Integration of clinical and genomic data for decision support in cancer

Yorgos Goletsis[1,2], Themis P. Exarchos[1,3], Nikolaos Giannakeas[1,4], Markos G. Tsipouras[1], and Dimitrios I. Fotiadis[1,5,*]

[1]Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Greece
[2]Dept. of Economics, University of Ioannina, Greece
[3]Dept. of Medical Physics, Medical School, University of Ioannina, Greece
[4]Laboratory of Biological Chemistry, Medical School, University of Ioannina, Greece
[5]Biomedical Research Institute - FORTH, Ioannina, Greece

[*]Corresponding author
Unit of Medical Technology and Intelligent Information Systems,
Department of Computer Science, University of Ioannina,
PO Box 1186, GR 451 10 Ioannina, GREECE
tel.: +30 26510 98803, fax: +30 26510 98889
e-mail: fotiadis@cs.uoi.gr

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 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) 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 (Louie, 2007). 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.

In this chapter we address decision support for cancer, by exploiting clinical data and identifying mutations on tumour suppressor genes. 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, real time conclusions can be drawn for early diagnosis, staging and more effective cancer treatment.

2. Background

Clinical Decision Support Systems are active knowledge systems which use two or more items of patient data to generate case-specific advice (Fotiadis, 2006). 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 (Mitchel, 2006) and data mining (intelligent systems) (Tan, 2005). CDSSs require the input of patient-specific clinical variables (medical data) and as a result provide patient specific recommendation.

Medical data are observations regarding a patient, including demographic details (i.e. age, sex), medical history (i.e. diabetes, obesity), laboratory examinations (e.g. creatinine, triglyceride), biomedical signals (ECG, EMG), medical images (i.e. MRI, CT) etc. Demographic details, medical history and laboratory data are the most easily obtained and recorded and, therefore, most commonly included in electronic patient records. On the other hand, biomedical signals and medical images require more effort in order to be acquired in a digital format and must be processed for useful feature extraction. Apart from these, several types of genomic data can be generated from laboratory examinations, i.e. gene DNA or protein sequences, gene expression data, microarray images etc. Genomic data can also be used for medical diagnosis, disease prevention and population genetics studies. Although, medical data are sufficient for the diagnosis of several diseases, recent studies have demonstrated the high information value of genomic data, especially in specific types of diseases, such as cancer diseases.

The great amount and the complexity of the available genetic data, complicates their analysis from conventional data analysis methods and requires higher order analysis methods such as data mining techniques. Lately, data mining has received much attention in bioinformatics and molecular biology (Cook, 2001). Data mining methods are usually applied in the analysis of data coming from DNA microarrays or mass spectrometry. Over the last few years, several scientific reports have shown the potential of data mining to infer clinically relevant models from molecular data and thus provide clinical decision support. The majority of papers published in the area of data mining for genomic medicine deals with the analysis of gene expression data coming from DNA microarrays (Walker, 2004, Jiang, 2005, Shah, 2007) consisting of thousands of genes for each patient, with the aim to diagnose types of diseases and to obtain a prognosis which may lead to therapeutic decisions. Most of the research works are related to oncology (Louie, 2007), where there is a strong need for defining individualized therapeutic strategies. Another area where data mining has been applied is the analysis of genes or proteins, represented as sequences (Exarchos, 2006); sequential pattern mining techniques are suitable for analyzing these types of data (Zaki, 2000).

Several CDSSs for cancer have been proposed in the literature. Most of the approaches are based solely on clinical data and a few methods exist that provide cancer decision support using microarray gene expression data. The cancer 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. A CDSS which is already in clinical use is PAPNET (Boon, 2001) which deals with cervical cancer. PAPNET uses 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 (Doornewaard, 1999, Nieminen, 2003). Colon cancer has also been studied, using clinical data and fuzzy classification trees (Chiang, 2005) or pattern analysis of gene expression levels (Alon, 1999). A CDSS that combines imaging data with pathology data for colon cancer has also been proposed (Slaymaker, 2006). CDSSs proposed for prostate cancer, employ prostate specific antigen (PSA) serum marker, digital rectal examination, Gleason sum, age and race (Remzi, 2003). Another approach for decision support in prostate cancer is based on gene expression profiles (Singh, 2002). Regarding bladder cancer, a CDSS has been developed based on proteomic data (Parekattil, 2003). Concerning breast cancer, the potential of microarray data has been analysed (Van't Veer, 2002). Also, a recent CDSS has been developed that integrates data mining with clinical guidelines towards breast cancer decision support (Skevofilakas, 2005). It should be noted that all CDSSs mentioned above are just research attempts and only PAPNET is in clinical use.

3. Clinical Decision Support using Clinicogenomic Profiles

3.1 Methodology

Conventional approaches for CDSS focus on a single outcome regarding 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 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 Fig. 1:

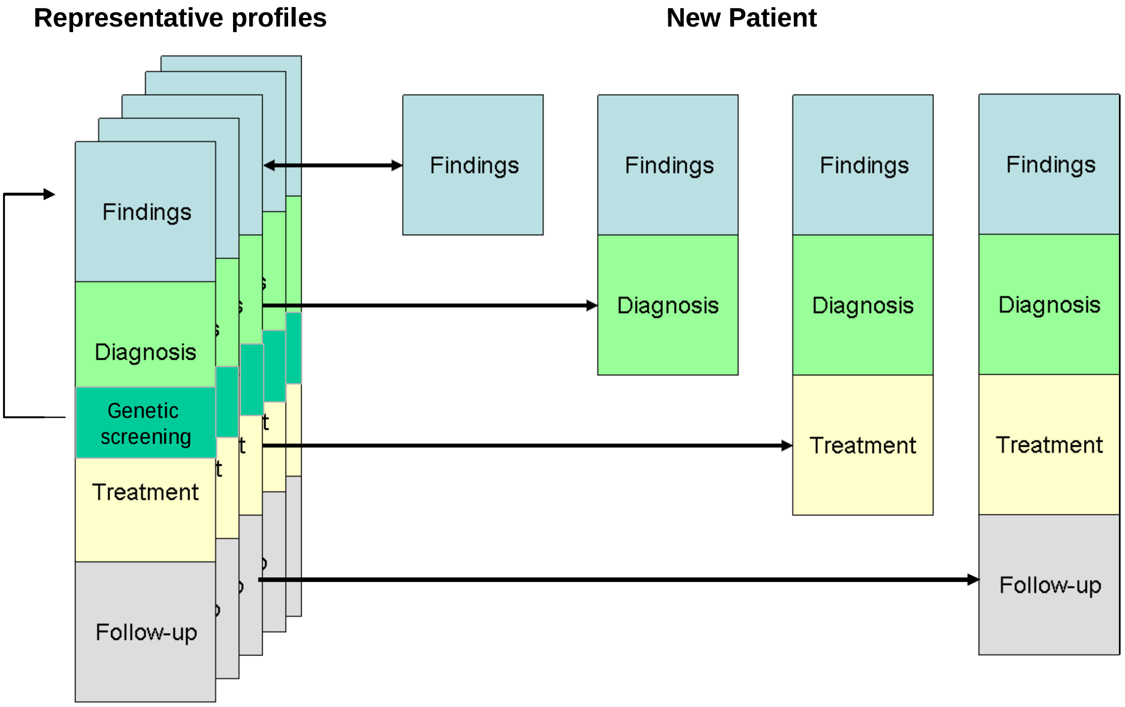

Fig. 1: Decision support based on profiles extraction. Unknown features of patient are derived by known features of similar cases.

3.2 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.

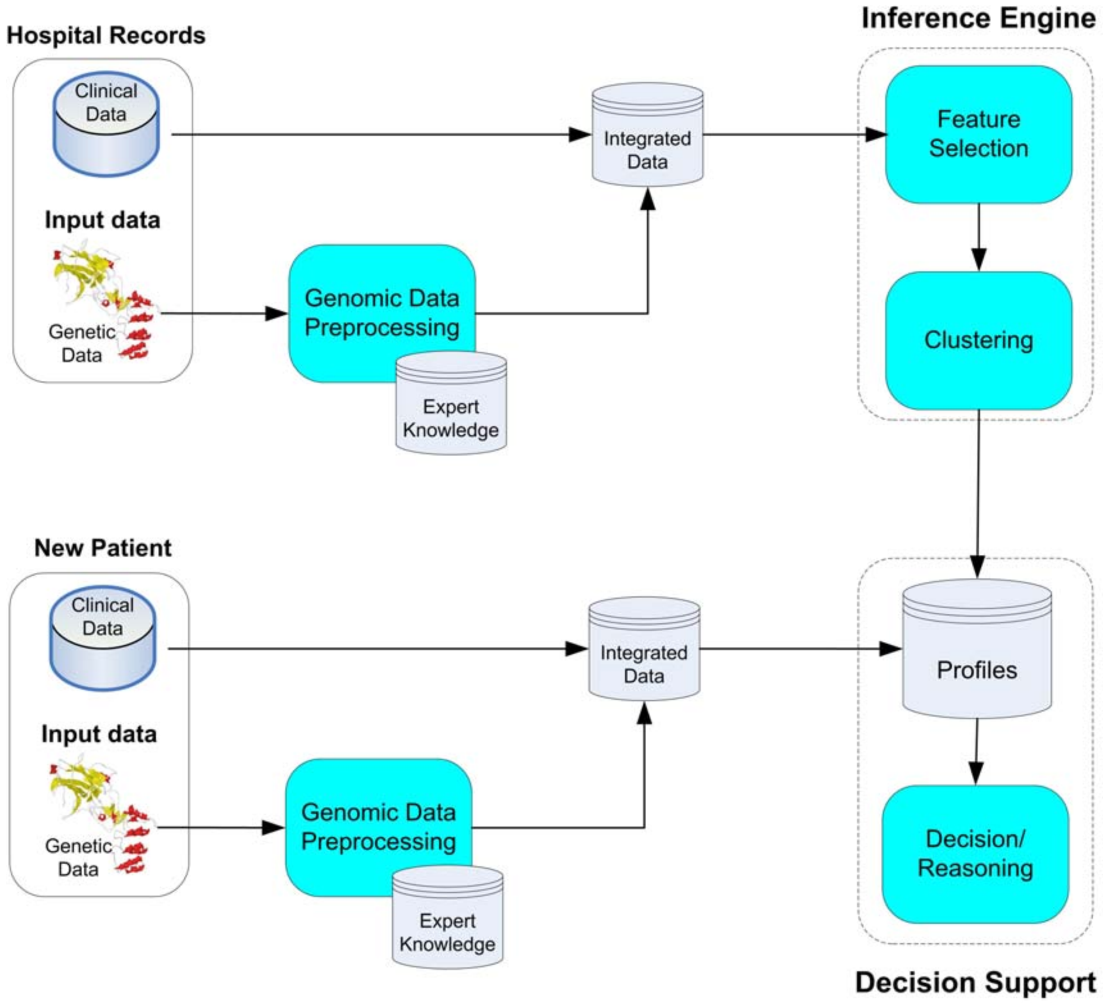

Fig 2. Representation of a general scheme for a CDSS, integrating clinical and genomic information.

3.2.1 Types of Data

Clinical data that are contained in electronic patient records (i.e. demographic details, medical history and laboratory data) are usually presented in a simple and structured format, thus simplifying the analysis. On the other hand, genomic data are not structured and, therefore, appropriate preprocessing is needed in order to transform them into a more structured format. Three different kinds of biological data may be available to clinicians: (i) genomic data, often represented by a collection of single nucleotide polymorphisms (SNPs), DNA sequence variations that occur when a single nucleotide in the genome sequence is altered; since each individual has many SNPs, their sequence forms a unique DNA pattern for that person; (ii) gene expression data, which can be measured with DNA microarrays to obtain a snapshot of the activity of all genes in one tissue at a given time or with techniques that rely on a polymerase chain reaction (PCR) and real-time PCR when the expression of only a few genes needs to be measured with better precision; (iii) protein expression data, which can include a complete set of protein profiles obtained with mass spectra technologies, or a few protein markers which can be measured with ad hoc essays.

3.2.2 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 between clinical and genomic data, differential diagnosis, treatment etc.

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. Finally, the reduced set of features is used by a clustering algorithm. K-means (MacQueen, 1967), fuzzy k-means (Bezdek, 1981) and expectation maximization (Dempster, 1977), are known approaches for clustering and can be involved for profile extraction. The profiles are the output of the clustering procedure (centroids). A deficiency of several clustering algorithms is that the number of centroids (profiles) must be predefined; this 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.

3.3 Application to colon cancer

Colon cancer includes cancerous growths in the colon, rectum and appendix. It is the third most common form 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 with colon cancer (Read, 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, 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). 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 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.

Table 1. Indicative SNPs transitions and positions in the TP53 gene, related with 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]

Some of the genes described above are acquired from the subjects and based on the SNP information regarding every acquired gene, such as SNPs in Table 1 for TP53 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 that are related with colon cancer are the input to the methodology and, after following the above described inference engine creation methodology, clinicogenomic profiles are generated. These profiles are able to provide advanced cancer decision support to new patients.

4. Future Trends

There should be no doubt that 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. According to this, medical informatics are combined with bioinformatics towards biomedical informatics. Biomedical Informatics is the emerging discipline that aims to put these two worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered. 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. Another challenge is the lack of terminological and ontological compatibility, which could be solved by means of a uniformed representation. Besides new data models, ontologies are/have to be developed in order to link genomic and clinical data. Furthermore, standards are required to ensure interoperability between disparate data sources.

5. 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 with genomic instability. The integration of clinical and genomic data makes the prospect for developing personalized healthcare, ever more realistic.

References

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

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, USA.

Boon, M.E., & Kok, L.P. (2001). Using artificial neural networks to screen cervical smears: How new technology enhances health care. Clinical applications of artificial neural networks, Cambridge University Press, Cambridge, 81–89.

Chan, I., Wells, W. 3rd, Mulkern, R.V., Haker, S., Zhang, J., Zou, K.H., Maier, S.E., & Tempany C.M.. (2003). Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2 weighted magnetic resonance imaging; a multichannel statistical classifier. Medical Physics, 30(9), 2390–2398.

Chiang, I.J., Shieh, M.J., Hsu, J.Y., & Wong, J.M. (2005). Building a medical decision support system for Colon Polyp screening by using Fuzzy Classification Trees. Applied Intelligence, 22(1), 61-75.

Chicurel, M. (2002). Bioinformatics: bringing it all together. Nature, 419, 751–757.

Cook, D.J., Lawrence, B.H., Su, S., Maglothin, R., & Jonyer, I. (2001). Structural Mining of Molecular Biology Data: A Graph-Based Tool for Discovering and Analyzing Biological Patterns in Structural Databases. IEEE Engineering in Medicine and Biology, 4, 67-74.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1), 1-38.

Doornewaard, H., van der Schouw, Y.T., van der Graaf, Y., Bos, A.B., Habbema, J.D., & van den Tweel, J.G. (1999). The diagnostic value of computerassisted primary cervical smear screening: A longitudinal cohort study. Modern Pathology, 12(11), 995-1000.

Exarchos, T.P., Papaloukas, C., Lampros, C., & Fotiadis, D.I. (2006). Protein Classification using Sequential Pattern Mining. In the proceedings of the 28th

Annual International Conference of the Engineering in Medicine and Biology Society, USA, 5814-5817.

Fotiadis, D.I., Goletsis, Y., Likas, A., & Papadopoulos, A. (2006). Clinical Decision Support Systems. Encyclopedia of Biomedical Engineering, Wiley.

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

Jiang, X.R., & Gruenwald, L. (2005). Microarray gene expression data association rules mining based on BSC-tree and FIS-tree. Data & Knowledge Engineering, 53(1), 3–29.

Enderle, J., Blanchard, S.M., & Bronzino, J. (2005). Introduction to Biomedical Engineering (2nd Edition). John Enderle Academic Press, USA

Kohane, I.S. (2000). Bioinformatics and clinical informatics: the imperative to collaborate. Journal of American Medical Informatics Association, 7, 512–516.

Kressner, U., Inganäs, M., Byding, S., Blikstad, I., Pahlman, L., Glimelius, B., & Lindmark, G. (1999). Prognostic value of p53 genetic changes in colorectal cancer. Journal Clinical Oncology, 17, 593-599.

Li, J., Wong, L., & Yang, Q. (2005). Data Mining in Bioinformatics

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.

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.

Mitchell, T. (2006). Machine Learning, Springer, McGraw-Hill Education (ISE Editions).

Nieminen, P., Hakama, M., Viikki, M., Tarkkanen, J., & Anttila, A. (2003). 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, 103(3), 422–426.

Parekattil, S.J., Fisher, H.A., & Kogan, B.A. (2003). 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, 169(3), 917–920.

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

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.

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

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., & Sellers, W.R. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2), 203-209.

Shah, S., & Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine, 37(2), 251-261.

Skevofilakas, M.T., Nikita, K.S., Templakesis, P.H., Birbas., K.N., Kaklamanos, I.G., & Bonatsos, G.N. (2005). 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, 2429-2432.

Slaymaker, M., Brady, M., Reddington, F., Simpson, A., Gavaghan, D., & Quirke, P. (2006). 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, 63-68.

Tan, P.N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining, Addison Wesley. USA.

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

Walker, P.R., Smith, B, Liu, Q.Y., Famili, A.F., Valdes, J.J., Liu, Z., & Lach, B. (2004). Data mining of gene expression changes in Alzheimer brain. Artificial Intelligence in Medicine, 31(2), 137-154.

Zaki, M.J. (2000). Sequence mining in categorical domains: Incorporating constraints. In the Proceedings of the 9th international conference on information and knowledge management, USA, 422–429.

Terms and Definitions

Clinical Decision Support Systems: they are entities that intend to support clinical personnel in medical decision-making tasks. In more technical terms, CDSSs are active knowledge systems that use two or more items of patient data to generate case-specific advice.

Data Integration: is the problem of combining data residing at different sources and providing the user with a unified view of these data. This important problem emerges in several scientific domains e.g. combining results from different bioinformatics repositories.

Data Mining: it is the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Cluster analysis: it is the task of decomposing or partitioning a dataset into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

Single Nucleotide Polymorphism (SNP): it is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species or between paired chromosomes in an individual.

Cancer Staging: knowledge of the extent of the cancer at the time of diagnosis. It is based on three components: the size or depth of penetration of the tumour (T), the involvement of lymph nodes (N), and the presence or absence of metastases (M).

Tumour Suppressor Gene: it is a gene that reduces the probability that a cell in a multicellular organism will turn into a tumor cell. A mutation or deletion of such a gene will increase the probability of the formation of a tumor.

Mutation: it is a change in the genetic material (usually DNA or RNA) of a living being. Mutations can happen for a lot of different reasons. They can happen because of errors during cell division, because of radiation, chemicals, etc.