Modeling Tumor Cellularity in Newly Diagnosed GBMs using MR Imaging and Spectroscopy

Alexandra Constantin[1,2], Sarah J. Nelson[2], and Ruzena Bajcsy[1]

[1] Electrical Engineering and Computer Science, University of California, Berkeley, USA
[2] Department of Radiology, University of California, San Francisco, USA
{alexacon,bajcsy}@eecs.berkeley.edu,nelson@radiology.ucsf.edu

Abstract. In this paper, we analyze the relationship between parameters of brain tumors obtained through *in vivo* magnetic resonance imaging (MRI), *in vivo* magnetic resonance spectroscopy (MRS), and *ex vivo* immunohistochemistry (IHC). The goal of our project is to provide a quantitative definition of tumor cellularity based on the *in vivo* parameters. Biopsy samples obtained from previously untreated patients with a diagnosis of GBM are used to find the link between imaging parameters at the specific biopsy locations and IHC parameters from the corresponding tissue samples. A functional tree (FT) model of tumor cellularity is learned from the *in vivo* parameters and the remaining histological parameters. The tumor cellularity model is then tested on examples which contain only *in vivo* parameters, by first estimating the remaining IHC parameters by applying the Expectation Maximization (EM) algorithm, and then using the complete parameter vector for classification.

1 Introduction

Current brain tumor research involves the analysis of multiple heterogeneous data sets, like various types of MR images and spectroscopy, as well as clinical and histopathology data. The successful integration of these images can provide insight into the status and development of brain tumors beyond what can be inferred from each single image. One of the most significant limitations of *in vivo* MRI is the underlying uncertainty about whether a voxel contains tumor. While MRI data can be used to detect abnormalities and provide good spatial information about different tissue properties, the true composition of the tissues can only be determined histologically. The histopathological characterization of brain tumors requires the use of small biopsy samples obtained by performing an invasive surgical procedure. The analysis of these samples provides parameters describing the density of the cancer cells (*i.e.*, the tumor cellularity score) and the proliferative and invasive capacity of the tumor. Based on these parameters, brain tumors are given a histopathological grade of malignancy. However, a biopsy may not accurately represent the real grade or proliferative capacity of a tumor as a whole, because of the heterogeneity which exists within the tumor. Many studies have been done to assess the correlation between *in vivo* MRI and MRS and various histological parameters [8, 9]. These studies focus on finding associations between pairs of *in vivo* and histological parameters, but do not attempt to obtain a tumor profile based on multiple data modalities. Other studies try to assess the utility of global features extracted from

in vivo MRI and MRS in determining the grade of a tumor [2, 5, 7, 10], therefore making an indirect link between the *in vivo* and histological parameters.

In this project, we explore the link between the histological parameters of biopsy samples and the *in vivo* MR and MRS parameters at the biopsy locations. The goal is to obtain an *in vivo* definition of tumor cellularity, which is a histological measure of the density of cancer cells in a biopsy specimen, and which represents the gold standard measurement of how much cancer is present. We use a functional tree model to obtain an approximation of tumor cellularity based on other histological parameters, as well as *in vivo* parameters. This model can then be used to obtain a tumor cellularity score based on the *in vivo* parameters alone, by first estimating the remaining histological parameters using the EM algorithm for multiple imputations, and then using the whole parameter vector for classification. The model thus provides a quantitative definition of tumor based on the *in vivo* parameters alone. This *in vivo* definition of tumor can be analyzed in all image regions showing abnormalities, thus providing better spatial resolution than biopsy samples. This non-invasive tumor definition could decrease the need for invasive biopsies, or be used to better select the biopsy locations.

2 Data Acquisition, Preprocessing, and Feature Extraction

The patients in this study received full MR examinations on a 3T scanner. The examination included three-dimensional gadolinium-enhanced T1-weighted (T1C), fluid attenuated inversion recovery (FLAIR), diffusion weighted and dynamic perfusion weighted images (see Figure 1 for illustrations). Apparent diffusion coefficient (ADC) and fractional anisotropy (FA) maps were calculated from the diffusion data. The perfusion weighted imaging (PWI) was modeled parametrically, using a modified gamma-variate function [4] yielding cerebral blood volume (CBV), peak height (PH), recirculation factor (RF), and percent recovery to baseline (RECOV), and non-parametrically, yielding peak height (PH) and percent recovery to baseline (RECOV). Lactate-edited MRS data were also acquired as part of the exam. The spectral amplitudes and line-widths of choline (CHO), creatine (CRE), N-Acetyl-Aspartate (NAA), lactate (LAC), and lipid (LIP) were estimated from the spectra. All images in the MR exam were rigidly registered to the T1C image.

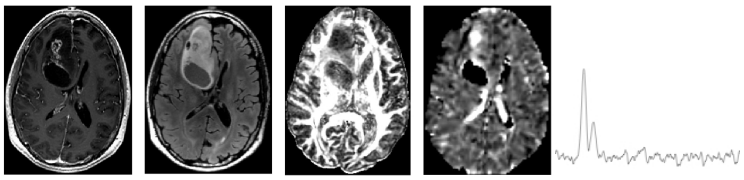

Fig. 1: From left to right: T1C, FLAIR, FA, NONPAR_RECOV, and one voxel of MRS data.

Tissue sample locations were selected based on surgically accessible areas with low ADC, elevated CHO to NAA ratio (CNI), or elevated PWI peak height and low recovery.

For histological analysis, the tissue samples were analyzed by a pathologist, who scored the samples for tumor cellularity, MIB proliferation index, and axonal integrity (SMI-31), and noted the presence of simple, complex, and delicate vasculature, hyperplasia, and CA9 gene expression.

The data set consists of seventy-eight tissue samples obtained from thirty-five GBM patients. Each tissue sample has ten associated immunohistochemistry (IHC) parameters. The tumor cellularity can be grouped into two categories - low (17 samples) and high (61 samples); or three categories - low (15 samples), medium (30 samples), and high (33 samples). The imaging parameters at the biopsy locations were obtained by creating a 5mm biopsy mask centered at the recorded biopsy location of each tissue sample. The median intensity values inside this mask in each of the MR and MRS images were used as the *in vivo* parameters. A total of seventeen *in vivo* parameters were used. Ten of the samples have missing perfusion data and twenty-one have missing spectral data.

3 Models

The problem of classifying data samples into low and high tumor cellularity samples is formulated as a supervised learning problem. The *in vivo* and immunohistochemistry parameters of each tissue sample are used to form a feature vector. The associated tumor cellularity class is the desired output. The task of the supervised learner is to determine the tumor cellularity class for any valid input vector, after having seen a number of training examples, by generalizing from the presented data to unseen situations in a reasonable way. The missing values in the training data are first estimated using the EM algorithm. An FT classifier is then used to learn a model of tumor cellularity based on the *in vivo* parameters, as well as the remaining histological parameters. The model is then used to predict the tumor cellularity of unseen examples, either using the full input vector, or only the *in vivo* parameters. In the latter case, the EM algorithm for multiple imputations is used to estimate the histological parameters, and the resulting input vector is used for classification.

3.1 EM Imputation of Missing Values

Missing values often hinder the analyses of multivariate data. In this study, even though the rate of missing values for each parameter is not very high, there are very few patients with complete data. Therefore, we use the EM algorithm to fill in missing data with plausible values, as well as to estimate the histological parameters when desiring to predict tumor cellularity based on the *in vivo* parameters alone. The EM algorithm provides an iterative approach to the problem of maximum likelihood parameter estimation in statistical models with latent variables. In the expectation (E) step, the values of the missing variables are estimated by calculating the probability of the latent variables given the observed variables and the current values of the model parameters. In the maximization (M) step, the parameters are adjusted based on the current estimates of the missing values. These steps are repeated until the sequence of parameters converges

to the maximum likelihood estimates that average over the distribution of missing values. The EM algorithm for multiple imputations is presented in detail by Schafer [6]. Below is a short description of the algorithm.

The dataset Y is assumed to be a matrix of n rows and p columns, with the rows corresponding to observations and the columns corresponding to variables. The complete dataset contains observed and missing portions: $Y = (Y_{obs}, Y_{mis})$. Let $y_{i,j}$, with $i \in 1 \dots n$ and $j \in 1 \dots p$, denote an individual element of Y and let y_i denote a row of the data matrix. The main model assumption is that $y_1 \dots y_n$ are independent realizations of a multivariate normal distribution with mean vector μ and covariance matrix Σ. When both μ and Σ are unknown, the conjugate prior distribution for the multivariate normal data model is the normal inverted-Whishart distribution [6]. Suppose that Σ is inverted-Wishart and that μ given Σ is assumed to be conditionally multivariate normal. Then the complete-data likelihood function is [6]:

$$L(\theta|Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\big\{\frac{n}{2} tr\Sigma^{-1}S\big\} \exp\big\{-\frac{n}{2}(\bar{y}-\mu)^T\Sigma^{-1}(\bar{y}-\mu)\big\}. \tag{1}$$

The complete-data posterior is normal inverted-Wishart. The prior distribution of μ is assumed to be uniform over the p-dimensional real space, resulting in the following complete-data posterior [6]:

$$\mu|\Sigma, Y \propto N(\bar{y}, n^{-1}\Sigma); \Sigma|Y \propto W^{-1}(n-1, (nS)^{-1}). \tag{2}$$

where the updated parameters are [6]:

$$\begin{aligned} &\tau' = \tau + n; m' = m + n; \mu_0' = \big(\frac{n}{\tau+n}\big)\bar{y} + \big(\frac{\tau}{\tau+n}\big)\mu_0; \\ &\Lambda' = \big[\Lambda^{-1} + nS + \frac{\tau}{\tau+n}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T\big]. \end{aligned} \tag{3}$$

The E-step calculates the conditional expectations of the missing variables conditioned on the observed variables and fixed model parameters. This amounts to calculating the expected complete log likelihood, and thus to finding the complete-data sufficient statistics over $P(Y_{mis}|Y_{obs}, \theta)$ for assumed value of θ. The sufficient statistics are of the form $\sum_i y_{i,j}$ and $\sum_i y_{i,j}y_{i,k}$ [6]. We thus need to find the expectations of $y_{i,j}$ and $y_{i,j}y_{i,k}$ over $P(Y_{mis}|Y_{obs}, \theta)$. The distribution $P(y_{i(mis)}|y_{i(obs)}, \theta)$ is a multivariate normal linear regression of $y_{i(mis)}$ and $y_{i(obs)}$ [6]. The parameters of this regression can be calculated by sweeping the θ-matrix on the positions corresponding to the variables in $y_{i(obs)}$, as described by Schafer [6]. The E-step consists of calculating and summing the expected values of y_{ij} and $y_{ij}y_{ik}$ for each j and k. Carrying out the M-step involves maximizing the expected complete log likelihood with respect to the parameters.

3.2 Functional Trees for Classification

Tree induction methods and regression models are popular techniques for supervised learning tasks, both for the prediction of nominal classes and numerical values. Regression methods fit a simple linear or logistic model to the data, resulting in low variance but potentially high bias estimates. Decision trees classify data into categories based on a series of questions or rules about attributes of the class. Decision tree classifiers are able to capture nonlinear patterns in the data, but are less stable and more prone to overfitting. These two schemes can be combined into functional trees, which are able to use

decision nodes with multivariate tests and leaf nodes that make predictions using linear or logistic functions. FTs were selected as the model of choice in this study because they are efficient to construct and easy to interpret, they can be used for either classification or numerical prediction, and they are formed through a constructive induction process that selects relevant attributes automatically, without changing the representation of the data. This property of FT models is important in the medical imaging field, where it is necessary to be able to interpret the role of different attributes in the prediction process, in order to gain new insights into the disease mechanism and treatment options.

Functional and logistic trees were introduced by Gama [1] and Landwehr et al. [3]. Next, we provide a short description of the algorithm used in this paper. A functional tree is built by starting at the root. The existing set of attributes is extended using a constructor function that fits regression functions using the LogitBoost algorithm [3]. This algorithm starts out with a simple linear regression model based on the most predictive attribute. In every iteration, it computes response variables that encode the error of the currently fit model on the training examples, and then tries to improve the model by adding another simple linear regression function fit by least-squared error. Because every multiple linear regression function can be expressed as a sum of simple linear regression functions, the general model does not change whether we use multiple or simple linear regression functions. LogitBoost is guaranteed to converge to the maximum likelihood estimate. However, if the method is stopped before convergence, this will result in the automatic selection of the most relevant attributes. Therefore, LogitBoost is stopped based on a cross-validation method: more iterations are performed (and therefore more attributes are included) only if this improves prediction accuracy over unseen instances. Once the set of attributes is extended, the attribute that maximizes the information gain ratio is selected as a splitting criterion. The child nodes are split recursively, by incrementally refining the regression models already fit at higher levels in the tree, thus taking into account the attributes that are only predictive locally. Tree growing stops if a node contains less than fifteen examples or if a particular split results in two subsets, one of which contains less than two examples. A linear model is only built at a node if that node contains at least five examples. Otherwise, a leaf with the majority class of the node is returned. Once a tree has been grown, it is pruned back using a bottom-up procedure. At each non-leaf node three possibilities are considered: performing no pruning, replacing the node with a leaf that predicts a constant, or replacing it with a leaf that predicts the value of the constructor function that was learned at the node during tree construction. The option that leads to the smallest error on a pruning data set is selected. Predicting a test instance using a functional tree is accomplished by traversing the tree in a bottom-down fashion. At each decision node the local constructor function is used to extend the set of attributes and the decision test determines the path that the instance will follow. Once a leaf is reached, the instance is classified using either the constant or the constructor function at that leaf.

4 Results

The data described in Section 2 was first preprocessed and the missing values were estimated using the EM algorithm. The log likelihood threshold for convergence was set

to 10^{-4}. The FT classifier was used to learn a model for classifying low and high tumor cellularity based on the remaining histological parameters and the *in vivo* parameters.

The model was first tested using complete input vectors. Table 1 shows that the full *in vivo* and histology model is able to distinguish between low and high tumor cellularity with 96 % accuracy. This model outperforms the model based on *in vivo* parameters alone. The binary tumor cellularity model is based on one regression node as indicative of high tumor cellularity:

```
2.21 −0.28FLAIR −0.65FA −0.6NONPAR_RECOV +0.41CHO +0.23CRE −0.39LIP
−0.31DV_PRES +0.38CA9 +0.88SMI31 +1.9MIB > 0.
```

The histology parameters are, in general, the most important in predicting tumor cellularity. This can be expected, because the imaging parameters come from an area around the estimated biopsy location, and might not exactly correspond to the tissue samples themselves. Among the *in vivo* parameters, the choline, lipid, recovery, and FA are important predictors of high cellularity. The best model uses only ten parameters, suggesting that some of the data is redundant. There is an imbalance between the number of low and high tumor cellularity examples in the data set, so a better model could be learned if more low cellularity data samples were used for training.

Table 1: Training and leave-one-out cross-validation accuracy results for binary and three-category tumor cellularity models using different sets of parameters

Features	**Binary Cellularity**		**Three-Categ Cellularity**	
	Training Accuracy	*Cross-Validation*	*Training Accuracy*	*Cross-Validation*
in vivo + hist	100%	96%	90%	77%
hist	92 %	87 %	81 %	69 %
MRS + hist	97 %	88 %	86 %	65 %
Perf + hist	95 %	89 %	86 %	68 %
Diffu + hist	95 %	90 %	83 %	70 %
Anat + hist	95 %	88 %	83 %	68 %
in vivo	87%	73%	68%	24 %

The *in vivo* only binary tumor cellularity model first divides the data into high and low choline. Among the low choline biopsies, the ones with low nonparametric recovery get classified as high in cellularity, and the ones with high recovery get classified as having low cellularity. The high choline biopsies are classified using a regression model. In this model, high T1C, nonlinear RF, choline and creatine, and low FA, CBV, PH, NAA and lipid, are indicative of high tumor cellularity.

Table 2 shows that the information obtained from the histology parameters can be used to improve tumor cellularity classification even when histological data is not available. Using the full model to predict tumor cellularity based on *in vivo* parameters and EM estimates of the "missing" histological parameters yields a significantly higher prediction accuracy compared to the *in vivo* model. Learning the structure of the model from histology and using the *in vivo* parameters to estimate the histological data gives better performance than trying to estimate tumor cellularity using the *in vivo* parameters directly.

Table 2: Accuracy of binary cellularity model with EM estimated histology.

Training Data	Test Data	Acc
in vivo + hist	*in vivo* + hist	96%
in vivo + hist	*in vivo* + hist EM estimates	82 %
in vivo	*in vivo*	73 %

Table 3: Confusion matrix for full data three-category cellularity model.

Class. 1	Class. 2	Class. 3	Actual
14	1	0	**1**
0	26	4	**2**
0	4	29	**3**

The three-category tumor cellularity model builds a regression function for each class. Below are the models for each class:

```
LOW:   -3.03 +0.8FA +0.76NONPAR_RECOV -0.41CHO +0.41LIP
       -1.57SMI31 -0.46SMPL -0.51CMPLX -3.16MIB>0
MEDIUM: 0.5 -0.17FLAIR +0.39NONLIN_NRF -0.19NONLIN_NPH -0.09NONPAR_RECOV +0.13CHO
       +0.09LIP -0.13CNI-0.17CCRI-0.2SMPL+0.32CMPLX-0.29MIB>0
HIGH:  -0.75 -0.69FA -0.64LIP +0.32CNI -0.4DV_PRES +1.69SMI31 +0.89SIMPL +0.7MIB>0
```

The accuracy of the three-category model is summarized in Table 1. The full model is able to predict tumor cellularity with 77% accuracy, given a three-category response. The confusion matrix for the full model is shown in Table 3. Most of the instances are classified correctly, and the classification errors that occur are always between two adjacent classes.

In addition to testing the accuracy of the FT model on examples with known classes, we also used the model to predict the cellularity of every voxel inside the manually defined abnormal FLAIR region of an MR exam. The results are illustrated in Figure 2. The model classifies a large part of the voxels in the center of the abnormal region as being high cellularity. After performing morphological operations on the resulting mask, we notice a center of high cellularity, with lower cellularity outwards. The amount of data available for estimating the joint distribution of *in vivo* and histological parameters was relatively small given the number of histological parameters that need to be estimated. Also, the data were not randomly selected and comes from a relatively limited distribution of parameters. This explains some of the noise in the original high cellularity mask. Obtaining more data from random locations in the abnormality region is necessary in order to be able to more accurately extrapolate tumor cellularity values beyond the biopsy locations.

5 Conclusions

In this project, we provided a framework for learning a quantitative *in vivo* definition of tumor cellularity. The model we proposed was able to accurately classify data instances into tumor cellularity categories using a set of *in vivo* and histological parameters, as well as the *in vivo* parameters and estimates of histological parameters. We were able to find *in vivo* parameters that were instrumental in predicting high tumor cellularity in all models, such as high choline and low recovery. The framework we described can be used to create predicted maps of tumor cellularity at the spatial resolution of anatomical images provided that training data covering the entire distribution of parameters in the region of interest is available. Building such a model could lead to a better biopsy

selection process or to the early prediction of tumor cellularity and grade before biopsies are acquired and processed.

Acknowledgments. This research was supported by grants NIH PO1 CA118816 and NIH RO1 CA127612. We would like to thank Susan Chang, Soonmee Cha, and Joanna Phillips for their invaluable help.

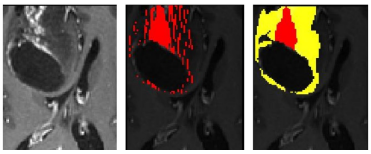

Fig. 2: Tumor cellularity predictions on a T1C image. Middle: the high cellularity mask in red. Right: tumor cellularity masks (red = high, yellow = low), corrected using morphological closing.

References

1. Gama, J.: Functional trees. Machine Learning 55(3), 219–250 (2004)
2. Garcia-Gomez, J., Tortajada, S., Vicente, J., Saez, C., Castells, X., Luts, J., Julia-Sape, M., Juan-Ciscar, A., Huffel, S.V., Barcelo, A., Arino, J., Arus, C., Robles, M.: Genomics and metabolomics research for brain tumour diagnosis based on machine learning. In: IWANN. pp. 1012–1019 (2007)
3. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Machine Learning 59(1-2), 161–205 (2005)
4. Lee, M., Cha, S., Chang, S., Nelson, S.: Partial-volume model for determining white matter and gray matter cerebral blood volume for analysis of gliomas. Magnetic Resonance Imaging 23, 257–266 (2006)
5. Metsis, V., Huang, H., Makedon, F., Tzika, A.: Heterogeneous data fusion to type brain tumor biopsies. In: AIAI. pp. 233–240 (2009)
6. Schafer, J.: Analysis of Incomplete Multivariate Data. Chapman and Hall, London, UK (1997)
7. Simonetti, A.W., Melssen, W.J., de Edelenyi, F.S., van Asten, J.J.A., Heerschap, A., Buydens, L.M.C.: Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification. Nuclear Magnetic Resonance in Biomedicine 18(1), 34–43 (2005)
8. Sugahara, T., Korogi, Y., Kochi, M., Ikushima, I., Shigematu, Y., Hirai, T., Okuda, T., Liang, L., Ge, Y., Komohara, Y., Ushio, Y., Takahashi, M.: Usefulness of diffusion-weighted MRI with echo-planar technique in the evaluation of cellularity in gliomas. Journal of Magnetic Resonance Imaging 9(1), 53–60 (1999)
9. Yerli, H., Agildere, A., Ozen, O., Geyik, E.and Atalay, B., Elhan, A.: Evaluation of cerebral glioma grade by using normal side creatine as an internal reference in multi-voxel 1H-MR spectroscopy. Diagnostic and Interventional Radiology 13(1), 3–9 (March 2007)
10. Zacharaki, E., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E., Davatzikos, C.: Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magnetic Resonance in Medicine (2009)