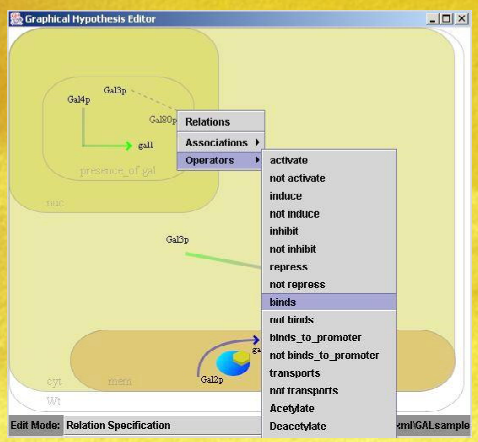
Robustness and Adaptation in "Information Ecosystems"

Stephen Racunas
Christopher Griffin

"Cross-Disciplinary Cross-Pollination"

- Bioinformatics work à Information Security
  - Managing hypotheses, evaluating scenarios
  - Inference based on observations *in context*
  - Integration of heterogeneous data to support inference
- Information Security à Bioinformatics
  - Identify, prove, discourage IP violations
  - Confidentiality vs. availability tradeoffs
  - Secure management for large and oft-used data sets
- Common Problems and Overlapping Interests
  - Inference without compromise
  - Provenance and tracking
  - Medical records || compartmentalized data

Compose and Evaluate Hypotheses



- Intuitive interface for composing and editing hypotheses
- Library of events
- Formal semantics specify hypothesized events *in context*

- Identify errors in hypotheses
- Multiple data types, sources
- Suggest refinements

hy1 = (ev0+ev1) and (ev2+ev3)
ev0 = Gal2p transports galactose in mem in wt
ev1 = galactose activate Gal3p in cyt in wt
ev2 = Gal3p Binds_to_promoter gal1 in nuc in wt
ev3 = Gal3p induce gal1 in nuc in wt in presence_of galatose
Total supp:9, Total contr:1

Directly Address End-User Concerns

- People worry about specific security threat scenarios

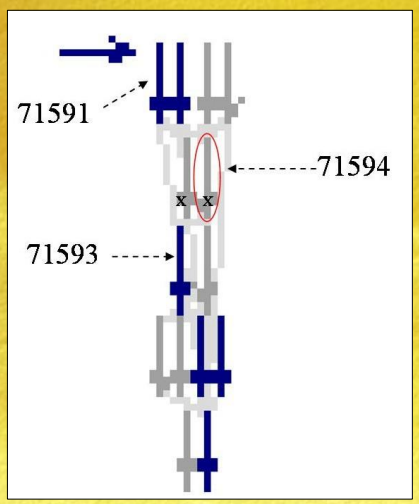
- These concerns can be used to form hypotheses about historical, ongoing, or potential attacks

- Hypotheses relate entities and intent
  - Do data from observed breaches support the hypothesis?
  - Do current observations disprove the hypothesis outright?
  - If there is little support for this particular hypothesis, what entity+intent combinations *are more highly* supported?
  - How are these highly supported combinations related?

Considerations of Intent

- Actions always happen in a context
- Attacks are caused by a certain set of root intents
  - Perhaps an intruder intends to probe the security of the US DoD infrastructure in order to garner sensitive information
  - Perhaps an intruder intends to cause economic damage to a specific target within a network or to the network itself
- Identification of intent
  - Help protect from future attacks
  - Dramatically reduce false alarms
  - Mitigate the effects of an attack in progress
- Help responders identify intent by identifying streams of events directed to a common goal
- Approach the problem with a method firmly rooted in the formal scientific method of hypothesize-and-test

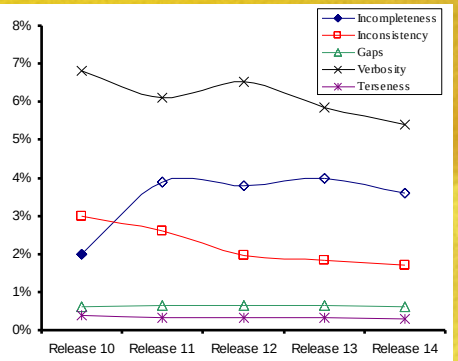
"Proofreading" Knowledge Sources



- Identify "holes" in knowledge
- Proofread for:
  - Inconsistencies
  - Incompleteness
  - Gaps
  - Well-formedness
  - Misleading Cycles
- Suggest "patches"
  - Internal consistency
  - Cross-validation with respect to similar resources

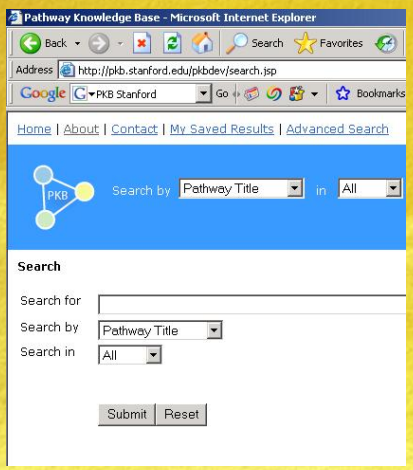
Sample KB Proofreading

Property	R- 10	R- 11	R- 12	R- 13	R- 14
Incompleteness	2 %	3.90 %	3.80 %	4.00 %	3.60 %
Inconsistency	3 %	2.60 %	1.95 %	1.84 %	1.70 %
Gaps	0.60 %	0.64 %	0.65 %	0.65 %	0.60 %
Verbosity	6.80 %	6.10 %	6.52 %	5.85 %	5.40 %
Terseness	0.40 %	0.32 %	0.32 %	0.32 %	0.30 %
90% Well formed	40 %	42 %	45 %	44 %	47 %
Self-loops	0 %	0 %	0 %	0 %	0 %



- Techniques and scripts used by KB staff for proofreading their future releases

Knowledge Base Unification

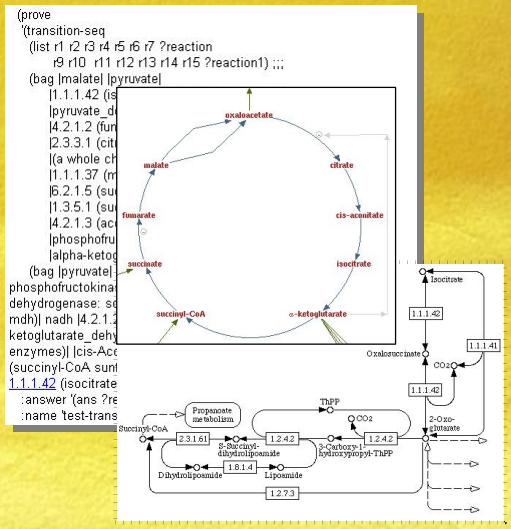


- Human (981 pathways)
  - Reactome (688 pathways)
  - BioCyc (185 pathways)
  - KEGG (108 pathways)
- E coli (494 pathways)
  - Reactome (184 pathways)
  - BioCyc (235 pathways)
  - KEGG (75 pathways)
- Yeast (548 pathways)
  - Reactome (313 pathways)
  - BioCyc (173 pathways)
  - KEGG (62 pathways)

Hypothesis-Based Querying

Given at least a "semi-formal" phrasing…

1. Select the best-match resource
2. Highlight conflicts
3. Present the "best completion" from the union of all other resources



Meta-Data and Testing for Intent

- What could someone actually *do* if they *could* gain access to a given information resource?
  - Game theory
  - Meta-data

  à Hypothesis space consisting of motives, potential future actions, and the contexts in which actions take place

- Analysis of hypothesis space
  - Past precedents
  - Online observations
  - Prune the potential outcome structure in real time
  - Provide a "least likely to reject" solution set

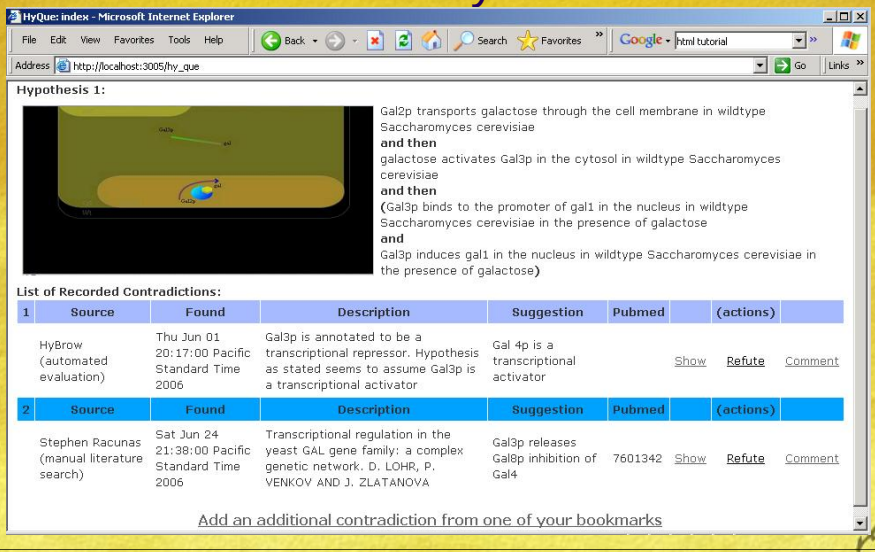
Biologists and Information Security

- Biologists worry about information security too…
- Unauthorized use of experimental results by others
  - Data (very bad)
  - Paradigm-shifting insights (much worse)
  - "Getting scooped" can ruin a career or a lab

All this without even broaching the issue of bio-*medical* data!

- How to be helpful without being overly vulnerable?
  - Help others to avoid following false leads
  - But don't allow others to steal your unpublished good ideas
  - Track what results contradict which hypotheses and how
  - But respect sensitive aspects of methodologies or raw data

Community Tools



“Information Ecosystem”

- Organizations and individuals produce, consume, and compete for information
  - Most require outside information to function
  - Primary objective is not to restrict information flow
- Objective is to *make an entity or organization* ***maximally competitive*** *in its information environment*
  - Such optimization is possible
  - Rich history in biological domain
- We seek new techniques for combining logical and numerical optimization in a contradiction-based way

The End

Thank you!