Informatics links between histological features and genetics in cancer

Kun Huang, PhD

Department of Biomedical Informatics
The Ohio State University Wexner Medical Center
Data Integration

- Integrative genomics / trans-omics approach

  Behavior, syndrome and clinical outcome (EMR)

  Morphology (sub-cellular, cellular, tissue, organ)

  Proteome (profiling, quantity, modification)

  Transcriptome (gene expression, non-coding RNA)

  Epigenetics (DNA methylation, histone modification, microRNA)

  Genotype (SNV, CNV, structural variation)
Team

Huang

Machiraju

Jeff Baumes, Kitware

Jie Zhang

David Manthey, Kitware
Leveraging

- NCI CPTAC Contract
  - Integrate proteomics data from CPTAC project
  - High performance computing (GPU and cluster)
Tasks

Aim 1 – Develop software libraries for integrative genomics in cancer research, specifically for integrating genomic, histological images and clinical data for cancer biomarker discovery and subtyping.

Aim 2 – Develop an integrative and expandable open source platform for managing, analyzing, and integrating multiple data types in integrative genomics for cancer with visual analytic capabilities for cancer biomarker discovery.

Aim 3 – Test the completed software platform with cancer systems biology studies and build an ecosystem based on the open source framework for integrative genomics and in particular for imaging genomics in cancer.
From Correlation to Integration

- **Pathology**
- **Proteome**
- **Transcriptome**

**Correlative Study**

- **Pathology**
- **Transcriptome**
- **Epigenome**

**Model-based Study**

- **Clinical Attributes**
- **Transcriptome**

**Integrative Clustering**

**Applications**

- **Triple Negative Breast Cancer**
- **Breast Cancer Proteomics**

- **Lung Adenocarcinoma**

**Breast Cancer Patient Stratification**

- **A two-step algorithm**
- **Molecular regularized algorithm**
Pipeline Overview

1. Whole Slide Images
2. Representative Patches
3. Image Preprocessing
4. SuperPixel Segmentation
5. LBP Features
6. Tissue Classification
7. Cell Segmentation
8. Feature Extraction

Epithelial Features
Stromal Features
Feature Extraction

Image Features

Epithelial Features
- Discovered Salient Features
  - Layer Features
  - Geometry Features

Stromal Features
- Texture Features
- Pathological Features
  - Haralick Features
  - Density
  - Fraction
  - Shape

E: epithelial features; S: stromal features

E-28; S-6
E-14; S-6
E-20; S-20
E-2; S-2
E-2; S-1
E-8; S-8
Image Analysis

Whole-slide image

Image tiles (40X magnification)

Classification map

Computation units
- CPU
- SSE
- CPU
- C/C++
- GPU
- Intel Xeon Phi

Assign classification labels
- Label 1
- Label 2
- background
- undetermined

Catalyurek
Patient Stratification on Morphology

Hierarchical clustering of breast cancer patients based on imaging features.

“High malignancy” group

“Low malignancy” group
Associated Protein Co-Expression Network

- Identified 124 protein co-expression networks
- 4 protein networks differentially expressed

<table>
<thead>
<tr>
<th>Protein Networks</th>
<th>Proteins in the Networks</th>
<th>Cytoband</th>
<th>pValue</th>
<th>Enriched Cytobands</th>
<th>Top GO Molecular Functions (pValue)</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MYOCD</td>
<td>17p11.2</td>
<td>3.880E-2</td>
<td>16p13 P-value = 2.170E-5</td>
<td>smooth muscle cell differentiation(8.184E-7); muscle cell differentiation(1.889E-4); regulation of transforming growth factor beta receptor signaling pathway.</td>
<td>Enriched in 16q13</td>
</tr>
<tr>
<td></td>
<td>MKL1</td>
<td>22q13</td>
<td>5.986E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MKL2</td>
<td>16p13.12</td>
<td>2.306E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLYWCH2</td>
<td>16p13.3</td>
<td>2.170E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CISD3</td>
<td>17q12</td>
<td>1.439E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HN1L</td>
<td>16p13.3</td>
<td>2.170E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STUB1</td>
<td>16p13.3</td>
<td>2.170E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SUCLG1</td>
<td>2p11.2</td>
<td>1.536E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DNA Copy Number Variance

Protein Networks

1. Proteins in the Networks
   - MYOCD
   - MKL1
   - MKL2
   - FLYWCH2
   - CISD3
   - HN1L
   - STUB1
   - SUCLG1

**MKL1**
- R square = 0.26018
- PCC = 0.51008

**MKL2**
- R square = 0.28307
- PCC = 0.53205

**STUB1**
- R square = 0.33351
- PCC = 0.5775

**CISD3**
- R square = 0.28342
- PCC = 0.53237
STUB1 gene as a hub gene
Prognosis Validation

Kaplan–Meier survival curves of prognostic model on multiple public breast cancer datasets. From left to right: Wang, Perou and NKI data respectively.

Wang

Perou

NKI
Molecular Subgroup 1

Molecular Subgroup 2

Molecular Subgroup 3

Sample $d_i$

Sample $d_j$

DENSE

NOT DENSE

Wang et al, Methods, 2013
The Algorithm

Algorithm 1: Molecular Regularized Consensus Patient Stratification

**Data:** Similarity Matrix $\tilde{S}$, Molecular Density Weight Matrix $W$, the number of clusters in final consensus $k$, MaxIter, precision $\epsilon$

**Result:** Cluster indicator matrix $U$.

initialize $\tilde{U}^{(1)} > 0$, $t = 1$, $\Delta = +\infty$;

while $t < $ MaxIter and $\Delta > \epsilon$ do

Update $\tilde{U}_{ij}^{(t+1)} \leftarrow \tilde{U}_{ij}^{(t)} \sqrt{\frac{[(W \circ S)\tilde{U}D]_{ij}}{[(W \circ \tilde{U}D\tilde{U}^T)\tilde{U}D]_{ij}}}$;

Update $D_{ij}^{(t+1)} \leftarrow D_{ij}^{(t)} \sqrt{\frac{[\tilde{U}^T(S \circ W)\tilde{U}]_{ij}}{[\tilde{U}^T(\tilde{U}D\tilde{U}^T \circ W)\tilde{U}]_{ij}}}$;

Compute $\Delta = \| \tilde{S} - W \circ (\tilde{U}D\tilde{U}^T) \|_F^2$;

$t = t + 1$;

end

Discretize $\tilde{U}$ to binary membership matrix.

**Algorithm 1:** Molecular Regularized Consensus Patient Stratification

Prognosis

Histology Type

Tumor Grade

Disease Stage

BCE

HGPA

CSPA

RCP

Visualization

Ding et al, BMC Bioinformatics, 2014
Interactive Patient Stratification

Ding et al, BMC Bioinformatics, 2014
Graphie – Visual Analytics of Imaging Features

Refined features

GRAPHIE: graph based histology image explorer
Hao Ding¹, Chao Wang¹, Kun Huang², Raghu Machiraju²
From 5th Symposium on Biological Data Visualization
Dublin, Ireland, 10-11 July 2015
Graphie – Feature Selection

Examine feature distinctiveness for groups of images.

- Student’s t-test
- Boxplots
- Re-generate with selected feature subset
Combining Graphie, iGPSE, & SUMO

Histology images

Genome

Epigenome

Proteome

Subtype Analysis
Integrative Biomarkers
Molecular behaviours

...
State of (O)SUMO

iGPSe: Interactive Genomics Patient Stratification explorer

Upload Dataset
Upload the Gene expression profile
Choose File No file chosen

Upload the MicroRNA expression profile
Choose File No file chosen

Upload the Clinical profile
Choose File No file chosen

Gene expression
null

microRNA
null

Clinical
null
SUMO Architecture

- **SUMO (workflow management, UI)**
  - iGPSe
  - Graphie

- **Girder (authentication, data management)**
  - Girder Worker (analysis execution)

- **Adding workflows**
  - JSON metadata for tasks
  - R/Python source

- **Source code snippet**
  ```python
  blur_image = {
    "inputs": [
      {
        "name": "blur_input", "type": "image", "format": "plt"},
      {
        "name": "blur_radius", "type": "number", "format": "number"}
    ],
    "outputs": [{
      "name": "blur_output", "type": "image", "format": "plt"},
      "script": ""
    }
  }
  from PIL import ImageFilter
  blur_output = blur_input.filter(ImageFilter.GaussianBlur(blur_radius))
  ```
Basic Workflow - Silhouette Plot

**Inputs**

Any Girder-uploaded dataset, access controlled

**Outputs**

May be static R charts or interactive D3 plots

Outputs are stored in user’s Girder space

---

**R script**

```r
# R script

# Basic Workflow - Silhouette Plot

# Inputs

# Any Girder-uploaded dataset, access controlled

# Outputs

# May be static R charts or interactive D3 plots

# Outputs are stored in user’s Girder space

```

---

**Input / output descriptions**

- **Inputs**
  - Key: `silhouette_input_path`
    - Name: `MicroRNA Expression Profile`
    - Description: "MicroRNA data.
    - Type: file
    - Subtype: data
    - Notes: "This must be a CSV file with one column per subject, one header row, and one data row per microRNA.
    - Notes: "Number of Clusters (k)"
      - Name: `Number of Clusters (k)`
      - Description: "The number (k) of clusters used in calculations.
    - Type: integer
    - Default: 2
    - Notes: "Clustering is performed via k-means.

- **Parameters**
  - Key: `target_threshold`
    - Name: `Target Threshold`
    - Type: integer
    - Default: 1

- **Outputs**
  - Key: `silhouette_plot`
    - Name: `Silhouette Plot`
    - Description: "Silhouette plot.
    - Type: image
    - Subtype: jpeg
    - Notes: "See Silhouette plot.
      - Notes: "Silhouette of pairs (k - 1) microRNAs, k = h.
    - Notes: "Average silhouette width.

---

**Outputs**

May be static R charts or interactive D3 plots

Outputs are stored in user’s Girder space
2-Stage Workflow - iGPSe

Stage 1 Inputs
Genetic and clinical datasets

Stage 1 Outputs
Two static R cluster heatmaps
One interactive D3 set selection that serves as Stage 2 input

Stage 2 Output
Survivability plot for selected sets

Workflow link between stages

JSON metadata for Stage 1
R source for Stage 1

JSON metadata for Stage 2
R source for Stage 2
SUMO

Initial application available at http://osumo.org

Plan for more machine learning algorithms (consensus learning, multiview learning) for data integration on SUMO

Pan-cancer study as driven application

We welcome

✧ Beta users
✧ Use cases and workflows to deploy
✧ Open-source contributions https://github.com/osumo/osumo
Demonstration

- https://www.youtube.com/watch?v=d96h4EnwRxY
More Comprehensive Workflow Proposed