Published in final edited form as:
Proc IEEE Int Symp Biomed Imaging. 2011 March 30; : 1624–1627. doi:10.1109/ISBI.2011.5872714.

MORPHOLOGICAL SIGNATURES AND GENOMIC CORRELATES IN GLIOBLASTOMA

Lee A.D. Cooper[*], **Jun Kong**[*], **Fusheng Wang**[*], **Tahsin Kurc**[*], **Carlos S. Moreno**[†,*], **Daniel J. Brat**[†], and **Joel H. Saltz**[*]

[*]Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322

[†]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA 30322

Abstract

Large multimodal datasets such as The Cancer Genome Atlas present an opportunity to perform correlative studies of tissue morphology and genomics to explore the morphological phenotypes associated with gene expression and genetic alterations. In this paper we present an investigation of Cancer Genome Atlas data that correlates morphology with recently discovered molecular subtypes of glioblastoma. Using image analysis to segment and extract features from millions of cells, we calculate high-dimensional morphological signatures to describe trends of nuclear morphology and cytoplasmic staining in whole-slide images. We illustrate the similarities between the analysis of these signatures and predictive studies of gene expression, both in terms of limited sample size and high-dimensionality. Our top-down analysis demonstrates the power of morphological signatures to predict clinically-relevant molecular tumor subtypes, with 85.4% recognition of the proneural subtype. A complementary bottom-up analysis shows that self-aggregating clusters have statistically significant associations with tumor subtype and reveals the existence of remarkable structure in the morphological signature space of glioblastomas.

Index Terms

bioinformatics; in silico; digital pathology; image analysis; microscopy

1. INTRODUCTION

The National Institute of Health's *The Cancer Genome Atlas* (TCGA)1 project is producing a large multimodal datasets containing pathology imaging, radiology, genomic, and clinical data for glioblastoma and other tumor types [1]. These datasets present a unique opportunity to perform correlative studies of tissue morphology and genomics to explore the morphological phenotypes associated with patient outcome, gene expression and genetic alterations. To investigate these relationships in glioblastoma brain tumors we have developed an imaging system to analyze millions of cells within hundreds of whole-slide pathology images to determine the survival, gene expression and genomic correlates of cellular morphology.

A study of TCGA molecular data has identified four clinically-relevant subtypes of human glioblastoma [2]. The *Proneural* (PN), *Neural* (NR), *Classical* (CL) and *Mesenchymal* (MS) tumor subtypes are each defined by characteristic gene expression profiles and genetic

[1]http://cancergenome.nih.gov/

alterations including mutations and chromosomal amplifications/deletions, and differ in response to treatment and survival expectations. This analysis also compared the gene expression profiles of tumor subtypes to those of ordinary neural cell types and found evidence suggesting that the glioblastoma subtypes are reminiscent of distinct neural cell types.

In a previous study we examined the links between nuclear morphology and these four tumor subtypes in glioblastoma [3]. In this paper we extend this analysis to include features describing cytoplasmic staining to produce an enhanced morphological signature describing both nuclear morphology and the surrounding cytoplasm. A top-down analysis is performed to determine effectiveness of enhanced signatures in predicting tumor subtype. A bottom-up self-aggregation of morphological signatures is also performed to examine the links between the natural clustering of signatures and tumor subtype.

2. MORPHOLOGICAL ANALYSIS

We have developed a system for the quantitative morphological analysis of microanatomy in whole-slide images. The system consists of a number of stages, as depicted in Figure 1. Nuclei are first segmented and the surrounding cytoplasmic spaces are identified. A set of features is extracted to describe the morphology and texture of each individual nucleus and the staining characteristics of its cytoplasm. The segmented objects and corresponding extracted features are stored in a database for further analysis. Using the database, a morphological signature is calculated for each slide using first and second order statistics. This workflow executes on a computing cluster and currently supports over 600 slides containing an estimated 254 million nuclei.

2.1. Segmentation and Feature Extraction

The first stage of our analysis segments individual nuclei using a combination of simple image processing operations. Color images are first thresholded to identify and remove blood and nontissue regions. The remaining areas are converted to grayscale and a morphological reconstruction is applied to remove debris. Overlapped nuclei are then separated using a watershed operation. Each region corresponding to a segmented nucleus is then dilated by a specific margin to identify the surrounding cytoplasmic space. Following segmentation, a collection of features is calculated for each segmented nucleus to represent characteristics of nuclear morphology and nuclear and cytoplasmic staining. A complete list of these features is available in [3]. A color deconvolution algorithm is first applied to the cytoplasmic space to isolate hematoxylin and eosin stain signals into separate channels prior to analyzing cytoplasmic intensity, texture, and gradients [4]. Features representing morphology are not calculated for the cytoplasmic space since the shape is strictly derived from the nucleus boundary.

2.2. Morphological Signature Calculation

The final stage of our analysis calculates a high-dimensional morphological signature for each whole-slide image. For each feature f_i, we calculate the first moment $\mu_i = E\{f_i\}$, and the second moment $C_{i,j} = E\{(f_i - \mu_i)(f_j - \mu_j)\}$ for each pair of features $(i, j) \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, N\}$ where N is the number of features calculated for each segmented entity. Second order statistics are necessary to represent the relationships between features describing nuclear morphology, nuclear morphology and nuclear staining, or nuclear morphology/staining and cytoplasmic staining within the whole-slide image. This produces an $N(N + 3)/2$-dimensional feature vector to represent the morphology of each slide in high dimensional space. Our current implementation uses $N = 74$ features resulting in a 2849-dimensional signature.

2.3. Pathology Analytical Imaging Standards Database

The scale of our derived morphological datasets requires a coordinated approach to data management to support virtual experiments like those presented in this paper. Currently we maintain a glioblastoma dataset containing more than 600 images with an average of 400 thousand nuclei per slide. Morphological characterizations produce 1.5GB/slide of metadata describing algorithm parameters, object boundaries, and features. To address this issue the Pathology Analytical and Imaging Standards (PAIS) model was developed to provide flexible representation of analysis and characterization data [5]. The PAIS model is implemented using a relational database to provide permanent management of analysis results and knowledge discovery through query support.

3. RESULTS AND DISCUSSIONS

We performed two experiments to explore the relationship between the molecularly-defined tumor subtypes and the morphological signatures derived from image analysis. In the first experiment, a top-down classification was applied to predict tumor subtype from morphological signature. In the second experiment, a bottom-up unsupervised clustering was performed to examine structure in the high-dimensional morphological feature space. The bottom-up clusters were further analyzed to determine if the tumor subtypes were overrepresented or underrepresented within each cluster.

To analyze morphology, 200 slides corresponding to 77 distinct patients were acquired from the National Cancer Institute's TCGA public data portal. These images are 20× magnification whole-slide scans of formalin-fixed paraffin embedded sections stained with hematoxylin and eosin. The slides were analyzed using the pipeline from Section 2 to calculate signatures of nuclear morphometry and cytoplasmic staining. The tumor subtypes for a subset of these patients were acquired from [2]. The subtype labels for the remaining samples not included in [2] were predicted using Prediction Analysis of Microarrays [6] with Affymetrix HTU133 gene expression data also acquired from the TCGA portal. In all, our dataset consisted of 49 PN samples, 38 NR samples, 61 CL samples and 52 MS samples.

3.1. Prediction of Molecular Tumor Subtype using Morphological Signatures

We performed a classification experiment to examine the power of morphometric signatures to predict molecular tumor subtype. This experiment is similar to a gene-expression study that aims to predict a given biological state or other external truth using gene expression profiles obtained from microarray experiments. The challenges encountered in predictive gene expression studies remain the same in our case. Sample size is limited and so the high-dimensional space is sparsely populated, making meaningful learning of structure in feature-space difficult. We addressed these challenges using techniques commonly found in predictive gene expression studies. Feature selection was used to reduce dimensionality and increase signal-to-noise ratio by eliminating correlated or irrelevant features. Classification results were validated using *K*-fold cross validation to provide evidence of classifier generality.

Tumor subtypes were used as ground truth to test and train a multiclass support vector machine (SVM). The multiclass SVM consisted of a collection of six binary SVMs trained to distinguish between each pair of tumor subtypes. Given a test sample, the collection of binary SVMs was evaluated and the test sample was assigned to the subtype according to majority vote. Exponential kernels were used for training SVMs to deal with nonlinearities with σ ranging from 1 to 3. For each binary SVM, the relevant features separating the tumor subtype pair were identified using a margin-maximizing optimization with L1-norm penalty to encourage sparsity in selection [7]. Classifier training and testing was performed using a

10-fold cross validation with stratification to preserve the subtype proportions. The validation was repeated 100 times with randomized fold assignments to report the mean and standard deviation of positive prediction for each class, defined as the ratio of true positives to the sum of true positives and false negatives.

Table 1 shows the positive prediction values for each of the binary subtype SVMs and the number of features selected for each. The multiclass SVM results are presented in Table 2. The mean and standard deviation of positive prediction are presented along with the aggregated confusion matrix. Most binary classifiers achieve 90%+ accuracy with only ≈ 1 – 2% of features selected in each case. The distinction between classical and mesenchymal tumor subtypes is the least accurate at 86.4%. The four-way multiclass classifier performs reasonably for proneural, classical, and mesenchymal subtypes, but has some problems with morphological classification of the neural subtype. From the confusion matrix we note that neural is most often mistaken for either proneural or classical subtypes, but not the mesenchymal subtype.

We note that in similar predictive studies of gene expression, selected features (genes) offer biological insight into the differences between sample classes. These selected features can be mined for biological associations using any number of tools based on ontologies of function, disease, or localization. Currently the biological interpretation of most morphological features is unclear. We recognize that providing biological insight is the central aim of correlative morphological studies and we are currently working on methods to annotate and describe signature features in the glioblastoma context.

3.2. Consensus Clustering and Subtype Enrichment

A consensus clustering procedure was applied to the morphological signatures to determine if natural morphological clusters exist. This method uses repeated application of K-means clustering to measure, for each pair of morphological signatures, the frequency with which they are clustered together [8]. Model selection on K was used to identify that $K = 4$ produced the highest cophenetic correlation (0.98) among the possibilities $K = 2, 3, 4, 5, 6$. The structure that corresponds to this exceptional correlation is apparent in Figure 2. Each row/column of Figure 2 contains the co-clustering frequencies between all samples. The strong block-diagonal structure indicates that the four clusters reliably form regardless of K-means initialization, and provides evidence that significant structure exists within the high-dimensional morphological signatures.

The content of the consensus clusters was further analyzed to determine relationships to the tumor subtypes. For each consensus cluster we asked the following question: Are any of the tumor subtypes significantly more or less frequent in this cluster than can be expected? The distribution of the entire dataset provides the expectation in terms of subtype proportions (24.5% PN, 19% NR, 30.5% CL, and 26% MS). Given these proportions, we used the hypergeometric distribution to calculate the probability of a given subtype being *over-represented* or *under-represented* in each consensus cluster. The hypergeometric probability mass function $p(k)$ with parameters (Ω, S, C) models the probability of finding k instances of a subtype in a consensus cluster containing C samples if S total instances out of Ω total samples are expected

$$p(k) = \frac{\binom{S}{k}\binom{\Omega - S}{C - k}}{\binom{\Omega}{C}}.$$

The significance of over or under representation was calculated by summing over $p(k)$ to determine the probabilities $p - O$, $p - U$ of finding an over or under representation that is at least as extreme as what is observed

$$p - O = \sum_{k=M}^{C} p(k), \quad p - U = \sum_{k=0}^{M} p(k),$$

where M is the observed number of instances of the tumor subtype in question within the consensus cluster.

The breakdown of the consensus clusters in terms of tumor subtype is presented in Table 3. Cluster one (39 samples) is enriched in proneural samples and neural samples are conspicuously absent. Cluster two (39 samples) is similarly enriched in proneural samples with classical samples underrepresented. Cluster three (32 samples) is enriched with classical subtype samples and the proneural subtype is significantly underrepresented in cluster four (92 samples). Mesenchymal samples are neither over or under represented in any cluster.

4. CONCLUSION

The paper presents a correlative analysis between high-dimensional morphological signatures and four molecularly defined subtypes of glioblastoma tumors. Top down classification results suggest that signatures of nuclear morphology and cytoplasmic staining have reasonable power to predict tumor subtype. A separate bottom-up clustering analysis, where morphological signatures are free to self-aggregate irrespective of tumor subtype, shows clear structure in the high-dimensional space of morphological signatures. In future work we plan to investigate this structure further to identify significant molecular correlates of the consensus clustering. We are currently focused on developing methods to provide biological insight into correlative morphological studies, including ways to better represent the morphologies of the heterogeneous cell populations encountered in whole-slide images.

Acknowledgments

This research is supported in part by by NCI Contract No. N01-CO-12400 and 94995NBS23 and HHSN261200800001E, by Grant Number R01LM009239 from the NLM, and by NSF CNS 0615155, 79077CBS10, and 0403342.

REFERENCES

1. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; vol. 455:1061–1068.
2. Verhaak RGW, Hoadley KA, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. Cancer Cell. 2009; vol. 17:98–110.
3. Cooper LAD, Kong J, et al. An integrative approach for in silico glioma research. Biomedical Engineering, IEEE Transactions on. 2010 oct.; vol. 57(no. 10):2617–2621.
4. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Analytical and Quantitative Cytology and Histology. 2001 aug.; vol. 23(no. 4):291–299. [PubMed: 11531144]
5. Wang F, Kurc T, Widener P, Pan T, Kong J, Cooper L, et al. High-performance systems for in silico microscopy imaging studies. Data Integration in the Life Sciences. Lecture Notes in Computer Science. 2010; vol. 6254/2010:3–18.

6. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS. 2002; vol. 99:6567–6572. [PubMed: 12011421]

7. Sun Y, Todorovic S, Goodison S. Local-learning-based feature selection for high-dimensional data analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010; vol. 32:1610–1626. [PubMed: 20634556]

8. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003; vol. 52(no. 1–2):91–118.

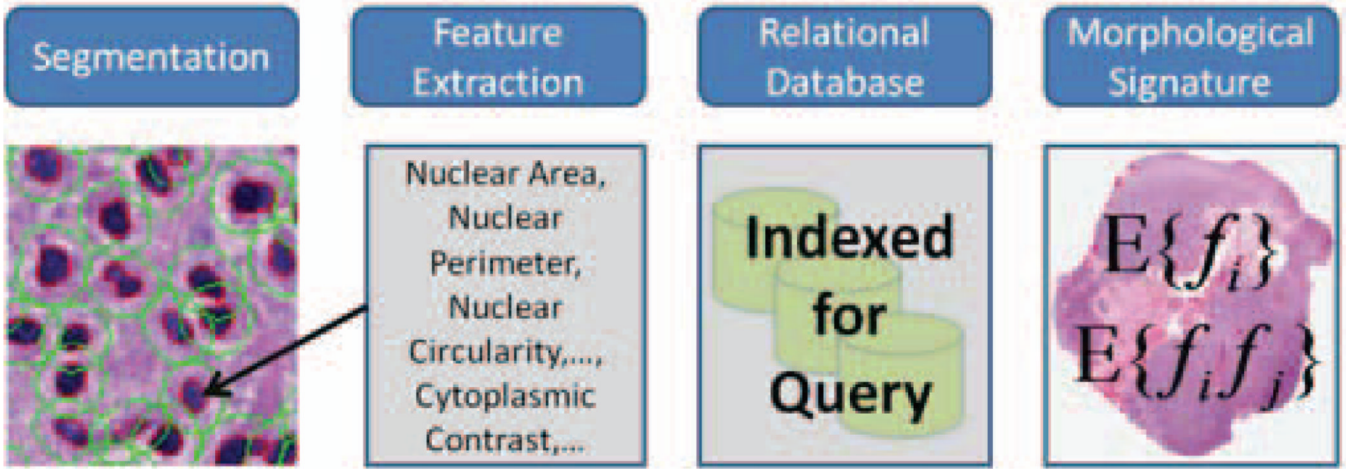

Fig. 1.
Morphological analysis. Characterizations of nuclear shape and cytoplasmic staining describing each cell are indexed to support calculation of whole-slide morphological signatures.

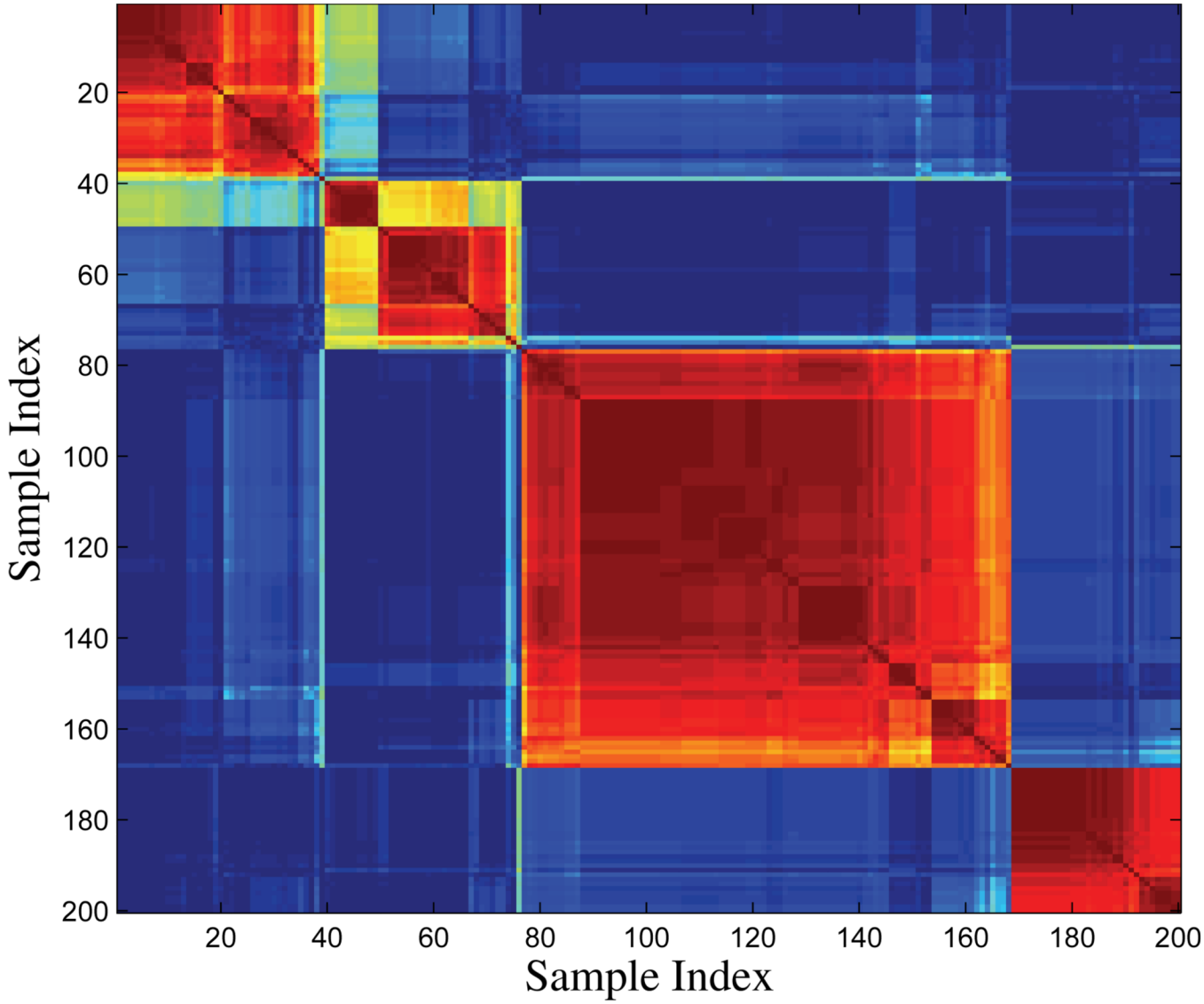

Fig. 2.
Consensus clustering. Bottom-up clustering of morphological signatures reveals four distinct clusters.

Table 1

One-versus-one binary SVM classifier accuracy and feature count. Mean and standard deviation of positive prediction are listed for 100 independent 10-fold cross validations.

	Positive Prediction	Selected Features
PN vs. NR	94.4±1.2%	20
PN vs. CL	91.4±1.0%	39
PN vs. MS	92.7±0.6%	32
NR vs. CL	90.3±1.2%	20
NR vs. MS	92.0±1.3%	27
CL vs. MS	86.4±1.2%	33

Table 2

Multiclass SVM accuracy and aggregate confusion.

	PN	NR	CL	MS	Positive Prediction
PN	4184	98	319	299	85.4±1.0%
NR	357	2748	536	132	72.3±2.5%
CL	245	189	5167	499	84.7±1.8%
MS	208	242	631	4119	79.2±2.2%

Table 3

Enrichment of tumor subtypes in bottom-up consensus clustering. The minimum p-value of either over/under representation is listed with an O or U. Statistically significant results ($p < 0.05$) are highlighted below.

Consensus Cluster	Proneural		Neural		Classical		Mesenchymal	
	% Samples	*p-O*/*p-U*	% Samples	*p-O*/*p-U*	% Samples	*p-O*/*p-U*	% Samples	*p-O*/*p-U*
1	46.2%	**8e-4** O	2.6%	**1.3e-3** U	33.3%	0.40 O	18.0%	0.14 U
2	46.0%	**1.3e-3** O	24.3%	0.24 O	2.7%	**5.8e-6** U	27.0%	0.51 O
3	15.6%	0.15 U	15.6%	0.40 U	43.8%	6.1e-2 O	25.0%	0.54 U
4	9.8%	**4.6e-6** U	25.0%	**3.5e-2** O	35.9%	8.6e-2 O	29.4%	0.20 O