

Analysis of Cancer Omics Data In A Semantic Web Framework

Matthew E. Holford¹, James P. McCusker², Kei-Hoi Cheung^{3,4,5}, and Michael Krauthammer^{1,2}

¹ Interdepartmental Program in Computational Biology & Bioinformatics,

² Department of Pathology,

³ Department of Computer Science,

⁴ Center for Medical Informatics

⁵ Department of Genetics

Yale University

New Haven, CT

Abstract. Our work concerns the elucidation of the cancer (epi)genome, transcriptome and proteome to better understand the complex interplay between a cancer cell's molecular state and its response to anti-cancer therapy. To study the problem, we have previously focused on data warehousing technologies and statistical data integration. In this paper, we present recent work on extending our analytical capabilities using Semantic Web technology. A key new component presented here is a SPARQL endpoint to our existing data warehouse. This endpoint allows the merging of observed quantitative data with existing data from semantic knowledge sources such as Gene Ontology (GO). We show how such variegated quantitative and functional data can be integrated and accessed in a universal manner using Semantic Web tools. We also demonstrate how Description Logic (DL) reasoning can be used to infer previously unstated conclusions from existing knowledge bases. As proof of concept, we illustrate the ability of our setup to answer complex queries on resistance of cancer cells to Decitabine, a demethylating agent.

1 Introduction

The Yale Specialized Program in Research Excellence (SPORE) in skin cancer is a large translational cancer project, which aims at rapidly moving biological insights from the “bench to bedside”. As part of the effort, the SPORE collects skin cancer samples from mostly malignant melanoma patients and performs a multitude of Omics studies, probing the melanoma genome, epigenome, transcriptome and proteome. The idea is to integrate this data with clinical outcome information to derive prognostic and predictive biomarkers, i.e. genomic markers that predict patient survival and drug therapy effectiveness, respectively. Conventionally, these markers are either derived statistically in an unbiased fashion [33], or by prior knowledge and candidate (gene) selection [17]. We are interested in combining these approaches, and are developing means for unbiased assessment of Omics data using existing knowledge on cellular processes that affect

drug effectiveness. In particular, we are employing Semantic Web technology to create RDF graphs that define the genomic state of cancer cells and the functional annotation of the cells' molecular entities (i.e. genes or proteins). We use SPARQL to query these graphs to better understand the molecular basis of drug resistance and sensitivity.

We start by retrieving quantitative data from a large relational database, a component of the Corvus architecture [19], storing melanoma Omics data. We do this by providing a new semantic component of Corvus, a SPARQL endpoint which relies upon Hibernate¹ for Object Relational Mapping (ORM). Through this endpoint, we can dynamically create RDF graphs of the data stored within. We then merge such graphs with SKOS-converted Gene Ontology (GO) [1] information to annotate genomic elements with functional data, such as their involvement in certain cellular processes.

As a case study, we used the new Corvus SPARQL endpoint to create an RDF graph with data representing drug response to Decitabine, a demethylating agent that has been shown to be clinically active in melanoma [16]. Using SPARQL, we queried Corvus for melanoma samples with information on promoter methylation status and gene expression before and after Decitabine treatment. The resulting graph is augmented with functional annotations from GO. It is then interrogated for the molecular mechanisms explaining why some samples have better response to Decitabine treatment than others.

2 Methods

To attain these goals, we needed to build a model that integrated quantitative Omics data with functional information. Our model incorporates gene expression and methylation data for seven melanoma cell lines [13]; it also contains Gene Ontology (GO) annotations for the whole of the human genome. Expressing this model as an RDF triple store affords us a number of advantages. First, it provides a way for others to borrow from and build upon our work. It allows us to use the standardized SPARQL interface to perform queries that bridge quantitative and functional knowledge. It also gives us the capability to infer previously unstated information by reasoning over the data with a Semantic Web aware Description Logic (DL) reasoner. We attempted wherever possible to borrow terms from well-established OBO ontologies [30]. Doing so places our work under the auspices of community defined best practice and allows our model to be used in conjunction with similarly designed knowledge bases. Building the model involved the use of a variety of cutting-edge Semantic Web technologies and required the creation of several novel tools. The work proceeded along two major lines: (i). Conversion of relational data from melanoma cell lines to RDF/OWL and (ii). Integration of specific gene annotations with the Gene Ontology.

The issue of integrating quantitative and functional biological information to infer relevant new information has been frequently explored. A notable example

¹ <http://www.hibernate.org>

is HyBrow, a tool for the generation and evaluation of biological hypotheses [24]. The user can derive hypotheses from HyBrow's knowledge base of functional biological information and test them against various high-throughput data sources. BioBIKE offers an environment for users to integrate a wide variety of experimental and genomic data to reach new conclusions [11]. Originally released as a LISP interactive library [18], the software is now web-based to accommodate users lacking in programming expertise. When combined with the BioDeducta module, it enables automated reasoning [28]. Although both HyBrow and BioBIKE make extensive use of ontologies, neither is Semantic Web enabled. Recent efforts by the National Cancer Institute as part of the caBIG initiative [12] have focused on addressing the integration issue through the use of an Extraction-Transform-Load (ETL) strategy. Notably, the caIntegrator2² project uses ETL to integrate quantitative Omics data from caArray [14] and functional biological data from caBio [9]. The Bio2RDF project is notable for providing normalized URIs for a wealth of identifiers and relationships from functional biology in the hopes of allowing easier integration of diverse data sets [4].

2.1 Quantitative Data From Melanoma Cell Lines

We examined data derived from seven melanoma cell lines (WW165, YUMAC, YUGEN8, YUSAC2, YUSIT1, YULAC and YURIF). These lines have been experimentally classified using IC50 values from dose-response analysis as being either sensitive to (YUMAC, YUSAC2, YULAC, YUSIT1, YUGEN8) or resistant to (WW165, YURIF) decitabine (5-Aza-2'-deoxy-cytidine, Aza), a DNA methyltransferase inhibitor. Specifically we looked at relative methylation values prior to administration of AZA and the ratio of gene expression following administration of AZA to before. The methylation values were obtained from a Nimblegen promoter array using the Methyl-DNA immunoprecipitation (MeDIP) technique [22, 23]. Gene expression ratios were obtained using a custom 2-channel Nimblegen array. Data from both arrays are available for download through ArrayExpress³. We used the Gene Element Ontology (GELO) to align the array probes to RefSeq identifiers [32].

Rationale for building a SPARQL endpoint These data were stored in a relational database component of Corvus, a data warehouse for experimental data, which currently holds over 4 million observations from diverse Omics experiments across melanoma cell lines. Presently, Corvus exists as a Java library with object-relational mapping (ORM) accomplished through Hibernate. Quantitative cancer omics data is stored in a standard database schema specified by the ORM. We present here a new semantic interface to Corvus which retrieves data in the form of RDF triples. Unfortunately, the sheer volume of data contained within our local Corvus database would result in a triple store of such size as to be untenable for the purposes of DL reasoning. What was needed

² <http://cabig.nci.nih.gov/tools/caIntegrator2>

³ <http://www.ebi.ac.uk/arrayexpress/>

instead was a way to retrieve a subset of the Corvus model containing only the information essential to the problem at hand. Ideally this could be accomplished in a dynamic fashion.

Integration of traditional relational databases with RDF has been extensively explored in recent years [27]. Typically the approach is to create a generic mapping between relational and RDF schema. This has been done either through automatic mappings, where relational tables correspond to RDFS classes and relational columns to RDF predicates [6], or with domain-specific semantics [26]. Some tools, such as d2rq [5], provide for both and allow user customization for complex cases such as when mappings are not one-to-one. Mappings may be stored in a variety of formats, ranging from XML configuration files to custom languages such as R2O [2]. These mapping artifacts can then be used to dynamically generate SQL queries to the relational database based upon queries expressed according to the RDF schema, usually using SPARQL.

We experimented directly with the d2rq framework, which allows a relational database to be queried like a triple store using SPARQL. Using a configuration file to map Corvus database fields to RDF properties, we were able to generate SPARQL queries that retrieved a manageable subset of the Corvus database. However, we found that the SQL generated by the tool to query the relational database was inefficient and data retrieval took longer than expected. We decided instead to leverage the Hibernate mappings already part of the Corvus model to interact with the relational database. We wrote a SPARQL interface to the Corvus model which interacts directly with the Java library, taking advantage of Hibernate's ability to optimize and cache relational queries. To the best of our knowledge, although the issue of mapping SPARQL to object oriented representations such as Hibernate has been discussed [7, 15], no tools for doing this have been released to the public. Our approach is to create wrapper classes around the Hibernate mapping classes which map the property getters to RDF predicates. Indirect mappings make possible situations in which the RDF and relational schemas do not correspond one to one. Though this approach is not necessarily a universal solution, we felt that given Corvus' ability to represent such a broad swathe of Omics data, the performance gain offered by these customized mappings more than justified the up-front expense of their creation.

Corvus model to RDF mapping We mapped fields from the Corvus database to classes and relationships from OBO ontologies. In particular, we employed terms from Information Artifact Ontology (IAO) ⁴ and Ontology for Biomedical Investigations (OBI) [8]. In addition to being actively developed, these ontologies are notable for building upon the foundation Basic Formal Ontology (BFO) ⁵ and the OBO Relation Ontology (RO) [31] which were specially designed to be extensible by any biomedical ontology. This allows our modeled Corvus data to be incorporated with other OBO ontologies with relative ease. It should be

⁴ <http://code.google.com/p/information-artifact-ontology>

⁵ <http://www.ifomis.org/bfo>

noted that we are simply borrowing terms from these ontologies, not incorporating them in their entirety as doing so would have a significant deleterious effect on reasoning performance. This does not pose a hindrance to our goals as we do not need to make inferences across the whole hierarchy of terms in these ontologies. By using the terms, however, we provide an entry point for others who may wish to explore this type of inferencing in the future. Quantitative data storage in the Corvus model is centered around the *Observation* class. Instances of this class represent individual data points in a collection of data, such as an array. They contain the numerical value of the data as well as pointers to other classes indicating the type and provenance of the data. These other classes include *Dataset*, which holds metadata on experimental conditions, *Measure*, which specifies details about the type of data being measured, *Sample*, which describes the cell line being measured and *Reporter*, the genomic feature (typically a gene) for which data is being reported. We mapped *Measurement* to the IAO class *measurement datum* and used the IAO data property *has measurement value* to associate numerical data values. *Dataset* was linked to the IAO class *data set*. Individual *Observations* can be specified as belonging to a *Dataset* using the RO property *part of*. Samples were declared as instances of the OBI class *cell culture*. Association of an *Observation* with a *Sample* was done using IAO's *is about* property. *Reporter* was linked to the *Genomic Region* class from the GELO ontology. This class is defined as a superclass of the OBO Sequence Ontology's (SO) [10] *biological region* class and it attaches properties to assign a reference location for a genomic element within the genome. For the purposes of our data, *Reporters* were made instances of SO's *transcript* class, as the reference sequence (RefSeq) was used. We used the Uniform Resource Identifiers (URIs) for RefSeq sequences provided by the Bio2RDF project. Using this normalized identifier allows us to easily link with other resources describing the same genes. To capture information from the Corvus *Measure* class, instead of mapping to an instance of a class, we forwarded two of *Measure*'s fields to properties in the domain of *measurement datum*. These were the IAO *is quality measurement of* property and the IAO *has measurement unit label* property. Finally, we used the Dublin Core [34] annotation properties *title* and *identifier* to assign names for *Samples*, *Datasets* and *Reporters* and reference identifiers to *Reporters*. A detailed view of this model is provided in figure 1.

Querying the Corvus SPARQL endpoint To retrieve a subset of our Corvus database that was sufficient for our ultimate querying purposes, we issued a SPARQL query that would retrieve all relevant information for the seven cell lines mentioned above. We used a SPARQL *DESCRIBE* query which simply returns all relevant properties for a type into an RDF graph. Our query retrieves all *Observations* associated with the cell lines and pulls in information on the lines and experimental conditions from the *Sample* and *Dataset* tables and all genes with values from the *Reporter* table. We issued the following SPARQL query for each of the seven cell lines:

```
PREFIX obo: <http://purl/obolibrary.org/obo/>
```

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ro: <http://www.obofoundry.org/ro/ro.owl#>
```

```
DESCRIBE ?rep ?obs ?data ?samp
WHERE {
  ?samp dc:title 'YUMAC' .
  # IAO_0000136 = 'is_about'
  ?obs obo:IAO_0000136 ?samp .
  ?obs ro:part_of ?data .
  ?obs obo:IAO_0000136 ?rep .
}
```

Retrieval of a populated RDF graph containing the approximately 120,000 observation for a cell line using our Hibernate-based mapping typically took between one and two minutes.

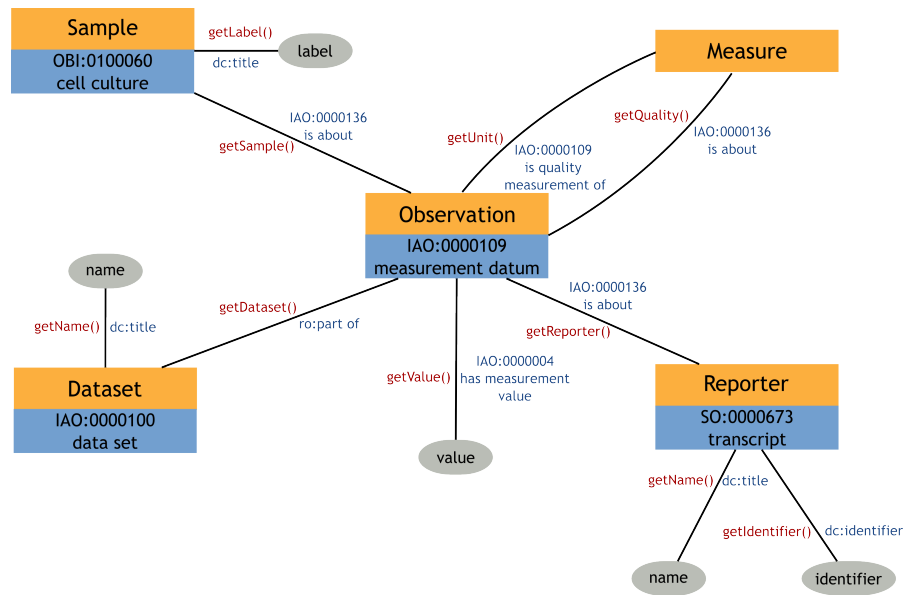


Fig. 1. Diagram showing Java classes in the Corvus model (orange boxes) next to their corresponding OWL classes (blue boxes). Data or annotation properties are shown as gray ellipses. Edge labels show the Java method used to call the Corvus model in red and the RDF property used in the semantic model in blue.

2.2 Annotated GO Terms

To include functional information about genes, we decided to incorporate the well-known Gene Ontology (GO). GO is presented in the OBO format, a simple model for expressing hierarchies of terms and the relationships between them. Although significantly less powerful for inferencing than a fully DL-compatible language like OWL, the OBO language makes it straightforward to declare relationships between classes of object. We found an effective compromise to be the use of the Simple Knowledge Organization System (SKOS) [21]. In this ontology, written in OWL, terms such as those in OBO taxonomies are expressed as instances of a *Concept* class. Class subsumption is handled through OWL object properties that describe *Concepts* as *broader* or *narrower* than other *Concepts*. In this system, properties can be assigned easily to class-like terms without violating the strictures of OWL-DL. This approach offers significant advantages for querying and reasoning, as the common alternative, creation of restrictions on classes, is computationally expensive while still requiring the creation of individual instantiations to infer properties. Using the OBO to SKOS conversion tools developed at University of Manchester ⁶, we created a GO-SKOS ontology which converts GO terms to instances of *Concept* and *is a* relationships to *broader* relationships.

We downloaded the standard human genome annotations provided by the Gene Ontology consortium. In order to easily merge with our Corvus graph, we converted the GO annotation file's HUGO symbols to RefSeq identifiers using conversion tables made available from Entrez ⁷ and used the Bio2RDF normalized URIs. In fitting with the Corvus model, we cast individual refseqs as instances of the SO:transcript class. We then used three basic relationships from RO to link the gene to its appropriate term in whichever of GO's three main hierarchies. Genes annotated with a Biological Process term were linked using *participates in*; those labeled as expressing a Molecular Function were linked using *has function* and genes marked as being located in a particular Cellular Component were linked using *part of*. We also wished for the properties assigned to genes to propagate up the chain of hierarchy. In other words, if a particular gene participates in a specific biological process, we wanted the reasoner to be able to infer that it also participates in the more generic process. For example, genes participating in apoptosis also participate in the more general process of cell death and in biological processes in general. To accomplish this, we used an OWL property chain, a new feature in OWL 2, to associate *participates in* with *broader*, stating that if A participates in B and C is a broader concept than B, then A participates in C as well. This type of inference is possible because the *is-a* (subsumption) relationship between SKOS concepts is a relationship between individuals rather than between classes. The relationship is illustrated in figure 2. We made the same declarations for the *has function* and *part of* properties.

With these declarations in place we were able to run the ontology through a DL reasoner and create a greatly expanded set of RDF triples with all infer-

⁶ <http://www.cs.man.ac.uk/~sjupp/skos>

⁷ <http://www.ncbi.nlm.nih.gov/Entrez>

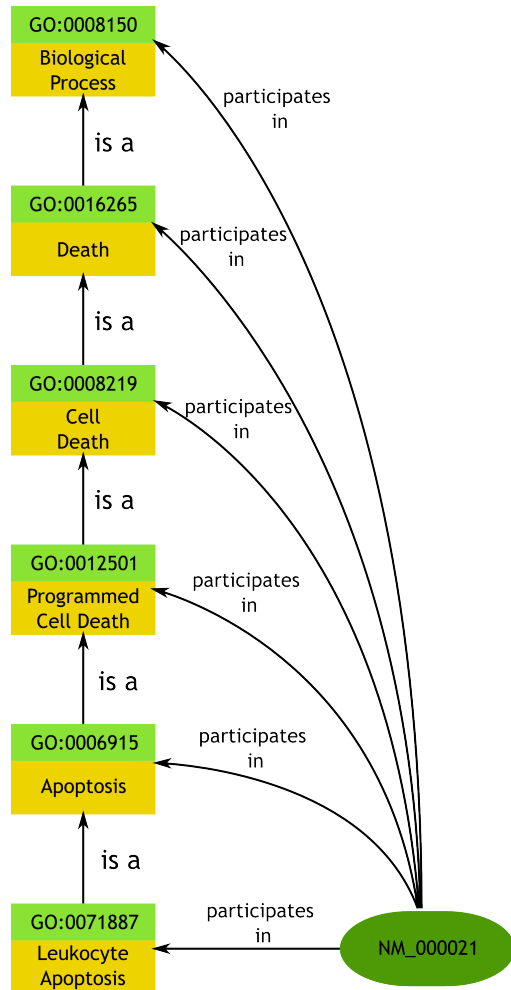


Fig. 2. Diagram showing the propagation of the participates in property up the class subsumption hierarchy. This inference is achieved by using an OWL 2 property chain associating the participates in property with the SKOS broader property.

ences spelled out (i.e. all annotation properties propagated along the hierarchy). There is a trade-off here as we gain faster query times by precomputing all inferences at the expense of additional storage space and less flexibility, as we need to recompile when the underlying data changes. Creation of the fully entailed GO annotation RDF graph took approximately five minutes on our Linux workstation using 8 GB of memory.

Merging of RDF graphs The GO annotation model could at this point be merged with the Corvus quantitative data model, the points in common being the instances of *SO transcript* representing individual RefSeqs/genes. Because we use identical URIs from the Bio2RDF namespace to describe these instances, we can assure that we are referring to the same gene in the two sources. This merged model could now be queried using SPARQL. The full architecture of our setup for creating an RDF graph from Corvus and merging it with the GO graph is shown in figure 3.

The Corvus SPARQL endpoint Application Programming Interface (API) was written in Java making extensive use of the Jena API for RDF manipulation and the closely related ARQ API for SPARQL processing⁸. The GO Annotation pre-processing was handled by a Java program making use of the OWLAPI OWL2 library [3] and the Pellet DL reasoner for Semantic Web data [29]. Merging of the ontologies was also handled by Java code using first the ARQ API to issue the SPARQL query on the relational Corvus store and then OWLAPI to perform the actual merge. The merged dataset was loaded into an instance of TDB, an RDF triple store employing the Jena libraries. It was then loaded into a running instance of Joseki, a web application allowing execution of SPARQL queries over HTTP. Joseki also uses the Jena libraries extensively. An endpoint for the merged dataset is available at <http://doppio.med.yale.edu:2020/sparql>.

3 Results and Discussion

We wanted to show that it was possible to use Corvus to execute arbitrarily complex queries incorporating information across varied knowledge domains. To this end, we tried to verify cell lines that were resistant or sensitive to Decitabine, a demethylating agent used for melanoma therapy. Our formulated query asks for genes involved in apoptosis with high methylation values prior to Decitabine administration and increased gene expression following. We use values from two datasets obtained from the Corvus SPARQL endpoint, relative methylation values prior to treatment and ratio of gene expression post- to pre-treatment. Apoptosis-related genes were found using the merged triples from the GO annotations. Our SPARQL query was as follows:

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ro: <http://www.obofoundry.org/ro/ro.owl#>
```

⁸ <http://jena.sourceforge.net>

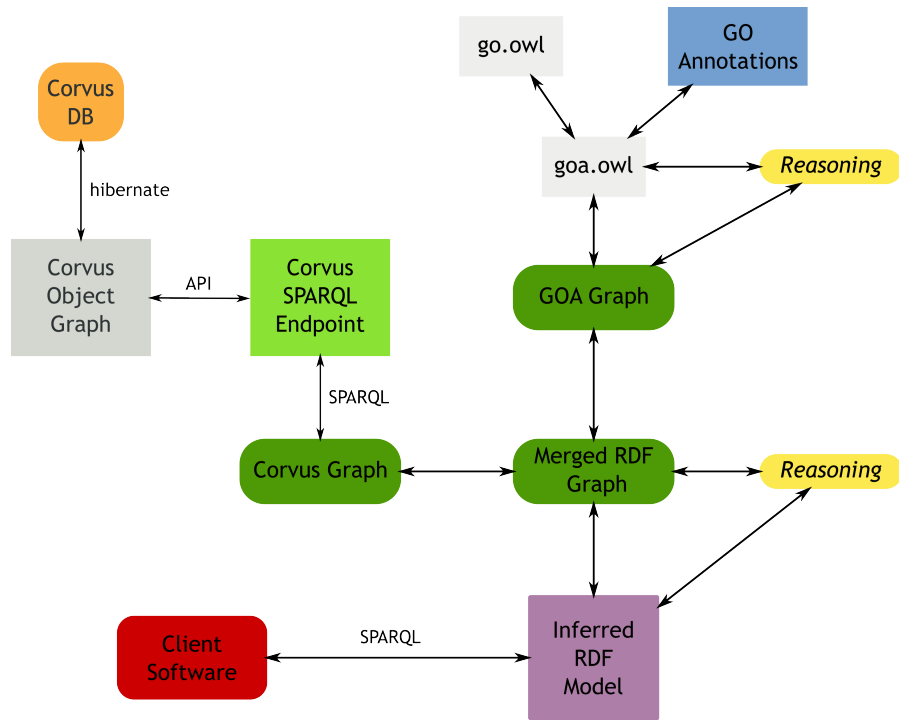


Fig. 3. Diagram showing the architecture of the integrated model we used to perform the queries in this paper.

```

PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX go: <http://purl.org/obo/owl/GO#>

SELECT distinct ?rep ?samp
WHERE {
  ?ds dc:title "Methylation Relative" .
  ?obs ro:part_of ?ds .
  # IAO_0000004 = 'has_measurement_value'
  ?obs obo:IAO_0000004 ?obsVal .
  # IAO_0000136 = 'is_about'
  ?obs obo:IAO_0000136 ?rep .
  ?obs obo:IAO_0000136 ?samp .
  # OBI_0100060 = 'cell culture'
  ?samp a obo:OBI_0100060 .
  ?ds2 dc:title "AZA Pre-Post Treatment Ratios" .
  ?obs2 ro:part_of ?ds2 .
  ?obs2 obo:IAO_0000136 ?rep .
  ?obs2 obo:IAO_0000136 ?samp .
  ?obs2 obo:IAO_0000004 ?obsVal2 .
  ?rep ro:participates_in go:0006915 .
  FILTER ( ?obsVal > 2 ) .
  FILTER ( ?obsVal2 > 1 )
}

```

This query returns the URIs of genes and cell lines that match the aforementioned criteria. Using features from the recently standardized SPARQL 1.1, we can aggregate genes by cell line to get a count of highly expressed genes per cell line. The slightly modified SPARQL query is:

```

PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ro: <http://www.obofoundry.org/ro/ro.owl#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX go: <http://purl.org/obo/owl/GO#>

SELECT (count(?rep) as ?repcount) ?samp
WHERE {
  ?ds dc:title "Methylation Relative" .
  ?obs ro:part_of ?ds .
  # IAO_0000004 = 'has_measurement_value'
  ?obs obo:IAO_0000004 ?obsVal .
  # IAO_0000136 = 'is_about'
  ?obs obo:IAO_0000136 ?rep .
  ?obs obo:IAO_0000136 ?samp .
  # OBI_0100060 = 'cell culture'
  ?samp a obo:OBI_0100060 .
  ?ds2 dc:title "AZA Pre-Post Treatment Ratios" .
  ?obs2 ro:part_of ?ds2 .
}

```

```

?obs2 obo:IA0_0000136 ?rep .
?obs2 obo:IA0_0000136 ?samp .
?obs2 obo:IA0_0000004 ?obsVal2 .
?rep ro:participates_in go:0006915 .
FILTER ( ?obsVal > 2 ) .
FILTER ( ?obsVal2 > 1 )
} GROUP BY (?samp)

```

We can compare these counts to what we know from experimental data regarding the level of sensitivity/resistance of various cell lines [13]. The results are shown in the following table: The sensitive cell lines with low IC50

Cell Line	Gene Count	IC50 (nM)
YUMAC	22	34
YUSAC	7	91
YULAC	9	110
YUSIT1	2	132
YUGEN8	6	139
WW165	2	239
YURIF	0	255

Fig. 4. Table showing the seven melanoma cell lines, the total number of apoptosis-related genes positively expressed that were formerly methylated and the IC50 value.

values (YUMAC, YUSAC and YULAC) had the three highest gene counts, whereas the two most resistant lines (WW165 and YURIF) had the lowest. As the mechanism of Decitabine action is demethylation of gene promoters, and (re)expression of the corresponding genes, these results give rise to the following hypothesis: Decitabine targets apoptosis-related gene promoters predominantly in Decitabine-sensitive cell lines, thus conveying its cytotoxic effect by activating the apoptosis pathway. The following validation steps are warranted to strengthen the hypothesis: First, one might want to independently test in vitro both the demethylation of the implicated gene promoters, as well as the re-expression of the corresponding genes. Also, the finding should be repeated in a larger cohort of melanoma samples. A current limitation of our SPARQL query is that we only interrogate for fold change after Decitabine treatment. As shown in prior work, the absolute change in expression values after treatment should also be taken into account [25].

4 Conclusion

Our proof of concept query illustrates how easily data from various sources can be integrated using the common framework of OWL/RDF. It reveals some of the power of Semantic Web reasoning and querying tools for inferring and elucidating discovered knowledge. It also shows the importance of customization in mapping

non-semantic data to RDF. While generic tools mapping relational data to RDF have recently emerged, our experience with `d2rq` has shown that there are still areas where direct mapping is significantly more efficient and flexible. Our work also makes a strong case for the benefits of using linked data, as use of the Bio2RDF normalized URI for RefSeqs made integration of the two branches of our ontology a breeze.

The flexibility of the Corvus model will allow us to incorporate quantitative Omics data from a variety of modalities. In the future, this could include cancer data from `caArray` or `caIntegrator` or data obtained directly from `ArrayExpress` using `MAGETab2RDF` [20]. Essentially, Corvus functions as a contextualized observation repository and we intend to incorporate information from other contexts including clinical data and generic provenance data. We hope to use the new semantic access point to Corvus to integrate this data with other types of information such as pathway and pharmacological data. The simplicity and elegance of the integrated Semantic Web approach also suggests its usefulness as an access point to making sense of variegated data for researchers unequipped with the programming or mathematical expertise to work with traditional data mining tools.

Acknowledgments. This work has been supported by the National Cancer Institute (Yale SPORE in skin cancer - 5P50CA121974) and the National Library of Medicine (Yale Biomedical Informatics Research Training Program - 5T15LM007056).

References

1. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al.: Gene ontology: tool for the unification of biology. *Nature genetics* 25(1), 25–29 (2000)
2. Barrasa, J., Corcho, Ó., Gómez-Pérez, A.: R2O, an extensible and semantically based database-to-ontology mapping language. In: SWDB. vol. 3372. Citeseer (2004)
3. Bechhofer, S., Volz, R., Lord, P.: Cooking the Semantic Web with the OWL API. *The SemanticWeb-ISWC 2003* pp. 659–675 (2003)
4. Belleau, F., Nolin, M., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics* 41(5), 706–716 (2008)
5. Bizer, C., Seaborne, A.: D2RQ-treating non-RDF databases as virtual RDF graphs. In: *Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*. Citeseer (2004)
6. Chen, H., Wu, Z., Zheng, G., Mao, Y.: RDF-based schema mediation for database grid. In: *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*. pp. 456–460. IEEE Computer Society (2004)
7. Corno, W., Corcoglioniti, F., Celino, I., Della Valle, E.: Exposing heterogeneous data sources as SPARQL endpoints through an object-oriented abstraction. *The Semantic Web* pp. 434–448 (2008)

8. Courtot, M., Bug, W., Gibson, F., Lister, A., Malone, J., Schober, D., Brinkman, R., Ruttenberg, A.: The owl of biomedical investigations. In: CEUR Workshop Proceedings. vol. 432. Citeseer (2009)
9. Covitz, P., Hartel, F., Schaefer, C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S., Buetow, K.: caCORE: a common infrastructure for cancer informatics. *Bioinformatics* 19(18), 2404 (2003)
10. Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R., Ashburner, M.: The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* 6(5), R44 (2005)
11. Elhai, J., Taton, A., Massar, J., Myers, J., Travers, M., Casey, J., Slupesky, M., Shrager, J.: BioBIKE: A Web-based, programmable, integrated biological knowledge base. *Nucleic Acids Research* (2009)
12. Fenstermacher, D., Street, C., McSherry, T., Nayak, V., Overby, C., Feldman, M.: The Cancer Biomedical Informatics Grid (caBIG; $\sup_i \text{TM}_i / \sup_i$). In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the. pp. 743–746. IEEE (2006)
13. Halaban, R., Krauthammer, M., Pelizzola, M., Cheng, E., Kovacs, D., Sznol, M., Ariyan, S., Narayan, D., Bacchiocchi, A., Molinaro, A., et al.: Integrative analysis of epigenetic modulation in melanoma cell response to decitabine: clinical implications. *PLoS One* 4(2), 4563 (2009)
14. Heiskanen, M., Lorenz, J., Bian, X., Madhavan, S., Gustafson, S., Muju, S., Neuberger, B., Tran, P., Settnek, S., Hartel, F., et al.: Cancer microarray informatics (caArray) data management and analysis tools at the National Cancer Institute (NCI) Center for Bioinformatics. *Proceedings of the American Association for Cancer Research* 2005(1), 7 (2005)
15. Hillairet, G., Bertrand, F., Lafaye, J.: Rewriting Queries by Means of Model Transformations from SPARQL to OQL and Vice-Versa. *Theory and Practice of Model Transformations* pp. 116–131 (2009)
16. Jabbour, E., Issa, J., Garcia-Manero, G., Kantarjian, H.: Evolution of decitabine development. *Cancer* 112(11), 2341–2351 (2008)
17. Koga, Y., Pelizzola, M., Cheng, E., Krauthammer, M., Sznol, M., Ariyan, S., Narayan, D., Molinaro, A., Halaban, R., Weissman, S.: Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome research* 19(8), 1462 (2009)
18. Massar, J., Travers, M., Elhai, J., Shrager, J.: BioLingua: a programmable knowledge environment for biologists. *Bioinformatics* 21(2), 199 (2005)
19. McCusker, J., Phillips, J., Beltrán, A., Finkelstein, A., Krauthammer, M.: Semantic web data warehousing for caGrid. *BMC bioinformatics* 10(Suppl 10), S2 (2009)
20. McCusker, J., McGuinness, D.: Provenance of High Throughput Biomedical Experiments. In: International Provenance and Annotations Workshop (2010)
21. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple knowledge organisation for the web. In: Proceedings of the International Conference on Dublin Core and Metadata Applications. vol. 5, pp. 12–15 (2005)
22. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F., Walker, M., Watson, D., Park, T., et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 351(27), 2817 (2004)
23. Pelizzola, M., Koga, Y., Urban, A., Krauthammer, M., Weissman, S., Halaban, R., Molinaro, A.: MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome research* 18(10), 1652 (2008)

24. Racunas, S., Shah, N., Albert, I., Fedoroff, N.: HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 20(Suppl 1), i257 (2004)
25. Rubinstein, J., Tran, N., Ma, S., Halaban, R., Krauthammer, M.: Genome-wide methylation and expression profiling identifies promoter characteristics affecting demethylation-induced gene up-regulation in melanoma. *BMC Medical Genomics* 3(1), 4 (2010)
26. Sahoo, S., Bodenreider, O., Rutter, J., Skinner, K., Sheth, A.: An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence. *Journal of biomedical informatics* 41(5), 752–765 (2008)
27. Sahoo, S., Halb, W., Hellmann, S., Idehen, K., Thibodeau Jr, T., Auer, S., Sequeda, J., Ezzat, A.: A survey of current approaches for mapping of relational databases to RDF. W3C RDB2RDF Incubator Group report (2009)
28. Shragar, J., Waldinger, R., Stickel, M., Massar, J.: Deductive biocomputing. *PLoS one* 2(4), 339 (2007)
29. Sirin, E., Parsia, B., Grau, B., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web* 5(2), 51–53 (2007)
30. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* 25(11), 1251–1255 (2007)
31. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., Rosse, C.: Relations in biomedical ontologies. *Genome biology* 6(5), R46 (2005)
32. Szpakowski, S., McCusker, J., Krauthammer, M.: Using Semantic Web Technologies to Annotate and Align Microarray Designs. *Cancer Informatics* 8, 65–73 (2009)
33. Van't, V., Laura, J., Hongyue, D., Van De Vijver, M., He, Y., Hart, A., et al.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871), 530–536 (2002)
34. Weibel, S.: The Dublin Core: a simple content description model for electronic resources. *Bulletin of the American Society for Information Science and Technology* 24(1), 9–11 (1997)