Modeling Biological Systems and Analyzing Large-Scale Data Sets

ilya shmulevich
TCGA Data Types

25* forms of cancer

- glioblastoma multiforme (brain)
- squamous carcinoma (lung)
- serous cystadenocarcinoma (ovarian)

Biospecimen Core Resource with more than 150 Tissue Source Sites

6 Cancer Genomic Characterization Centers

3 Genome Sequencing Centers

7 Genome Data Analysis Centers

Data Coordinating Center

Multiple data types

- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic report/images
- Tissue anatomic site
- Surgical history
- Gene expression/RNA sequence
- Chromosomal copy number
- Loss of heterozygosity
- Methylation patterns
- miRNA expression
- DNA sequence
- RPPA (protein)
- Subset for Mass Spec
TCGA Research Network
Heterogeneous data
Clinical variables contributing to tumor aggressiveness


<table>
<thead>
<tr>
<th>Clinical variables</th>
<th>Less Aggressive</th>
<th>More Aggressive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distant Metastasis</td>
<td>M0=No</td>
<td>M1=Yes</td>
</tr>
<tr>
<td>Tumor Stage</td>
<td>Early (I-II)</td>
<td>Late(III-IV)</td>
</tr>
<tr>
<td>Fraction Lymph Nodes Positive by H &amp; E</td>
<td></td>
<td>0 – 100 %</td>
</tr>
<tr>
<td>Lymphatic Invasion Present</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Vascular Invasion Present</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Histological Type</td>
<td>Mucinous</td>
<td>Non-mucinous</td>
</tr>
</tbody>
</table>

[Diagram showing DNA mutations, copy-number and structural variations, gene expression (mRNA), DNA methylation, and microRNA expression]
Voltage-gated Na+ channel SCN5A is a key regulator of a gene transcriptional network that controls colon cancer invasion.


Vesteinn Thorsson, Dick
Web-based Apps

http://explorer.cancerregulome.org
The **Regulome Explorer** is an interactive web application that allows the user to explore multivariate relationships in data.

explorer.cancerregulome.org
RF-ACE, a multivariate statistical inference method based on ensembles of decision trees, which seeks to uncover significant associations between features in the input data matrix.
RF-ACE has high predictive power and is resistant to over-fitting.

Computational challenges:
- mixed data types: continuous, discrete, and categorical
- tens of thousands of features x tens or hundreds of samples
- non-linear, noisy, and multivariate relationships
- correlated features
- missing data

http://code.google.com/p/rf-ace/

RF-ACE features:
- handles mixed variable types
- does not require imputation of missing values
- random subsampling rather than combinatorial search
- statistical testing removes redundant features
- “importance” p-value for each candidate predictor
- fast, portable implementation in C++
A Master MicroRNA Network for EMT in OvCa
A multilevel pan-cancer view: from genes to hallmarks

Theo Knijnenburg
Mutational investment

- Gene
- Pathway
- Hallmark

- Sustaining Proliferative signalling
- Neuronal plasticity
- DNA repair
- Hallmarks

- Reprogramming, energy metabolism
- Sustained Angiogenesis
- Tumor-promoting inflammation
- Evading immune destruction
- Cancer immunoediting
- Sustaining Proliferative signaling
- Reacting cell death
Billions of Associations!

explorer.cancerregulome.org
Motivating questions

• Repurposing
  – Which existing cancer drugs may be therapeutic in which other cancers?
  – Which inhibitors with no current cancer indications may be therapeutic in certain cancers?

• Opportunity
  – TCGA primary tumor data may serve as the basis for guided investigation of these open questions
Guiding principle

• The direct protein target for most inhibitors is not the sensitizing aberrated protein itself
  – e.g., AKT1 inhibitors are most effective against cell lines with PTEN mutations

Song et al. (2012)
Proof of concept:
Associations between drug targets (e.g., AKT1) and sensitizing aberrations (e.g., PTEN) also evident in TCGA
Proof of concept:
Associations between drug targets (e.g., AKT1) and sensitizing aberrations (e.g., PTEN) also evident in TCGA
Approach

- Create large heterogeneous graph of associations from TCGA data, literature, databases, …
  - [Billions of edges, Terabytes of data]

- Query on Cray YarcData uRiKA graph analytics appliance
  - No locality of reference, graphs hard to partition
  - [Minutes rather than hours per query]

- Identify aberrated gene → target → drug relationships for drugs with and without known efficacy in cancer
Integrating multiple data sources into a (big) graph

Genomic aberrations | Therapeutic targets | Candidate inhibitors

- TCGA
- Pathway Commons
- RNAi
- DrugBank
- FDA
- NLM
- Domine

Institute for Systems Biology
Revolutionizing Science. Enhancing Life
Graph Data Model:
Resource Description Framework (RDF)
Example SPARQL Query

```
SELECT DISTINCT ?name ?p ?o WHERE {
  { ?name a nlm:Concept . } UNION {
    ?name a nlm:NamedEntity . }
  FILTER (STR(?o) = "Seed Gene List")
}
OPTIONAL { ?drug a nlm:NamedEntity . }
}
```

**Literature**
- **Seed Gene List**
- **TCGA Database**
- **Associated Genes**
- **Small Molecules**
- **Cancer Type**

**cancer.gov approved drugs**
Example Result: PTEN associations in UCEC

Genomic aberrations

Candidate targets

Candidate inhibitors
Example Result: PTEN associations in UCEC

- Genomic aberrations
- Candidate targets
- Candidate inhibitors

PTEN

PIK3R1/PIK3CA

Wortmannin

PTEN mutation status

PDB id 3hhm
Repurposing existing cancer drugs in other cancers

Genomic aberrations

Candidate targets

Candidate inhibitors

Existing cancer indication

New cancer indication

Target

Cancer Drug A
Example Result

- **TP53** is frequently mutated in most tumor types
- **ABCG2**, also known as Breast Cancer Resistance Protein (BCRP), is associated with TP53 mutation in TCGA breast cancer data
- **Nelfinavir**, an HIV protease inhibitor, also binds ABCG2 and many other proteins
- High-throughput cell line screening of breast cancer cells recently identified Nelfinavir as a selective inhibitor. “It can be brought to HER2-breast cancer treatment trials with the same dosage regimen as that used among HIV patients.” [Shim et al. JNCI 2012]
Understanding behavior of massive multicellular systems: *BioCellion*

Source: http://www.sjrcd.org/soilhealth/soilagg.html

Ductal Carcinoma model:
Nicholas Flann, Utah State Univ.

Source: http://www.theregister.co.uk

Acknowledgments

Brady Bernard, Ryan Bressler, Andrea Eakin, Timo Erkkilä, Lisa Iype, Seunghwa Kang, Theo Knijnenburg, Roger Kramer, Richard Kreisberg, Kalle Leinonen, Jake Lin, Yuexin Liu, Michael Miller, Sheila Reynolds, Hector Rovira, Vesteinn Thorsson, Da Yang, Wei Zhang