Intelligent Patient Profiling for Diagnosis, Staging and Treatment Selection in Colon Cancer

Yorgos Goletsis, *Member, IEEE,* Themis P. Exarchos, *Student member, IEEE,* Nikolaos Giannakeas, *Student member, IEEE* and Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*—The selection of a personalized treatment plan for a patient with cancer can be of critical importance for his health or even survival. A Decision Support Platform that can associate the patient clinical situation with the patient DNA Single Nucleotide Polymorphisms (SNPs) can provide the oncologist with a better understanding of the personalized conditions of every single patient. In this paper we present the MATCH plarform 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. The core of the platform is based on clustering techniques which provide profiles of patients with similar clinical features and genetic predispositions to cancer. The patients which share the same profile should probably have similar treatment plan and follow up. Through the integration of the clinical and genetic data of a patient, real time conclusions can be drawn for his early diagnosis, staging and more effective colon cancer treatment. Intelligent components are designed and developed which identify single nucleotide polymorphisms (SNPs) from the gene sequences and combine them with the clinical situation of the patient. The produced clinico-genomic profiles are used as a decision support tool for newly sequenced patients.**

I. Introduction

The latest breakthroughs of the technology in the biomolecular sciences and applications have shifted the research application and need from the production of genetic or biological data to the efficient analysis of these data in a meaningful way for clinical decision support. New technologies allow for high volume affordable production and collection of information on biological sequences, gene expression levels and protein structure, on almost every aspect of the molecular architecture of living organisms. 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 [1].

Most of the already developed Clinical Decission Support Systems (CDSSs) are based solely on clinical data. Alternatively, a few methods that exist for cancer decision support employ only microarray gene expression data. The cancer specific CDSSs concern several different types of cancer and employ various techniques for their development. The majority of systems are still in a research level and only a few are being used in clinical practice. PAPNET [2] is a CDSS already in use, which deals with cervical cancer. PAPNET uses artificial neural networks (ANNs) to extract abnormal cell appearances from vaginal smear slides and describe them in histological terms. Other CDSSs for cervical cancer concentrate on the evaluation of the benefits of the PAPNET system [3,4]. Colon cancer has also been studied, using clinical data and fuzzy classification trees [5] or pattern analysis of gene expression levels [6]. A combination of imaging data with pathology data for colon cancer has also been proposed [7]. CDSSs proposed for prostate cancer, employ prostate specific antigen (PSA) serum marker, digital rectal examination, Gleason sum, age and race [8]. Another approach for decision support in prostate cancer is based on gene expression profiles [9]. Regarding bladder cancer, a CDSS has been developed based on proteomic data [10]. Concerning breast cancer, the potential of microarray data has been analysed [11]. Also, a recent approach has been developed that integrates data mining with clinical guidelines towards breast cancer decision support [12]. It should be noted that all CDSSs mentioned above are just research attempts and only PAPNET is in clinical use. Moreover, all these systems make use either of clinical data or biological data for patient characterization; according to our knowledge, there exists no system which integrates clinical with biological data for more accurate decision support.

Manuscript received July 5, 2008.

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).

Y. Goletsis is with the Department. of Economics, University of Ioannina, Ioannina, Greece. (e-mail: goletsis@cc.uoi.gr).

T. P. Exarchos is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, with the Dept. of Medical Physics, Medical School, University of Ioannina, Ioannina, Greece and with the Institute of Biomedical Technology, CERETETH, Larissa, Greece. (e-mail: exarchos@cc.uoi.gr).

N. Giannakeas is with the Laboratory of Biological Chemistry, Medical School, University of Ioannina, Ioannina, Greece and with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece (e-mail: me01310@cc.uoi.gr).

D.I. Fotiadis is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110 and with the Institute of Biomedical Technology, CERETETH, Larissa, Greece. (0030-26510-98803; fax: 0030-26510-97092; e-mail: fotiadis@cs.uoi.gr).

In this paper we present MATCH, a platform for intelligent patient profiling for the decision support of cancer treatment, by exploiting clinical and genomic data. The platform performs data integration between medicine and molecular biology, by appropriately combining clinical and genomic features in order to handle cancer diseases. The constitution of such a decision support platform 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. Additionally, special tools provide visual analysis of the genetic data, visual representation of statistical properties of the data as well as statistical analyses.

II. Materials And Methods

The overall idea of the decision support functionalities of MATCH, is shown in Fig. 1. The approach followed in MATCH 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 decision support 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 while the ones that share similar finding and diagnosis should be treated in the same way and so on. 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 decision support; 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 profiles. The matching profiles are examined in order to indicate potential diagnosis (the term diagnosis here refers mainly to the identification of cancer sub-type). In this sense, MATCH offers a more detailed and extensive staging procedure. If the diagnosis is confirmed, genetic screening may be proposed to the subject and then, the profiles are further examined, in order to make a decision regarding the preferred treatment and follow-up.

The architecture of the MATCH platform 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 platform's functionality are as follows:

A. 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 analysis and decision support components. The patient data component is responsible for the collection, integration, storage and presentation/output of the data.

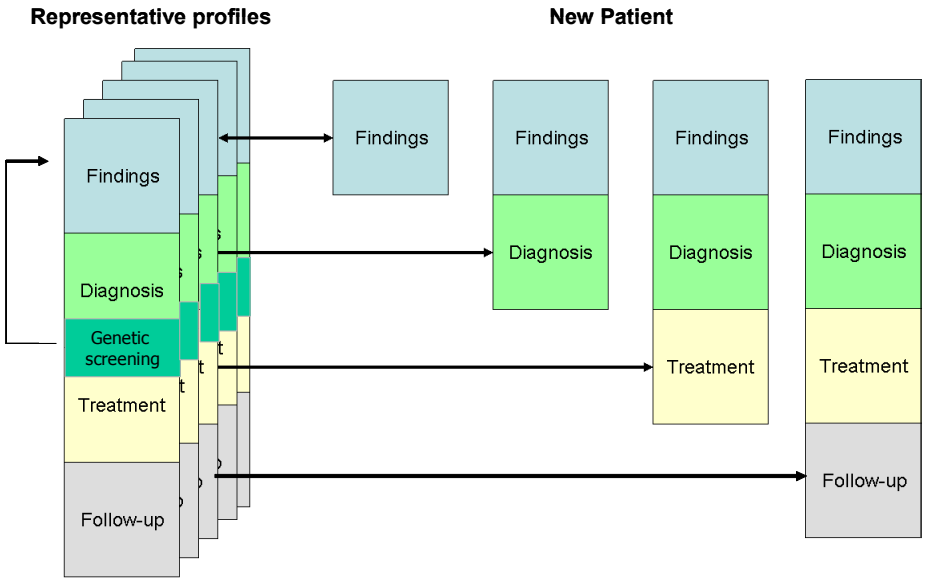

Figure 1. Decision support based on profiles extraction. Unknown features of patient are derived by known features of similar cases.

B. Pattern Analysis component.

The Pattern Analysis component uses the acquired data from the patient data component. Pattern analysis is based on the k-means algorithm, which classifies patients into clusters (profiles), 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 the city-block distance for the discrete data. A deficiency of the k-means algorithm is that the number of 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 [13]. More specifically, the available data are divided into ten folds. K-means clustering is performed in each of the folds iteratively, by increasing each time the value of k by 1. The sum of squared errors (SSE) is computed in every iteration and when SSE stops decreasing or it is stabilized, the corresponding value of k is chosen as the optimal one. Moreover, through the interfaces of the pattern analysis, the user has a variety of choices including the features to be used for profiling, subgrouping criteria and automated or manual input of the number of profiles to be extracted.

C. 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 a 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).

D. 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 platform. This is done by performing comparisons to the profiles obtained by the Pattern Analysis Component and computing the distance (normalized Euclidean distance) of new patients with all the profiles that were extracted from 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 communicated to the user interfaces, part of which uses visualization tools for the representation of molecular structures. The methodology for creating the profiles for decision support is shown schematically in fig. 3.

E. Other functionalities

Visualization: A visualization tool provides the users of the platform with visual functionalities over some parts of the patient clinico-genomic profile (geneSNPVista application, [14]). A Statistical Analysis Tool over the profile for specific patient profiles that are of specific interest to the users (doctors, bioinformaticians, biomedical researchers) is also part of the platform web based front-end.

The visualization platform contains the two aforementioned applications that cover the needs of the users

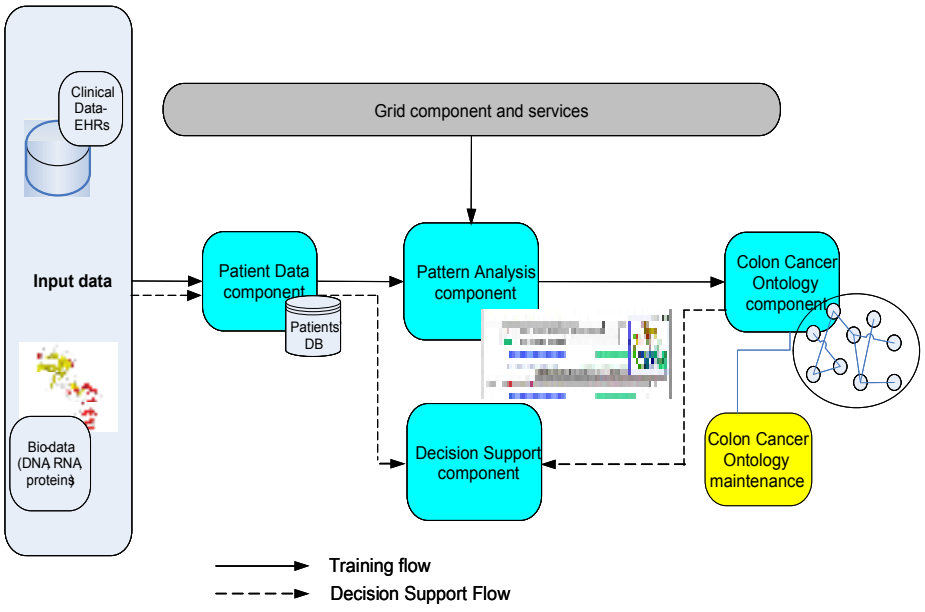

Figure 2: The architecture of MATCH platform.

for the creation of an environment where both visualization and statistical analysis can take place. Both applications were fully customized in order to meet the users' needs for visualization of exclusive genomic information.

SNPVista is an open-source, graphic tool that allows the user to visualize genomic information (SNPs and sequences), identify SNP positions inside the sequences of the genes (tp53, tp63, tp73 for the time being, others can be added later on) and associate this information with the overall process of MATCH prognosis, treatment and follow-up process.

MATCH statistical tool is a graphical tool which supports statistical analysis; it allows the user to visualize information extracted from the patient's dataset within statistical measure plots (histograms and scatter plots) over specific profiles (as these profiles are generated by the clustering process). It associates the information that is included in one or more profiles with the calculation of specific statistical measures over the profiles. It provides a user-friendly environment to visualize the requested information (per profile and per feature).

Proteomic Information Retrieval: The role of this tool is to search, collect and present information related to genetic information based on queries submitted by users. The tool can provide information either for specific genes or broad phenotypes.

The Proteomic Information Retrieval tool facilitates automated periodical updates in the background and downloads a number of documents. These documents are downloaded from the Genetic Association Database and contain information and bibliography about genes. These resources are used both as primary source to answer queries and as a map assisting in locating additional information in the web. Based on the submitted queries, the application accesses the following web based databases: OMIM, PUBMED, ENTREZ and ENSEMBLE. The data collected from these databases are bibliographical, gene related sequences (mRNA, Protein) and gene and broad phenotype specific information.

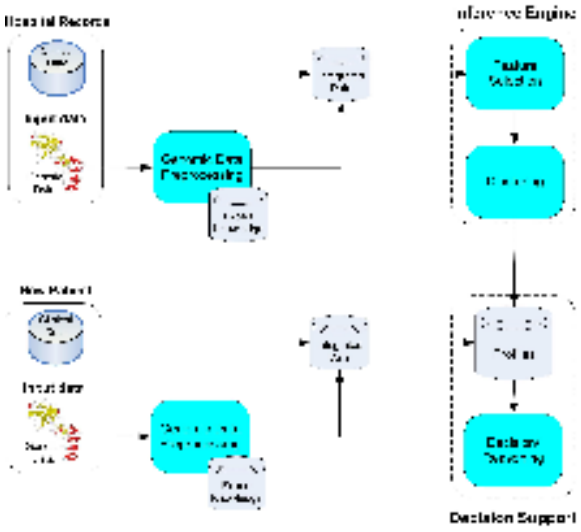

Figure 3: Creating profiles using clinical and genomic data using clustering techniques and application of the profiles for decision support.

III. Application

The first experimental setup of the MATCH platform focuses on 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 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 [15]: 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, currently the genes TP53, TP63 and TP73 are employed.

An efficient way to process the above genes is to detect Single Nucleotide Polymorphisms (SNPs) [16]. SNPs data are qualitative data providing information about the genomic at a specific locus of a gene. 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. Table 1 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 the 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.

The above genes 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

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

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

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.

For the creation of the database and the creation of the profiles 202 patients with diagnosed colon cancer were employed. Fig. 4 presents the four main steps for the decision support and reasoning for a new patient. Specifically, in Fig. 4(a) the medical experts select the features to be considered for the creation of the profiles. The features could be clinical and laboratory related as well as genetic (SNPs). Fig. 4(b) presents the similarity at the new patient to the created profiles. For each of the profiles the user can view the profile data (Fig. 4(c) – right) and compared them to the patient data (Fig. 4(c) – left). Statistical analysis of patients belonging to the same profile, shown in Fig. 4(d), further support the reasoning process.

IV. Discussion and Conclusions

The MATCH platform 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

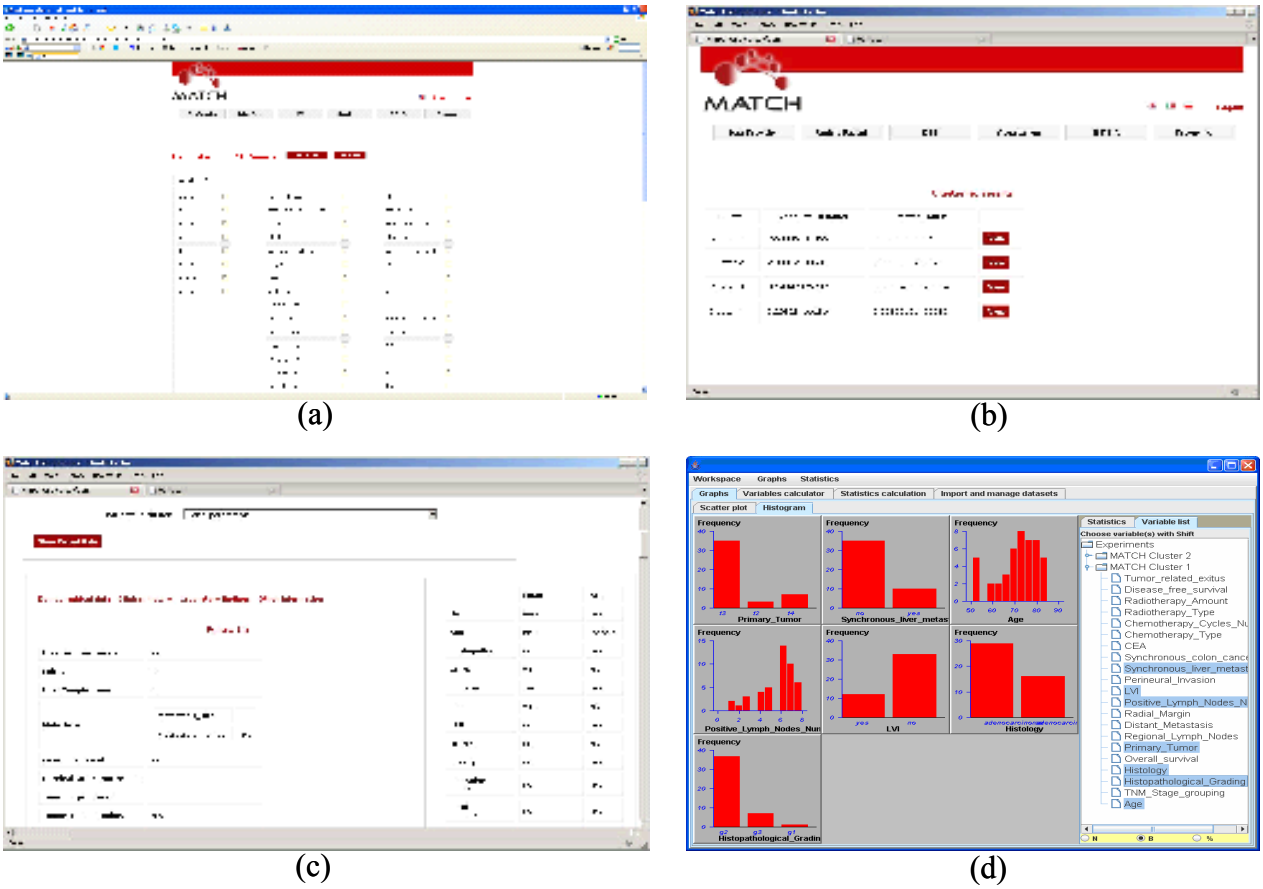

Figure 4. The main steps of MATCH for the decision support: (a) Feature selection (b) The similarity between a new patient and an extracted profile. (c) Profile data versus patient data and (d) Statistical analysis.

diagnosis process. In this way it provides an integrated platform 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, visualization tools for visualizing the SNPs in the DNA sequences are used. Currently, the decision support takes into consideration the SNPs existing in the genes TP53, TP63 and TP73. The decrease of the price of sequencing will have as a result the employment of additional genes for the more accurate characterization of the patients. Moreover, the use of mRNA or mass spectrometry data will be also considered.

REFERENCES

[1] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy and P. Tarczy-Hornoch, "Data integration and genomic medicine," *Journal of Biomedical Informatics*, vol. 40, pp. 5–16, 2007.

[2] M. E. Boon and L. P. Kok, "Using artificial neural networks to screen cervical smears: How new technology enhances health care," *Clinical applications of artificial neural networks*, Cambridge University Press, Cambridge, pp. 81–89, 2001.

[3] H. Doornewaard, Y. T.van der Schouw, Y. van der Graaf, A. B.Bos, J. D. Habbema, and J. G.van den Tweel, "The diagnostic value of computerassisted primary cervical smear screening: A longitudinal cohort study," *Modern Pathology*, vol. 12, no. 11, pp. 995-1000, 1999.

[4] P. Nieminen, M. Hakama, M. Viikki, J. Tarkkanen, and A. Anttila, "Prospective and randomised public-health trial on neural network-assisted screening for cervical cancer in Finland: Results of the first year," *International Journal of Cancer*, vol. 103, no. 3, pp. 422–426, 2003.

[5] I. J. Chiang, M. J. Shieh, J. Y. Hsu, and J. M. Wong, "Building a medical decision support system for Colon Polyp screening by using Fuzzy Classification Trees," *Applied Intelligence*, vol. 22, no. 1, pp. 61-75, 2005.

[6] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Cell Biology*, vol. 96, pp. 6745–6750, 1999.

[7] M. Slaymaker, M. Brady, F. Reddington, A. Simpson, D. Gavaghan, and P. Quirke, "A prototype infrastructure for the secure aggregation of imaging and pathology data for colorectal cancer care," *In the Proceeding of IEEE Computer Based Medical Systems*, USA, pp. 63-68, 2006.

[8] Remzi, M., Anagnostou, T., Ravery, V., Zlotta, A., Stephan, C., Marberger, M., & Djavan, B. (2003). An artificial neural network to predict the outcome of repeat prostate biopsies. Urology, 62(3), 456–460.

[9] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behaviour," *Cancer Cell*, vol. 1, no. 2, pp. 203-209, 2002.

[10] S. J.Parekattil, H. A. Fisher, and B. A. Kogan, "Neural network using combined urine nuclear matrix protein-22, monocyte chemoattractant protein-1 and urinary intercellular adhesion molecule-1 to detect bladder cancer," *The Journal of Urology,* vol. 169, no.3, pp. 917–920, 2003.

[11] L. J. Van 'T Veer, H. Dai, M. J. Van De Vijner, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Wintteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[12] M. T. Skevofilakas, K. S. Nikita, P. H. Templakesis, K. N. Birbas, I. G. Kaklamanos and G. N. Bonatsos, "A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines" *In the Proceedings of the 27th Annual Conference of IEEE Engineering in Medicine and Biology*, China, pp. 2429-2432, 2005.

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

[14] N. Shah, M.V. Teplitsky, S. Minovitsky, L.A. Pennacchio, P. Hugenholtz, B. Hamann, I. Dubchak, "SNP-VISTA: An interactive SNP visualization tool," *BMC Bioinformatics*, vol. 6, pp. 292, 2005.

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

[16] S. Sielinski, "Similarity measures for clustering SNP and epidemiological data" *Technical report of university of Dortmund*, 2005.