Genomics and Metabolomics Research for Brain Tumour Diagnosis Based on Machine Learning

Juan M. García–Gómez[1], Salvador Tortajada[1], Javier Vicente[1], Carlos Sáez, Xavier Castells[2], Jan Luts[3], Margarida Julià–Sapé[2], Alfons Juan–Císcar[4], Sabine Van Huffel[3], Anna Barceló[2], Joaquín Ariño[2], Carles Arús[2], and Montserrat Robles[1]

[1] ITACA-IBIME, Universidad Politécnica de Valencia, Spain
[2] Universitat Autònoma de Barcelona, Spain
[3] Katholieke Universiteit Leuven, Dept. of Electrical Engineering, ESAT-SCD(SISTA)
[4] DSIC, Universidad Politécnica de Valencia, Spain

Abstract. The incorporation of new biomedical technologies in the diagnosis and prognosis of cancer is changing medicine to an evidence-based diagnosis. We summarize some studies related to brain tumour research in Europe, based on the metabolic information provided by in vivo Magnetic Resonance Spectroscopy (MRS) and transcriptomic profiling observed by DNA microarrays. The first result presents the improvement in brain tumour diagnosis by combining Long TE and Short TE single voxel MR Spectra. Afterwards, a mixture model for binned and truncated data to characterize and classify MRS is reviewed. The classification of Glioblastomas Multiforme and Meningothelial Meningiomas using single-labeling cDNA-based microarrays was studied as proof of principle in the incorporation of genomic information to clinical diagnosis. Finally, we present a Decision Support System for in-vivo classification of brain tumours were the best inferred classifiers are deployed for their clinical use.

1 Introduction

New biomedical technologies may allow the interpretation of the origin of the illnesses, moving from the diagnosis and treatment of the patients to the evidence-based medicine paradigm. In brain tumour research, biomedical data coming from different biological levels are analysed under a systemic paradigm to understand the behavior, origin and discrimination of the types of tumours. The increase of the available data and their complexity make convenient to use automatic techniques based on Information and Communication Technology (ICT) to assist clinicians during the decision making for the diagnosis and treatment of new patients.

Nowadays, one of the most important goals of applied Machine Learning (ML) research in brain tumours is to create an automatic Decision Support System (DSS) for classify brain tumours using in-vivo data. A DSS would facilitate evidence-based clinical decision-making using, for example, ^{1}H Magnetic Resonance Spectroscopy (^{1}H MRS) and Imaging (MRI) data. This new paradigm would also include new criteria such as genetic-based tumour classification.

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 1012–1019, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Currently, the integration of heterogenous sources of biomedical data for cancer research is the main challenge in the analyze of the behavior, the origin, and the discrimination of the tumour types. ^{1}H MRS, MRI, in-vitro Spectroscopy (HR-MAS), DNA microarrays, histopathological analysis, and clinical management of the patients can provide complementary points of view in this research.

Therefore, clinicians, histopathologists, epidemiologists and data-mining groups join efforts in the acquisition and analysis of biomedical data in major health problems such like brain tumors. The quick and easy access to the most recent ICT developments will provide translational results for clinical application.

The contributions of ML to brain tumour research are focused on two practical outputs. The first one is to provide engines which can offer an objective solution to specific tumour discrimination problems. These engines (i.e. classifiers) are linked to a DSS in order to provide an easy and user-friendly access for clinicians. The second purpose is to obtain knowledge from brain tumours, providing new information about the origin of the illness and/or providing new taxonomies. The ML discipline could contribute in the generation of profiles or relations among biomedical data or among different brain tumour classes.

^{1}H MRS is slowly becoming an additional accurate non-invasive technique for initial examination of brain masses [1]. This is due to its capability to provide useful chemical information of different metabolites for characterizing brain tumors and its complementary role to MRI [2]. The complexity of the ^{1}H MRS signals and their underlying mixture of biological compounds, the use of Pattern Recognition (PR) methods, in combination with Biomedical Signal Processing, is suitable to obtain clinically usable knowledge from the acquired biomedical data. The predictive tools should be able to solve routine questions quickly to reinforce other evidences. On the other hand, when uncommon cases are analyzed, radiologists may use the DSS as a research support tool for comparing and analyzing the results and relevant features provided by the models of bilateral discrimination questions already solved on available data [3].

The improvement of diagnosis and prognosis of tumours may come from the better understanding of their origin. Hence, genes involved in the activated pathways of the tumoral tissues and the identification of molecular subtypes within the currently defined types of tumours may incorporate the molecular diagnosis in the clinical environment. For this objective, bioinformatic methods should be tuned to analyze the genomic data coming from the patients; i.e. microarray data should be corrected for artificial artifacts to measure the expression of the transcribed genes in a tissue.

The advances in automatic analysis of brain tumours are directly applicable in clinical environment, where the experts are interested in the latest improvements in the analysis of both routine and special cases. This can be provided by a DSS [4].

During the development of the DSS for brain tumours, many approaches are applied to fit the ML-based decision engines. Next sections summarize four different methodologies related to the brain tumour research using ML techniques. First, we present the study of feature selection and comparison of linear and non-linear models when combining Long TE and Short TE Single Voxel (SV) MR Spectra in the brain tumour diagnosis. Afterwards, we summarize the development made in [5] to obtain a useful representation of the metabolic information included in the ^{1}H MRS by a mixture for

binned and truncated data. In the study of the molecular biology of the brain tumours, we were inerested on developing a proof of principle for the tumour diagnosis in clinical environments based on DNA microarrays. Therefore, the classification of Glioblastomas Multiforme and Meningothelial Meningiomas by single-labeling cDNA-based microarrays was successfully obtained by the application of bioinformatics and ML methods. Finally, we summarize the last designs of a DSS for brain tumours that provides the access to the in vivo inferred classifiers in a distributed architecture around hospitals.

2 Combination of Long TE and Short TE SV MR Spectra for Brain Tumour Diagnosis

Different protocols are typically applied in the acquisition of ^{1}H MRS for neuroradiological classification of brain tumours. Short TE (20-35 ms) ^{1}H MRS data allows to observe several metabolites which are considered useful for tumour classification. On the other hand, Long TE (135-136 ms) ^{1}H MRS signals are less informative than Short TE signals, but it is easier to extract relevant information for classification from the former ones [1].

In order to statistically characterize the improvement of the automatic classification of brain tumours when two times of echo are used, three different predictive models were prepared. The first one was fitted using features from Long TE spectra, the second using Short TE spectra and the third using both Long TE and Short TE spectra. Hence, we analyzed the improvement of the classification using features from both spectra (Long TE + Short TE) compared to the classification using either Long TE or Short TE spectra.

Majos in [6] showed that when both STE and LTE are used, clinical classification accuracy in brain tumour diagnosis improves. Tortajada et al [7] showed an improvement of the descriptive discrimination of Aggressive (AGG), Low Grade Glial (LGG) and Meningioma (MEN) tumors [8].

185 brain tumour cases grouped in three super-classes of diagnosis: aggressive tumours -glioblastomas and metastasis- (AGG), meningiomas (MEN) and low-glial mixture (LGG) of Astrocytomas grade II, Oligodendrogliomas and Oligoastrocytomas were used to generate predictive models based on pattern recognition techniques. Spectra were provided by six international centers in the framework of the INTERPRET [9] project, acquired at 1.5T according to consensus protocols [8]; LTE: PRESS (1598-2020ms/135-136ms/1000, 2500Hz/512, 2048) (TR/TE/SW/points), STE: PRESS or STEAM (1600-2020ms/20-32ms/1000, 2500Hz/512, 2048). The signal processing was performed automatically following the protocols defined in [10] and the analyzed spectral range was [4.1, 0.5] ppm.

We performed pairwise classification among AGG, MEN and LGG super-classes, handling the points in the spectral range of interest in STE and LTE as a multivariate space like in [3]. In the combined approach, variables from both spectra were included in the models. Models generated by each approach were evaluated by an external 150-Random Sampling Training Test in order to avoid bias in the estimation of the error during the model selection. Relief-F feature selection was applied to retrieve interesting

features from the original spectra. Least-Squares Support Vector Machines (LS-SVMs) were the classification methods because of their regularization property, which is useful when the number of features used in the classifier increases.

We found significant differences in classifier performance using the different subsets (LTE, STE or both LTE and STE). For discriminating between MEN and the other two super-classes, both STE and LTE are needed to obtain the best results; In the MEN vs. AGG task, the combination of the full spectra produced the best model (95% of accuracy). In the MEN vs. LGG task, the combination of the 21 variables from LTE and 29 from STE was the best model (98% of accuracy). On the other hand, if we aim to distinguish between AGG and LGG, acquisition of STE should be enough. In this task, the best model was composed by only 10 variables (93%) and the equivalent model of the combined approach was composed by only STE variables.

3 Modeling of Magnetic Resonance Spectra Using Mixtures for Binned and Truncated Data

The extraction of relevant information for classification from the biomedical signals is a major challenge when complex data are analyzed. ^{1}H MRS provides the biochemical composition of a tissue under study, and is typically used for in-vivo diagnosis of brain tumours.

Prior knowledge of the relative position of the organic compound contributions in the ^{1}H MRS inspired the development of a probabilistic mixture model and its EM-based Maximum Likelihood Estimation for binned and truncated data. The parametric space of the mixture model estimated by the EM algorithm can be used as feature space of a classification model. This approach is aimed to achieve a dimensionality reduction of the spectra composed by informative variables.

The definition and estimation of the probabilistic model based on binned and truncated data to fit ^{1}H MRS was introduced in [5]. Under the hypothesis that the model summarizes the information of the metabolites observed in the spectra, the estimated parameters for each spectrum were used as input features of linear models to classify the tumours.

We assumed that the samples have a common probability density function, irrespective of their originating bins. This density function is a parametric C-component mixture that correspond to the C resonances of the metabolites,

$$p_{\mathbf{\Theta}}(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \, p_{\mathbf{\Theta}'}(\mathbf{x} \mid c) \tag{1}$$

where $\mathbf{\Theta} = (\pi, \mathbf{\Theta}')$ is the parameter vector of the mixture; $\pi = (\pi_1, \ldots, \pi_C)$ is the vector of mixture coefficients, subject to $\sum_c \pi_c = 1$, and $\mathbf{\Theta}'$ includes the parameters required to define each mixture component $p_{\mathbf{\Theta}'}(\mathbf{x} \mid c)$, $c = 1, \ldots, C$. Two parametric forms were studied:

1. PF1: Normal densities of unknown means μ_c, and independent variances $\sigma_1^2, \ldots, \sigma_C^2$; that is, for all $c = 1, \ldots, C$ metabolites

$$p_{\mathbf{\Theta}'}(\mathbf{x} \mid c) \sim N(\mu_c, \sigma_c^2) \tag{2}$$

2. PF2: Normal densities of means known up to a *global shift* μ_0, and independent variances $\sigma_1^2, \ldots, \sigma_C^2$; that is, for all $c = 1, \ldots, C$ metabolites

$$p_{\mathbf{\Theta}'}(\mathbf{x} \mid c) \sim N(\mu_0 + \delta_c, \sigma_c^2) \tag{3}$$

The EM-based feature extraction methodology was applied to 147 Short TE spectra of the Interpret project [8]: 77 Glioblastoma Multiforme (GBM), 50 Meningioma (MEN) and 20 Astrocytoma grade II (AS2). The objective of this study was to obtain pairwise classifiers of the individual classes instead of classifiers of super-classes. Hence, less prevalent classes of the dataset were discarded.

The vector of mixture coefficients $\hat{\pi}$ estimated by PF1 or PF2 composed the parametric space of the linear classifiers. To initialize the models the $\mu_c, \forall c = 1, \ldots, C$ were set to the well-known chemical shift of the metabolites. Pairwise classifiers were produced for the diagnosis of GBM, MEN and AS2 based on LDA. These classifiers were compared to a Principal Component Analysis-based classifier. The estimation of the error was carried out by cross validation with 10 stratified partitions. PF2 achieved the best performance when classifying AS2 from GBM (94.8% of accuracy) and AS2 from MEN (92.9% of accuracy). In the discrimination of MEN from GBM, both PF1 and PF2 estimations were considerably better than the PCA model (89.8% for PF1, 85.8% for PF2 and .82.7% for PCA).

This probabilistic mixture model for binned and truncated data may be extended to incorporate patterns of the types of tumours instead of being individually applied to each patient. These models would be also composed by mixtures to represent heterogenous types, like the Glioblastoma Multiforme type.

4 Glioblastomas Multiforme and Meningothelial Meningiomas Discrimination by Single-Labeling cDNA-Based Microarrays

Clinical research aims to incorporate the genomic information in the diagnosis and management procedure of cancer. In the study of genetic origin of brain tumours, ML techniques are applied to search the differentially expressed genes and to discriminate brain tumours using microarray data.

Different technologies have been developed to study the gene expression at the transcriptomic level. Single-labeling cDNA microarrays are a cheap technology for researchers and the manufacturing of these microarrays is more flexible than any commercial product. However, the single-labeling in cDNA is relatively new and it has not been applied in many studies, therefore, before studying molecular diagnosis by means of the definition of molecular subtypes of histopathological diagnosis, it is necessary to demonstrate the capability of the technology combined with the high-level analysis.

As a proof of principle, we decided to apply ML on microarray data to discriminate between GBM and MEN tumours. These two types of tumours are histological and pathologically the most different among the primary brain tumours. Hence, ML methods are expected to show robustness in their discrimination.

A training database consisting of ten GBM, and eleven MEN, was used to fit and evaluate ML classifiers for the two diagnoses. The test database consisted of fourteen blind and totally independent samples.

Two approaches for fitting, selection and evaluation of the automatic classifiers were studied. Random feature selection and stepwise feature selection were applied in the set of genes that were previously selected by the Mann-Whitney test. The selected genes were used as input for linear classifiers to discriminate GBM and MEN.

Both, random selection and stepwise selection procedures were evaluated by a random sampling train test with 200 repetitions, that used 70 % of the dataset (15 samples) for training and 30% for validation (6 samples). Both approaches achieved a mean accuracy above 94%. Besides, every sample of the independent test set was correctly classified by both approaches.

In our experiments, ML techniques on microarray data showed robustness in the generalization capability of the classification between GBM and MEN types. Characterization of gene signatures for both GBM and MEN provided meaningful information about biological processes underlying each tumour type. This biological information allowed verifying that the gene signature of each tumour type was concordant with the pathological characteristics of such tumours. Finally, validation of microarray-based gene expression values was performed using real time polymerase chain reaction (RT-PCR).

The results of this study achieved the proof of principle we established to demonstrate the utility of the single-labeling cDNA microarrays combined with Machine Learning procedures to discriminate brain tumours. Therefore, we can go ahead with more complex classifications and the definition of molecular subtype based on the developed methodology. The knowledge developed in the genetic level may be useful to redefine the classification carried out with in vivo data.

5 Distributed Decision Support System for Brain Tumour Classification Based on Machine Learning

As a result of the studies based on ML, models to predict the diagnosis of new patients are obtained. In order to make useful them, they should be accessible to clinicians in a simple and direct way. This is done by means of Decision Support Systems, that are aimed to reduce the gap between ICT and the medical experts in cancer research.

A prototype of the DSS was developed to help the clinicians through ML classifiers. This prototype offers a classifier based on LDA with Long and Short TE MR features as input [7]. Finally, a GUI was designed to allow user-friendly human-machine interaction.

The graphical user interface (see Figure 1) shows the estimated tumour class corresponding to the tested sample. The visualization of the latent space provided by the LDA classifier is useful for clinicians to decide whether the tested case is routine or a special one.

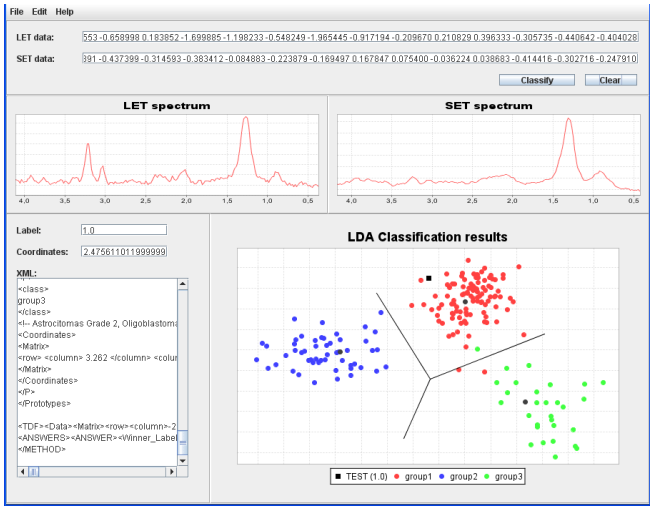

Fig. 1. Prototype to classify among AGG (Group 1), MEN (Group 2), and LGG (Group 3). The main part of the screenshot corresponds to the LDA latent space.

.

6 Conclusions

The use of ML techniques for diagnosis of brain tumours has showed a great performance and a very promising future. Nevertheless, in order to obtain a robust evaluation of the accuracy of the classifiers, larger datasets are needed for training and testing purposes.

Scant research has been devoted to the robust evaluation of the classifiers in future patients. For this purpose, an evaluation of brain tumour classifiers developed by data acquired during the Interpret project will be tested by a totally independent and multicentric test set acquired afterwards.

The special character of the biomedical environment leads to the research of classifiers combining data in dynamic Decision Support Systems. In further research activities, evaluation methodologies in distributed DSS are being developed to audit the classifiers during their use in a distributed network for the help in the brain tumour diagnosis and prognosis.

Acknowledgements. This work was partially funded by the European Commission: eTUMOUR (contract no. FP6-2002-LIFESCIHEALTH 503094), HealthAgents (contract no. FP6-2005-IST 027213); the Spanish Ministerio de Sanidad y Consumo, INBIOMED, PI052245; and the Programa de Apoyo a la Investigación y Desarrollo, PAID-00-06 UPV; We thank INTERPRET partners for providing data, in particular C. Majós (IDI-Bellvitge), A. Moreno (Centre Diagnostic Pedralbes), John Griffiths (SGUL), Arend Heerschap (RU), Witold Gajewicz (MUL), and Jorge Calvar (FLENI).

References

1. Howe, F.A., Opstad, K.S.: 1H MR spectroscopy of brain tumours and masses. NMR Biomed 16(3), 123–131 (2003)
2. Galanaud, D., Nicoli, F., Chinot, O., Confort-Gouny, S., Figarella-Branger, D., Roche, P., Fuentes, S., Le Fur, Y., Ranjeva, J.P., Cozzone, P.J.: Noninvasive diagnostic assessment of brain tumors using combined in vivo MR imaging and spectroscopy. Magn. Reson Med. 55(6), 1236–1245 (2006)
3. Tate, A.R., Underwood, J., Acosta, D.M., Julia-Sape, M., Majos, C., Moreno-Torres, A., Howe, F.A., van der Graaf, M., Lefournier, V., Murphy, M.M., Loosemore, A., Ladroue, C., Wesseling, P., Luc Bosson, J., Cabanas, M.E., Simonetti, A.W., Gajewicz, W., Calvar, J., Capdevila, A., Wilkins, P.R., Bell, B.A., Remy, C., Heerschap, A., Watson, D., Griffiths, J.R., Arús, C.: Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed 19(4), 411–434 (2006)
4. Arús, C., Celda, B., Dasmahapatra, S., Dupplaw, D., González-Vélez, H., van Huffel, S., Lewis, P., Lluch i Ariet, M., Mier, M., Peet, A., Robles, M.: On the design of a web-based decision support system for brain tumour diagnosis using distributed agents. In: IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops), Hong Kong, pp. 208–211 (2006)
5. Garcia-Gomez, J.M., Robles, M., Huffel, S.V., Juan-Císcar, A.: Modelling of magnetic resonance spectra using mixtures for binned and truncated data. In: Springer-Verlag (ed.) Proceedings of the 1st Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), Girona, Spain. Lecture Notes in Computer Science Series, Springer, Heidelberg (2007)
6. Majos, C., Julia-Sape, M., Alonso, J., Serrallonga, M., Aguilera, C., Acebes, J.J., Arus, C., Gili, J.: Brain tumor classification by proton MR spectroscopy: comparison of diagnostic accuracy at short and long TE. AJNR Am. J Neuroradiol 25(10), 1696–1704 (2004)
7. Tortajada, S., García-Gómez, J.M., Vidal, C., Arús, C., Juliá-Sapé, M., Moreno, A., Robles, M.: Improved classification by pattern recognition of brain tumours combining long and short echo time 1h-mr spectra. In: SpringerLink. (ed.) Book of Abstracts ESMRMB 2006, - Supplement 1, Journal Magnetic Resonance Materials in Physics, Biology and Medicine. vol. 19, pp. 168–169 (2006)
8. Julia-Sape, M., Acosta, D., Mier, M., Arus, C., Watson, D.: A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. Magn. Reson Mater Phy. 19(1), 22–33 (2006)
9. INTERPRET Consortium: Interpret web site. (Accessed: 27 January 2007) `http://azizu.uab.es/INTERPRET/`
10. García-Gómez, J.M., Vidal-Fernández, C., Robles-Viejo, M.: Preliminary choice of pattern recognition techniques and methodology, discussion on implementation details. Technical report, BET-IM, Univerdidad Politécnica de Valencia (2005)