Automated Hypothesis Generation Based on Mining Scientific Literature

Scott Spangler*,1, Angela D. Wilkins*,3, Benjamin J. Bachman3, Meena Nagarajan1, Tajhal Dayaram3, Peter Haas1, Sam Regenbogen3, Curtis R. Pickering2, Austin Comer2, Jeffrey N. Myers2, Ioana Stanoi1, Linda Kato1, Ana Lelescu1, Jacques J. Labrie1, Neha Parikh3, Andreas Martin Lisewski3, Lawrence Donehower3, Ying Chen1, Olivier Lichtarge3

1IBM Research
San Jose, California
2The University of Texas MD Anderson Cancer Center
3Baylor College of Medicine
Houston, Texas

McNair Medical Institute, DARPA
DATA OVERFLOW

A mismatch between raw data and our analytic abilities
Overall

• 50 million scientific papers
• 1 million more per year
• 2 new papers per minute

Biomedical research

• $10^3$ to $10^5$ papers per topic areas
• Over 70,000 papers on p53 (a tumor suppressor)

A fundamental bottleneck: we cannot keep up with discovery
Most findings are seldom cited and have low influence.

Self-referential and large communities can skew the distribution.
Most findings are seldom cited and have low influence.

Self referential and large communities can skew the distribution.

Too little time to read and to learn.
A KNOWLEDGE INTEGRATION TOOLKIT: KnIT

KNOWLEDGE ACQUISITION
- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entities as a points in feature space: these features are coordinates that form an aggregate “text signature” of the entity

KNOWLEDGE REPRESENTATION
- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors’ functional properties

REASONING
- Diffuses information among linked entities
- Rank order candidates for further experimentation of novel annotation predictions.
- The domain expert can evaluate the rankings, the supporting evidence, and choose which to pursue experimentally.

To accelerate scientific progress by integrating mining, visualization, and analytics
p53 is the “first responder” to cellular stress, as a result it is linked to a plethora of brain diseases.
A KNOWLEDGE INTEGRATION TOOLKIT: KnIT

KNOWLEDGE ACQUISITION
• Survey relevant text
• Extracts relevant entities (human proteins called kinases)
• Model each entities as a points in feature space: these features are coordinates that form an aggregate “text signature” of the entity

To accelerate scientific progress
by integrating mining, visualization, and analytics
Biological Entity Similarity Network
A KNOWLEDGE INTEGRATION TOOLKIT: KnIT

KNOWLEDGE ACQUISITION
- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entity as a points in feature space: these features are coordinates that form an aggregate “text signature” of the entity

KNOWLEDGE REPRESENTATION
- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors’ functional properties

To accelerate scientific progress by integrating mining, visualization, and analytics
DISCOVERY END TO END WORKFLOW

System T Annotators
Custom Java Annotators

Annotators

Annotator Registry

Normalizers

Normalizers Registry

Big Insights Based Map Reduce Framework

Filter
Annotate
Normalize
Transform And Index

Index
Entity Graph

Content Browsing Services
Entity Analytics Services
Custom User Interfaces

Curated Documents

Document Registry

Dictionaries

Dictionary Registry

Annotator Registry

Normalizers Registry

Managing the Workflow
Kinase Similarity Tree
Algorithm

Algorithm 1 Create an n-ary similarity tree from a set of entities

Input: entities, n
Output: n-ary similarity tree
mostTypicalFV = average(entities)
root = closestTo FV(entities, mostTypical FV)
entities.remove(root)
candidates = {root}
while not entities.isEmpty()
    (e, c) = closestPair(entities, candidates)
c.addChild(e)
    if c.numChildren() == n then
        candidates.remove(c)
    end if
    candidates.add(e)
entities.remove(e)
end while
Return: root
A KNOWLEDGE INTEGRATION TOOLKIT: KnIT

KNOWLEDGE ACQUISITION
- Survey relevant text
- Extracts relevant entities (human proteins called kinases)
- Model each entity as a point in feature space: these features are coordinates that form an aggregate “text signature” of the entity

KNOWLEDGE REPRESENTATION
- A graph represents similarity relationship among entities.
- Helps visualize hidden literature connections between entities.
- Coloring may reveal sub-graphs of clusters of interest.
- Critically, a sub-graph may contain unexpected entities
- Hypothesis: perhaps these unexpected entities share their neighbors’ functional properties

REASONING
- Diffuses information among linked entities
- Rank order candidates for further experimentation of novel annotation predictions.
- The domain expert can evaluate the rankings, the supporting evidence, and choose which to pursue experimentally.

To accelerate scientific progress
by integrating mining, visualization, and analytics
RETROSPECTIVE CONTROL

GRAPH INFORMATION DIFFUSION OF P53 KINASE LABELS KNOWN PRIOR TO 2003 RECOVERS P53 KINASES DISCOVERED AFTER 2003

Example of this type of diffusion: Lisewski et al. *CELL* (2014)
LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

RECOVERY OF KNOW P53 KINASES

![ROC Curve](chart)

- Literature-based, AUC=0.691
- Improved Dictionary, AUC=0.751
- TFIDF, AUC=0.800
- Random
LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

RECOVERY OF KNOW P53 KINASES

COMPARISON TO SEQUENCE ANALYSIS
LEAVE-ONE-OUT EXPERIMENTS ALSO SUGGEST THAT THIS APPROACH CAN PREDICT WHICH KINASES TARGET A GIVEN PROTEIN

RECOVERY OF KNOW P53 KINASES

COMPARISON TO SEQUENCE ANALYSIS

KINASE PREDICTIONS IN 45 OTHER PROTEINS
BONA FIDE EXPERIMENTAL VALIDATION

A.

(Darker = more phosphate)

p53

p53 Phosphorylation Level

0.94  0.08  0.00  0.01  0.02  1.00

NEK2  PKN1  PDGFRβ  TNK2  INSRR  CHK1

(B) CONDITIONS

NEK2 Present?  +  -  +  -
PKN1 Present?  -  +  -  +

RESULTS

Detecting p53

Detecting NEK2

Detecting PKN1

Lane 1  2  3  4

p53

Negative Control Antibody
A KNOWLEDGE INTEGRATION TOOLKIT: KnIT

- Laboratory support for p53 kinases predicted from text mining
- Proof of principle for a strategy to predict some unknown fact from the scientific literature
- A first step to predict new connections based on everything else that is known.
- Future: more work needed to
  - Broaden the scope of proteins and functions
  - Comprehensive networks of interactions
  - To gather a more complete understanding of the mechanisms behind disease
  - Translate this into clinical impact.
  - Test this approach of mining literature to identify hidden relationships beyond cancer and beyond biology to other areas of human thought where text is a bottleneck.
- Such acceleration of discovery is not only desirable, but also indispensable for human flourishing.