A FRAMEWORK FOR CANCER DECISION SUPPORT BASED ON PROFILING BY INTEGRATING CLINICAL AND GENOMIC DATA: APPLICATION TO COLON CANCER

T.P. EXARCHOS
Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece

N. GIANNAKEAS
Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece

Y. GOLETSIS
Dept. of Economics, University of Ioannina, GR 45110 Ioannina, Greece
University of Ioannina, GR 451 10 Ioannina, Greece

C. PAPALOUKAS
Dept. of Biological Applications and Technology, University of Ioannina, GR 45110 Ioannina, Greece

D.I. FOTIADIS
Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, GR 45110 Ioannina, GREECE

In this paper we present a general framework for decision support in cancer diseases. The framework is based on profiling patients with similar characteristics and based on these profiles, advanced decision support is provided to new patients. The novel feature of the proposed work is the integration of clinical with genetic data, in the most prominent way so as to maximize the information related to the status of a patient. The paper also presents a specific implementation that integrates sequence data and Single Nucleotide Polymorphisms, related to colon cancer, with clinical data, recommended by the experts to play a vital role in colon cancer. The outcome of the specific implementation is a set of clinico-genomic profiles, which are employed for decision support, autometed diagnosis, prognosis, treatment and follow-up in colon cancer.

1. 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 of the technology in the 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. For this reason, bioinformatics is asked to provide tools for biological information processing, representing today's key in understanding the molecular basis of physiological and pathological genotypes. The exploitation of bioinformatics for medical diagnosis appears as an emerging field for the integration of clinical and genomic features, maximizing the information regarding the patient's health status and the quality of the computer aided diagnosis.

Cancer is one of the prominent domains, where this integration is expected to bring significant achievements. As genetic features play significant role in the metabolism and the function of the 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 depend on the 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 the development of fully personalised treatment plan almost impossible. The use of clinical decision support systems (CDSSs) can assist in the processing of the available information and provide accurate staging, personalised treatment selection and prognosis. The development of electronic patient records and of 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 the data management. The integration of the patient data (quantitative) that are hard to be processed by a human decision maker (the clinician) further imposes the use of CDSSs in personalized medical care [1]. The future vision - but current need - will not include generic treatment plans according to some naive reasoning, but totally personalised treatment based on the clinicogenomic profile of the patient

2. Clinical Decision Support using Clinicogenomic Profiles

Clinical Decision Support Systems are active knowledge systems which use two or more items of patient data to generate case-specific advice [2]. CDSSs are used to enhance diagnostic efforts and include computer based programs that,

based on information entered by the clinician, provide extensive differential diagnosis, staging (if possible), treatment, follow-up, etc. CDSSs consist of an inference engine that is used to associate the input variables with the target outcome. This inference engine can be developed based either on explicit medical knowledge, expressed in a set of rules (knowledge based systems) or on data driven techniques, such as machine learning [3] and data mining (intelligent systems) [4]. CDSSs require the input of patient-specific clinical variables (medical data) and as a result provide patient specific recommendation.

2.1. *Description of the Methodology*

Conventional approaches for CDSS focus on a single outcome related to 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 CDSS 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 (successfully or unsuccessfully) treated patients, with clear follow-up description, are used to create the profiles. These profiles constitute the core of the CDSSs; 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 (the term diagnosis here refers mainly to the identification of cancer sub-type). If the diagnosis is correct, genetic screening may be proposed to the subject and then, the clusters are further examined, in order to make a decision about the preferred treatment and follow-up. The above decision support idea is shown schematically in Fig. 1.

2.2. *Description of the system*

Known approaches for the creation of CDSSs are based on the analysis of clinical data using machine learning techniques. This scheme can be expanded to include genomic information, as well. In order to extract a set of profiles, the integration of clinical and genomic data is first required. Then, data analysis is realized in order to discover useful knowledge in the form of profiles. Several techniques and algorithms can be used for data analysis such as neural approaches, statistical analysis, data mining, clustering and others. Data analysis is a two stage procedure: (i) creation of an inference engine (training stage) and (ii) use of this engine for decision support. The type of analysis to be used

greatly depends on the available information and the desired outcome. Clustering algorithms can be employed in order to extract patient clinico-genomic profiles. An initial set of records, including clinical and genomic data along with all diagnosis/treatment/follow-up information, must be available for the creation of the inference engine. The records are used for clustering and the centroids of the generated clusters constitute the profiles. These profiles are then used for decision support; new patients with similar clinical and genomic data are assigned to the same cluster, i.e. they share the same profile. Thus, a probable diagnosis, treatment and follow-up, is selected.

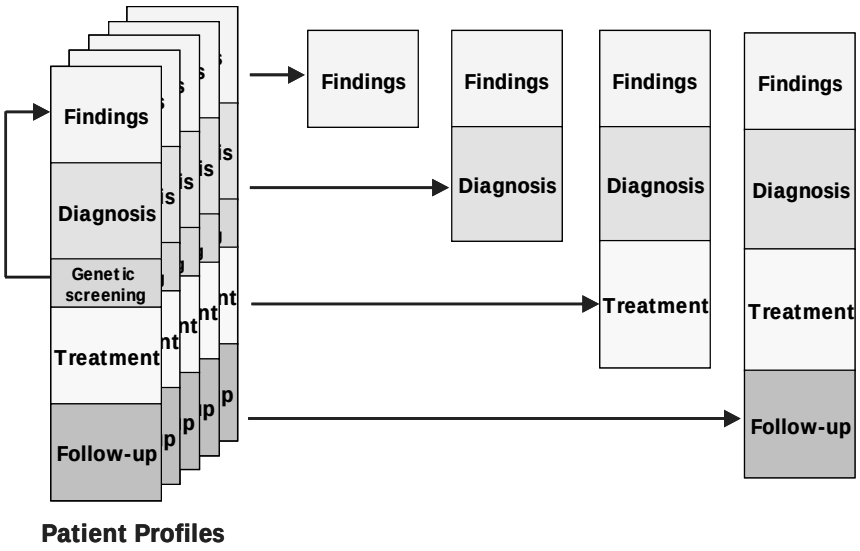

Figure 1: Decision support based on profiles. Unknown features (diagnosis, treatment, follow-up) of a new case, described only by findings and maybe the diagnosis, are derived by known features of similar cases.

2.2.1. Data processing

Depending on the type of the available biological data, different preprocessing steps should be performed in order to derive structured biological information, while expert knowledge could favor the preprocessing steps. The processing stage is necessary in order to transform the genomic data into a more easy-to-analyse form, allowing their integration along with the clinical data into the data analysis stage. Also, the genomic data processing might take advantage of expert knowledge, i.e. known genomic abnormalities. Finally, the integrated data (clinical and genomic) are analysed in order to discover useful knowledge

that can be used for decision support purposes. This knowledge can be in the form of associations among clinical data, genomic data, diagnosis, treatment and follow-up.

The initial dataset (clinical or genomic) is defined by the experts and includes all features that according to their opinion are highly related with the domain under discussion (clinical disease). After acquiring the integrated data, a feature selection technique is applied in order to reduce the number of features and remove irrelevant or redundant ones. Since the proposed scheme for decision support focuses on several outcomes, a supervised feature selection technique cannot be employed. For this reason, a method based on principal component analysis is used to reduce the number of features [5]. Finally, the reduced set of features is used by a clustering algorithm. k-means algorithm [6] is a promising approach for clustering and can be employed for profile extraction. k-means handles 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. Furthermore, k-means algorithm is an order independent algorithm, since for a given initial distribution of clusters generates the same partition of the data at the end of the partitioning process, independently of the order in which the samples are presented to the algorithm. 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 [7]. This technique divides the data into ten sets and performs clustering in each of them. Initially, k is set to 2 and the mean value of the sum of squared errors over the ten sets is computed. k is increased until the mean value of the sum of squared errors is stabilized or is higher than the previous value of k (k-1).

2.3 Application to colon cancer

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 his 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 to colon cancer [8]: 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, the Zollinger-Ellison syndrome and the Gardner's syndrome are related to colon cancer.

In the context of genomic data related to colon cancer, malignant changes of the large bowel epithelium are caused by mutations of specific genes among which we can differentiate [9]:

- 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) [10]. SNPs data are qualitative data providing information about the genomic at a specific locus of a gene. An SNP is a point mutation present in at least 1 % of a population. A point mutation is a substitution of one base pair or a deletion, which means, the respective base pair is missing, or an addition of one base pair exists. 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 for the two chromosomes, e.g. A/A, A/G, G/G, if the most frequent variant is adenine and the single nucleotide polymorphism is an exchange from adenine to guanine.

According to previous medical knowledge, there are several SNPs with known relation to colon cancer. Some indicative SNPs already related to colon cancer according to several sources in the literature, identified in TP53 gene are presented in Table 1. The expert knowledge contains information about the position of the SNPs in the gene sequence (i.e. exon, codon position and amino acid position), the transition of the nucleotides and the translation of the mRNA to protein. Based on the list of known SNPs related to colon cancer, appropriate genomic information is derived, revealing the existence or not of these SNPs in the patient's genes.

Some of the genes described above are acquired from the subjects and based on the SNP information concerning each acquired gene, such as SNPs in Table 1 for TP53 gene, new features are derived, each one containing information related to the existence or not of these SNPs in the patient's gene sequence. The derived features along with the aforementioned clinical data which are related with colon cancer are the input to the methodology and the output are the

generated clinicogenomic profiles. These profiles are able to provide advanced cancer decision support for new patients.

Table 1. Indicative SNPs transitions and positions in the TP53 gene, related to colon cancer.

Region	mRNA pos.	Codon pos.	Amino acid pos.	Function	Transition	Protein residue transition
Exon_10	1347	1	366	nonsynonymous	G/T	Ala [A]/Ser [S]
Exon_10	1266	1	339	nonsynonymous	A/G	Lys [K]/Glu[E]
Exon_9	1242	3	331	synonymous	A/G	Gln [Q]/Gln[Q]
Exon_8	1095	1	282	nonsynonymous	T/C	Trp [W]/Arg[R]
Exon_8	1083	1	278	nonsynonymous	G/C	Ala [A]/Pro[P]
Exon_8	1069	2	273	nonsynonymous	A/G	His [H]/Arg[R]
Exon_7	1021	2	257	nonsynonymous	A/T	Gln [Q]/Leu[L]
Exon_7	998	3	249	nonsynonymous	T/G	Ser [S]/Arg[R]
Exon_7	994	2	248	nonsynonymous	A/G	Gln [Q]/Arg[R]
Exon_7	984	1	245	nonsynonymous	A/G	Ser [S]/Gly[G]
Exon_7	982	2	244	nonsynonymous	A/G	Asp [D]/Gly[G]
Exon_7	973	2	241	nonsynonymous	T/C	Phe [F]/Gly[G]
Exon_5	775	2	175	nonsynonymous	A/G	His [H]/Arg[R]
Exon_5	702	1	151	nonsynonymous	A/T/C	Thr [T]/Ser[S]/Pro [P]
Exon_5	663	1	138	nonsynonymous	C/G	Pro [P]/Ala [A]
Exon_5	649	2	133	nonsynonymous	C/T	Thr [T]/Met [M]
Exon_4	580	2	110	nonsynonymous	T/G	Leu [L]/Arg [R]
Exon_4	466	2	72	nonsynonymous	G/C	Arg [R]/Pro [P]
Exon_4	390	1	47	nonsynonymous	T/C	Ser [S]/Pro [P]
Exon_4	359	3	36	synonymous	A/G	Pro [P]/Pro [P]
Exon_4	353	3	34	synonymous	A/C	Pro [P]/Pro [P]
Exon_2	314	3	21	synonymous	T/C	Asp [D]/Asp [D]

3 Conclusions

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 types of cancer are related to genomic instability. The integration of clinical data with genetic data makes the prospect for developing personalized healthcare, even more real.

Acknowledgments. This research is part funded by the European Commission as part of the project MATCH (Automated diagnosis system for the treatment of

colon cancer by discovering mutations on tumor suppressor genes, IST-2005-027266).

References

1. Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. Journal of Biomedical Informatics, 40, 5–16.
2. Fotiadis, D.I., Goletsis, Y., Likas, A., & Papadopoulos, A. (2006). Clinical Decision Support Systems. in M. Akay, ed., Encyclopaedia of Biomedical Engineering, Wiley.
3. Mitchell, T. (2006). Machine Learning, Springer, McGraw-Hill Education (ISE Editions).
4. Tan, P.N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining, Addison Wesley. USA.
5. Webb, A.: Statistical Pattern Recognition. Arnold, New York, USA (1999).
6. MacQueen, J.B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.
7. Witten, I.H., and Frank, E.: Data Mining: Practical machine learning tools and techniques with java implementations. Morgan Kaufmann, California, USA (2005).
8. Read, T.E., & Kodner, I.J. (1999). Colorectal cancer: risk factors and recommendations for early detection. American Family Physician, 59(11), 3083-3092.
9. Houlston, R.S., & Tomlinson, I.P.M. (1997). Genetic prognostic markers in colorectal cancer. Journal Clinical Pathology: Molecular Pathology, 50, 281-288.
10. Sielinski, S. (2005). Similarity measures for clustering SNP and epidemiological data. Technical report of university of Dortmund.