Accelerating Biomedical Discovery

Lawrence Hunter, Ph.D.
Professor and Director
Computational Bioscience Program
University of Colorado School of Medicine

Larry.Hunter@uchsc.edu
http://compbio.uchsc.edu/Hunter
How to Understand Gene Sets?

- There is no “gene” for any complex phenotype; gene products function together in dynamic groups.

- A key task is to understand why a set of gene products are grouped together in a condition, exploiting all existing knowledge about:
  - The genes (all of them)
  - Their relationships ($|\text{genes}|^2$)
  - The condition(s) under study.
The amount of information relevant to the task

1,000 genomes project will create 1,400GB next year
http://1000genomes.org

Genbank growth rate

PubMed growth rate

\[ y = e^{0.0418x} \]
\[ R^2 = 0.99 \]

\[ y = e^{0.031x} \]
\[ R^2 = 0.95 \]
Yet Still Not Enough!

- Experimental coverage of interactions and pathways is still sparse, especially in mammals

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>PROTEINS</th>
<th>INTERACTIONS</th>
<th>EXPERIMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila melanogaster</td>
<td>7066</td>
<td>21004</td>
<td>21106</td>
</tr>
<tr>
<td>(fruit fly)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>4920</td>
<td>18272</td>
<td>23238</td>
</tr>
<tr>
<td>(baker's yeast)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>1847</td>
<td>7427</td>
<td>9065</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>2643</td>
<td>4037</td>
<td>4082</td>
</tr>
<tr>
<td>Homo sapiens (Human)</td>
<td>1124</td>
<td>1648</td>
<td>2423</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>710</td>
<td>1425</td>
<td>1425</td>
</tr>
<tr>
<td>Mus musculus (house mouse)</td>
<td>264</td>
<td>374</td>
<td>527</td>
</tr>
<tr>
<td>Rattus norvegicus (Norway rat)</td>
<td>102</td>
<td>132</td>
<td>196</td>
</tr>
</tbody>
</table>
Exponential knowledge growth

- 1,170 peer-reviewed gene-related databases in 2009 NAR db issue
- 804,399 PubMed entries in 2008 (> 2,200/day)
- Breakdown of disciplinary boundaries makes more of it relevant to each of us
- “Like drinking from a firehose” – Jim Ostell
Knowledge-based data analysis

- **Goal:** Bring all of this information (and more!) to bear on analyzing experimental results.

- **How? 3R systems**
  - Integrate multiple databases (using the semantic web)
  - Extract knowledge from the literature
  - Infer implicit interactions
  - Build knowledge networks
    - Nodes are fiducials, like genes or ontology terms
    - Arcs (relations) are qualified (typed) and quantified (with reliability)
  - Deliver a tool for analysts to use knowledge networks to understand experiments and generate hypotheses
The best source of knowledge is the literature

OpenDMAP is significant progress in concept recognition in biomedical text

Even simple-minded approaches are powerful

- Gene co-occurrence widely used
- Thresholded co-occurrence fraction is better

\[
CF(C_X, C_Y) = \frac{f(C_X, C_Y)}{\min\{f(C_X), f(C_Y)\}}
\]
OpenDMAP extracts typed relations from the literature

- **Concept recognition tool**
  - Connect ontological terms to literature instances
  - Built on Protégé knowledge representation system
  - New project to hook to NCBO ontologies dynamically

- **Language patterns associated with concepts and slots**
  - Patterns can contain text literals, other concepts, constraints (conceptual or syntactic), ordering information, etc.
  - Linked to many text analysis engines via UIMA

- **Best performance in BioCreative II IPS task**

- **>500,000 instances of three predicates (with arguments) extracted from Medline Abstracts**

Reasoning in knowledge networks

**Ddc; MGI:94876**

- **BP** carboxylic acid metabolic process (GO:0019752)
- **BP** catecholamine biosynthesis process (GO:0042423)
- **BP** response to toxin (GO:00009636)

**Cadps; MGI:1350922**

- **BP** catecholamine secretion (GO:0050432)
- **BP** protein transport (GO:0015031)
- **BP** vesicle organization (GO:0016050)

---

[Bada & Hunter, 2006]
Inferred interactions

- Dramatically increase coverage…
- But at the cost of lower reliability
- We apply new method to assess reliability without an explicit gold standard
- [Leach, et al., 2007; Gabow, et al., 2008]

### Top 1,000 Craniofacial genes
(1,000,000 possible edges)

<table>
<thead>
<tr>
<th>Source</th>
<th># edges</th>
<th>Consensus Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affinity Chromatography</td>
<td>3</td>
<td>0.91</td>
</tr>
<tr>
<td>Competitive Binding</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>Crosslinking</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>Immunoprecipitation</td>
<td>11</td>
<td>0.33</td>
</tr>
<tr>
<td>Yeast 2 hybrid</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>DMAP transport relations</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>Literature co-mention</td>
<td>89</td>
<td>0.3</td>
</tr>
<tr>
<td>PreBIND</td>
<td>4</td>
<td>0.33</td>
</tr>
<tr>
<td>PreMod</td>
<td>2718</td>
<td>0.19</td>
</tr>
<tr>
<td>Co-KEGG</td>
<td>1195</td>
<td>0.04</td>
</tr>
<tr>
<td>Co-InterPro</td>
<td>4470</td>
<td>0.01</td>
</tr>
<tr>
<td>Co-Phenotype</td>
<td>12298</td>
<td>0.01</td>
</tr>
<tr>
<td>Co-GO:BP</td>
<td>21203</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Co-GO:MF</td>
<td>38774</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Co-GO:CC</td>
<td>44974</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Co-ChEBI</td>
<td>15542</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
• Combine diverse sources…
  – Databases of interactions
  – Information extracted from the literature (CF or DMAP)
  – Inference of interactions
• … Into a unified knowledge summary network:
  – Every link gets a reliability value
  – Combine multiple links for one pair into a single summary
    • More sources → more reliable
    • Better sources → more reliable
    • “Noisy Or” versus “Linear Opinion Pool”
• Summaries allow for effective use of noisy inferences
  – [Leach PhD thesis 2007; Leach et al., 2007]
Knowledge-based analysis of experimental data

• High-throughput studies generate their own interaction networks tied to fiducials
  – E.g. Gene correlation coefficients in expression data

• Combine with background knowledge by:
  – Averaging (highlights already known linkages)
  – Hanisch (ISMB 2002) method (emphasizes data linkages not yet well supported by the literature)

• Report highest scoring data + knowledge linkages, color coding for scores of average, logistic or both.
The Hanalyzer: 3R proof of concept

• [Leach, Tipney, et al., PLoS Comp Bio 2009]
  http://hanalyzer.sourceforge.org

• Knowledge network built for mouse
  – NLP only CF and DMAP for three relationships from PubMed abstracts

• Simple reasoning (co-annotation, including ontology cross-products)

• Visualization of combined knowledge / data network via Cytoscape + new plugins
Knowledge Network

External sources
- Ontology annotations
- Medline abstracts
- Gene database_1
- Gene database_2
- Gene database_n

Reading methods
- Biomedical language processing
- Parsers & Provenance tracker

Reasoning methods
- Ontology enrichment
- Co-annotation inference
- Literature co-occurrence
- Co-database inference
- Semantic integration
- Reliability estimation

Reporting methods
- Experimental data
- Data Network
- Knowledge Network
- Network integration methods
- Visualization & Drill-down tool

Integration methods
- Literature co-occurrence
- Co-annotation inference
- Semantic integration
- Reliability estimation
First application: Craniofacial Development

- NICHD-funded study (Rich Spritz; Trevor Williams) focused on cleft lip & palate
- Well designed gene expression array experiment:
  - Craniofacial development in normal mice (control)
  - Three tissues (Maxillary prominence, Fronto-nasal prominence, Mandible)
  - Five time points (every 12 hours from E10.5)
  - Seven biological replicates per condition (well powered)
- >1,000 genes differentially expressed among at least 2 of the 15 conditions (FDR<0.01)
The Whole Network

Craniofacial dataset, covering all genes on the Affy mouse chip.

Graph of top 1000 edges using AVE or HANISCH (1734 in total). Edges identified by both.

Focus on mid-size subnetwork
Co-occurrence in abstracts: PMID:16407395...
R = 0.0105
DMAP transport relation
R = 0.1034
Shared GO molecular functions: GO:3705...
R = 0.0284
Shared GO cell component: GO:5667...
R = 0.0190
Shared GO biological processes: GO:6139...
R = 0.0172
Shared knockout phenotypes: MP:5374 ...
R = 0.018
Shared interpro domains: IPR:11598...
R = 0.0438
Premod_M interaction: Mod074699
R = 0.1005
Inferred link through shared GO/ChEBI: ChEBI:16991
R = 0.01

$P_{knowledge} ≡ 1 - \prod_{i} (1 - R_i) \approx 0.305$

$P_{average} ≡ \frac{P_{knowledge} + P_{data}}{2} \approx 0.393$

$P_{logit} ≡ \frac{(1 - e^{-s(P_{knowledge} - v)})^{-1} + (1 - e^{-s(P_{data} - v)})^{-1}}{2} \approx 0.49996$
Strong data and background knowledge facilitate explanations

- Goal is abductive inference: why are these genes doing this?
  - Specifically, why the increase in mandible before the increase in maxilla, and not at all in the frontonasal prominence?
Exploring the knowledge network
Scientist + aide + literature → explanation: tongue development

The delayed onset, at E12.5, of the same group of proteins during mastication muscle development.

Myoblast differentiation and proliferation continues until E15 at which point the tongue muscle is completely formed.

Myogenic cells invade the tongue primodia ~E11
On to Discovery

- Add the strong data, weak background knowledge (Hanisch) edges to the previous network, bringing in new genes.
- Four of these genes not previously implicated in facial muscle development (1 almost completely unannotated)
Prediction validated!

Zim1, E12.5

E43rik, E12.5

ApoBEC2, E11.5

HoxA2, E12.5
Transforming biomedical research with 3R systems?

- Deeper connections to the literature
  - NLP on full texts of journal articles & textbooks
  - Stay current, be aware of priority & citations
- Abductive QA (provide evidence, explanation)
- Better user experience in reporting
  - Integration with an analyst’s notebook
  - More and better sense-making approaches
  - Different types of data (e.g. GWAS)
  - Automated focus on “interesting” material
“In contrast to clathrin-coated and COPI- or COPII-coated vesicles, caveolae are thought to invaginate and collect cargo proteins by virtue of the lipid composition of the calveolar membrane, rather than by the assembly of a cytosolic protein coat. Caveolae pinch off from the plasma membrane and can deliver their contents either to endosome-like compartments or (in a process called transcytosis, which is discussed later) to the plasma membrane on the opposite side of a polarized cell.” etc!
KR&R poses new challenges

• Need many on-the-fly terms (cross-products!)
  – Not all cross-products are valid: caveoli of muscle cells work, but not all CCs are in all cells (e.g. axons)

• Need many new relationships:
  – has-function, is-realization-of, occurs-in, precedes, results-in-formation-of, results-in-transport-to…

• Need to integrate multiple ontologies: e.g. cell from CL (muscle) and cell from CC (caveoli)

• Non-logical inference! CAV3 annotated to caveoli; can’t logically infer caveoli of muscle cells.
To find out more...

- [http://hanalyzer.sourceforge.net](http://hanalyzer.sourceforge.net)
- [http://www.youtube.com/watch?v=jAegU3aZbWI](http://www.youtube.com/watch?v=jAegU3aZbWI) (or just search YouTube for “hanalyzer”)
- Presentation at ISMB Highlights track
- See also our Ontology Quality Assurance talk at ISMB (Verspoor, et al.)
Preview: Univocality in GO

• Univocality (Spinoza, 1677)
  “a shared interpretation of the nature of reality”

• For GO/OBO, consistency of expression

• Transformation-based method detects failures:
  - GO:0052387 -- induction by organism of symbiont apoptosis
  - GO:0052351 -- induction by organism of systemic acquired resistance in symbiont
  - GO:0021861 -- radial glial cell differentiation in the forebrain
  - GO:0021846 -- cell proliferation in forebrain
  - GO:0000282 -- cellular bud site selection
  - GO:0000918 -- selection of site for barrier septum formation
  - GO:0043247 -- telomere maintenance in response to DNA damage
  - GO:0042770 -- DNA damage response, signal transduction

• See Verspoor, et al., in ISMB proceedings…
Acknowledgements

- Sonia Leach (Design, first implementation)
- Hannah Tipney (Analyst)
- Bill Baumgartner (UIMA, Software engineer)
- Philip Ogren (Knowtator)
- Mike Bada (Ontologist)
- Helen Johnson (Linguist)
- Kevin Cohen (NLP guru)
- Lynne Fox (Librarian)
- Aaron Gabow (Programmer)

- NIH grants
  - R01 LM 009254
  - R01 LM 008111
  - R01 GM 083649
  - G08 LM 009639
  - T15 LM 009451

- MIT Press for permission to use *Being Alive* for doing science
Opportunities at one of the best Computational Bioscience Programs

• Top faculty, great research, serious education
• Institutional Training Grant from NLM
  – Generous graduate and postdoctoral fellowships
• Grad school application deadline January 1
• Currently open faculty positions & postdocs
• More info at http://compbio.uchsc.edu
• Ask me for details