

Pattern Analysis and Decision Support for Cancer through Clinico-Genomic Profiles

Themis P. Exarchos¹, Nikolaos Giannakeas¹, Yorgos Goletsis^{1,2,*}, Costas Papaloukas^{1,3}, and Dimitrios I. Fotiadis¹

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

*Indicate corresponding author

me01238@cc.uoi.gr, me01310@cc.uoi.gr, goletsis@cc.uoi.gr, papalouk@cc.uoi.gr, fotiadis@cs.uoi.gr

Abstract. 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 with genomic instability. In this paper we present a novel system for pattern analysis and decision support in cancer. The system integrates clinical data from electronic health records and genomic data. Pattern analysis and data mining methods are applied to these integrated data and the discovered knowledge is used for cancer decision support. Through this integration, conclusions can be drawn for early diagnosis, staging and cancer treatment.

Keywords: Cancer, decision support systems, pattern analysis, profiles, data integration.

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 developments in the biomolecular sciences have as a result an explosive growth of biological data available to the scientific community. Bioinformatics provides tools for biological information processing, representing today's key in understanding the molecular basis of physiological and pathological genotypes [1]. The exploitation of bioinformatics for medical diagnosis is an emerging field for the integration of clinical and genomic features and 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 [2] depends 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 fully personalised treatment plan almost impossible. The use of clinical decision support systems (CDSSs) [3] 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 [4]. The future vision - but current need - will not include generic treatment plans according to some naive reasoning, but totally personalised treatment based on the clinico-genomic profile of the patient [2].

In this paper we address decision support for cancer, by exploiting clinical data and genomic data. The goal is to perform data integration between medicine and molecular biology, by developing a framework where, clinical and genomic features are appropriately combined in order to handle cancer diseases. The constitution of such a decision support system is based on a) cancer clinical data and b) biological information that is derived from genomic sources. Through this integration, conclusions can be drawn for early diagnosis, staging and effective cancer treatment.

2 Clinical Decision Support using Clinico-genomic Profiles

2.1 Description of the Methodology

Most of the proposed approaches for clinical decision support focus on a single outcome regarding their domain of application. A different approach is to generate profiles associating the input features (e.g. findings) with several outcomes. These profiles include clinical and genomic data along with specific diagnosis, treatment and follow-up recommendations. 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 the cancer sub-type). 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 (Fig. 1).

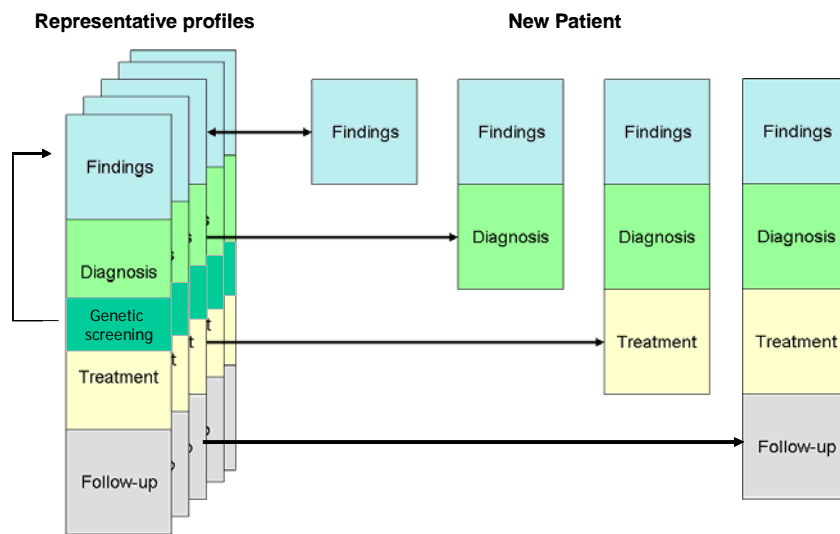


Fig. 1. Decision support based on profiles. Unknown features (diagnosis, treatment, follow-up) of a new case, described only with findings are derived by known features of similar cases.

2.2 General 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. Both, the creation of the inference engine and the decision support procedure are presented in Fig. 2.

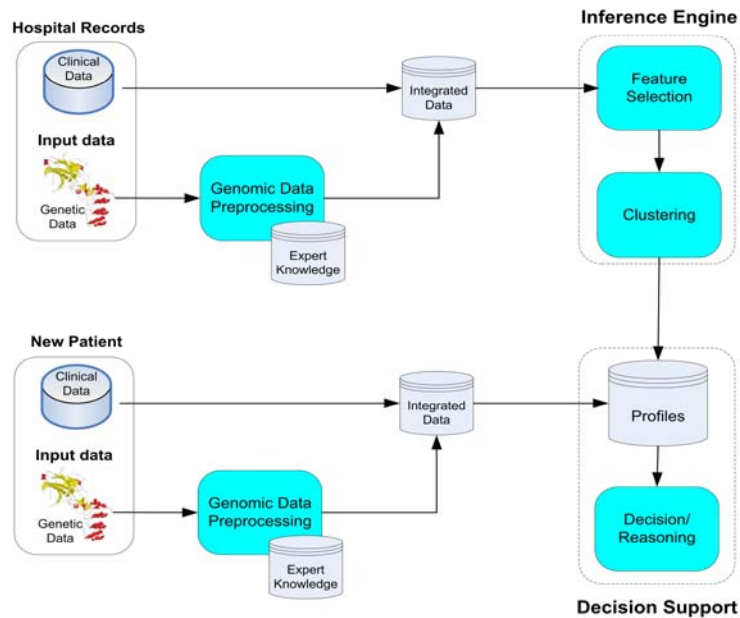


Fig. 2. Representation of a general scheme for the creation of an inference engine of a profile based CDSS, integrating clinical and genomic information. A new case (patient) is applied to the profiles for decision support and reasoning.

2.2.1 Types of data

Clinical data such as demographic details, medical history and laboratory data are usually presented in a structured format, making their analysis an easy task. The proposed method combines the clinical data with biological data, and more specifically, with gene sequence data. An efficient way to process the above gene sequences is to detect Single Nucleotide Polymorphisms (SNPs) [5]. 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. 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 form of categories denoting the combinations of base pairs for the two chromosomes, e.g. A/G, if the most frequent variant is adenine and the single nucleotide polymorphism is an exchange from adenine to guanine.

2.2.2 Data processing

Since the gene sequence data are not structured, appropriate preprocessing is needed in order to transform them into a more structured format. Gene sequences are acquired from the subjects and based on the SNP information regarding every acquired gene new features are derived. Each of these features contains 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 input to the inference engine, in order to generate clinico-genomic profiles. These profiles are able to provide advanced cancer decision support to new patients.

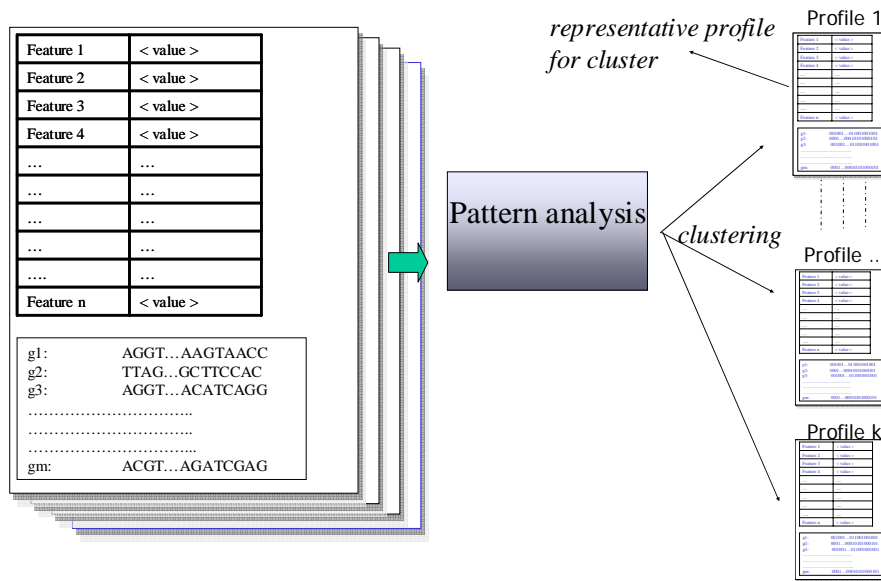


Fig. 3. Schematic representation of clustering integrated clinical and genomic data. The outputs are the representative clinico-genomic profiles for each generated cluster.

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 at hand (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 can not be employed. For this reason, a method based on principal component analysis is used to reduce the number of features [6]. Finally, the reduced set of features is used by a clustering algorithm. k-means algorithm [7] is a promising approach for clustering and can be involved 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 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, irrespective 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 [8]. This technique divides the data into 10 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 10 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). A schematic representation of clustering integrated clinical and genomic features and extracting a set of profiles is shown in Fig. 3.

3 Conclusions and Future work

Several challenges remain, regarding clinical and genomic data integration to facilitate clinical decision support. The opportunities of combining these two types of data are obvious, as they allow obtaining new insights concerning diagnosis, prognosis and treatment. A limitation of this combination is that although data exist, usually their enormous volume and their heterogeneity constitute their analysis and association a very difficult task. The lack of terminological and ontological compatibility, which could be solved by means of a uniform representation is another future challenge. Besides new data models, ontologies are/have to be developed in order to link genomic and clinical data and standards are required to ensure interoperability between disparate data sources.

Acknowledgments. This research is partly funded by the European Commission as part of the project MATCH (IST-2005-027266).

References

1. Campbell, A.M., Heyer, L.J.: *Discovering Genomics, Proteomics and Bioinformatics*. Benjamin Cummings, CA, USA (2007).
2. Goletsis, Y., Exarchos, T.P., Giannakeas, N., Tsipouras, M.G., Fotiadis, D.I.: 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).
3. Fotiadis, D.I., Goletsis, Y., Likas, A., & Papadopoulos, A.: *Clinical Decision Support Systems*. *Encyclopedia of Biomedical Engineering*, Wiley (2003).
4. Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P.: Data integration and genomic medicine. *Journal of Biomedical Informatics* 40 (2007) 5–16.
5. Sielinski, S. (2005). Similarity measures for clustering SNP and epidemiological data. Technical report of university of Dortmund.
6. Webb, A.: *Statistical Pattern Recognition*. Arnold, New York, USA (1999).
7. Tan, P.N., Steinbach, M., & Kumar, V.: *Introduction to Data Mining*, Addison Wesley. USA (2005).
8. Witten, I.H., and Frank, E.: *Data Mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann, CA, USA (2005).