Towards an integrated knowledge system for capturing gene expression events

Aravind Venkatesan[†] Vladimir Mironov[†*] and Martin Kuiper
Department of Biology, NTNU, 7491 Trondheim, Norway

ABSTRACT

Transcriptional regulation of gene expression is an important mechanism in many biological processes. Aberrations in this mechanism have been implicated in cancer and other diseases. Effective investigation of gene expression mechanisms requires a system-wide integration and assessment of all available knowledge of the underlying molecular networks. This calls for a method that effectively manages and integrates the available data. We have built a semantic web based knowledge system that constitutes a significant step in this direction: the Gene Expression Knowledge Base (GeXKB). The GeXKB encompasses three application ontologies: the Gene Expression Ontology (GeXO), the Regulation of Gene Expression Ontology (ReXO), and the Regulation of Transcription Ontology (ReTO). These three ontologies, respectively, integrate gene expression information that is increasingly more specific, yet decreasing in coverage, from a variety of sources. The system is capable of answering complex biological questions with respect to gene expression and in this way facilitates the formulation or assessment of new hypothesis. Here we discuss the architecture of these ontologies and the data integration process and provide examples demonstrating the utility thereof. The knowledge base is freely available for download and can be queried through a SPARQL endpoint (http://www.semantic-systems-biology.org/apo/).

1 INTRODUCTION

Research in the Life Sciences is supported by a plethora of databases (see overview at www.pathguide.org). Moreover, the continuing advancements in functional genomics technologies make it possible to create an overwhelming amount of data in a single experiment. The many hypotheses that can be derived from such experiments must be assessed against a multitude of information and knowledge bases, often represented in a variety of formats. Scientists therefore become increasingly dependent on sophisticated computer technologies to integrate and manage all the available information. Furthermore, the drastic increase in the available information and a lack of adhering to accepted formal representations across all disparate knowledge bases allows only a fraction of the knowledge to be easily considered in the analysis of new data, or causes a user to query many databases individually, sometimes even without the support of ontology terms that would warrant a common semantics of queries in different databases. As discussed by Antezana et al. (2009), application ontologies can facilitate the query process itself as the ontology ensures a uniform semantics across all data.

1.1 Need for an integrated resource that captures gene expression knowledge

Transcriptional gene expression and its regulation depend on a large variety of cellular processes that control the timing and level of transcription of an individual gene, often in a cell- or condition specific manner. Regulation of the expression of protein coding genes is extensively studied. Gene expression falls into two main phases, *i.e.* transcription and translation. During the process of transcription, proteins called transcription factors bind to specific DNA sequence motifs (binding sites) of a gene, playing a key role in initiating or inhibiting the formation of an active RNA Polymerase II transcription complex. Active transcription produces pre-mRNAs which are subsequently processed (removal of introns, and polyadenylation of the transcript) upon which mature mRNAs are transported from the nucleus to the cytoplasm where the mRNA is translated into a protein. Regulatory processes of gene expression occur at different levels, enabling the cell to adapt to different conditions by controlling its structure and function. Furthermore, the process of gene expression may also be influenced at the epigenetic level, where nucleotide or protein modifications can cause heritable changes in expression of otherwise identical gene sequences. Abnormalities in the regulation of gene expression can cause diseases such as the occurrence of malignant cell proliferation.

The knowledge required to decipher the various processes involved in gene expression continues to grow. However,

[*] To whom correspondence should be addressed: mironov@nt.ntnu.no

[†] These authors contributed equally

for a systems-wide understanding of gene regulation, there is a need for efficiently capturing knowledge of this domain in its entirety and to further facilitate efficient querying of this data. For instance, the complex one-to-many relationships of a transcription factor like Myc includes thousands of target genes, representing a wide variety of functions and processes. An ontology-driven approach would best solve the issue of knowledge querying, representation and management. Previously, attempts have been made to model the gene regulation process; resulting in the Gene Regulation Ontology (GRO) (Beisswanger et al., 2008). GRO provides a conceptual model to represent common knowledge about the gene regulation domain. However, it was primarily built as a scaffold for knowledge intensive natural language processing (NLP) tasks and lacks the granularity in concepts much needed for advanced querying and hypothesis generation.

We have built a system that integrates existing ontologies relevant for the domain of gene expression to support the discovery of new scientific knowledge. We have named this knowledge system: the **Ge**ne E**x**pression **K**nowledge **B**ase (GeXKB). This system is conceived as part of the Semantic Systems Biology (SSB) (http://www.semantic-systems-biology.org) initiative and comprises at the current stage three application ontologies that capture the knowledge about gene expression, namely the **Ge**ne E**x**pression **O**ntology (GeXO), **Re**gulation of Gene E**x**pression **O**ntology (ReXO) and the **Re**gulation of **T**ranscription **O**ntology (ReTO).

2 GEXKB OBJECTIVES AND DESIGN PRINCIPLES

GeXKB is designed to provide the molecular biologist with a knowledge system that captures knowledge on a variety of aspects of the gene expression process. To this end it should be able to provide answers to questions like:

- *'Which are the proteins that act as chromatin remodeling proteins and as modulators of transcription factor activity?'*
- *'Which are the proteins that participate in two successive regulatory pathways?'*.
- *'Which are the transcription factors (Human) that are located in the cytoplasm?'*.

The following design principles were followed in the process of GeXKB development:

- 'is a' completeness
- 'all-some' semantics
- only classes used for modelling of the domain of discourse (see Table 1)
- maximal flexibility both for users and for future extensions

3 GEXKB ARCHITECTURE AND CONSTRUCTION

The core of the three ontologies is built of terms from a number of well established biomedical ontologies, first of all GO (Ashburner et al., 2000) and Molecular Interactions ontology (Kerrien et al., 2007), The core is used to integrate data from GOA (Barrell et al., 2009), IntAct database (Kerrien et al., 2007), KEGG (Kanehisa and Goto, 2000), UniProtKB (Magrane and Uniprot consortium, 2011) and NCBI Gene (Wheeler et al., 2005). In the subsequent sections we describe the architecture and the main features of the ontologies.

3.1 Data integration pipeline

The ontologies were built using an automated pipeline implemented with the use of the library ONTO-PERL (Antezana et al., 2008).

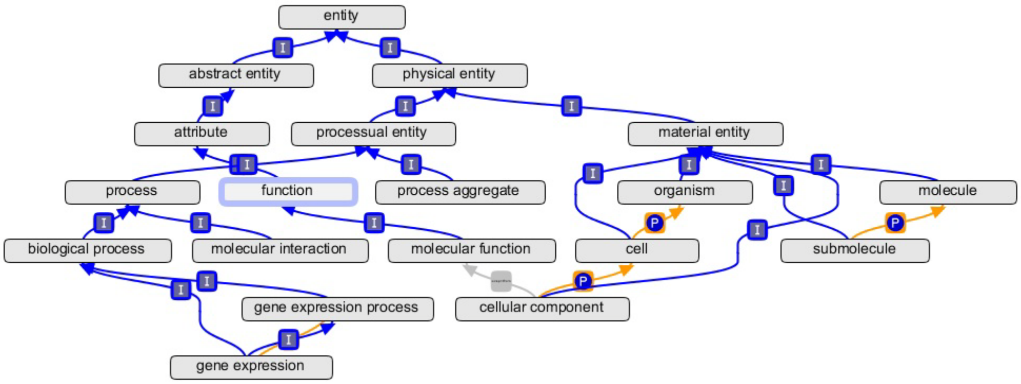

Figure 1: The figure illustrates the seed ontology of GeXO.

3.1.1 Building seed ontologies:

GeXO, ReXO and ReTO share a common Upper Level Ontology (ULO), which provides a general scaffold for data integration. It was developed on the basis of the Science Integrated Ontology (SIO) (http://code.google.com/p/semanticscience/wiki/SIO) with the addition of few terms from other ontologies. The origin of the terms is preserved in external references. The ULO is generated on the fly by the pipeline and does not exist as an individual artifact. The upper level term IDs are of the form ‘SSB:nnnnnnn’.

The ULO is then merged with GO (domain specific fragments of Biological Process, complete Cellular Component, complete Molecular Function), MI ('interaction type' branch), and the Biorel ontology (Blondé et al. 2011). This yields three ontologies referred to as seed ontologies. To be more specific, in order to build the seed ontology for GeXO, the term ‘gene expression’ (GO:0010467) and all its descendants are imported. For ReXO and ReTO the corresponding GO terms are: 'regulation of gene expression' (GO:0010468) and 'regulation of transcription, DNA dependent' (GO:0006355). We refer to these three terms as sub-roots. Each of them is connected to the ULO as a subclass of 'biological process'. To ensure 'is a' completeness, each of the ontologies is complemented with an auxiliary term - (‘gene expression process’ (GeXO:0000001), 'process of regulation of gene expression’ (ReXO:0000001), ‘process of regulation of DNA-dependent transcription’ (ReTO:0000001)), which becomes the parent of all the terms that did not have an 'is a' path to the sub-root. Apart from this, the three seed ontologies are structurally identical (Figure 1).

3.1.2 Building species specific intermediate ontologies:

The GeXKB ontologies support three model organisms: *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*.

The corresponding three species-specific intermediate ontologies were developed in the following steps:

(1) For each species GOA annotations are used to extract all the associations involving domain specific Biological Process terms incorporated in the previous phase. The corresponding proteins are added as child terms to the upper level term ‘protein’ (SSB:0001211) and referred to as 'core proteins' hereafter.

(2) From the IntAct database all the interactions involving at least one of the core proteins are retrieved and incorporated into the knowledge base along with their pertinent information. This results in a further extension of the set of proteins in the KB.

3.1.3 Building the complete ontologies:

This is the final phase in the generation of the ontologies which proceed as follows:

(1) The species specific ontologies (from the previous step) are merged together.

(2) From the KEGG database all the pathways involving at least one of the core proteins are extracted and incorporated in the KB along with the pertinent information. The pathway terms become children of the term 'SSB:0011221' ('pathway', 'BioPAX:Pathway'). The corresponding KEGG orthology groups are incorporated as children of the term 'protein cluster' (SSB:0001122). This step results in a second extension of the set of proteins.

(3) Putative orthology relationships were computed with the use of the high-performance library TurboOrtho (Ekseth et al., 2010), a multi-threaded C++ implementation of the OrthoMCL algorithm (Li et al., 2003). The relations including core proteins are added to the KB, leading to the final extension of the set of proteins.

(4) The set of proteins in the GeXKB was finally augmented with:

- GOA annotations for Cellular Components and Molecular Functions,
- Additional information (e.g. protein modifications) from UniProtKB,
- The corresponding genes along with the pertinent information from NCBI.

The final result is the three ontologies in the OBO (Smith et al., 2007) format.

3.1.4 Enhancing the utility of the ontologies:

(1) Transitive closures were constructed with the use of the library ONTO-PERL for the following relation types: 'is a', 'part of', ‘regulates’.

(2) The ontologies were exported in a number of formats: RDF, OWL, XML, and DOT.

(3) The RDF exports were used to populate a triple store, refer Table 2 (Virtuoso Open Link).

Ontology	No. of classes	No. of relations	No. of instances
GeXO	168417	15	0
ReXO	152962	15	0
ReTO	141095	15	0

Table 1: An overview of the ontologies in GeXKB

3.2 GeXKB and the Semantic Web

The Semantic Web (Berners-Lee and Hendler, 2001) is an extension of the WWW which aims at building a web of data accessible both by computers and human beings. This new technology is increasingly gaining momentum, in particular in the domain of Life Sciences (Antezana et al., 2009).

In order to make use of these new technologies, the RDF versions of the ontologies have been loaded into Open Link Virtuoso (http://virtuoso.openlinksw.com) and can be accessed via a SPARQL query page (http://www.semantic-systems-biology.org/apo/queryingcco/sparql). In contrast to other Semantic Web formalisms, such as OWL, RDF enables handling of large amounts of knowledge due to its simple and flexible syntax, making querying tractable. However, on the downside the low expressivity of RDF/RDFS imposes limitations on the inferencing over the knowledge base. To overcome this limitation, Blondé et al. (2011) have developed a novel approach for semi-automated reasoning on RDF stores with the use of the SPARUL update language (http://www.w3.org/TR/sparql11-update/). This allows for pre-computing the inferences supported by the store, thus making implicit knowledge explicit and available for querying. In order to provide maximum flexibility for querying, two graphs are available for each of the ontologies - with or without closures (e.g. GeXO-tc and GeXO, 'tc' standing for 'total closure').

The most convincing evidence of the success of the Semantic Web is the quick expansion of the Linked Data cloud (Heath and Bizer, 2011). In the course of the design of GeXKB a number of decisions were made to facilitate the migration of GeXKB eventually to the Linked Data cloud. For instance, we have re-used original IDs as much as possible. If the original IDs include a name-space (e.g. GO, MI) they were adopted without any modifications, otherwise the IDs were prepended with a name-space (for example UPKB for UniProtKB or NCBIgn for NCBI Gene), separated by a colon from the original ID (the colons are replaced with underscores in the RDF renderings). The re-use of the IDs benefits as well the users due to faster query execution and the familiarity of the IDs. Furthermore, in compliance with the Linked Data recommendations we minted the URIs in our own common name-space: http://www.semantic-systems-biology.org/ and have consistently used *rdfs:label* properties to aid human readability of the results.

RDF graphs	GeXO	GeXO-tc	ReXO	ReXO-tc	ReTO	ReTO-tc
No. of triples	~3.3 million	~23 million	~3 million	~19.9 million	~2.8 million	~19.1 million

Table 2: Shows the number of triples in the individual graphs of GeXKB

4 QUERYING GEXKB

In this section we demonstrate the utility of GeXKB with the help of a few example SPARQL queries. These queries are available as a part of a list of sample queries provided on the query page (http://www.semantic-systems-biology.org/apo/queryingcco/sparql). To query GeXKB, the base URI and the prefixes are set and the SELECT block specifies the variables to be part of the solution. The RDF triple pattern queried is defined in the WHERE block. The queries are as follows:

Q1: (see Table 3)

Biological question: *Which proteins can act as chromatin remodeling proteins and as modulators of transcription factor activity?*

SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX graph1: <ReXO>
PREFIX graph2: <ReTO-tc>

SELECT distinct ?protein_id ?protein_name
WHERE {
 GRAPH graph1: {
  ? protein_id ssb:is_a ssb:SSB_0001211 .
  ?b_process ssb:is_a ssb:GO_0040029 .
  ?b_process ssb:has_participant ? protein_id .
  ? protein_id ssb:has_source taxon: .
 }
 GRAPH graph2: {
  ssb:GO_0034401 ssb:has_participant ? protein_id .
  ? protein_id rdfs:label ?protein_name .
 }
}
LIMIT 4
```

Q2:

Biological question: *Which proteins participate in both the JAK/STAT signaling pathway and Apoptosis?*
SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX pathway1: <SSB#KEGG_ko04630>
PREFIX pathway2: <SSB#KEGG_ko04210>
PREFIX graph: <GeXO>

SELECT distinct ?protein
WHERE {
 GRAPH graph: {
  ?prot_id ssb:is_a ssb:SSB_0001211 .
  ?prot_id ssb:is_member_of ?cluster .
  pathway1: ssb:has_agent ?cluster .
  ?prot_id ssb:has_source taxon: .
 }
 GRAPH graph: {
  ?prot_id ssb:is_member_of ?cluster .
  pathway2: ssb:has_agent ?cluster .
  ?prot_id rdfs:label ?protein .
 }
}
```

Q3:

Biological question: *Which are the transcription factors (Human) that are located in the cytoplasm?*
SPARQL query:

```
BASE <http://www.semantic-systems-biology.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb: <SSB#>
PREFIX taxon: <SSB#NCBItx_9606>
PREFIX location: <SSB#GO_0005737>
PREFIX graph: <ReTO-tc>

SELECT distinct ?protein ?protein_name
WHERE {
 GRAPH graph: {
  ?protein ssb:is_a ssb:SSB_0001211 .
  ?protein rdfs:label  ?protein_name .
  ssb:GO_0006355 ssb:has_participant ?protein .
  ?protein ssb:has_function ?function .
  ?function ssb:is_a  ssb:GO_0003700 .
  location: ssb:contains ?protein .
  ?protein ssb:has_source taxon: .
 }
}
```

These queries offer just a glimpse of the repertoire of biological question that can be addressed to the knowledge system. In addition, users could also query the knowledge base in combination with other complementary semantic web resources to formulate advanced queries for hypothesis generation. This could be performed through the query federation features that are included in the latest version of SPARQL (ver. 1.1) and will be explored in the future.

Protein ID	**Protein Name**
http://www.semantic-systems-biology.org/SSB#UPKB_Q9NS37	ZHANG_HUMAN
http://www.semantic-systems-biology.org/SSB#UPKB_P14373	TRI27_HUMAN
http://www.semantic-systems-biology.org/SSB#UPKB_Q62158	TRI27_MOUSE
http://www.semantic-systems-biology.org/SSB#UPKB_P17947	SPI1_HUMAN

Table 3: The table shows the results for Q1

5 CONCLUSION

The drastic increase in the amount of data generated in the field of molecular biology and biomedicine requires efficient knowledge management practices. Ontologies certainly provide a robust method to integrate data and efficiently represent specific (sub) domain knowledge. With the creation of GeXKB, we have built a knowledge system that specifically supports researchers focusing on various aspects of gene expression. The three ontologies provide the user with the flexibility of choosing an ontology depending on the breadth and specificity of information needed. Further flexibility is afforded by a range of available formats for knowledge representation (OBO, RDF, OWL), data exchange (XML), and visualisation (DOT).

The presented examples demonstrate the utility of our knowledge base with respect to answering realistic domain specific questions, and this utility is expected to grow with its further development. The primary goal will be to augment the knowledge base with additional high quality, curated sources of information with documented transcription factor function and relations between transcription factors and their target genes.

ACKNOWLEDGEMENTS

This work is funded by the Norwegian University of Science and Technology (NTNU), Norway. AV was funded by Faculty of Natural Science and Technology and VM was funded by FUGE Mid-Norway.

REFERENCES

Antezana, E., Egaña, M., De Baets, B., Kuiper, M., and Mironov, V. (2008). ONTO-PERL: an API for supporting the development and analysis of bio-ontologies. Bioinformatics. Mar 15;24(6):885-7.

Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management: the emerging role of the Semantic Web technologies. Brief Bioinform., 10(4): 392-407.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski , K. et al., (2000). Gene ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-9.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Research* 37: D396-D403.

Beisswanger, E., Lee, V., Kim, J. J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. Stud Health Technol Inform. 2008;136:9-14.

Berners-Lee, T. and Hendler, J. (2001). 'Publishing on the semantic web'. *Nature,* 410**,** 1023-4.

Blondé, W., Mironov, V., Venkatesan, A., Antezana, E., De Baets, B., and Kuiper M. (2011). Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics*, Jun 1;27(11):1562-8.

Ekseth, O., Lindi, B., Kuiper, M., and Mironov, V. TurboOrtho – a high performance alternative to OrthhoMCL. *European Conference on Computaional Biology*: *September 2010*; *Ghent.*

Heath, T., and Bizer, . (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res,* 28**:**27-30.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod N, Bader, G. D., Xenarios, I. et al. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. 9;5:44.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M. et al. (2007). IntAct - open source resource for molecular interaction data. *Nucleic Acids Res*, 35:D561-565.

Li, L., Stoeckert, C. J. Jr., Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13:2178-2189.

Magrane M. and the UniProt consortium UniProt Knowledgebase: a hub of integrated protein data Database, 2011

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25(11), 1251–1255.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D.M., DiCuccio, M. et al. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005, 33:D39-45.