Biomedical Data Science for Precision Medicine

Vanathi Gopalakrishnan, PhD
Associate Professor of Biomedical Informatics, Intelligent Systems and Computational & Systems Biology
Director, PRoBE Laboratory for Pattern Recognition from Biomedical Evidence
Department of Biomedical Informatics
School of Medicine
University of Pittsburgh

PRoBE Lab: Research Interests
Overall Goal: To Accelerate Biomedical Knowledge Discovery

PRoBE Lab: Research Interests
Overall Goal: To Accelerate Biomedical Knowledge Discovery

“BIG” DATA IN BIOMEDICINE

• “Omic”-driven
  – Genotyping
  – Gene expression
  – Next-gen sequencing
  – Epigenomic
  – Proteomic
  – Metabolomic
  – Exposomic
  – Lipidomic

• EHR-driven
  – Demographics
  – Medical history
  – Medication
  – Allergies
  – Lab test results
  – Radiology images
  – Billing information

Source: http://www.slideshare.net/TWilckens/inn-ventis-precision-medicine2014
DATA WE WORK WITH

- Proteomic
  - Mass Spectral (large/wide/deep)
  - Immunoassay
- Genomic
  - GWAS/SNPs (very large)
  - DNA Methylation
  - Gene Expression
  - microRNA
- Images
  - Cardiac MRI
  - Brain fMRI

From biomarker* discovery studies for early detection of:
- Amyotrophic Lateral Sclerosis
- Alzheimer’s
- Lung Cancer
- Breast Cancer
- Esophageal Cancer
- Inflammatory Bowel Diseases
- Ulcerative Colitis, Crohn’s
- Coronary Artery Disease
- Pediatric Cardiomyopathy

*The term “biomarker” or “biological marker”, refers to “a broad subcategory of medical signs—that is, objective indications of medical state observed from outside the patient—which can be measured accurately and reproducibly. Medical signs stand in contrast to medical symptoms, which are limited to those indications of health or illness perceived by patients themselves.” - Kyle Strimbu, Jorge A. Tavel. What are Biomarkers? Curr Opin HIV AIDS. 2010 Nov 5(6):463–466.

Biomedical Data Science

Source: https://www.samsi.info/workshop/interdisciplinary-approaches-biomedical-data-science-challenges-samsi-innovations-lab-july

Source: https://deltamodelling.com/consumer_research

TODAY’s TALK

Case Study 1

Case Study 2

'omic' data

Image-derived data

Same underlying predictive modeling process, different pre-processing of data/images/text

Classification Rules

IF-condition

THEN-consequent
Biomarker Discovery

- **Goal:**
  - Find a panel of measurable markers that accurately predict or classify disease (health) state
- **Why?**
  - Early detection of disease -> effective treatment
  - Identifying disease progression -> effective intervention
  - Elucidating disease mechanism -> new treatment options
- **Key Challenges:** Insufficient, Noisy data

Outline of strategies for biomarker discovery through utilization of emerging technologies


What the data looks like...

<table>
<thead>
<tr>
<th></th>
<th>Marker 1</th>
<th>Marker 2</th>
<th>Marker 3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>9.1</td>
<td>16.2</td>
<td>100.5</td>
<td>Cancer</td>
</tr>
<tr>
<td>Sample 2</td>
<td>6.2</td>
<td>8.3</td>
<td>50.7</td>
<td>Control</td>
</tr>
</tbody>
</table>


ProB: Prognostic Biodata Engine applied to cancer patient management. ProB is an automated analysis of serum for the prediction of disease stage, disease, or outcomes. Once the disease has been diagnosed and the patient’s treatment is confirmed, a protein modification pattern recognition searching tool can be used to identify individual protein candidates that are associated with the specific state of the patient in the network. The candidate therapy or bioactive factors are then tested in the individual patient.
Case Study Problem Overview: Clinical Proteomics

- Major goal:
  - Identify disease-specific biomarker panels
  - from mass spectral analyses of body fluids
  - to facilitate early detection of disease

- Body fluids:
  - serum, cerebrospinal fluid (CSF)

- Searching for:
  - Important proteins and peptides associated with disease states

Overview: Process and Challenges

- Biological samples from cases & controls
- Sample processing/mass spectral analyses
- Signal/Image Processing Analytics
- Predictive Modeling/Data Analytics
- Disease early detection test
- Actionable Knowledge
- Human-understandable models

Lung Cancer Early Detection

- Clinical History
- CT screen Results
- Blood Test

RISK PREDICTION MODEL

Accuracy of early detection
- # Unnecessary biopsies

Multiplexed serologic
Quantitative immunoassays
Luminex xMAP® technology

Ten marker panel: JTO paper

- Marker measurements for 139 controls and 66 cases (replicates averaged)

RL

- 10-marker panel

RISK PREDICTION MODEL

Accuracy of early detection
- # Unnecessary biopsies

Multiplexed serologic
Quantitative immunoassays
Luminex xMAP® technology

Luminex® assays on blinded
set of 30 cases, 30 controls + RL modeling

Classifier with sensitivity 73.3%, specificity 93.3%
Best performance on stage I and II cases
85% SN, Balanced Accuracy 89.2%
(3 mistakes out of 20)
Our Lab has developed:

- **RL-Wrap**
  - Wrapper based rule learner
- **EPO-KB**
  - Empirical Proteomics Ontology Knowledge Base
- **EBD**
  - Efficient Bayesian Discretization
- **BRL**
  - Bayesian Rule Learning
- **TRL**
  - Transfer Rule Learning
- **PAIFE**
  - Partitioning-based Adaptive Irrelevant Feature Eliminator

**Markers and their discretized ranges**

<table>
<thead>
<tr>
<th>Marker</th>
<th>LOW</th>
<th>MEDIUM</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL5</td>
<td>≤ 11.853</td>
<td>&gt; 11.853 &amp; ≤ 36.702</td>
<td>&gt; 36.702</td>
</tr>
<tr>
<td>CRP</td>
<td>≤ 27.430</td>
<td>&gt; 27.430</td>
<td></td>
</tr>
<tr>
<td>HGF</td>
<td>≤ 0.282</td>
<td>&gt; 0.282 &amp; ≤ 0.819</td>
<td>&gt; 0.819</td>
</tr>
<tr>
<td>MIF</td>
<td>≤ 4.897</td>
<td>&gt; 4.897</td>
<td></td>
</tr>
<tr>
<td>PRL</td>
<td>≤ 8.305</td>
<td>&gt; 8.305 &amp; ≤ 17.376</td>
<td>&gt; 17.376</td>
</tr>
<tr>
<td>SAA1</td>
<td>≤ 29.238</td>
<td>&gt; 29.238</td>
<td></td>
</tr>
<tr>
<td>SERPINE1</td>
<td>≤ 220.274</td>
<td>&gt; 220.274</td>
<td></td>
</tr>
<tr>
<td>TTR</td>
<td>≤ 108.75</td>
<td>&gt; 108.75 &amp; ≤ 175.947</td>
<td>&gt; 175.947</td>
</tr>
</tbody>
</table>

**TRL model for lung cancer case-control discrimination: New Data Set, 10 markers as seed**

1. IF((PRL = HIGH)) THEN (Class = Case)
   - CF = 0.993, P = 0.0, TP = 142, FP = 1
2. IF((PRL = LOW)) THEN (Class = Control)
   - CF = 0.966, P = 0.0, TP = 287, FP = 10
3. IF((TTR = HIGH)) THEN (Class = Control)
   - CF = 0.899, P = 0.0, TP = 142, FP = 16
4. IF((CCL5 = LOW)) THEN (Class = Case)
   - CF = 0.88, P = 0.0, TP = 22, FP = 3
5. IF((CRP = HIGH)) THEN (Class = Case)
   - CF = 0.925, P = 0.0, TP = 37, FP = 3
6. IF((SA1 = HIGH) AND (TTR = LOW)) THEN (Class = Case)
   - CF = 0.972, P = 0.0, TP = 35, FP = 1
7. IF((MF = HIGH) AND (TTR = LOW)) THEN (Class = Case)
   - CF = 0.939, P = 0.0, TP = 46, FP = 3
8. IF((MF = HIGH) AND (SA1 = HIGH)) THEN (Class = Case)
   - CF = 0.938, P = 0.0, TP = 12, FP = 1
9. IF((MF = HIGH) AND (PRL = MEDIUM)) THEN (Class = Case)
   - CF = 0.923, P = 0.0, TP = 123, FP = 14
10. IF((ERBB2 = HIGH) AND (SERPINE1 = HIGH)) THEN (Class = Control)
    - CF = 0.898, P = 0.0, TP = 123, FP = 14
11. IF((ERBB2 = MEDIUM) AND (TTR = LOW)) THEN (Class = Case)
    - CF = 0.868, P = 0.0, TP = 33, FP = 5

**Markers discretized ranges**

<table>
<thead>
<tr>
<th>Marker</th>
<th>LOW</th>
<th>MEDIUM</th>
<th>HIGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCL5</td>
<td>≤ 11.853</td>
<td>&gt; 11.853 &amp; ≤ 36.702</td>
<td>&gt; 36.702</td>
</tr>
<tr>
<td>CRP</td>
<td>≤ 27.430</td>
<td>&gt; 27.430</td>
<td></td>
</tr>
<tr>
<td>HGF</td>
<td>≤ 0.282</td>
<td>&gt; 0.282 &amp; ≤ 0.819</td>
<td>&gt; 0.819</td>
</tr>
<tr>
<td>MIF</td>
<td>≤ 4.897</td>
<td>&gt; 4.897</td>
<td></td>
</tr>
<tr>
<td>PRL</td>
<td>≤ 8.305</td>
<td>&gt; 8.305 &amp; ≤ 17.376</td>
<td>&gt; 17.376</td>
</tr>
<tr>
<td>SAA1</td>
<td>≤ 29.238</td>
<td>&gt; 29.238</td>
<td></td>
</tr>
<tr>
<td>SERPINE1</td>
<td>≤ 220.274</td>
<td>&gt; 220.274</td>
<td></td>
</tr>
<tr>
<td>TTR</td>
<td>≤ 108.75</td>
<td>&gt; 108.75 &amp; ≤ 175.947</td>
<td>&gt; 175.947</td>
</tr>
</tbody>
</table>

**TRL performance**

<table>
<thead>
<tr>
<th>Model</th>
<th>SN %</th>
<th>SP %</th>
<th>BACC %</th>
<th>ACC %</th>
<th>Model size</th>
<th>Variables used</th>
<th>Variables retained</th>
<th>Rules retained after transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL</td>
<td>91.6</td>
<td>96.3</td>
<td>93.908</td>
<td>94.561</td>
<td>12</td>
<td>ERBB2, CRP, CCL5, MIF, TTR, SAA1, SERPINE1, PRL</td>
<td>7 of the 8 markers were in JTO model</td>
<td>8/12 Prior rules are reported as p# in TRL model</td>
</tr>
<tr>
<td>TRL</td>
<td>91.6</td>
<td>96.3</td>
<td>93.908</td>
<td>94.561</td>
<td>12</td>
<td>ERBB2, CRP, CCL5, MIF, TTR, SAA1, SERPINE1, PRL</td>
<td>7 of the 8 markers were in JTO model</td>
<td>8/12 Prior rules are reported as p# in TRL model</td>
</tr>
</tbody>
</table>
Transfer Learning

- [http://www.youtube.com/watch?v=osTXwmyxBVM](http://www.youtube.com/watch?v=osTXwmyxBVM)

**Our Hypothesis**

Transfer Learning of classification rules between related source and target data sets will improve performance on target data, compared to baseline learning without transfer.

**Illustration of the TRL algorithm**

- First application of transfer learning to biomarker discovery
- Previous evidence from reproducibility studies suggests that it is possible to find common information
- Past transfer learning methods typically produce classifiers that are difficult to understand (or use all variables)
  - TRL transfers knowledge in the form of interpretable, modular rules or rule structure
  - Variable selection is embedded in the learning algorithm
Whole-rule and Rule Structure Transfer

Whole-rule transfer: impose target discretization on source

<table>
<thead>
<tr>
<th>Source data set</th>
<th>Target data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A High</td>
<td>A High</td>
</tr>
<tr>
<td>B Middle</td>
<td>B High</td>
</tr>
<tr>
<td>C Low</td>
<td>C Low</td>
</tr>
<tr>
<td>A Low</td>
<td>A Low</td>
</tr>
<tr>
<td>B Middle</td>
<td>B Low</td>
</tr>
<tr>
<td>C High</td>
<td>C High</td>
</tr>
</tbody>
</table>

RULE STRUCTURE TRANSFER
Example prior rule:
IF (MZ_7.23 = High) THEN (Group = Cancer)

The rule is then instantiated from the target data discretization as:
IF (MZ_7.23 = High) THEN (Group = Cancer)
IF (MZ_7.23 = Low) THEN (Group = Cancer)
IF (MZ_7.23 = High) THEN (Group = Normal)
IF (MZ_7.23 = Low) THEN (Group = Normal)

Useful because:
• measurements might not correspond for the variable between source & target
• prior rule can come from literature or other sources

Questions
• What is a biomarker?
• What are some examples of biomedical big data?
• How can predictive models be developed from these data?
• What are some challenges in biomedical data science and how has the PRoBE lab overcome these?

Another novel solution

Bayesian Rule Learning (BRL)
• Quantifies uncertainty in the validity of rule models
• Learns parsimonious models, interpretable rules
• Hybrid system combines learning of constrained Bayesian networks (BNs) with rule-based inference
• BN model uncertainty in data well, though inference can be intractable
  • Dr. Cooper and colleagues developed K2 score for evaluating goodness of fit of BN model to data

Hypothesis: BRL system is useful for disease classification and biomarker discovery

Bayesian rule learning for biomedical data mining
Vanathi Gopalakrishnan*, Jonathan L. Lustgarten, Shyam Visweswaran and Gregory F. Cooper
Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue Suite M-163, Pittsburgh, PA 15260, USA

Data and text mining

Modified in BRL
**Approach**

- Develop BRL algorithms and extensions for disease classification and biomarker discovery
- Apply BRL methods to biomarker discovery datasets to obtain models for classifying cases and controls for diverse diseases
- Validate BRL models on new sets of retrospectively obtained case-control samples

---

**Impact:**

**BRL models for EAC prediction**

**Provisional Patent Application Filing**

  - Discovery set of 32 patients, Validation set of 67
  - BRL predictive model (Disc->Val) AUROC = 86%
  - BRL combined data ten-fold cross-val AUROC = 93%

---

**BRL applied to Breast Cancer Case-Control discrimination**

Discovery of biomarkers for breast cancer case-control discrimination from LC-MS-MS analyses of samples from the Walter Reed National Military Medical Center using two in-house developed methods, NBR and BRL.

### Table 3: Functional analysis of 8 of the 11 genes

<table>
<thead>
<tr>
<th>Functions Annotation</th>
<th>P-value</th>
<th>Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholangiocarcinoma</td>
<td>4.27E-06</td>
<td>A1BR.A1M1.DPRKNG1</td>
</tr>
<tr>
<td>Liver cancer</td>
<td>6.08E-04</td>
<td>A1BR.A1M1.DPRKNG1,G1N1</td>
</tr>
<tr>
<td>Digestive organ tumor</td>
<td>1.45E-03</td>
<td>A1BR.A1M1.DPRKNG1,G1N1,1FT1H4.KNG1</td>
</tr>
<tr>
<td>Prostatic intraepithelial neoplasia</td>
<td>2.40E-03</td>
<td>G1PX3,1SERP1FNG1</td>
</tr>
<tr>
<td>Serous ovarian carcinoma process</td>
<td>3.18E-03</td>
<td>G1PX3,1SERP1FNG1</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>1.06E-02</td>
<td>G1PX3,1FT1H4,1SERP1FNG1</td>
</tr>
<tr>
<td>Mucinous ovarian cancer</td>
<td>1.53E-02</td>
<td>G1PX3</td>
</tr>
<tr>
<td>Clear-cell ovarian carcinoma</td>
<td>1.60E-02</td>
<td>G1PX3</td>
</tr>
<tr>
<td>Incidence of liver tumor</td>
<td>1.74E-02</td>
<td>1FT1H4</td>
</tr>
<tr>
<td>Acinar-cell carcinoma</td>
<td>1.85E-02</td>
<td>A1ZIPI</td>
</tr>
<tr>
<td>Renal cancer</td>
<td>2.22E-02</td>
<td>G1PX3</td>
</tr>
<tr>
<td>Pancreatic duct adenocarcinoma</td>
<td>3.54E-02</td>
<td>A1ZIPI,1FT1H4</td>
</tr>
<tr>
<td>Carcinoma</td>
<td>3.93E-02</td>
<td>C1Q8,1G1PX3,1FT1H4,1SERP1FNG1</td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>4.03E-02</td>
<td>KNG1</td>
</tr>
</tbody>
</table>

Outreach: BRL applied to pediatric cardiomyopathy classification from cardiac MRI-derived markers


BRL Project: Impact on nextGen

What next: Ensemble Classification with BRL (ecBRL)


Ensemble classification with BRL aims to increase prediction accuracy via selection and combination of multiple base classifiers as opposed to a single “best” base classifier. Scientific Challenges Include: how to combine predictions and model average heterogeneous types of models, such as linear and non-linear. Also, how to retain interpretability.

Literature mining for putative biomarkers

- “Semi-automated Literature Mining to Identify Putative Biomarkers of Disease from Multiple Biofluids”- Rick Jordan, Shyam Visweswaran, Vanathi Gopalakrishnan
- Looked at 5.3 million abstracts from PubMed to identify putative biomarkers of lung and breast cancer across multiple (14 separate) biofluids.
- The frequency of occurrence of biomarker-disease-biofluid association in these abstracts will be used as informative priors for BRL/ecBRL.

TCGA Data Modeling – New Work with doctoral student Arturo Lopez Pineda

Novel application of Junction Trees to the Interpretation of Epigenetic Differences among Lung Cancer Subtypes

Arturo López Pineda, MS, Vanathi Gopalakrishnan, PhD

Won 2015 Marco Ramoni Distinguished Paper Award at AMIA TBI
Acknowledgements

PRoBE Lab Graduates, Members and Collaborators, particularly:
- Jonathan L. Lustgarten, PhD
- Philip Ganchev, PhD
- Jeya Balasubramanian, MS
- Henry Ogoe, MS, ABD
- Shyam Visweswaran, MD, PhD
- Gregory F. Cooper, MD, PhD
- William L. Bigbee, PhD

[Logos of National Institutes of Health, National Library of Medicine, and National Science Foundation]