Chapter 37
Enabling Heterogeneous Data Integration and Biomedical Event Prediction Through ICT: The Test Case of Cancer Reoccurrence

Marco Picone, Sebastian Steger, Konstantinos Exarchos, Marco De Fazio, Yorgos Goletsis, Dimitrios I. Fotiadis, Elena Martinelli, and Diego Ardigò

Abstract Early prediction of cancer reoccurrence constitutes a challenge for oncologists and surgeons. This chapter describes one ongoing experience, the EU-Project NeoMark, where scientists from different medical and biology research fields joined efforts with Information Technology experts to identify methods and algorithms that are able to early predict the reoccurrence risk for one of the most devastating tumors, the oral cavity squamous cell carcinoma (OSCC). The challenge of NeoMark is to develop algorithms able to identify a "signature" or bio-profile of the disease, by integrating multiscale and multivariate data from medical images, genomic profile from tissue and circulating cells RNA, and other medical parameters collected from patients before and after treatment. A limited number of relevant biomarkers will be identified and used in a real-time PCR device for early detection of disease reoccurrence.

Keywords Cancer informatics · Experimental medicine and analysis tools · Image processing in medicine and biological sciences · Biological data mining and knowledge discovery · Biological data integration and visualization

1 Introduction

Malignant neoplasms are – as a whole – the second cause of death in the western world. The main mechanism causing progressive inability and death in cancer patients is represented by the occurrence of loco-regional relapses and distant metastases whose incidence remains still high despite the recent development of several effective treatments. Reoccurrence is mainly due to the persistence of tumor cells after treatment without any clinical, laboratory, and imaging evidence of residual disease. The identification of new, reliable biomarkers of disease, discriminating

M. Picone (✉)
MultiMed s.r.l, Cremona, Italy
e-mail: marco.picone@multi-med.it

H.R. Arabnia and Q.-N. Tran (eds.), *Software Tools and Algorithms for Biological Systems*, Advances in Experimental Medicine and Biology 696,
DOI 10.1007/978-1-4419-7046-6_37, © Springer Science+Business Media, LLC 2011

which patients are at highest risk of relapses, is therefore of primary importance in cancer research to allow focus follow-up efforts and limit adjuvant chemotherapy only to high-risk patients. In addition to a better risk prediction, the identification of biomarkers for early diagnosis of relapses would have the potentiality to improve patient's survival. Oral squamous cell carcinoma (OSCC) represents about 5% of all cancers and provides a prototypical example of this issue. OSCC has a reoccurrence rate of about 25–50% over a period of 5 years, leading to severe consequences on physical appearance and ability to eat and speak, invasive and disabling surgical interventions, and death [1]. A strict follow-up is usually undertaken, and additional treatments are often planned to reduce the risk of reoccurrence even in the presence of disease remission (an approach called "adjuvant"). Adjuvant chemo- and radiotherapy treatments have important side effects, and a large group of patients would probably not require them to decrease the risk of reoccurrence. However, it is currently almost impossible to identify the high-risk subjects to be candidated to an aggressive treatment.

In this chapter, we describe the prototypical test case of the NeoMark project aiming to provide IT support to candidate biomarker identification from clinical, imaging, and molecular biology data. In NeoMark, we investigate an innovative strategy to identify relevant biomarkers of cancer reoccurrence risk and presence, integrating high-throughput gene expression analysis in tumor and blood cells, and IT-assisted imaging with traditional staging and follow-up protocols, to improve the stratification of reoccurrence risk and the earlier identification of loco-regional relapses. The idea behind NeoMark is that by analyzing a sufficient set of different types of data (clinical, biomedical, genomic, histological, from digital imaging, from surgery evidence, etc.) of patients affected by OSCC before treatment and at the time of remission, a set of relevant biomarkers appearing only in the presence of the disease might be identified. The reoccurrence of the same biomarker phenotype during post-remission follow-up may precede the clinical manifestation of the relapse thus allowing earlier intervention.

To pursue this ambitious aim, we need to collect, store, integrate, and analyze heterogeneous clinical and biomedical data, merge health care and data mining, and introduce molecular biology into clinical practice. The NeoMark integrated platform addresses these issues and enables the medical doctor jointly analyze, revise, and exploit traditional clinical data in conjunction with image analysis tools and gene expression data in better supported clinical decisions during the follow-up of OSCC patients in remission phase. The work presented in this chapter is co-funded by the European Union under the 7th Framework Program, Information and Communication Technologies (EU-FP7-ICT-2007-2-22483-NeoMark).

2 System Description

The versatile user requirements and especially the integration of heterogeneous input data required a careful design of the NeoMark system. Our goal was to integrate as much functionality as possible in a single unified service-oriented

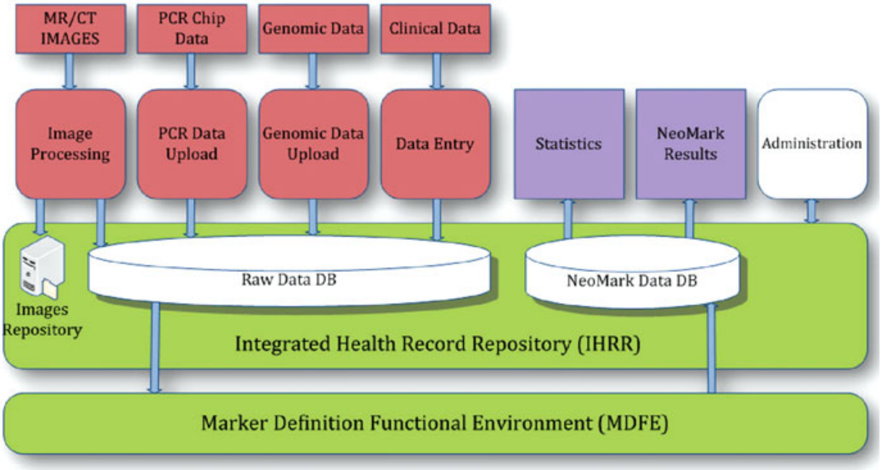

Fig. 1 NeoMark system overview

system, achieving great flexibility and usability. These properties increase the user acceptance and may decrease human error. The basic scheme of the implemented system can be seen in Fig. 1. Most of the user interaction is done via the web interface. The physician can manage patients, enter clinical data, and view all features and the NeoMark results. The clinician can upload genomic data and researchers can view anonymous statistics, which could serve as a base for future research on oral cancer. Furthermore, entering imaging data and uploading PCR chip data can be performed using standalone tools (see Sects. 2.1 and 2.3 for details). Heterogeneous NeoMark data (general information about the patient, Clinical, filtered, and cleaned Genomic and Imaging data as well as OSCC reoccurrence prediction, patient-specific, and disease-specific risk factors) are stored in a single database – the Integrated Health Record Repository (IHRR) – on a central NeoMark server. The participating hospitals are connected to the very same NeoMark server, allowing the data-driven training algorithm to incorporate patient data from all participating hospitals at the same time and enabling users to perform inter-hospital comparisons of patient data.

The NeoMark server contains an Apache-based web server, a MySQL-based database, an FTP server for the storage of medical images, and the *marker definition functional environment (MDFE)*, a data analysis module which is the core of the system. Based on the heterogeneous input data, this module estimates the likelihood of a relapse and identifies OSCC risk factors (for details, see Sect. 3).

In order to protect the patient's right of privacy, none of the data (e.g., the name) that allows to uniquely identify the patient is stored on the central NeoMark server. Only a unique NeoMark ID can identify a patient. Those IDs are connected to the patients via individual databases that are located within each hospital. A JAVA-based tool – *the sensitive data tool* – that can be started directly from the web interface provides a user interface which allows to manage that kind of patient data.

2.1 Image Processing and Image Feature Extraction

Medical images cannot directly be analyzed by the NeoMark MDFE because the amount of information is too high while not being meaningful due to the lack of semantic enrichment. An experienced radiologist uses the *image processing tool* to extract meaningful numeric features of tumors and suspicious lymph nodes from CT, CT with contrast, MR T1 TSE, and MR T2 TSE images.

All images are acquired before treatment and then every 6 month during follow-up. The high resolution (1 mm slice thickness) CT images cover the entire head and neck region, whereas the MR images only cover the tumors and significant lymph nodes.

The further processing steps [2] consist of image registration, tumor/lymph node segmentation, and feature extraction (FE).

In order to extract features from several images jointly, they have to be transformed to a common coordinate system. This process, called image registration, has to be performed on each of the images. Due to the rigidity of the human skull, it is sufficient to only allow translations and rotations (rigid registration). However, to achieve good registration results having only those six degrees of freedom, the position of the head relative to the spine has to be as similar as possible in all images. A coarse prealignment is performed manually by the radiologist. Then the registration is automatically refined by maximizing a suitable similarity measurement. For images from different modalitites (MR/CT), mutual information [3,4] has proven to achieve good results.

In the next step, the regions of interest (ROIs, here tumors and suspicious lymph nodes) need to be segmented. (Semi-) Automated lymph node segmentation has been studied exhaustively [5–8]. The image processing tool uses a mass spring model [9] to segment lymph nodes. This approach balances internal forces (i.e., forces that preserve the shape) and external forces (i.e., forces that push the model toward the lymph node borders) that are applied to an initial spheric lymph node model that has been placed in the center of a lymph node by the user.

Tumors in the oral cavity, however, cannot be segmented automatically with sufficient accuracy because of the high variance in shape and appearance and their similarity to surrounding tissue, which requires the knowledge of an experienced radiologist. Therefore, a user-friendly graphical user interface allows the radiologist to delineate the tumor in each image slice.

Once the boundaries of the ROIs are known, geometric and texture-based features can be extracted. A selection of them is:

- *The volume*. It can be easily computed by simply counting all voxels in the ROI and multiplying that by the cuboid volume of one voxel.
- *The axes*. Assuming that the ROI is convex, the axes of ROI coincide with axes of the enclosing minimal volume bounding box. An approximation of that box can be computed by performing a main component analysis on the points of the surface of the ROI [10]. The directions of the axes are then given by the eigenvectors. This approach for computing the axes is rotation invariant and not affected

by subjectivity in contrast to current clinical practice, where the radiologist delineates the axes manually in the axial plane only.

- *The contrast take up rate*. This texture-based feature is computed by measuring the mean squared error of the CT and the CT with contrast image. This feature, as well as the water content, is extremely sensitive to accurate registration because it directly compares voxel values in different images.
- *The water content*. Water appears bright in T2 MR images and dark in T1 MR images. Therefore, a measure of the water content is the difference between T2 voxels and the average T2 voxel multiplied with the difference between the average T1 voxel and the T1 voxels.

In addition to the automated extracted features, the radiologist enters further properties like the location and the amount of infiltration of the surrounding tissue. This enables a semiautomatic staging [11] of the tumor.

2.2 Genomic Data Cleaning and Filtering

Gene expression data are usually coming from FE files. A FE file is usually a tab-delimited text file that contains all the data extracted from the experiment by the FE software. The file contains the Log2-ratio value as well as (very) raw intensity data, background information, meta-data on the experiment and on the scanning settings, annotation data to identify genes, etc. In order to integrate into a study database, the expression values of each sample should be extracted from FE file and uploaded into the data matrix after being assigned to the correct patient or sample. The relevant information that is stored in the database are Feature Name, Probe Name Gene Name, Systematic Name, Description, and Log2-ratio. For this reason, we have a specific tool that analyzes *control features*, *duplicate features*, *filtering of genes based on low data quality*, and *filtering of genes with high number of missing values* and generates as output a cleaned file with a small dimension that contains only relevant information. This cleaned file can be uploaded from a specific page of NeoMark Web Application into the database. For the genomic data, there are two kinds of tables. The first one contains the cleaned relevant information with the Log2-ratio value and in addition also the normalized one calculated with the information that is already stored in the database. The other table is designed to store genomic data expression following the group ontology.

2.3 PCR Tool

A novel qRT-PCR platform is under development in STMicroelectronics to obtain quantitative information about the PCR amplification of the targeted genes. It is a portable, real-time, integrated analytical system based on qRT-PCR performed in an array of silicon microchambers. The small size of the components and its low power requirements make this system an ideal candidate for further miniaturization into a hand-held, point-of-care device. The qRT-PCR lab-on-chip is disposable and relatively inexpensive to make this method of analysis economically viable. The excellent thermal conductivity of silicon makes it ideal in applications requiring rapid cycles of heating and cooling. The silicon core of the qRT-PCR microchip is fabricated using photolithographic techniques: heaters are fabricated directly onto the surface of the chip, along with the thermal sensors monitoring the temperature and providing feedback to the temperature controller, while cooling is achieved via forced air using a fan. Designs for the qRT-PCR chips range from a single reaction chamber to arrays of microchambers of varying sizes and depths for multiple simultaneous reactions. The PCR reaction dynamics can be monitored locally in real-time, using a dedicated optical system recording the fluorescence intensity at each thermal cycle. The real-time PCR portable device analyzes a set of predefined genes (up to 20) and reports their expression value in relation with a housekeeping gene. In the long-term approach, this system could be the low cost alternative to the microarray devices. The purpose of the system is to speed up and lower the cost for the gene expression analysis for a particular set of genes of particular relevance in the oncology of oral cancer. During a first evaluation phase, the returned values of gene expression are correlated to data reported from the microarray tool. Results must be comparable in terms of returned value for gene expression, which is usually correlated to a reference gene of known expression. In the final release, however, the gene expression values could be directly sent to the genomic data repository system of the Neomark Database. The upload is integrated in the instruments' GUI, and reports to the Neomark Database. Any cleaning and filtering of RAW data are done in the system GUI, and uploaded data will be clean and with only significant values. Up to now significant results have been achieved in successfully amplifying both DNA and RNA, with quantitative results comparable with the one from the same sample amplified on a standard lab thermal cycler.

3 Data Analysis

The analysis of the heterogeneous data constitutes the cornerstone of the NeoMark artificial intelligence component. The aim of this component is twofold: (1) to assess the risk of reoccurrence in the very early stages of treatment, i.e., as soon as the patient reaches remission, and (2) to efficiently and effectively model the disease evolution during the whole follow-up period based on a multitude of heterogeneous data, thus monitoring the patient's therapeutic progression. As described in the clinical scenario of the NeoMark project, for each patient who has been diagnosed with oral cancer a wide range of heterogeneous data is collected and analyzed. Specifically, due to the complex nature of the disease, a holistic approach is performed which integrates a great multitude of clinical, imaging, and genomic data to "frame" every possible aspect related to the onset and progression of oral cancer.

3.1 Early Risk Assessment

For this purpose, an initial “snapshot” of all the patient’s attributes is acquired which consists of clinical, imaging, and gene expression data. Later, certain basic preprocessing steps (i.e., outlier detection and missing values handling) are used which aim to ameliorate the quality of the available input set. Next, the wrapper feature selection algorithm [12] is used to discard redundant and non-informative features and maintain the most discriminatory ones. The reduced set of attributes, maintained after the feature selection, is then provided as input to a classification algorithm; in the final integrated NeoMark server, the user may choose among a selection of algorithms which have been trained carefully and yield the best results [13]. The objective of this classification is to stratify the patients into two classes, i.e., remittents and non-remittents, based only on their initial clinical profile.

3.2 Disease Evolution Monitoring

In the present study, we use Dynamic Bayesian Networks to early identify potential relapses of the disease, during the follow-up. As described in the clinical scenario, a snapshot of the patient’s medical condition is acquired during every predefined follow-up with the doctor. By exploiting the information of history snapshots, we aim to model the progression of the disease in the future. The proposed prognostic model is based on DBNs, which are temporal extensions of Bayesian Networks (BNs) [14]. A BN can be described as where is a directed acyclic graph, where the nodes correspond to a set of random variables $X = \{x_1, x_2, \ldots, x_N\}$, and P is a joint probability distribution of variables in X, which factorizes as:

$$P(X) = \prod_{i=1}^{N} P[x_i | \pi_G(x_i)] \tag{1}$$

where $\pi_G(x)$ denotes the parents in G. A DBN can be defined as a pair $DB = (B_0, B_{\text{trans}})$ where B_0 is a BN, defining the prior $P(X_0)$, and B_{trans} is a two-slice temporal BN (2TBN) which defines $P(X_t|X_{t-1})$. The semantics of a DBN can be defined by “unrolling” the 2TBN until we have T time-slices. The resulting joint distribution is given by:

$$P(X_1, X_2, \ldots, X_T) = \prod_{i=1}^{T} \prod_{i=1}^{N} P[x_i^t | \pi_G(x_i^t)] \tag{2}$$

In order to build a model that successfully evaluates the current state or predicts a state in the future (next time slice), we need to finetune both the intra- and the inter-slice dependencies of the DBN network, using both expert knowledge as a prior model and experimental data to get a more accurate posterior model. After the training procedure, we are able to conjecture about the probability of any variable for every time slice, including of course the probability for reoccurrence.

4 Conclusion

We have presented a novel ICT-enabled cancer reoccurrence prediction method and have described the system implementing this idea. In addition to the great innovation of collecting and jointly interpreting such an enormous amount of heterogeneous data, the development of the NeoMark system led to further innovations:

- The data analysis component predicts not only the probability of a relapse overall but also the probability at a given time. All predictions are updated upon retrieval of follow-up input data.
- For the first time genomic data obtained from a PCR chip will eventually replace the expensive and complex laboratory-based genomic data extraction.
- The innovative semiautomatic multimodal image FE algorithms extract imaging features of tumors and lymph nodes that are well suited for further processing by the data analysis component due to their numeric manner and robustness.

Currently the system is in a late stage of development, and the first patient's data are about to be entered. Once enough NeoMark data are collected, the system can be trained and first results can be obtained and evaluated. We believe not only to be able to predict the reoccurrence of OSCC but also to obtain yet unknown risk factors of the disease.

In any case, the system will provide a large database customized to oral cancer which enables to perform a variety of clinical studies. Furthermore, the system may support the radiologist in diagnosing by providing a semantically enriched image database.

References

1. B. Boyle, P. Levin, editor. *World Cancer Report*. International Agency for Research on Cancer, 2008.
2. S. Steger, M. Erdt, G. Chiari, and G. Sakas. Feature extraction from medical images for an oral cancer reoccurrence prediction environment. In *World Congress on Medical Physics and Biomedical Engineering*, 2009.
3. F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging*, 16(2):187–198, 1997.
4. W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Med Image Anal*, 1(1):35–51, Mar 1996.
5. D. M. Honea, G. Yaorong, W. E. Snyder, P. F. Hemler, and D. J. Vining. Lymph node segmentation using active contours. volume 3034, pages 265–273. SPIE, 1997.

6. D. Maleike, M. Fabel, R. Tetzlaff, H. von Tengg-Kobligk, T. Heimann, H-P. Meinzer, and I. Wolf. Lymph node segmentation on CT images by a shape model guided deformable surface method. volume 6914, page 69141S. SPIE, 2008.
7. J. Rogowska, K. Batchelder, G. S. Gazelle, E. F. Halpern, W. Connor, and G. L. Wolf. Evaluation of selected two-dimensional segmentation techniques for computed tomography quantitation of lymph nodes. *Invest Radiol*, 31(3):138–145, Mar 1996.
8. G. Unal, G. Slabaugh, A. Ess, A. Yezzi, T. Fang, J. Tyan, M. Requardt, R. Krieg, R. Seethamraju, M. Harisinghani, and R. Weissleder. Semi-automatic lymph node segmentation in ln-mri. In *Proc. IEEE International Conference on Image Processing*, pages 77–80, 8–11 Oct. 2006.
9. J. Dornheim, H. Seim, B. Preim, I. Hertel, and G. Strauss. Segmentation of neck lymph nodes in CT datasets with stable 3d mass-spring models segmentation of neck lymph nodes. *Acad Radiol*, 14(11):1389–1399, Nov 2007.
10. G. Barequet and S. Har-peled. Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *J Algorithms*, 38:82–91.
11. R. V. P. Hutter, M. Klimpfinger, L. H. Sobin, C. Wittekind, F. L. Greene, editor. *TNM Atlas*. Springer, Berlin, 5th edition, 2007.
12. G. H. John and R. Kohavi. Wrappers for feature subset selection. *Artif Intell*, 97:273–324, 1997.
13. V. Kumar, P.-N. Tan, M. Steinbach. Introduction to data mining. Pearson Addison Wesley, Boston, 1st edition, 2006.
14. K. P. Murphy. Dynamic bayesian netoworks: Representation, inference and learning. University of California, 2002.