Inferring the hierarchical structure and function of a cell from millions of biological measurements

Michael Yu (UC San Diego)

mikeyu@ucsd.edu

Ontology Summit

April 5, 2017
Outline

1. Infer the hierarchical *structure* of a cell from millions of biological measurements. Result: data-driven gene ontology

2. Use this knowledge of cell *structure* to predict cell *function*. Result: functionalization of the ontology
A tsunami of biological data

Measurements of genes and gene products

DNA → RNA → Protein

Networks of gene-gene interactions

Genetic interactions

Protein-protein interactions

Model systems and species

Cost per Genome

This talk will focus on data and models for budding yeast (Saccharomyces cerevisiae)
The Goal: turn data into biological models
Inferring a hierarchy of cellular subsystems from biological data

- Genes are hierarchically clustered into cellular subsystems based on their functional similarity observed in data.
- The resulting hierarchy is called a “data-driven gene ontology”
Alignment of data-driven gene ontology to the curated Gene Ontology (GO)

- Identify which data-driven terms are new and which recapitulate existing knowledge in GO
- Suggest names for data-driven terms
Data-driven gene ontology recapitulates GO but also discovers new cellular subsystems

How many data-driven terms are seen in GO?  
How many GO-terms are seen in the data-driven ontology?

Data-driven gene ontology

![Venn diagram showing the overlap between GO terms and data-driven terms.](image)

![Bar chart showing the fraction of GO terms aligned to data-driven ontology.](image)
New subunits of the proteasome complex revealed in data-driven gene ontology

New subunits are defined by dense sets of gene-gene interactions
Refined molecular understanding of the autophagy (a process in which the cell recycles internal parts)
Interactive visualization of the hierarchical structure (atgo.ucsd.edu)
Outline

1. Infer the hierarchical *structure* of a cell from millions of biological measurements. 
   Result: data-driven gene ontology

2. Use this knowledge of cell *structure* to predict cell *function*. 
   Result: functionalization of the ontology
Translate genotype to phenotype
A central question of genetics

Genotype
(DNA sequence)

Phenotype
- Disease risk/state/treatment
- Physical attributes (e.g. height)
- Cell behaviors (e.g. growth rate)
Translation of genotype to phenotype by a hierarchy of cell subsystems

- Genotype
- Ontotype (Hierarchy)
- Phenotype

- Small complexes, subunits, reactions
- Large complexes, signaling pathways
- Organelles, broad processes

>3 million genotype-phenotype measurements
>12 million genotype-phenotype predictions
2,503 new measurements in DNA repair & nuclear lumen
Genetic Interactions
(A particular type of genotype-phenotype relation)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Cell Growth</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wildtype $wt$</td>
<td>1.00</td>
<td>$\varepsilon = 0$ Expected: No Interaction</td>
</tr>
<tr>
<td>Single-Gene Deletion $a\Delta$</td>
<td>0.70</td>
<td>$\varepsilon &lt; 0$ Slower-than Expected: “Negative” Interaction</td>
</tr>
<tr>
<td>Double-Gene Deletion $a\Delta b\Delta$</td>
<td>0.49, 0.70</td>
<td>$\varepsilon &gt; 0$ Faster-than-Expected: “Positive” Interaction</td>
</tr>
</tbody>
</table>

- Genetic interactions measured for ~5 million pairs of genes in *S. cerevisiae*
- From Costanzo et al. *Science*, 2010. (One of the largest genotype-phenotype datasets)
- Given a pair of genes, predict if there is a genetic interaction.
A hierarchical interpretation of genetic interactions

Genetic interactions will “bundle” together at higher Levels of the hierarchy

>400 within-term & >1,300 between-term enrichments

Enrichments explain 36.7% of genetic interactions
A supervised machine learning framework for predicting phenotype from genotype
The "ontotype" is a multi-scale representation of the cell inbetween genotype and phenotype. It serves as a highly informative set of features for supervised learning of phenotype.
Predictions correspond to measurements

Our hierarchical model outperforms ...
1. Same model using a random hierarchy
2. Previous, non-hierarchical methods
3. Same model using a cut-down hierarchy

Hierarchical model of genetic interactions outperforms non-hierarchical methods
A “functionalized” gene ontology (combines cell structure and function)
Zooming into DNA Repair

Validation with further experiments

Accurate prediction of new genetic interaction measurements for 2,503 gene pairs impacting DNA repair and nuclear lumen
Zooming into the nuclear lumen
Summary

1. The structure of hierarchical systems like the cell can be accurately inferred from data
   1. Enabled by a deluge of experimental measurements
   2. Complements manually curated knowledge (e.g. GO)

2. This structure can be “functionalized” to make biological predictions
   1. Accurate translation of genotype to phenotype
   2. Machine learning guided by ontologies
Trey Ideker’s Group
• Michael Kramer
• Janusz Dutkowski
• Jianzhu Ma
• Rohith Srivas
• Fan Zheng
• Mitchell Flagg
• Keiichiro Ono
• Eric Sage
• Kate Licon
• Jason Kreisberg
• Cherie Ng
• Ideker Lab Members
• Trey Ideker

Collaborators
• Roded Sharan
• Hannes Braberg
• Nevan Krogan

Funding
• San Diego Center for Systems Biology
• Genetics Training Program
• Bioinformatics and Systems Biology Training Grant
References


