Molecular Correlates of Tumor Signatures from a Large Cohort
Molecular Basis of Heterogeneity and Morphometric Subtypes in Glioblastoma Multiforme (GBM)

Bahram Parvin
Imaging and Bioinformatics Lab
Lawrence Berkeley National Laboratory
Berkeley, Ca 94720
Molecular Correlates of Tumor Signatures from a Large Cohort

i) From a cohort (of whole slide images) to pathway analysis

ii) Tumor heterogeneity
Goals of The Cancer Genome Atlas (TCGA)

- To characterize genomic changes in cancer
- To accelerate understanding of molecular basis of cancer through large datasets
  — Case in point Glioblastoma Multiforme (GBM)
Goals of TCGA

Single mode or Correlative analysis

- Significantly mutated genes?
- Recurrently altered copy numbers?
- Altered pathways?
- Subtypes?
Example

Diagnostic:

- frequent microvasculature
- mild pleomorphism
- high cellularity
Example (continued)

Diagnostic:

- Pleomorphism
Example (continued)

Diagnostic:

- high pleomorphism
- low cellularity
Example (histology is heterogeneous)

Diagnostic:

• transition to necrosis
• high cellularity
Example (histology is heterogeneous)

Diagnostic:
- zonal necrosis
- moderate pleomorphism
- high cellularity
Example (continued)

Diagnostic:
• hemorrhage
• foreign objects
Barriers to the analysis of TCGA datasets

- Histology sections are not purely from solid tumors
  - Stromal, normal, and tumor region may be present
- Batch effect
  - No single lab can produce a large cohort.
    - In addition to biological heterogeneity, technical variations (e.g., staining and fixation) are persistent
A section in KIRC

- **Stroma**
- **Fat**
- **Tumor**
- **Lymphocytes**
Outline

• Motivation?
• Background
• Approach
• Results
• Current efforts
Background

• Three categories of research (Gurcan et al IEEE TBE 2009—a review)
  — Nuclear segmentation and multidimensional representation of tumor cells and association to recurrence, prognosis, … (Demir et al 2009 and others)
    • Issues?
  — Patch-based analysis
    • Kong et al (Saltz Lab) 2010
    • Bilgin et al (Yener Lab)
    • Issues?
  — Recruitment of lymphocytes
    • Madabushi’s Lab, 2010
Approach: from whole mount tissue sections to representation

- Tumor center ontology for managing computational processes
- Multidimensional representation
- Patch level classification
- Nuclear segmentation and organization
- Whole mount tissue sections
Use case and approach

- **Use case:** Glioblastoma multiforme (GBM)
  - Curated by removing tissue sections with artifacts (e.g., fold in tissue, pen mark, scanning anomaly)
  - Sample size
    - 380 whole mount tissue sections selected out of 447
    - 146 patients selected out of 152

- **Challenges?**
  - Technical and biological variations, very large datasets

- **Approach**
  - Development of robust and efficient image analysis algorithms
  - Computing morphometric features and meta-features
  - Subtyping based on selected features
  - Molecular association with morphometric subtypes
Primary analytical components and infrastructure

- A computational workflow for image analysis
- A schema for maintaining complex relationships at multiple scales
- Visualization tools for validation
- Techniques for subtyping and molecular association
- High performance computing — Clusters/cloud
Organization of analytical components

- Remote FTP site
- Consistency Checking between remote and local repositories
- Data Processing modules
  - Multiresolution extraction of strips and blocks from large scale tissue sections
  - JobManager
  - Image analysis
  - Lawrencium and CITRIS clusters
  - Local cluster
  - FeatureImporter
- Visualization modules
  - Multiresolution titling for visualization
  - Apache Image repository based on OMEIS
- Clinical data
- Image registry
- Job queue
- PostgreSQL
  - Multidimensional profiling
  - Tumor specific normalization
  - Tumor subtyping
- Tomcat web server
Image analysis at cellular and patch level of analysis

- Nuclear segmentation
- Patch level classification
  - Creating overlay maps for gated queries
  - Prognostic value
Cell-by-cell segmentation

• Issues
  — How to handle technical variations
    • Fixation, staining, …
  — How to handle biological heterogeneity
    • Cell types, cell size, secretion into background

• Approach
  — Off line:
    • Construct a diverse set of validated reference images and curate them
    • Represent each reference image in shape and RGB space as a Gaussian Mixture models
Cell-by-cell segmentation

• Approach
  — On line processing
    • Normalize each image against all reference images
    • Optimize against the reference dataset
      — Construct a data fitness term
        » Compute a global fitness term against the prior
        » Construct a local probability map
      — Construct a smoothness term
        » Geodesic constraint
    • Validate and separate touching cells
New algorithm enhances nuclear segmentation in the presence of technical variations

Chang et al, ISBI2012, TMI (in revision)
Approach (details)

• Offline:
  — Construct a diverse set of validated reference images and curate them
  — Represent each reference image in RGB and LoG space as a Gaussian Mixture model

Reference 1 → Blue Ratio → LoG

Reference N → Blue Ratio → LoG

Models for Nuclei/Background In RGB Space

Models for Nuclei/Background In LoG Space

Models for Nuclei/Background In RGB Space

Models for Nuclei/Background In LoG Space
Approach (details)

• On line processing
  — Normalize each image against all reference images
  — Optimize against the reference dataset
    • Construct a data fitness term
      — Compute a global fitness term against the prior
      — Construct a local probability map
    • Construct a smoothness term
      — Geodesic constraint
  — Validate and separate touching cells
Approach (on-line processing)
Approach (reducing color dimensionality)

- **Blue Ratio**
  - Reduce complexity for integrating LoG
  - Accentuate nuclear dye

\[
BR(x, y) = \frac{100 \times B(x, y)}{1 + R(x, y) + G(x, y)} \times \frac{256}{1 + B(x, y) + R(x, y) + G(x, y)}
\]

where \(R(x, y)\), \(G(x, y)\), and \(B(x, y)\) are intensity values of pixel \((x, y)\) in R, G, B channels. \(BR(x, y)\) is the blue ratio.

(a) Original Image
(b) Blue ratio
(c) Color Decomposition by [2]

Seed detection provides shape signature and local statistics
Approach (local fitness term)

- Seeds Detection
  - LoG on Blue Ratio
  - Local statistics for both nuclei and background in RGB space
- Local Fitness

1. Initial nuclear seeds: Negative Peaks of LoG
2. Initial background seeds: Positive Peaks of LoG
3. Refinement:
   1. $|\text{LoG}| > \text{Minimum Conservative Threshold}$
   2. Global background threshold, established by global histogram of blue ratio
   3. Local background threshold, established by following figure
Global and local fitness terms

- Data fitness term
  - Global fitness term

\[ E_{gf}(x_p = i) = - \sum_{k=1}^{N} \lambda^k \log(p_i^k(f^k(p))) \]

- Local fitness term

\[ E_{lf}(x_p = i) = -\gamma \log(p_i(f(p))) \]
Approach (separating touching nuclei)

- Geometric Reasoning [3]


Journal of Pattern Recognition, in press
Cell-by-cell segmentation result
Cell-by-cell segmentation result
Validation: multi-reference graph cut (MRGC) performs better

### Binarization

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRGC (Multi-Scale LoG)</td>
<td>0.77</td>
<td>0.82</td>
<td>0.794</td>
</tr>
<tr>
<td>MRGC</td>
<td>0.79</td>
<td>0.78</td>
<td>0.785</td>
</tr>
<tr>
<td>MRGC (Color Feature Only)</td>
<td>0.72</td>
<td>0.83</td>
<td>0.771</td>
</tr>
<tr>
<td>MRGC (Global Fitness Only)</td>
<td>0.80</td>
<td>0.71</td>
<td>0.752</td>
</tr>
<tr>
<td>Our Previous approach</td>
<td>0.78</td>
<td>0.65</td>
<td>0.709</td>
</tr>
<tr>
<td>MCV</td>
<td>0.69</td>
<td>0.75</td>
<td>0.719</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.59</td>
<td>0.76</td>
<td>0.664</td>
</tr>
</tbody>
</table>

### Segmentation

<table>
<thead>
<tr>
<th>Approach</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRGC</td>
<td>0.75</td>
<td>0.85</td>
<td>0.797</td>
</tr>
<tr>
<td>Our previous approach</td>
<td>0.63</td>
<td>0.75</td>
<td>0.685</td>
</tr>
</tbody>
</table>
Delineation of nuclei and vasculature
Cellularity is quantified

Chang et al
BMC Bioinformatics
2011
Patch level analysis

• Evolution
  — 1) Train with texture and color, represent with GMM, and enforce continuity by graph cut
  — 2) Train and classify
    • Compressible sensing, KDA, SVM (ISBI 2011)
  — 3) Subspace learning (ISBI 2012)
Classification through spatial grouping of low level features

Chang et al, ISBI2010
Classification through kernel and other sparse coding

Han et al, ISBI2011

<table>
<thead>
<tr>
<th></th>
<th>Sparse</th>
<th>KDA</th>
<th>SVM linear</th>
<th>SVM quad</th>
<th>SVM rbf</th>
<th>SVM poly</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-class</td>
<td>90.05</td>
<td>99.12</td>
<td>94.27</td>
<td>91.89</td>
<td>91.44</td>
<td>88.22</td>
</tr>
<tr>
<td>4-class</td>
<td>85.87</td>
<td>79.65</td>
<td>90.03</td>
<td>83.47</td>
<td>81.84</td>
<td>82.58</td>
</tr>
</tbody>
</table>
Independent subspace analysis

• Learn invariant kernels from underlying spatial distribution
  — A variant of Convolution Networks (ISBI 2012, LeCun)
• SVM
Independent Space Analysis

ICA: independent component analysis

ISA: independence between groups of components

Group sparse coding: \( \sqrt{.} \)

Sparse coding: \( (. )^2 \)

\[ P_1 \quad \quad P_2 \]
Representation at multiple scales

Structural features

Cell-by-cell measurement

A vector

A multidimensional distribution

Normalization across all tissue sections
Coarse level information (summary)

Tissue Pathology: cellularity, apoptosis rate, ...
Fine level information (cell-by-cell)

Chang et al BMC Bioinformatics, 2011
Han et al IEEE TCBB 2010
Parvin et al Systems Biology 2005
What are subtypes based on cellularity and nuclear size at the patient level

- 2 clusters
- 3 clusters
- 4 clusters
- 5 clusters
- 6 clusters
Methods for molecular association

• Two approaches:
  — univariate analysis
    • Does not (i) incorporate interactions among genes, (ii) incorporate prior knowledge, …
  — Multivariate analysis
    • Sparse association
      — With and without prior knowledge
    • Random forest
Linking subtypes to molecular data through sparse regression

Basic Regression Model

$$\min_{T \in \mathbb{R}^{N \times M}} \|X T - Y\|_F^2 + \lambda \|T\|_1$$
What is the distribution of each subtype and how well each subtype predicts survival as a function of treatment?

Subtyping based on cellularity

Chang et al, BMC Bioinformatics, 2011

Subtyping based on total DNA content is also predictive

Subtype 2

Chang et al, BMC Bioinformatics, 2011
Molecular association
Enrichment analysis

• Pathway enrichment
  — STAT Signaling
• Subnetwork enrichment

Can tumor composition be characterized?

- Since tumor is heterogeneous, can we query for subtypes at the **block levels** and learn about tumor composition?
What are the tumor histology subtypes?

Subtype 1

Subtype 2

Subtype 3

Subtype 4
Does heterogeneity play a role in survival as a result of a more intense therapy?

Loosely defined semantics of high and low!

- **Low cellularity**
  - High heterogeneity

- **High cellularity**
  - Low heterogeneity

![Graphs showing survival ratio over time for more intensive and less intensive therapies.](image-url)
Another view: Are cellularity and nuclear size predictive?

High cellularity and low nuclear size are better predictive of a more aggressive therapy.
Molecular basis of heterogeneity (FDR <0.15 with p value 0.05)

Han et al, in preparation

Han et al, ISBI 2012
Molecular basis of heterogeneity (FDR <0.15 with p value 0.05)

Zhu, T., et al., Endothelial cells create a stem cell niche in Glioblastoma by providing NOTCH ligands that nurture self-renewal of cancer stem-like cells. Cancer Research

Summary

• Large cohorts can be predictive in terms of outcome and relevant pathways

• There are many ways to slice through the data and metadata
  — Cellularity, nuclear size, cytoplasm feature
  — Heterogeneity (angiogenesis)

• Different indices lead to alternative subtyping
  — Alternative biological interpretation is possible

• Web site: tcga.lbl.gov
  — “Google map” like viewing of tissue sections with segmentation results overlaid
Acknowledgements

• Parvin Lab
  — Cemal Bilgin
  — Hang Chang
  — Ge Cong
  — Gerald Fontenay
  — Ju Han
  — Joel Hyman
  — Mina Lee
  — Leo Loss
  — Vivek Menon
  — Jeff Wang
  — Kai Zhang

• Collaborators
  — Paul Spellman (OHSU)
  — Rick Baehner (GHI)
  — Alexander D. Borowsky (UCD)
  — Joe Gray (OHSU)
  — TCGA

• Cluster computing
  — LBNL IT Division

• Funding
  — NIH/NCI (U24 CA1437991)
  — NIH/NCI (RO1 CA140663)
2013 IEEE International Symposium on Biomedical Imaging: From Nano to Macro
April 7-11, 2013 • San Francisco, California

Call for Papers

Bioimaging has continued to evolve across a wide spectrum of applications from diagnostics to personalized therapy to mechanistic understanding of biological processes. The field continues to be challenged with ever increasing demand on more robust methods and integration with clinical and molecular data.

The 2013 IEEE International Symposium on Biomedical Imaging will be held April 7-11, at the Westin Hotel in San Francisco, CA. ISBI is a joint initiative of the IEEE Signal Processing Society (SPS) and IEEE Engineering in Medicine and Biology Society (EMBS).

ISBI has a strong tradition of fostering integration between different imaging communities and of contributing to an integrative imaging approach accessing all scales of observation. We are soliciting all aspects of biological and medical imaging from mathematical image reconstruction for medical applications to integration of image-based modeling with genomic data for hypothesis generation and network inference.

The meeting will feature preconference tutorials, followed by a strong scientific program of plenary talks and special sessions as well as oral and poster presentation of peer-reviewed contributed papers.

For more information: www.biomedicalimaging.org or isbi2013-info@ieee.org

INVITED SPEAKERS

SIMON CHERRY
University of California, Davis

DAVID DONOHO
Stanford University

JOE GRAY
Oregon Health Sciences Institute

JOHN SEDAT
University of California, San Francisco

STEPHEN WONG
Methodist Hospital Research Institute

General Chair: Rahim Parvin, Lawrence Berkeley National Laboratory • Program Chairs: Michael Lustig, University of California, Santa Barbara; Lawrence Snell, Yale University • Special Sessions: Robert F. Murphy, Carnegie Mellon University • Tutorials: Jitka Kosecovic, Carnegie Mellon University • Finance Chair: In Han, LBNL, Local Arrangements: Hang Chung, Lawrence Berkeley National Laboratory