Machine Reading for Cancer Panomics

Hoifung Poon
Overview

Disease Genes
Drug Targets

Cancer Systems Modeling

High-Throughput Data
Overview

High-Throughput Data

Extract Pathways from PubMed

Disease Genes
Drug Targets

Grounded Semantic Parsing

KB

Overview
Panomics

… ATTCGGATATTAAAGGC …

Genome  Transcriptome  Epigenome
Genotype → Phenotype

High-Throughput Data

Disease Genes
Drug Targets
Precision Medicine
Vemurafenib on BRAF-V600 Melanoma

Before Treatment 15 Weeks
Vemurafenib on BRAF-V600 Melanoma

Before Treatment

15 Weeks

23 Weeks
Cancer

- Hundreds of mutations
- Most are “passenger”, not driver
- Can we identify likely drivers?

... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTAAAGCC ...

Tumor cells

Normal cells
Traditional Biology

Targeted Experiments

One hypothesis

Discovery
Genomics

High-Throughput Experiments

Many hypotheses

Discovery
Genomics

High-Throughput Experiments

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

Discovery
Example: Tumor Molecular Board

- 10-20 highly trained specialists
- Tens of hours on each patient
- Problem: Hard to scale
  
  **U.S. 2014: 1.6 million new cases, 585K deaths**
- Wanted: Decision support for clinical genomics
Pathway Knowledge

Genes work synergistically in pathways
Why Hard to Identify Drivers?

Complex diseases ← Perturb multiple pathways

Hanahan & Weinberg [Cell 2011]
Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

EphA2  
EphB2  
Ovarian Cancer
Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

EphA2 \( \times \) EphB2

Ovarian Cancer
Cancer Systems Modeling

**Gene A**

- **Transcription**
  - DNA → mRNA

- **Translation**
  - mRNA → Protein

- **Activation**
  - Protein → Protein Active

**Mutation effect**

- **Drug Target**

**Functional activity**

... ATTCGGATATTTAAGGC ...

...
Knowledge → Model

Gene A

Gene B

Gene C

DNA → mRNA → Protein → Protein Active

Transcription Factor

Protein Kinase

Knowledge → Model
Knowledge → Model

Gene A
DNA → mRNA → Protein → Protein Active

Gene B
DNA → mRNA → Protein → Protein Active

Gene C
DNA → mRNA → Protein → Protein Active
Knowledge → Model

Gene A
- DNA → mRNA → Protein → Protein Active

Gene B
- DNA → mRNA → Protein → Protein Active

Gene C
- DNA → mRNA → Protein → Protein Active

Transcription Factor
Protein Kinase

... ATTCGATATTTAGGC ...

Knowledge → Model
Knowledge → Model

Gene A
DNA → mRNA → Protein → Protein Active

Gene B
DNA → mRNA → Protein → Protein Active

Gene C
DNA → mRNA → Protein → Protein Active

Transcription Factor
Protein Kinase
Approach: Graph HMM

Gene A

DNA → mRNA → Protein → Protein Active

Gene B

DNA → mRNA → Protein → Protein Active

Gene C

DNA → mRNA → Protein → Protein Active

Transcription Factor

Protein Kinase
Extract Pathways from PubMed

High-Throughput Data

KB

Disease Genes
Drug Targets

... ATTCGGATATTTAAGGC ...
... ATTCGGGTATTTAAGCC ...

......
PubMed

- 24 millions abstracts
- Two new abstracts every minute
- Adds over one million every year
VDR+ binds to SMAD3 to form

... JUN expression is induced by SMAD3/4...

PMID: 123

PMID: 456
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1...
Long Tail of Variations

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins. BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

......

negative regulation

532 inhibited, 252 inhibition, 218 inhibit, 207 blocked, 175 inhibits, 157 decreased, 156 reduced, 112 suppressed, 108 decrease, 86 inhibitor, 81 Inhibition, 68 inhibitors, 67 abolished, 66 suppress, 65 block, 63 prevented, 48 suppression, 47 blocks, 44 inhibiting, 42 loss, 39 impaired, 38 reduction, 32 down-regulated, 29 abrogated, 27 prevents, 27 attenuated, 26 repression, 26 decreases, 26 down-regulation, 25 diminished, 25 downregulated, 25 suppresses, 22 interfere, 21 absence, 21 repress ......
Bottleneck: Annotated Examples

- GENIA (BioNLP Shared Task 2009-2013)
  - 1999 abstracts
  - MeSH: human, blood cell, transcription factor
- Challenge for “supervised” machine learning
- Can we breach this bottleneck?
Free Lunch #1: Distributional Similarity

- Similar context $\rightarrow$ Probably similar meaning
- **Annotation as latent variables**
  Textual expression $\rightarrow$ Recursive clusters
- **Unsupervised semantic parsing**

Recursive Clustering

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins. BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

......
TP53 inhibits BCL2. Tumor suppressor P53 down-regulates the activity of BCL-2 proteins. BCL2 transcription is suppressed by P53 expression. The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

......
Recursive Clustering

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins. BCL2 transcription is suppressed by P53 expression. The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

……
TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

......
Free Lunch #2: Existing KBs

- Many KBs available
  - Gene/Protein: GeneBank, UniProt, …
  - Pathways: NCI, Reactome, KEGG, BioCarta, …

- Annotation as latent variables
  - Textual expression → Table, column, join, …

- Grounded semantic parsing
\textbf{Relation Extraction}

\begin{tabular}{|c|c|c|}
\hline
\multicolumn{3}{|c|}{NCI-PID Pathway KB} \tabularnewline
\hline
\textbf{Regulation} & \textbf{Theme} & \textbf{Cause} \\
\hline
Positive & A2M & FOXO1 \\
Positive & ABCB1 & TP53 \\
Negative & BCL2 & TP53 \\
\ldots & \ldots & \ldots \\
\hline
\end{tabular}

TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

BCL2 transcription is suppressed by P53 expression.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 ...

……..
TP53 inhibits BCL2.

Tumor suppressor P53 downregulates BCL2 transcription.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 is mediated by...

......
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
GUSPEE

- Generalize distant supervision to extracting nested events
- Prior: Favor semantic parse grounded in KB
- Outperformed 19 out of 24 participants in GENIA Shared Task [Kim et al. 2009]

(POS-REG, BCL, (NEG-REG, IL-10, RFLAT))
(NEG-REG, TP53, (POS-REG, BCL, IL-2))
(POS-REG, AKT2, (POS-REG, IL-4, ERBB2))
(NEG-REG, (POS-REG, BCL, IL-2), BRAF)
Tree HMM

(POS-REG, BCL, (NEG-REG, IL-10, RFLAT))
(NEG-REG, TP53, (POS-REG, BCL, IL-2))
(POS-REG, AKT2, (POS-REG, IL-4, ERBB2))
(NEG-REG, (POS-REG, BCL, IL-2), BRAF)

BCL stimulates inhibition of RFLAT by IL-10.
Tree HMM

\[
P_\theta(z, t) = \prod_m P_{\text{EMIT}}(t_m \mid z_m, \theta) \cdot P_{\text{TRANS}}(z_m \mid z_{\pi(m)}, \theta)
\]

BCL stimulates inhibition of RFLAT by IL-10.
Expectation Maximization

\[ \theta^* = \arg \max_{\theta} \log P_{\theta}(T|K) \]

\[ = \arg \max_{\theta} \sum_{t \in T} \log \sum_{z} P_{\theta}(z, t) \cdot \phi_K(z) \]

Key challenge: non-local evidence

Virtual Evidence
The ability of IL-10 to block RFLAT requires BCL.
The ability of IL-10 to block RFLAT requires BCL.
The ability of *IL-10* to block *RFLAT* requires *BCL*.
# Best Supervised System

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>76.4</td>
<td>81.5</td>
<td>78.8</td>
</tr>
<tr>
<td>Transcription</td>
<td>49.4</td>
<td>73.6</td>
<td>59.1</td>
</tr>
<tr>
<td>Catabolism</td>
<td>65.6</td>
<td>80.0</td>
<td>74.4</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>73.9</td>
<td>84.5</td>
<td>78.9</td>
</tr>
<tr>
<td>Localization</td>
<td>74.6</td>
<td>75.8</td>
<td>75.2</td>
</tr>
<tr>
<td>Binding</td>
<td>48.0</td>
<td>50.9</td>
<td>49.4</td>
</tr>
<tr>
<td>Regulation</td>
<td>32.5</td>
<td>47.1</td>
<td>38.6</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>38.7</td>
<td>51.7</td>
<td>44.3</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>35.9</td>
<td>54.9</td>
<td>43.9</td>
</tr>
<tr>
<td><strong>Total Event F1</strong></td>
<td><strong>50.2</strong></td>
<td><strong>62.6</strong></td>
<td><strong>55.7</strong></td>
</tr>
</tbody>
</table>
## Preliminary Results

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>50.8</td>
<td>41.9</td>
<td>45.9</td>
</tr>
<tr>
<td>Transcription</td>
<td>18.3</td>
<td>14.0</td>
<td>15.9</td>
</tr>
<tr>
<td>Catabolism</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>36.2</td>
<td>43.6</td>
<td>39.5</td>
</tr>
<tr>
<td>Localization</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Binding</td>
<td>24.0</td>
<td>42.6</td>
<td>30.7</td>
</tr>
<tr>
<td>Regulation</td>
<td>2.5</td>
<td>5.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>11.4</td>
<td>21.4</td>
<td>14.9</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>4.4</td>
<td>16.4</td>
<td>6.9</td>
</tr>
<tr>
<td><strong>Total Event F1</strong></td>
<td>19.1</td>
<td>29.4</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Prototype-Driven Learning

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>55.3</td>
<td>88.3</td>
<td>68.0</td>
</tr>
<tr>
<td>Transcription</td>
<td>50.0</td>
<td>39.1</td>
<td>43.9</td>
</tr>
<tr>
<td>Catabolism</td>
<td>52.4</td>
<td>100.0</td>
<td>68.9</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>61.7</td>
<td>82.9</td>
<td>70.7</td>
</tr>
<tr>
<td>Localization</td>
<td>52.8</td>
<td>100.0</td>
<td>69.1</td>
</tr>
<tr>
<td>Binding</td>
<td>20.2</td>
<td>92.7</td>
<td>33.2</td>
</tr>
<tr>
<td>Regulation</td>
<td>24.1</td>
<td>64.0</td>
<td>35.0</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>17.4</td>
<td>63.8</td>
<td>27.4</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>8.4</td>
<td>52.8</td>
<td>14.5</td>
</tr>
<tr>
<td>Total Event F1</td>
<td>27.9</td>
<td>72.2</td>
<td>40.2</td>
</tr>
</tbody>
</table>
Outperformed 19 out of 24 supervised participants

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>55.3</td>
<td>88.3</td>
<td>68.0</td>
</tr>
<tr>
<td>Transcription</td>
<td>50.0</td>
<td>39.1</td>
<td>43.9</td>
</tr>
<tr>
<td>Catabolism</td>
<td>52.4</td>
<td>100.0</td>
<td>68.9</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>61.7</td>
<td>82.9</td>
<td>70.7</td>
</tr>
<tr>
<td>Localization</td>
<td>52.8</td>
<td>100.0</td>
<td>69.1</td>
</tr>
<tr>
<td>Binding</td>
<td>20.2</td>
<td>92.7</td>
<td>33.2</td>
</tr>
<tr>
<td>Regulation</td>
<td>24.1</td>
<td>64.0</td>
<td>35.0</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>17.4</td>
<td>63.8</td>
<td>27.4</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>8.4</td>
<td>52.8</td>
<td>14.5</td>
</tr>
<tr>
<td>Total Event F1</td>
<td>27.9</td>
<td>72.2</td>
<td>40.2</td>
</tr>
</tbody>
</table>

http://literome.azurewebsites.net
PubMed-Scale Extraction

- Preliminary pass:
  - 1.5 million instances
  - 13,000 genes, 838,000 unique regulations

- Applications:
  - UCSC Genome Browser, MSR Interactions Track
  - Expression profile modeling
  - Validate *de novo* pathway prediction
  - Etc.

Machine Science

Machine Science

Big Data
Machine Science

Big Data

Rich Knowledge
Machine Science

Big Data

Deep Model

Rich Knowledge

KB
Machine Science

Big Data

Deep Model

Hypotheses

Rich Knowledge

KB
Machine Science

Big Data

Deep Model

Hypotheses

Experiments

Rich Knowledge

KB
Machine Science

Deep Model

Hypotheses

Experiments

Big Data

Rich Knowledge

KB
Roadmap

- **Extract richer knowledge:**
  - Cell type, experimental condition, …
  - Hedging, negation, …

- **Formulate coherent models:**
  - Supporting evidence, contradiction, …
  - Intellectual gaps, hypotheses, …

- **Integrate w. data & experiments:**
  - Cancer panomics → Driver genes / pathways
  - Single-drug response → Drug combo prioritization
Personalized medicine approach to treating AML

The Leukemia & Lymphoma Society (LLS) and the Knight Cancer Institute at Oregon Health & Science University are leading a pioneering collaboration to develop a personalized medicine approach to improve outcomes for patients with acute myeloid leukemia (AML), a particularly devastating cancer of the blood and bone marrow. LLS provided $8.2 million to fund Beat AML, and here is how the collaboration will work:

1. In coordination with the Knight Cancer Institute, Stanford University, UT Southwestern Medical Center and Huntsman Cancer Institute will collect data from 900 AML patient samples within 3 years.

2. Illumina will perform genetic sequencing to identify mutations in the patient samples collected.

3. Intel will work with Knight Cancer’s bioinformatics team to apply its technology to accelerate computational analysis of the mutation data collected.

4. Drug and biotech companies will work with the collaboration to test drug compounds that target mutations suspected of driving disease progression. Array BioPharma will be first to test a therapeutic.
Personalized medicine approach to treating AML

The Leukemia & Lymphoma Society (LLS) and the Knight Cancer Institute at Oregon Health & Science University are leading a pioneering collaboration to develop a personalized medicine approach to improve outcomes for patients with acute myeloid leukemia (AML), a particularly devastating cancer of the blood and bone marrow. LLS provided $8.2 million to fund Beat AML and here is how the collaboration will work:
Decision Support for Clinical Genomics

- Raw Reads
- Variant Call RNA-Seq
- Clinical Observation
- Decision Support
- Knowledge Graph
- Literature
- Clinicians
Decision Support for Clinical Genomics

- Raw Reads
- Variant Call RNA-Seq
- Clinical Observation
- Decision Support
- Knowledge Graph
- Literature
- Clinicians
- NLP
## Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene</th>
<th>Mutation</th>
<th>Patient</th>
<th>Importance</th>
<th>Rationalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NPM1</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.0324689095154108</td>
<td>Graph</td>
</tr>
<tr>
<td>2</td>
<td>DNMT3A</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.0257080227869105</td>
<td>Graph</td>
</tr>
<tr>
<td>3</td>
<td>FLT3</td>
<td>No</td>
<td>TCGA-AB-2802</td>
<td>0.017949534996551</td>
<td>Graph</td>
</tr>
<tr>
<td>4</td>
<td>IDH1</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.0136209036068989</td>
<td>Graph</td>
</tr>
<tr>
<td>5</td>
<td>MT-ND5</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.0114896297732125</td>
<td>Graph</td>
</tr>
<tr>
<td>6</td>
<td>PTPN11</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.0103995735969948</td>
<td>Graph</td>
</tr>
<tr>
<td>7</td>
<td>TCHHL1</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.00957090470732164</td>
<td>Graph</td>
</tr>
<tr>
<td>8</td>
<td>MT-CYB</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.00934740227065732</td>
<td>Graph</td>
</tr>
<tr>
<td>9</td>
<td>TET2</td>
<td>No</td>
<td>TCGA-AB-2802</td>
<td>0.00899470604346163</td>
<td>Graph</td>
</tr>
<tr>
<td>10</td>
<td>C20orf24</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.00774601884004982</td>
<td>Graph</td>
</tr>
<tr>
<td>11</td>
<td>ANKRD30A</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.00774604597324909</td>
<td>Graph</td>
</tr>
<tr>
<td>12</td>
<td>TBX15</td>
<td>Yes</td>
<td>TCGA-AB-2802</td>
<td>0.00774313597737079</td>
<td>Graph</td>
</tr>
<tr>
<td>13</td>
<td>TP53</td>
<td>No</td>
<td>TCGA-AB-2802</td>
<td>0.00725453517242265</td>
<td>Graph</td>
</tr>
</tbody>
</table>
Drug Combo Analysis for FORETINIB / DOVITINIB
We Have Digitized Life
Next: Digitize Medicine

Knock down genes A, B, C → Cure

Summary

- Precision medicine is the future
- **Cancer systems modeling**
  Graphical model: Pathways + Panomics data
- **Extract pathways from PubMed**
  Machine reading by grounded semantic parsing
- **Literome**: KB for genomic medicine
Collaborators

- **Chicago**: Andrey Rzhetsky, Kevin White
- **OHSU**: Brian Drucker, Jeff Tyner
- **Berkeley AMP Lab**: David Patterson
- **Wisconsin**: Mark Craven, Anthony Gitter
- **UCSC**: Max Haeussler, David Haussler
- **Microsoft Research**: Chris Quirk, Kristina Toutanova, David Heckerman, Scott Yih, Lucy Vanderwende, Bill Bolosky, Ravi Pandya
Summary

High-Throughput Data

Disease Genes
Drug Targets