Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems, Evanston, IL, June 2016

Advances in Cognitive Systems (2016) Submitted 4/2016; published 6/2016

Automated Hypothesis Testing with Large Scientific Data Repositories

Yolanda Gil GIL@ISI.EDU
Daniel Garijo DGARIJO@ISI.EDU
Varun Ratnakar VARUNR@ISI.EDU
Rajiv Mayani MAYANI@ISI.EDU
Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey CA, 90292, USA

Ravali Adusumilli RAVALI@STANFORD.EDU
Hunter Boyce HBOYCE @STANFORD.EDU
Parag Mallick PARAGM @STANFORD.EDU
Stanford School of Medicine, Canary Center for Early Cancer Detection, Stanford University
1265 Welch Road, Stanford CA 94305, USA

Abstract

The automation of important aspects of scientific data analysis would significantly accelerate the pace of science and innovation. Although there has been a lot of work done towards that automation, the hypothesize-test-evaluate discovery cycle is still largely carried out by hand by researchers. This introduces a significant human bottleneck, which leads to inefficiencies, potential errors, and incomplete explorations of the hypothesis and data analysis space. We introduce a novel approach to automate the hypothesize-test-evaluate discovery cycle with an intelligent system that a scientist can task to test hypotheses of interest against a data repository. Our approach captures three types of data analytics knowledge: 1) common data analytic methods represented as semantic workflows; 2) meta-analysis methods that aggregate those results, represented as meta-workflows; and 3) data analysis strategies that specify for a type of hypothesis what data and methods to use, represented as lines of inquiry. Given a hypothesis specified by a scientist, appropriate lines of inquiry are triggered, which lead to retrieving relevant datasets, running relevant workflows on that data, and finally running meta-workflows on workflow results. The scientist is then presented with a level of confidence on the initial hypothesis, a revised hypothesis, and possibly with new hypotheses. We have implemented this approach in the DISK system, and applied it to multi-omics data analysis.

1. Introduction

The rate of data collection has vastly surpassed our ability to analyze it. In science, massive amounts of data are already available in repositories, waiting to be analyzed [Tomczak et al 2015, Rudnick et al. 2016]. As these repositories are constantly growing, an analysis performed today may give different results when performed in the future. Data analytics expertise is not easily disseminated, and institutions have more data than experts to analyze it. For example, in a recent survey of reviewers of Science magazine (which could be considered to be at the top of their field) a majority of respondents said that their lab did not have the necessary expertise to analyze

© 2016 Cognitive Systems Foundation. All rights reserved.

the data they already have [Science 2011]. The situation is likely worse for the vast majority of scientists, and in less privileged institutions and sectors. Indeed, data analytic processes are currently carried out by hand by investigators, introducing a significant human bottleneck that can lead to erroneous and incomplete explorations, and hampers reproducibility [Begley and Ellis 2012].

The automation of important aspects of scientific discovery would significantly accelerate the pace of science and innovation. Although scientific discovery may involve complex representational and paradigm changes [Kuhn 1962], AI researchers have automated important aspects of discovery such as experiment design and testing [Kulkarni and Simon 1988; King et al 2009] and induction of laws from given datasets [Langley et al 1987; Valdes-Perez 1997; Todorovski et al 2000; Schmidt and Lipson 2009]. Once a representation is chosen, discovery processes often involve searching through the space of possible hypotheses and models.

Our goal is to develop an intelligent system able to conduct hypothesis-driven data analysis of science data repositories. In many science domains, comprehensive data repositories are being developed with large amounts of diverse data. Given the necessary knowledge and methods, an intelligent system could autonomously analyze the data in a systematic, comprehensive, and efficient manner [Buchanan and Waltz 2009; Gil et al 2014]. Automating data analyses would also enforce consistency, as they would follow processes recognized by experts in the field of study. In addition, automation would facilitate inspectability and reproducibility of results.

In this paper we introduce a novel approach to automate the hypothesize-test-evaluate discovery cycle by capturing data analytics expertise and applying it to automatically test given hypotheses against existing data repositories. Our approach captures three types of data analytics knowledge: 1) common data analytic methods that can each be run on different types of data, represented as *semantically-enriched computational workflows*; 2) meta-analysis methods that aggregate disparate results from different data analysis methods, represented as *meta-workflows*; and 3) data analysis strategies that specify for a given type of hypothesis what kind of data to retrieve from the repository and methods (workflows, meta-workflows) to use, represented as *lines of inquiry*. Given a hypothesis specified by a scientist, appropriate lines of inquiry are triggered, which lead to retrieving relevant datasets, running appropriate workflows on that data, and finally running meta-workflows on workflow results. The final report generated includes a revised (possibly new) hypothesis based on the data and methods applied, and a detailed provenance record of the results that can be used to reproduce the work.

We have implemented this approach in the DISK framework, and used it for multi-omics data analysis. A key capability of DISK is the ability to capture of data analysis expertise in semantic workflows that can be automatically elaborated in the existing WINGS intelligent workflow system [Gil et al 2011a; Gil et al 2011b]. For a given domain, DISK is given a data catalog that describes the data repository in terms of metadata of its individual datasets using domain ontologies. In addition, DISK is pre-populated with workflows, meta-workflows, and lines of inquiry for the domain at hand, all described using ontologies that are consistent with the metadata in the data catalog. DISK automatically captures the provenance of the results, which can be used to generate appropriate explanations of new findings to scientists and to reproduce the results.

Major research contributions of this paper include a representation of common lines of inquiry that test hypotheses against data repositories, semantic workflows to capture common data analysis methods, and meta-reasoning strategies to aggregate data analysis results.

The paper begins with an overview of our approach, followed by a description of its implementation in the DISK system. We then present an initial evaluation with a functional prototype for the multi-omics domain. We conclude with a discussion of future lines of work.

2. An Opportunity in Multi-Omics Data Repositories for Cancer

Despite great advances in our ability to discover genomic abnormalities, the relationship between the genome and the functioning or malfunctioning of cells remains poorly understood [Ritchie et al 2015]. Multi-omic analysis enables the study of the genome (genomics data), its products which include expressed RNAs and proteins (transcriptomics and proteomics data respectively), and how those products interact amongst themselves and with the genome to drive cell behavior (phenotypic data). Understanding these relationships is crucial to uncover the mechanisms that lead to cancer and other diseases.

Projects like The Cancer Genome Atlas (TCGA) [Tomczak et al 2015] and the associated Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Rudnick et al. 2016] are creating large repositories of omics data that are rapidly approaching more than a petabyte of data. The data is collected in a defined, and relatively uniform way at dozens of sites for thousands of patients with different types of cancer. The data include diverse omics data, such as multiple types of genomic data (DNA sequencing, RNA transcriptomics, epigenetic) and proteomics, as well as pathologic data from biopsy (H+E), radiomic data (CT, MRI), and extensive clinical annotations. Though efforts have been made to minimize experimental variation, each type of data may be collected by a variety of approaches and from site-to-site on different types of instruments and through different (but standardized) protocols. For example, protein expression data may be collected in some cases through mass spectrometry (which is more expensive but more accurate) and in others through fluorescence and microarray experiments (more inexpensive but less accurate). Likewise, mass spectrometric datasets may be collected using TMT approaches at one center and spectral-counting approaches at another. An important feature of TCGA and CPTAC is that metadata regarding how samples were processed and what measurement modalities were used is available. The availability of these fairly complete and well-annotated datasets is a key enabler for investigating automating discovery from data repositories.

Although comprehensive data repositories such as TCGA and CPTAC are the target of a wide range of omics research, the analysis is occurring slowly and piecemeal. Different labs have expertise in specific types of data (e.g., one lab in genomics, another in proteomics), and consequently they each analyze narrow slices of the data available. In addition, analyses that use several types of omics data are infrequent, as they involve several labs and take several years to complete (e.g., [Zhang et al 2014; TCGA 2014]). Each type of omics data requires the use of several interconnected software tools, and each of them may require substantial domain knowledge. For example, a tool for clustering genomics data may have a parameter whose value is set based on the error rates of the microarray instrument used to collect the data. An analysis method for proteomics data collected through mass spectrometry may involve several tools, each with dozens of such constraints. The situation is exacerbated as new approaches published in the literature constantly expand the possible methods for data analysis and offer alternative approaches to gather evidence from data. Each lab uses some subset of the available algorithms and methods given their resources and expertise.

A major issue is the limited reproducibility of omics studies. Each lab has their own methods and associated software stack and infrastructure to carry out the analyses. Studies are not easily reproducible based on what is described in published articles [Ioannidis et al 2009]. In addition, it is not unusual that different analytic methods lead to vastly different results (e.g., [O'Rawe et al 2013]). This suggests a lack of consistency on how methods are applied, and that reproducibility and cross-comparison of methods should be commonplace.

The seminal multi-omics analysis in [Zhang et al 2014], mentioned above, was a study of colon cancer through genomics and proteomics data that confirmed the role of some known proteins and uncovered additional ones. These kinds of studies offer a watermark for our work, as we can develop intelligent systems that aim to emulate such discoveries by capturing substantial amounts of analytic knowledge that is not easily harness to conduct new analyses. It would also be exciting to the cancer omics community to see such analyses become transparent and reproducible.

3. A Framework for Automated Hypothesis Testing with Data Repositories

Our aim is to design a framework for automated hypothesis testing based on the kinds of knowledge that experts express in exploring, testing, and revising hypotheses.

Our approach is to capture general data analysis strategies that scientists would follow to test a given hypothesis. This includes knowledge for finding appropriate datasets in the data repository that are relevant to the hypothesis at hand, the data analysis methods that should be applied to all the datasets available, and meta-analysis methods that examine those results and synthesize an overall assessment of the initial hypothesis. These general data analysis strategies tend to be very prescriptive in terms of the major steps to be carried out, with the details being sorted out to suit the data and hypothesis at hand. Many aspects of our approach can be seen as a form of skeletal planning, where major steps of the process are explicitly stated at a high level and constraint reasoning is used to specialize the steps to suit the given problem [Friedland and Iwasaki 1985].

A key novel contribution of our work is to capture computationally the end-to-end scientific methods to test classes of hypotheses. While there are many machine learning methods to analyze a given dataset, they do not address how to find the data to be analyzed. Also, each machine learning method is one approach to do analysis, and a significant part of the end-to-end analysis process is to figure out which machine learning methods should be applied to the data at hand.

We describe our approach in two stages. First we introduce basic concepts and types of knowledge captured in our framework, giving examples from multi-omics data analysis. Then we describe the core algorithm and the role of each type of knowledge in its reasoning steps. In the next section, we describe our implementation of this approach in the DISK system and after that we show preliminary results and examples in multi-omics data analysis.

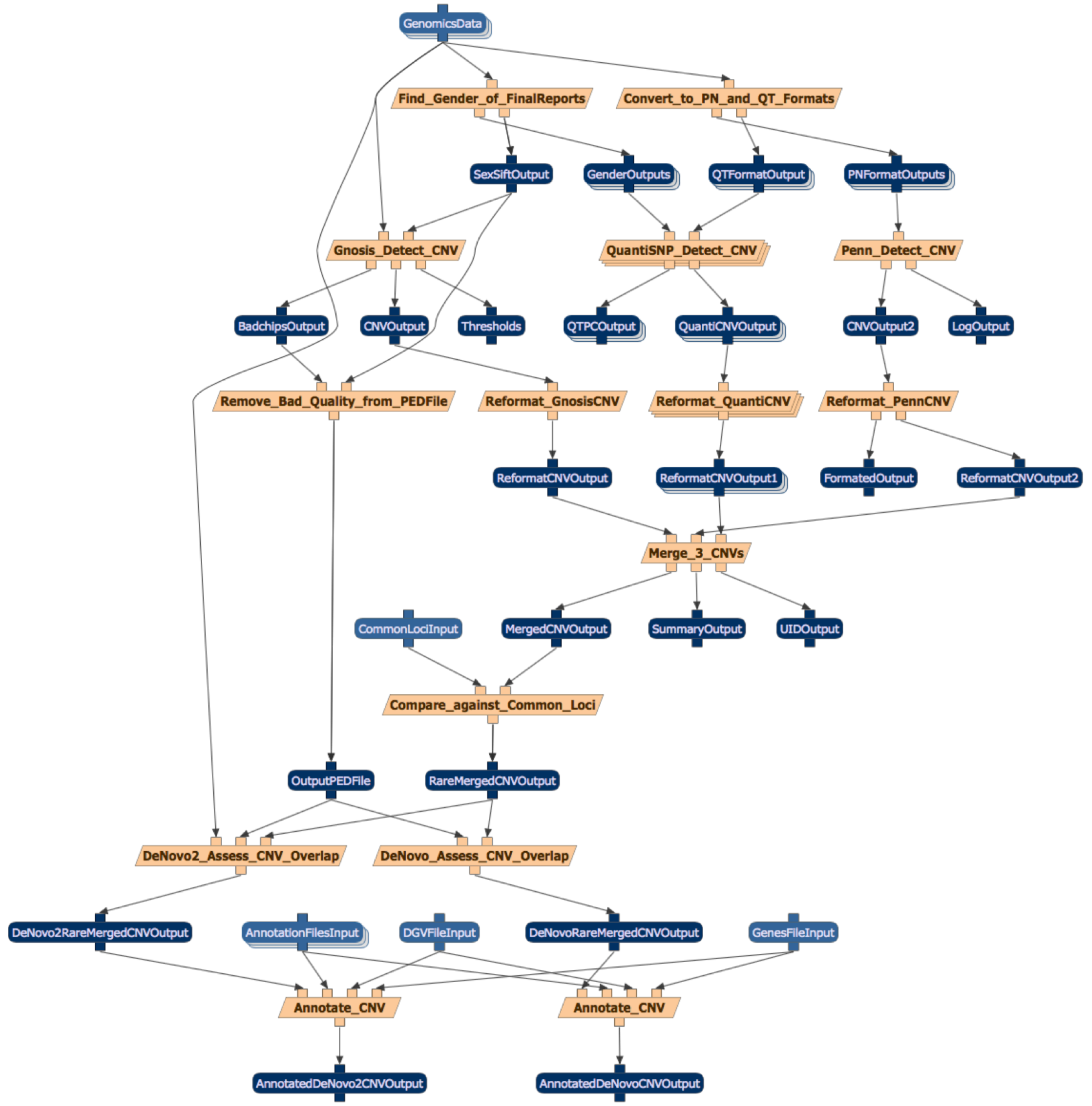


Figure 1. An example of a genomics workflow for copy number variant detection (from the work in [Gil et al 2013]). Data is shown in ovals (inputs in lighter color), and computational analytic steps are shown as rectangles.

We first introduce some terminology that will be used through the paper. A *computational workflow*, or workflow for short, defines the interconnected tasks that are needed to carry out a computational experiment [Taylor et al 2007]. Figure 1 shows an example of a workflow to detect copy number variants, from the work in [Gil et al 2013]. A workflow has *inputs*, which can be data or parameters, and produces intermediate results and *outputs*. The workflow in the figure processes genomics data and uses an ensemble of three well-known CNV detection algorithms

(Gnosis, PennCNV, and QuantiSNP), then compares the findings with known CNVs and annotates those that are novel. Workflow inputs are variables that can be given *bindings* to specific datasets in order to create an *execution-ready workflow*. For example, the workflow in Figure 1 can be used for data from a specific genome (e.g., a patient) to create an execution-ready workflow that can be run. *Unification* is a process to match ground literals against variabilized expressions, resulting in a set of bindings for the variables to the constants in the ground literals.

3.1 Representing Hypotheses

A *hypothesis* h consists of:

1. A *hypothesis statement*, which is a set of assertions about entities in the domain. For example, they express assertions such as "The mutant form of Protein ABCD is associated with colon cancer."
2. A *hypothesis qualifier*, which qualifies the veracity of the hypothesis based on the data and the analyses done so far. A typical qualifier is a numeric confidence value. For example, for the hypothesis statement above we could have a confidence value of 0.7.
3. A *hypothesis provenance*, which is a record of the analyses that were carried out to test the hypothesis statement. For example the provenance may include an analysis of mass spectrometry data for 25 patients with colon cancer and 25 healthy controls followed by clustering, cluster metrics and binary hypothesis testing.
4. A *hypothesis history*, which points to prior hypotheses that were revised to generate the current one. In our example, a prior hypothesis could have a statement such as "Protein ABCD is associated with cancer."

An appropriate language to represent each of these constituents should be chosen to fit the domain. We describe in the next section our work on representing hypothesis in DISK for the multi-omics domain.

3.2 Representing Lines of Inquiry

A *line of inquiry* represents potential analyses that can be pursued to test a type of hypothesis. A line of inquiry consists of:

1. A *hypothesis pattern*, which represents the type of hypotheses that can be explored with this line of inquiry. An example of a hypothesis pattern is "Protein ?p is associated with cancer ?c". This hypothesis pattern must be expressed in the same language as hypothesis statements, so that they can be matched against a user's hypothesis.
2. A set of *query patterns*, representing the kinds of data relevant to testing the hypothesis pattern as a set of templates of queries to a data repository. Several kinds of data may be relevant, so there may be several query patterns. For example, "Retrieve data of mass spectrometry experiments of tumor samples from patients with cancer ?c".
3. A set of *workflows*, which are procedures that capture data analysis methods that are applied to the data retrieved by the query patterns in order to test the hypothesis pattern.

For example, a workflow to analyze mass spectrometry data could include steps such as matching proteins from either a patient's custom database or on a reference human proteomic database to tandem mass spectrometry data.

4. A set of *workflow mappings*, indicating how the query patterns and the datasets retrieved should be used to instantiate the workflows above to create execution-ready workflows.
5. A *meta-workflow*, which describes how to aggregate the results of each of the workflow executions done to analyze the data, and return a revised hypothesis statement as a result. For example, suppose that one of the workflows in the line of inquiry is for analyzing protein mass spectrometry data for n patients finding evidence for the protein ?p with a p-value p1 and another workflow is for protein fluorescence data for m patients giving a p-value p1, then the meta-workflow would indicate how to combine that evidence into a joint confidence value.
6. A *meta-workflow mapping*, which describe how the results of the workflow executions are to be used to generate bindings for the inputs of the meta-workflow.

Our work makes two important assumptions. First, we assume that the lines of inquiry are in a total order. This is necessary so that for any hypothesis statement we are guaranteed to be able to select only one line of inquiry to pursue, which is the one ranked highest. If this assumption is relaxed and we allow a partial order, further meta-reasoning mechanisms would be needed to manage several lines of inquiry.

A second assumption that we make is that there is only one meta-workflow in each line of inquiry, and its output is a revised hypothesis statement with an associated qualifier (e.g., a confidence value). If this assumption is relaxed and we allow more than one meta-workflow, then our approach would need to be extended so that it would be clear how the results of different workflows would be combined to revise the initial hypothesis.

Lines of inquiry capture important knowledge to select data, analyze it, and synthesize a result. They orchestrate data retrieval, data analysis, and meta-reasoning processes. The incorporation of these types of knowledge and processes is a major novel aspect of our framework.

3.3 An Algorithm for Autonomous Hypothesis Testing

Table 1 gives a high-level overview of our algorithm for automated hypothesis testing with a data repository. Given a hypothesis, a knowledge base composed of lines of inquiry, a data repository, and a set of computational resources that limit the amount of analysis that is possible, the algorithm returns a revised hypothesis that may change the initial hypothesis statement (e.g., by making it more specific) or change its initial confidence value.

Initially, the hypothesis provided can just contain a hypothesis statement, which gets refined by the algorithm, and can be refined again over new iterations of the algorithm as new data or methods become available. Our framework is designed so that this algorithm can be iterated over time with a set of standing hypotheses of interest. As new data becomes available in the data repository, new analyses can be run which result in updates to the confidence value and the evidence associated with each hypothesis.

Table 1. High-Level algorithm for autonomous hypothesis testing against a data repository.

```
Given:
  - A hypothesis $H = \{H_{stat}, H_{qual}, H_{prov}, H_{hist}\}$
  - A ranked list of lines of inquiry, each as $LOI = \{H_{patt}, Q_{patts}, W, W_{mapps}, M, M_{mapps}\}$
  - A data repository DR

Return: $H' = \{H'_{stat}, H'_{qual}, H'_{prov}, H'_{hist}\}$ as a revised hypothesis for H

Do:
  1. Match $H_{stat}$ in H against $H_{patt}$ in every LOI, and of those that match select the one with the highest
     ranking, return a set of bindings $B^H$ for the variables in $H_{patt}$ to the constants in $H_{stat}$ of the selected LOI
  2. Instantiate the query patterns $Q_{patts}$ using the bindings $B^h$, returns a set of instantiated queries {Q}
  3. Query the data repository DR using {Q}, return a collection of datasets {D} for each query as {Q, D}
  4. For the input data to each workflow in W use $W_{mapps}$ and {Q, D} to generate bindings $B^W$, return a
     set of execution-ready workflows {WE}
  5. Run all workflows in {WE}, add the provenance records $P^W$ to $H'_{prov}$
  6. For the input data to each meta-workflow in M use $M_{mapps}$, $H_{patt}$, and $P^W$ to generate bindings $B^M$,
     return an execution-ready meta-workflow ME
  7. Run the execution-ready meta-workflow ME, add the provenance records $P^M$ to $H'_{prov}$
  8. Using the results of ME create $H'_{stat}$ and $H'_{qual}$, link H to $H'_{hist}$, return a revised hypothesis
     $H' = \{H'_{stat}, H'_{qual}, H'_{prov}, H'_{hist}\}$
```

4. DISK: Automated Hypothesis Testing with Large Data Repositories

We have developed the DISK (Automated DIscovery of Scientific Knowledge) system, which implements our approach. Figure 2 gives an overview of the components of DISK. In this section, we describe the representation of hypotheses, and their evaluation through lines of inquiry that include analytic workflows and meta-workflows. The workflows have semantic constraints that are used in the WINGS workflow system to automatically configure the workflows to the data at hand. There is an Interactive Discovery Agent component that that generates explanations for the user about the findings of the system.

4.1 Hypotheses in DISK

We use graphs of entities and relations to represent hypotheses. Graphs represent the hypothesis statement, qualifier, provenance, and history, as well as the interconnections among them. We use OWL and RDF [W3C 2014], both W3C semantic web standards widely used in biomedical research. These representation languages enable nested graphs so that a provenance graph can be attached to a hypothesis statement graph, as can the hypothesis qualifiers and history graphs.

Figure 3 illustrates our hypothesis representation with an example. The hypothesis statement indicates that the PRKCDBP protein is expressed in a patient's sample, one of the many samples that have been analyzed in the lab with results available in TCGA. The hypothesis was tested with two workflows (W1 and W2), using the dataset TCGA-AA-3561-01A-22 obtained from the patient's sample and described with the assertions shown at the bottom of the figure. The results of W1 and W2 were analyzed with meta-workflow M1, and that becomes the provenance of the hypothesis. The analysis concluded that the confidence on the hypothesis is 0.2, which becomes part of its qualifiers. The history is not illustrated in the figure, but it is also expressed as a graph.

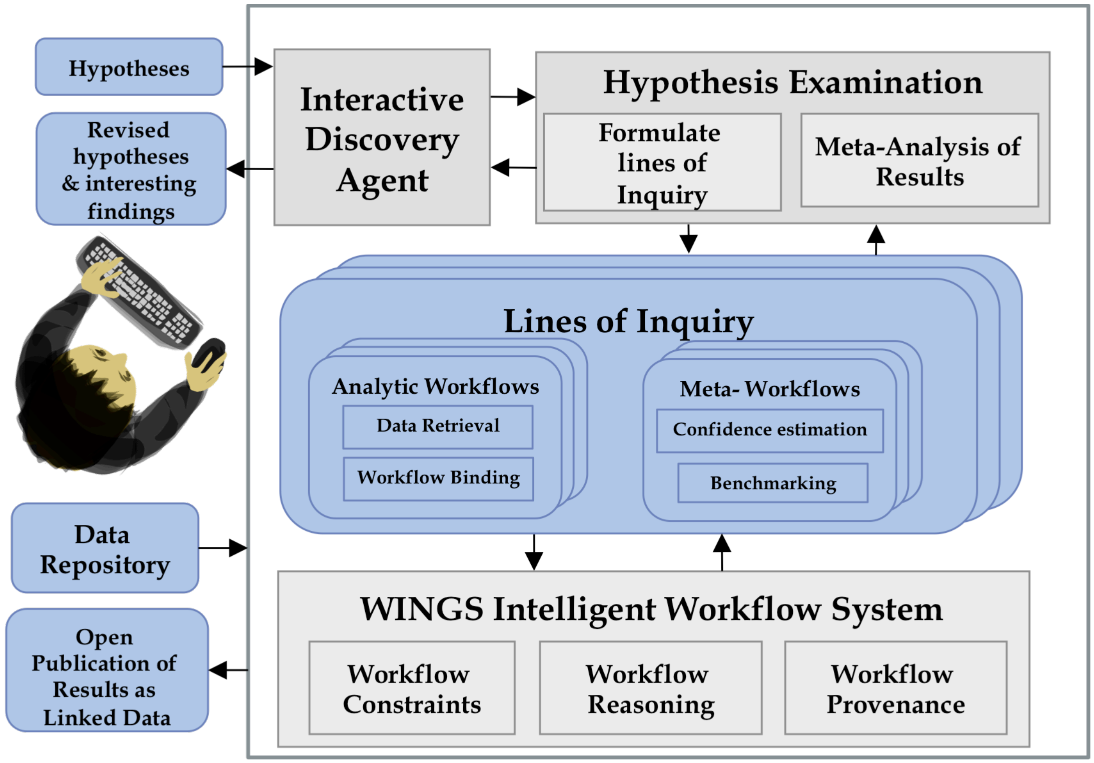


Figure 2. Major components of DISK

We use the W3C PROV standard [Moreau et al 2013] to capture how the hypothesis is supported by the analysis. For example, we use it to express that a workflow *used* some data as input that a workflow execution was *used* by a meta-workflow, that a confidence report *was generated by* some meta-workflow, etc. We also use W3C PROV to represent the hypothesis history, which captures the chain of events leading to a hypothesis revision. For example, we use it to express that a hypothesis was *a revision of* another, a value *was generated by* some activity, a workflow *used* some data as input, etc.

4.2 Workflows in DISK

DISK uses WINGS [Gil et al 2011] to represent data analytic workflows. WINGS is an intelligent workflow system that uses semantic representations to describe the constraints of the data and computational steps in the workflow. WINGS can reason about these constraints, propagating them through the workflow structure and use them to validate workflows. WINGS can run workflows on the Pegasus/Condor [Deelman et al 2005] or Apache OODT [Mattmann et al 2006] execution environments that can handle large-scale distributed data and computations. An example of a workflow in WINGS was shown in Figure 1.

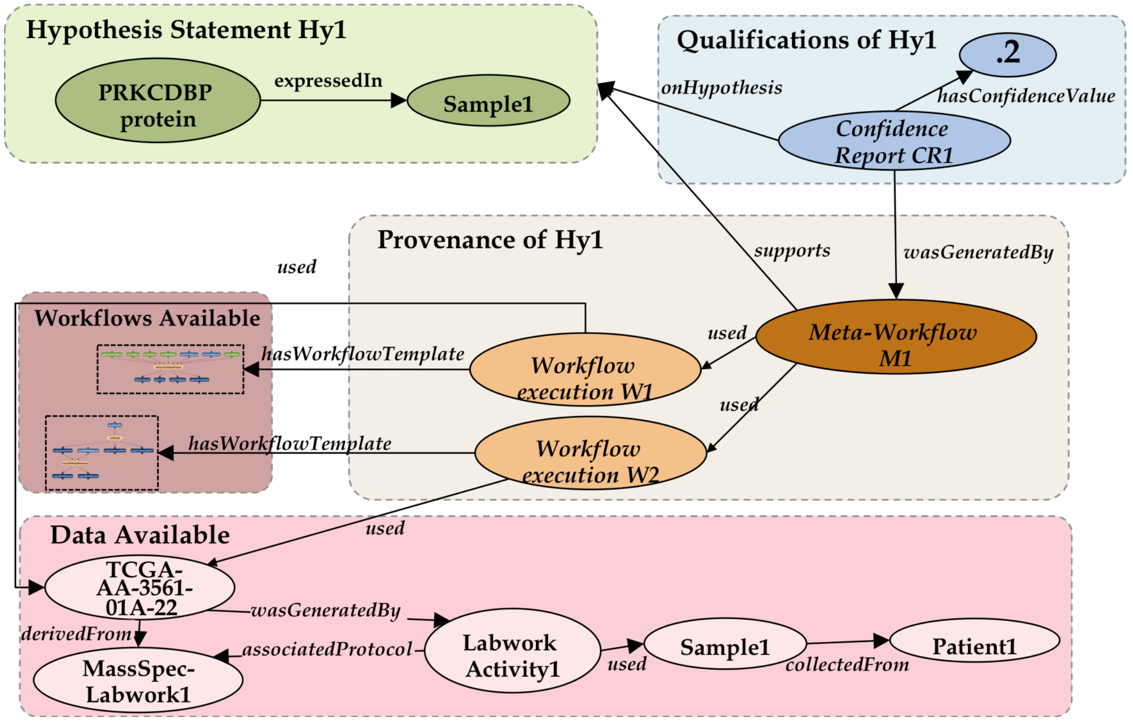


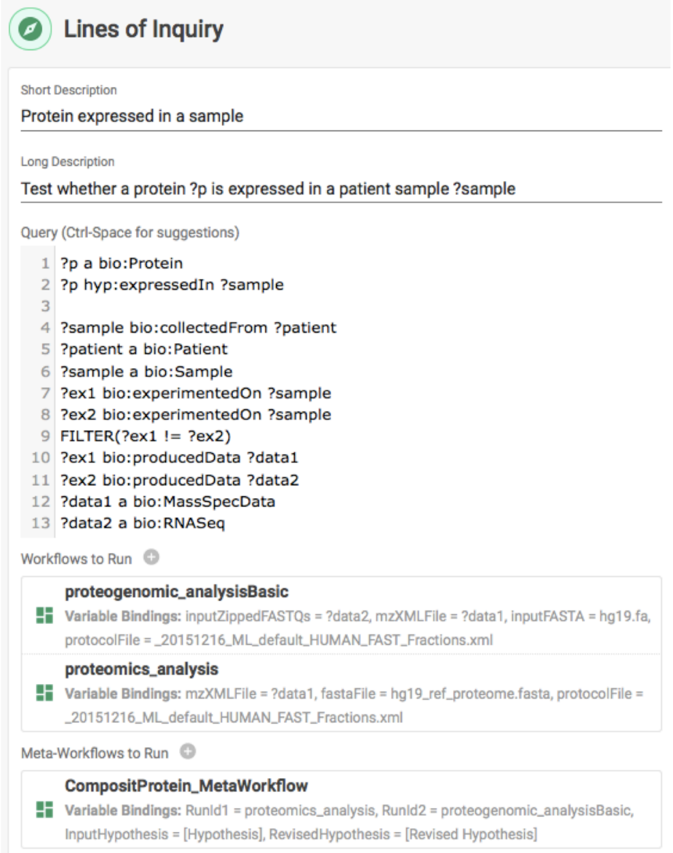
Figure 3. Representing hypotheses in DISK. The hypothesis graph is annotated with a qualifiers graph, a provenance graph. A history graph, not shown here, annotates hypothesis revisions.

4.3 Meta-Workflows in DISK

DISK uses WINGS also to represent meta-workflows. A meta-workflow has as inputs several workflow executions and compares them to generate an overall confidence value for the hypothesis. Our initial meta-workflows use simple weighted combinations of workflow results. We are creating meta-workflows that use benchmark datasets to learn how to assign these weights. Combining results from multiple sources of evidence and multiple methods is a very challenging aspect of multi-omics data analysis.

4.4 Lines of Inquiry in DISK

Lines of inquiry represent general strategies for exploring a hypothesis. Figure 4 shows an example of a line of inquiry. This line of inquiry is for testing whether a protein ?p is expressed in a sample ?sample. This is shown in line 1 of the query in the figure. To test this, a query to the data repository must be submitted to ask for experiments ?ex1 and ?ex2 done with ?sample that produced mass spectrometer data and RNASeq data respectively. The results of this query would be ?data1 and ?data2.



Lines of Inquiry

Short Description

Protein expressed in a sample

Long Description

Test whether a protein ?p is expressed in a patient sample ?sample

Query (Ctrl-Space for suggestions)

```
?p a bio:Protein
?p hyp:expressedIn ?sample

?sample bio:collectedFrom ?patient
?patient a bio:Patient
?sample a bio:Sample
?ex1 bio:experimentedOn ?sample
?ex2 bio:experimentedOn ?sample
FILTER(?ex1 != ?ex2)
?ex1 bio:producedData ?data1
?ex2 bio:producedData ?data2
?data1 a bio:MassSpecData
?data2 a bio:RNASeq
```

Workflows to Run

proteogenomic_analysisBasic
Variable Bindings: inputZippedFASTQs = ?data2, mzXMLFile = ?data1, inputFASTA = hg19.fa, protocolFile = _20151216_ML_default_HUMAN_FAST_Fractions.xml

proteomics_analysis
Variable Bindings: mzXMLFile = ?data1, fastaFile = hg19_ref_proteome.fasta, protocolFile = _20151216_ML_default_HUMAN_FAST_Fractions.xml

Meta-Workflows to Run

CompositProtein_MetaWorkflow
Variable Bindings: RunId1 = proteomics_analysis, RunId2 = proteogenomic_analysisBasic, InputHypothesis = [Hypothesis], RevisedHypothesis = [Revised Hypothesis]

Figure 4. A line of inquiry in DISK to test hypothesis about whether a protein is expressed in a patient's sample.

Those datasets are then used to run two workflows as the figure indicates: a proteogenomic analysis and a basic proteomics analysis. The proteogenomic analysis workflow will use both ?data1 and ?data2, and in addition requires comparisons with the reference human protein sequence database. The line of inquiry specifies that the workflow should use the latest builds of that database, indicated as hg19.fa and _20151216_ML_default_HUMAN_FAST_Fractions.xml. These all provide bindings for the workflow. The basic proteomics analysis only uses ?data1, the mass spectrometer data. Finally, a meta-workflow (shown at the bottom) analyzes the results of those two workflow runs, RunId1 for the proteomics workflow and RunId2 for the proteogenomics workflow.

4.5 Generating Explanations from Provenance Records

Provenance records are crucial to generate explanations to the scientist, and are a key component of the DISK Interactive Discovery Agent. DISK captures detailed aspects of provenance. Provenance is captured for hypothesis statements, as was shown in Figure 3, recording what workflows and meta-workflows were executed to generate confidence values. Provenance is also recorded in terms of hypothesis evolution, where DISK records new versions of the hypothesis that result when new data (or new workflows) become available. Finally, provenance is also recorded for workflow executions. Workflow provenance is captured automatically by WINGS and stored in an open repository as linked data [Garijo and Gil 2011], which makes the workflow accessible to others as a collection of web objects inputs, selected parameters, software components, and final results so they can be reproduced by others.

All the concepts and relationships necessary for the representation of hypotheses and lines of inquiry are represented in ontologies, and are available at a persistent site [Garijo et al 2016].

5. Preliminary Evaluation

Our preliminary evaluation focuses on the following hypothesis about our approach: Automated hypothesis testing will get comparable results to manual analyses done by experts. The results will be comparable because our framework captures expert knowledge to select data and methods for analysis and to combine the results.

To evaluate this hypothesis, we replicated significant portions of the comprehensive analyses performed by the CPTAC team and reported in [Zhang et al 2014]. In that article, genomics and proteomics data were used to find evidence of specific proteins appearing in colon cancer patients. Multi-omic analysis is often able to reveal mechanisms that cannot be detected from gene sequencing data alone as most cellular behavior is controlled by proteins.

The initial hypothesis given to DISK was that Protein kinase C delta-binding protein (PRKCDBP) is expressed in a specific patient sample. More advanced hypotheses involving multiple proteins and multiple patient samples require large-scale computational resources to run the analysis workflows. This simpler hypothesis allowed us to demonstrate that our lines of inquiry, workflows, and meta-workflows can replicate the results of the original study.

The formal representation of the hypothesis in DISK is:

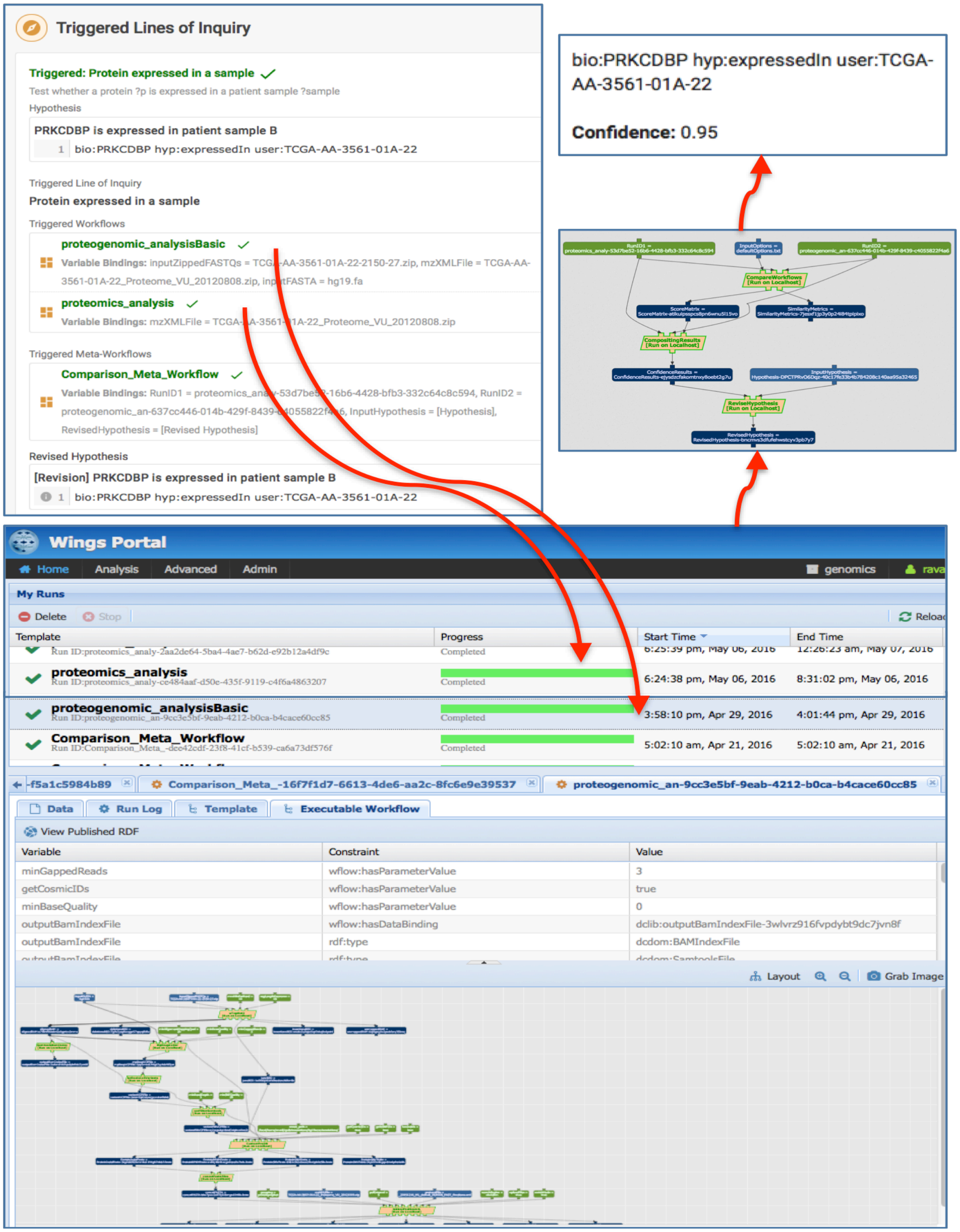
<bio:PRKCDBP hyp:expressedIn tcga:TCGA-AA-3561-01A-22>

The prefixes indicate that the entities are described in different ontologies: expressedIn is in the hypothesis ontology, the sample data is in the TCGA ontology, and the PRKCDBP protein is in the general biology ontology.

The hypothesis was matched against the available lines of inquiry. The most specific line of inquiry matched is the one that is shown in Figure 4. When our hypothesis matched the line of inquiry shown above, the query to the data repository returned a binding of ?data1 to TCGA-AA-3561-01A-22_Proteome_VU_20120808.zip and of ?data2 to TCGA-AA-3561-01A-22-2150-27.zip.

Figure 5 shows the triggered line of inquiry on the top left, showing the matched hypothesis, the bindings for the two workflows and the meta-workflow. The right side of the figure shows the proteogenomics workflow executed in WINGS, and a small diagram of the meta-workflow run. Notably, the first workflow initially found the hypothesis to be false. However, the second workflow supported a revised hypothesis in which a mutant form of PRKCDBP is present in the sample. The associated confidence value is higher in the revised hypothesis. The fundamental difference between these workflows was in the use of patient-specific sequencing data in evaluating a hypothesis. The proteomics workflow only uses a reference database, whereas the proteogenomics workflow uses patient-specific data. Notably, the inclusion of this additional data was able to validate a revised hypothesis, whereas the standard proteomics workflow would have invalidated the hypothesis.

We wanted to demonstrate that DISK is able to achieve comparable results to the analyses done by the original authors. To quantify similarity or difference, we looked across a range of levels. First, we verified that when running the same tools as the manuscript on identical data sources that results are identical. Next we investigated how subtle variations on tools (e.g. replacing tools with synonymous tools X! vs Myrimatch) impact results. Specific points for evaluation included peptide identifications, protein quantifications, and proteomic/transcriptomic correlations. As expected, synonymous tool substitution led to a non-trivial change in results and



Triggered Lines of Inquiry

Triggered: Protein expressed in a sample ✓

Test whether a protein ?p is expressed in a patient sample ?sample

Hypothesis

PRKCDBP is expressed in patient sample B

bio:PRKCDBP hyp:expressedIn user:TCGA-AA-3561-01A-22

Triggered Line of Inquiry

Protein expressed in a sample

Triggered Workflows

proteogenomic_analysisBasic ✓

Variable Bindings: inputZippedFASTQs = TCGA-AA-3561-01A-22-2150-27.zip, mzXMLFile = TCGA-AA-3561-01A-22_Proteome_VU_20120808.zip, inputFASTA = hg19.fa

proteomics_analysis ✓

Variable Bindings: mzXMLFile = TCGA-AA-3561-01A-22_Proteome_VU_20120808.zip

Triggered Meta-Workflows

Comparison_Meta_Workflow ✓

Variable Bindings: RunID1 = proteomics_analy-53d7be52-16b6-4428-bfb3-332c64c8c594, RunID2 = proteogenomic_an-637cc446-014b-429f-8439-c4055822f4a6, InputHypothesis = [Hypothesis], RevisedHypothesis = [Revised Hypothesis]

Revised Hypothesis

[Revision] PRKCDBP is expressed in patient sample B

bio:PRKCDBP hyp:expressedIn user:TCGA-AA-3561-01A-22

bio:PRKCDBP hyp:expressedIn user:TCGA-AA-3561-01A-22

Confidence: 0.95

RunID1 = proteomics_analy-53d7be52-16b6-4428-bfb3-332c64c8c594
InputOptions = defaultOptions.txt
RunID2 = proteogenomic_an-637cc446-014b-429f-8439-c4055822f4a6
CompareWorkflows [Run on Localhost]
ScoreMatrix = ScoreMatrix-atikuipsspcs8pn6wnu5l15vo
SimilarityMetrics = SimilarityMetrics-7jesxf1jp3y0p24i84tpipixo
CompositingResults [Run on Localhost]
ConfidenceResults = ConfidenceResults-ejysdzcfakomtnxy8oebt2g7u
InputHypothesis = Hypothesis-DPCTPRvO6Dqz-40c17fe33b4b784208c140aa95a32465
ReviseHypothesis [Run on Localhost]
RevisedHypothesis = RevisedHypothesis-bncmvs3dfufehwstcyv3pb7y7

Template	Progress	Start Time	End Time
Run ID:proteomics_analy-2aa2de64-5ba4-4ae7-b62d-e92b12a4df9c	Completed	6:25:39 pm, May 06, 2016	12:26:23 am, May 07, 2016
✓ **proteomics_analysis** Run ID:proteomics_analy-ce484aaf-d50e-435f-9119-c4f6a4863207	Completed	6:24:38 pm, May 06, 2016	8:31:02 pm, May 06, 2016
✓ **proteogenomic_analysisBasic** Run ID:proteogenomic_an-9cc3e5bf-9eab-4212-b0ca-b4cace60cc85	Completed	3:58:10 pm, Apr 29, 2016	4:01:44 pm, Apr 29, 2016
✓ **Comparison_Meta_Workflow** Run ID:Comparison_Meta_-dee42cdf-23f8-41cf-b539-ca6a73df576f	Completed	5:02:10 am, Apr 21, 2016	5:02:10 am, Apr 21, 2016

-f5a1c5984b89
Comparison_Meta_-16f7f1d7-6613-4de6-aa2c-8fc6e9e39537
proteogenomic_an-9cc3e5bf-9eab-4212-b0ca-b4cace60cc85
Data
Run Log
Template
Executable Workflow
View Published RDF

Variable	Constraint	Value
minGappedReads	wflow:hasParameterValue	3
getCosmicIDs	wflow:hasParameterValue	true
minBaseQuality	wflow:hasParameterValue	0
outputBamIndexFile	wflow:hasDataBinding	dclib:outputBamIndexFile-3wlvrz916fvpdybt9dc7jvn8f
outputBamIndexFile	rdf:type	dcdom:BAMIndexFile
outputBamIndexFile	rdf:type	dcdom:SamtoolsFile

Figure 5. A line of inquiry triggered in DISK to test a hypothesis, showing two workflow runs (bottom) and a meta-workflow's revision of the hypothesis confidence (top right).

in support for various hypotheses. These findings further suggest that in accepting a hypothesis, it may be important to explore a variety of possible workflows within a given line of inquiry.

From a biological perspective, among the most exciting capabilities of the system is to learn novel relationships from the data. In the analysis published in [Zhang et al 2014], several key observations emerged regarding genome mutations that were detectable in expressed proteins. As discussed above for one example, we were able to verify that those conclusions could have been made using DISK.

A detailed record of the ontologies, lines of inquiry, workflows, results, and provenance described here are available at a permanent site [Garijo et al 2016].

6. Related Work

Cognitive scientists and philosophers of science have documented discovery processes (e.g., [Kuhn 1962; Craver and Darden 2013; Chandrasekaran & Nersessian 2015]), but these processes have not been automated to date. Specific aspects of discovery such as mining laws from a pre-selected homogeneous dataset have been automated (e.g., [Langley et al 1987; Valdes-Perez 1997; Todorovski et al 2000]), and such approaches could be steps of the analytic methods in our proposed approach. There is some pioneering research on autonomous discovery systems that define a broader search for hypotheses [Lenat 1977; Lindsay et al 1980]. None of that work combines complex methods that analyze multiple kinds of data and combine the results as we aim to do in our work. The most related work is on automating the experimentation cycle [King et al 2009]. This work includes hypothesis formulation, hypothesis testing through physical experiments, and revision of hypothesis from observation of the experiment results. That work focuses on the physical execution of experiments but uses simpler data analysis than DISK does.

Several workflow systems are used for scientific applications [Taylor et al 2007]. They focus on capturing low-level mechanics of how to run at each step, rather than on reasoning.

Other relevant research includes representations of hypotheses extracted from the published literature, as they use a graph-based representation. These include EXPO [Soldatova and King 2006] and nanopublications [Groth et al 2010]. However, these models assume that a hypothesis is a static entity as it is in a published article. In our work, a hypothesis is a dynamic element that may have multiple revisions and may be composed of multiple assertions each with their own provenance. Our representation introduces terms to describe evolving hypotheses and evidence.

7. Conclusions and Future Work

As data repositories continue to grow, this disparity between data collection and data analysis rates will continue to worsen unless innovative approaches are taken. Our approach is to automate the hypothesize-test-evaluate discovery cycle with an intelligent system that captures complex data analytic knowledge as data analytic workflows, result aggregation meta-workflows, and hypothesis-relevant lines of inquiry. Given a hypothesis provided by a scientist, relevant lines of inquiry are triggered which specify what data are relevant, what analytic methods to run, and the methods to aggregate those results. A revised hypothesis is presented back to the scientist, including a level of confidence based on the data and methods applied as well as a detailed provenance record that can explain the findings. We have implemented this approach in the

DISK framework, and used it in cancer multi-omics to reproduce the results of a seminal article. We are studying the use of our approach in other domains of interest.

Another important area of future work is to address the combinatorial nature of the exploration space. Many lines of inquiry may be triggered, many datasets may be relevant, and many data analysis methods may be possible. Given limited computational resources, lines of inquiry could be extended with prioritization strategies based on domain heuristics. The automatic analysis of large data repositories will raise many such scalability challenges.

Acknowledgements. We gratefully acknowledge support from the Defense Advanced Research Projects Agency with award W911NF-15-1-0555, and from the National Institutes of Health under awards 1U01CA196387 and 1R01GM117097. We also acknowledge support from the Canary Foundation. We would like to thank Suzanne Pierce and John Gentle for valuable feedback on this work.

References

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature* 483, 531–533.

Buchanan, B. G., & Waltz, D. (2009). Automating Science. *Science*, 324(5923):43-44.

Chandrasekharan, S., & Nersessian, N.J. (2015). Building Cognition: The Construction of Computational Representations for Scientific Discovery. *Cognitive Science*, 39:8.

Craver C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. The University of Chicago Press.

Deelman, E., Singh, G., Su, M., Blythe, J., Gil, Y., et al. (2005) Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming*, 13.

Friedland, P., & Iwasaki, Y. (1985). The Concept and Implementation of Skeletal Plans. *J. Autom. Reasoning* 1(2):161-208.

Garijo, D., & Gil, Y. (2011). A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. *Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science (WORKS-11)*, held in conjunction with the 2011 ACM/IEEE Supercomputing Conference.

Garijo, D., Gil, Y., Ratnakar, V., Mayani, R., Mallick, P., Adusumilli, R., & Boyce, H. (2016). *Records of the analysis reported in the ACS-2016 article*. Available from http://www.w3id.org/disk/acs2016

Gil, Y., Gonzalez-Calero, P. A., Kim, J., Moody, J., & Ratnakar, V. (2011). A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs. Journal of Experimental and Theoretical Artificial Intelligence, 23(4).

Gil, Y., McWeeney, S., & Mason, C. E. (2013). Using Semantic Workflows to Disseminate Best Practices and Accelerate Discoveries in Multi-Omic Data Analysis. *Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics using AI (HIAI)*, held in conjunction with the Conference of the Association for the Advancement of Artificial Intelligence (AAAI), Bellevue, WA.

Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014) Amplify scientific discovery with artificial intelligence. *Science*, 346(6206):171-172.

Groth, P., Gibson, A., & Velterop, J. (2010) The anatomy of a nanopublication. *Information Services and Use*, 30(1-2):52-56.

Ioannidis J.P., Allison D.B., Ball C.A., Coulibaly I, Cui X., Culhane A.C., Falchi M, Furlanello C., et al. (2009). Repeatability of Published Microarray Gene Expression Analyses. *Nature Genetics*, 41(2).

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M. et al. (2009). The Automation of Science. *Science*, 324(3).

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Deepak Kulkarni, D., & Simon, H. A. (1988). The Processes of Scientific Discovery: The Strategy of Experimentation. *Cognitive Science,* 12(2):139-175.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT Press, Cambridge, MA.

Lenat. D. B. (1977) The Ubiquity of Discovery. *Artificial Intelligence* 9(3):257-285.

Lindsay, R. K., B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg. (1980). *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*. McGraw-Hill.

Mattmann, C. A., Crichton, D. J., Medvidovic, N., & Hughes, S. (2006). A software architecture-based framework for highly distributed and data intensive scientific applications. *Proceedings of the 28th international conference on Software engineering (ICSE)*.

Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., et al. (2013). PROV-DM: The PROV Data Model. *World Wide Web Consortium (W3C) Recommendation*.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., & Lyon, G. L. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* 5(28).

Ritchie, M. D, Holzinger, E. R., Li, R., Pendergrass, S.A, & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16,85–97.

Rudnick P. A. , Markey S. P., Roth J., Mirokhin Y., et al. (2016). A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J. Proteome Res*, 15(3).

Schmidt M., & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923):81-85.

Science Staff. (2011). Challenges and Opportunities: Introduction to Science Special Issue on Dealing with Data, *Science*, 331(6018):692-693.

Soldatova, L. N. & King, R. D. (2006). An Ontology of Scientific Experiments. *Journal of the Royal Society Interface*, Vol. 3, Issue 11.

Taylor, I., Deelman, E., Gannon, D., Shields, M., (Eds). (2007). *Workflows for e-Science*. Springer Verlag.

The Cancer Genome Atlas (TCGA) Network. (2014). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337.

Todorovski, L., Dzeroski, S., Srinivasan, A., Whiteley, J. P., & Gavaghan, D. (2000). Discovering the Structure of Partial Differential Equations from Example Behaviour. *International Conference on Machine Learning*.

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*. 19(1A): A68-A77.

Valdés-Pérez, R. E. (1995) Machine Discovery in Chemistry: New Results. *Artificial Intelligence,* 74(1).

World Wide Web Consortium (W3C). (2014). *Semantic Web Languages, Vocabularies, Inference, Linked Data, Query, and Vertical Applications*. Available from http://www.w3.org/standards/semanticweb/. Last retrieved April 12, 2014.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* 513,382–387.