trials in the area of biochemical and pharmaceutical research. MATCH goes further than previous attempts in the field by adding the genetic dimension in the diagnosis process. In this way it provides an integrated system for management, prognosis, diagnosis and treatment of colon cancer. Moreover, semantic information is integrated, using ontologies since data are structured based on a specially developed ontology module. In addition, profiles of patients are extracted using efficient clustering algorithms. Furthermore, grid computing for faster processing of the information and visualization components for visualizing the SNPs in the DNA sequences is used.

The exploitation of the genetic data and the association of clinical and genetic data produce the system's scientific added value. MATCH uses information directly from the DNA sequence of a patient in different stages of the cancer. Intelligent components are designed and developed that identify SNPs from the sequences and create patient profiles. The profiles of the patients are based on both clinical and genetic information. Clustering techniques are applied for patient profiling and discovery of new knowledge. Furthermore, the clinico-genomic profiles are used as a decision support tool for newly sequenced patients.

In conclusion, advances in genome technology are playing a growing role in medicine and healthcare. With the development of new technologies and opportunities for large-scale analysis of the genome, genomic data have a clear impact on medicine. Cancer prognostics and therapeutics are among the first major test cases for genomic medicine, given that all cancer is related with genomic instability. The integration of clinical data with genomic data makes the prospect for developing personalized healthcare, even more real.

Acknowledgments

This work is part funded by the European Commission (project MATCH: Automated Diagnosis System for the Treatment of Colon Cancer by Discovering Mutations on Tumour Suppressor Genes IST-2005-027266).

References

Campbell, A.M., Heyer, L.J. (2007). Discovering Genomics, Proteomics and Bioinformatics, Benjamin Cummings, CA, USA.

Goletsis, Y., Exarchos, T.P., Giannakeas, N., Tsipouras, M.G., Fotiadis, D.I. (2007) Integration of clinical and genomic data for decision support in cancer, in Wickramasinghe N., Geisler E. (eds), Encyclopedia of Healthcare Information Systems, Idea Group Publishing, USA, (to be published).

Houlston, R.S., & Tomlinson, I.P.M. (1997). Genetic prognostic markers in colorectal cancer. Journal Clinical Pathology: Molecular Pathology, vol. 50, 281-288.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. Journal of Biomedical Informatics, vol. 40, 5–16.

Read, T.E., & Kodner, I.J. (1999). Colorectal cancer: risk factors and recommendations for early detection. American Family Physician, vol. 59(11), 3083-3092.

Sielinski, S. (2005). Similarity measures for clustering SNP and epidemiological data. Technical report of university of Dortmund.

Witten, I.H., and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques with java implementations. Morgan Kaufmann, CA, USA.