Reasoning about Scientific Knowledge with Workflow Constraints: Towards Automated Discovery from Data Repositories

Yolanda Gil

Information Sciences Institute
and Department of Computer Science
University of Southern California
http://www.isi.edu/ikcap
gil@isi.edu

Ontology Summit, April 2017

With Daniel Garijo, Rajiv Mayani, and Varun Ratnakar (USC/ISI) and Parag Mallick, Ravali Adusumilli, and Hunter Boyce (Stanford)
Lack of Access to Data Analytics Expertise

Science, Dec 2011

Do you have the necessary expertise in your lab or group to analyze your data in the way you want?

- 26.5% YES
- 34.4% YES, through collaborators
- 23.0% NO
- 16.1% No special skills needed

“The next few years [particularly in medicine] the volume of data we need to analyze will expand exponentially.”
Cancer Multi-Omics

Proteogenomic characterization of human colon and rectal cancer


Affiliations | Contributions | Corresponding author

Nature 513, 382–387 (18 September 2014) | doi:10.1038/nature13438
Received 19 September 2013 | Accepted 02 May 2014 | Published online 20 July 2014

https://doi.org/10.1038/nature13438
Cancer Multi-Omics:
The Cancer Genome Atlas (TCGA)

Data Types

- Clinical Data
- Images
- Microsatellite Instability (MSI)
- DNA Sequencing
- miRNA Sequencing
- Protein Expression
- mRNA Sequencing
- Total RNA Sequencing
- Array-based Expression
- DNA Methylation
- Copy Number
Scientific Data Analysis Today: Inefficient, Incomplete, Irreproducible

- Data analysis is time consuming
- Not systematic
- Not updated when new data/methods become available
- Hard/impractical to reproduce prior work
- Overall process is manually done: inefficient and error-prone
- Analytic knowledge is compartmentalized
**DISK: Automated Discovery of Scientific Knowledge**

- **Long-Term Goal:** Human directs automated intelligent system to explore hypotheses of interest
  - Hypothesis-driven data analysis and discovery
    - Systematic and reproducible analyses
  - Report of findings with explanations

- **Approach:** Intelligent system that captures common data analysis strategies used by scientists in a domain
  - Build on WINGS intelligent workflow system that can adapt data analysis given the constraints of algorithmic steps
Capturing Data Analysis Strategies through *Lines of Inquiry*

**Hypothesis**

**Revised hypothesis**

**Lines of Inquiry**
Specify relevant analytic methods (workflows), type of data needed, and how to combine results

**Query to Retrieve Data**

**DATA CATALOG**

**Analytic Workflows**
- Workflow Library

**Meta-Workflows**
- Confidence Estimation
- Benchmarking
Related Work: Computational Scientific Discovery

- [Lenat 1976]
- [Lindsay et al 1980]
- [Langley 1981]
- [Falkenhainer 1985]
- [Kulkarni and Simon 1988]
- [Cheeseman et al 1989]
- [Zytkow et al 1990]
- [Simon 1996]
- [Valdes-Perez 1997]
- [Todorovski et al 2000]
- [Schmidt and Lipson 2009]
Related Work: Philosophy of Science

THE STRUCTURE OF SCIENTIFIC REVOLUTIONS

Computation, Causation, & Discovery

In Search of Mechanisms
Discoveries across the Life Sciences

Edited by Clark Glymour & Gregory F. Cooper
1 Representing Hypotheses and Lines of Inquiry

- **Hypothesis**
- **Revised hypothesis**

**Lines of Inquiry**
Specify relevant analytic methods (workflows), type of data needed, and how to combine results

**Query to Retrieve Data**

**Analytic Workflows**
- Workflow Library

**Meta-Workflows**
- Confidence Estimation
- Benchmarking
1) Initial Hypothesis

Hypothesis Statement Hy1

PRKCDBP protein expressedIn Patient AA-3561-01A
II) Lines of Inquiry

- Captures how to set up potential analyses that can be pursued to test a certain type of hypothesis

Hypothesis pattern

- Hypothesis Statement
  - Protein ?x
    - expressedIn
    - Patient ?y
  - onHypothesis

Qualifications

- Confidence
  - Report CR?
  - v1
    - hasConfidenceValue

Query to data repositories

- DataFile ?d
  - type
  - producedData
- RNASeq
- MassSpecData
- Experiment ?e
  - experimentedOn
- Sample ?s
  - collectedFrom
  - Patient ?p

Data Analysis Workflows

Updated hypothesis

Meta-workflows to analyze workflow results
III) Finding Relevant Data and Methods

Hypothesis Statement Hy1

- PRKCDBP protein

Line Of Inquiry

- hasWorkflowBindings

Workflows Available

Data Available

- ProtData1
- MassSpec-Labwork1

Labwork Activity1

- tissue1

Patient1

- associatedProtocol
- derivedFrom
- wasGeneratedBy
- used
- collectedFrom
IV) Running Workflows on the Data
V) Meta-Reasoning about Workflow Results

Hypothesis Statement Hy1
- PRKCDBP protein expressedIn Patient AA-3561-01A

Qualifications of Hy1
- Confidence Report CR1
  - hasConfidenceValue 0.8

Line Of Inquiry
- hasWorkflowBindings

Provenance of Hy1
- Meta-Workflow M1
  - supports
  - wasGeneratedBy
  - hasWorkflowTemplate
    - Workflow execution W1
      - used
    - Workflow execution W2
      - used

Data Available
- Protdata1
  - derivedFrom MassSpec-Labwork1

- Labwork Activity1
  - used tissue1 collectedFrom Patient1

Workflows Available
- hasWorkflowTemplate

- hasMetaworkflow Bindings
  - hasWorkflowTemplate

- hasData Query
DISK Hypothesis Ontology
http://disk-project.org/ontology/disk/

The DISK Ontology

Release 2016-09-02

This version:
http://disk-project.org/ontology/disk/disk-1.

Latest version:
http://disk-project.org/ontology/disk/

Revision:
1.0.0

Authors:
Varun Ratnakar, (Information Sciences Institute)
Daniel Garijo, (Information Sciences Institute)
Yolanda Gil, (Information Sciences Institute)

Our requirements:
- Hypothesis statement
- Hypothesis provenance
- Hypothesis evolution
- Hypothesis attribution

Classes

<table>
<thead>
<tr>
<th>Bundle</th>
<th>Confidence report</th>
<th>Data</th>
<th>Execution</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line of Inquiry</td>
<td>Metaworkflow</td>
<td>Plan</td>
<td>Statement</td>
<td>Triggered line of inquiry</td>
</tr>
<tr>
<td>Variable</td>
<td>Variable binding</td>
<td>Workflow</td>
<td>Workflow binding</td>
<td></td>
</tr>
</tbody>
</table>
Related Work: Hypothesis Vocabularies

- SWAN [Gao et al 2006]
- EXPO [Soldatova and King 2006]
- Nanopublications [Groth et al 2010]
- Ovopublications [Callahan and Dumontier 2013]
- Micropublications [Clark et al 2014]
- LSC (Linked Science Vocabulary)
- BEL (Biological Expression Language)

Our requirements:
- Hypothesis statement
- Hypothesis provenance
- Hypothesis evolution
- Hypothesis attribution
Representing Analytic Methods as Workflows

- **Hypothesis**
- **Revised hypothesis**
- **Lines of Inquiry**
  - Specify relevant analytic methods (workflows), type of data needed, and how to combine results

---

**Query to Retrieve Data**

**Analytic Workflows**
- Workflow Library

**Meta-Workflows**
- Confidence Estimation
- Benchmarking

**DATA CATALOG**
Multi-Omics Workflows in WINGS
Semantic Workflows in WINGS

- WINGS represents algorithms with same function as subclasses
- WINGS represents characteristics of each algorithm with semantic constraints
  - E.g.: xInteract is not compatible with MyriMatch
Generating Multiple Workflow Variants from a High-Level Workflow in WINGS

Automatic generation of valid candidates

Rules out invalid candidates given step constraints and input data characteristics

1. Peptide_Search-4-e5fb0c7b-700c-47e1-be36-2ad02ff5b305
   - comet: 1 job
   - xInteract: 1 job

2. Peptide_Search-4-03af0909-341f-48b8-b7df-cbdd5d6ba590
   - xTandem: 1 job
   - xInteract: 1 job

3. Peptide_Search-4-277a05b0-2298-41d6-97c1-cc14b83043cc
   - msgfPlus: 1 job
   - xInteract: 1 job

4. Peptide_Search-4-d291b1ec-a1e3-44bd-b082-c91787aaba47
   - sequest: 1 job
   - xInteract: 1 job

5. Peptide_Search-4-13ec7c1a-3b96-4fdf-8486-8774090c7e49
   - mascot: 1 job
   - xInteract: 1 job
Reasoning about Workflow Constraints [Gil et al JETAI 2011]

Work with P. Gonzalez (UCM) and Jihie Kim (ISI)

Workflows developed with Y. Liu (USC) and C. Mattmann (JPL)

Seed workflow from request
find input data requirements
parameter selection
workflow instantiation
workflow grounding
workflow mapping
workflows ranked
executable workflows

unified well-formed req.
seeded workflows
binding-ready workflows
bound workflows
configured workflows
workflow instances
ground workflows
top-k workflows
executable workflows

Online LDA
Parallel LDA
WINGS Workflow Reasoners: Propagating Constraints

?Model5 dcdom:isDiscrete true
?Model6 dcdom:isDiscrete true
?Model7 dcdom:isDiscrete true
?TestData dcdom:isDiscrete true
?Dataset4 dcdom:isDiscrete true
?Dataset3 dcdom:isDiscrete true
?Dataset4 dcdom:isDiscrete true
Assembling Analytic Results with Meta-Workflows

Hypothesis

Revised hypothesis

Lines of Inquiry
Specify relevant analytic methods (workflows), type of data needed, and how to combine results

Query to Retrieve Data

Analytic Workflows

Meta-Workflows

Workflow Library

Confidence Estimation

Benchmarking

DATA CATALOG
Meta-Workflows to Combine Multi-Omics Results

After running the workflows, meta-workflows analyze their results and generate a confidence value.

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Workflow 0</th>
<th>Workflow 1</th>
<th>Workflow 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow 0</td>
<td>1.00</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>Workflow 1</td>
<td>0.96</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>Workflow 2</td>
<td>0.72</td>
<td>0.84</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Provenance Recording and Explanation

**Lines of Inquiry**
Specifies relevant analytic methods (workflows), type of data needed, and how to combine results

**Analytic Workflows**
Workflow Bindings

**Meta-Workflows**
Confidence Estimation
Benchmarking

**Interactive Explanation**
Revised hypotheses & interesting findings
Representing Provenance: W3C PROV and OPM

PROV

[Moreau et al 2013]

OPM

[Moreau et al 2011]
The W3C PROV Ontology Standard
https://www.w3.org/TR/prov-o/

PROV-O: The PROV Ontology

W3C Recommendation 30 April 2013

This version:
http://www.w3.org/TR/2013/REC-prov-o-20130430/

Latest published version:
http://www.w3.org/TR/prov-o/

Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/

Previous version:
http://www.w3.org/TR/2013/PR-prov-o-20130312/

Editors:
Timothy Lebo, Rensselaer Polytechnic Institute, USA
Satya Sahoo, Case Western Reserve University, USA
Deborah McGuinness, Rensselaer Polytechnic Institute, USA

Contributors:
(In alphabetical order)
Khalid Belhajjame, University of Manchester, UK
James Cheney, University of Edinburgh, UK
David Corsar, University of Aberdeen, UK
Daniel Garjo, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Stian Soiland-Reyes, University of Manchester, UK
Stephan Zednik, Rensselaer Polytechnic Institute, USA
Jun Zhao, University of Oxford, UK
Workflow Ontologies

Generic Provenance

Plan Execution

Workflow execution

Plan Definition

Workflow template

Execution of

OPM, PROV

P-Plan

OPMW
Plan Ontology Based on PROV: P-PLAN

[Moreau et al 2013]

[Garrio and Gil 2013]
P-PLAN:
http://vocab.linkeddata.es/p-plan/

The P-PLAN Ontology

Release 12 March 2014

This version:
http://vocab.linkeddata.es/p-plan/version/03062014/

Latest version:
http://purl.org/net/p-plan#

Previous version:
http://vocab.linkeddata.es/p-plan/version/13032014/

Revision
1.3

Authors:
Daniel Garijo, Ontology Engineering Group, Universidad Politécnica de Madrid
Yolanda Gil, Information Sciences Institute, University of Southern California, US

Extended Ontologies:
PROV-O: The PROV Ontology
Workflow Ontology: OPMW

[Garijo et al 2012; Garijo et al 2017]
The OPMW-PROV Ontology

Release 22 December 2014

This version:  
http://www.opmw.org/model/OPMW_20141222

Latest version:  
http://www.opmw.org/model/OPMW

Previous version  
http://www.opmw.org/model/OPMW_20140711

Revision  
Revision 6

Authors:  
Daniel Garijo, Universidad Politécnica de Madrid  
Yolanda Gil, Information Sciences Institute, University of Southern California, US

Imported Ontologies:  
The Open Provenance Model Ontology (OPMO)  
The Open Provenance Model Vocabulary (OPMV) (indirect import)  
The PROV-O Ontology  
The P-plan Ontology
Abstraction Heterogeneity: Mapping Across Models Through Queries

Returning P-Plan from OPMW objects:

CONSTRUCT{
  ?activity2 p-plan:isPrecededBy ?activity. }
WHERE{
  ?activity2 opmw:uses / opmw:isGeneratedBy ?activity. }

Returning PROV from OPMW objects:

CONSTRUCT{
  ?activity a prov:Activity.
  ?activity2 a prov:Activity.
  ?activity2 prov:used ?u1.
  ?u1 prov:wasGeneratedBy ?activity. }
WHERE{
  ?u1 opmv:wasGeneratedBy ?activity. }
Automating Scientific Data Analysis

Knowledge about scientific processes, data, models, analytic tools, hypotheses, theories...

Meta-Strategies and Lines of Inquiry
- Determine what data is needed
- Choose analytic approach
- Combine results from diverse data

Semantic Workflow Reasoning
- Select among similar algorithms
- Select parameters based on data
- Automated workflow refinement

Workflow Mapping & Execution Management
- Map workflow tasks to resources
- Optimize for performance, reliability
- Manage new data products

Distributed Resource Management
- Large scale computations
- Remote job submission
- Management of job dependencies

Strategies for large-scale execution and knowledge about resources

Tools:
- Spark
- OODT
- Pegasus
- Hadoop
- Condor
Conclusions

■ Integrative studies of multi-source data are very rare
  • DISK represents data analytics expertise across different areas as lines of inquiry, workflows, and meta-workflows

“Automated Hypothesis Testing with Large Scientific Data Repositories.”

■ Data resources are constantly growing but studies are done once
  • DISK automatically reassesses hypotheses as new data becomes available

“Towards Continuous Scientific Data Analysis and Hypothesis Evolution.”

■ Reproducibility and explanation are key when multiple collaborators
  • DISK automatically generates data narratives from workflows