Structuring mined knowledge for the support of hypothesis generation in molecular biology

Marco Roos[1], M. Scott Marshall[1], Andrew P. Gibson[2], Pieter W. Adriaans[1]

[1] Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
[2] Swammerdam Institute for Life Sciences, University of Amsterdam
Amsterdam, The Netherlands
{roos, marshall, adriaans}@science.uva.nl, a.p.gibson@uva.nl

Abstract. Hypothesis generation in the life sciences is an empirical process in which obtaining and structuring knowledge from literature plays a significant role. Text mining and Information Extraction techniques are seen as key for programmatically accessing the knowledge captured in the form of free text. We describe progress towards an application that supports the task of generating a hypothesis about biomolecular mechanisms using Semantic Web technologies and a workflow to carry out text mining in a service-oriented architecture. The output is a semantic model with putative biological relationships that have been extracted from literature, with each relationship linked to the corresponding evidence. We present preliminary data that extends a model for chromatin (de)condensation. The methodology can be used to bootstrap the process of human-guided construction of semantically rich biological models using the results of knowledge extraction processes.

Keywords: Knowledge extraction, Hypothesis support, Molecular biology, Chromatin, Web service, Workflow, Semantic Web, OWL

1 Introduction

Conceiving or improving a hypothesis about a biomolecular mechanism usually implies integration of various types of information and distillation into a comprehensible model. This includes information from literature, our own knowledge, and interpretations of experimental data. Many Web resources such as Entrez PubMed[1] provide such information. However, the difficulty of information retrieval from literature reveals the scale of today's information overload: over 17 million biomedical documents are now available from PubMed. Support for extracting information from these resources is therefore a general requirement, with many scientists finding it increasingly challenging to ensure that all potentially

[1] http://www.ncbi.nlm.nih.gov/pubmed/

relevant facts are considered whilst forming a hypothesis. Developments in the area of information extraction promise to deliver applications that will more directly support the task of hypothesis generation. The general approach requires retrieving relevant documents, recognizing named entities (e.g. proteins) and their relationships, and storing results for later inspection [6, 10].

In this study, we address the question of how the results of a knowledge extraction procedure should be stored to best support hypothesis conception for experimental biology. In particular, we focus on epigenetics and chromatin research, where typical examples are qualitative hypothetical models that attempt to explain the role of various proteins in changing the level of condensation of DNA as a means to regulate transcription (see for instance [12]). To support the linking of a knowledge extraction process to this type of modelling, we present an approach that extracts information from text and populates an OWL-based knowledge base with the extraction results.

2 Methods and tools for knowledge extraction

Knowledge extraction was performed by web services from the Adaptive Information Discovery Application (AIDA) toolbox, a set of web services and infrastructure being developed for knowledge extraction and knowledge management in a virtual laboratory for e-science[1]. It contains services for document retrieval based on Lucene[2] [7], entity and relation recognition applying conditional random fields [5], and access to Sesame [1], a RDF repository that serves as our knowledge base. Ontologies were created in Protégé and conform to the OWL1.1 specification.

The general steps of the knowledge extraction process [6, 10] were implemented as a workflow in Taverna [3]. We added steps to provide a likelihood score, cross references to biological databases, and tabular results (Fig. 1). The likelihood of finding a document with query (q) and discovery (d) was calculated by: $-\log\left(\frac{QD_{\exp}}{QD}\right), QD_{\exp} = \left(\frac{Q}{D}\right)/N$, in which Q, D, and QD are the frequencies of documents containing q, d, and q *and* d; QD_{exp} is the expected frequency of documents containing q and d assuming independence of Q and D; N is the total number of documents in MedLine. The workflow further contains a web service for adding protein name synonyms to the original query and providing UniProt identifiers for human, rat, and mouse that we also used to filter false positives. This service, kindly provided by Martijn Schuemie, wraps components from the text analysis tool Anni2.0 [4]. At each

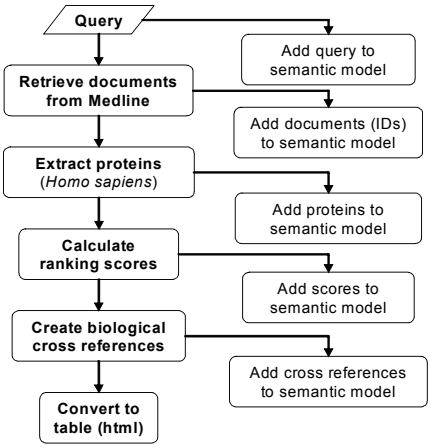

Fig. 1 - Workflow to extract proteins from literature and store them in a knowledge base.

[1] http://adaptivedisclosure.org
[2] http://lucene.apache.org/

step in the workflow, the results are converted into OWL instance statements in RDF format in order to populate the ontologies pre-loaded in our knowledge base.

References to our scientific research objects (ontologies, workflows, AIDA services) are stored as a pack on myExperiment.org that is available for download upon request (http://www.myexperiment.org/packs/27).

3 Model Representation in OWL

3.1 Different types of knowledge

In order to represent our biological hypothesis, we would like an OWL ontology of the relevant biological domain entities and their biological relationships. The purpose of our knowledge extraction procedure is to populate this model with instances. We would also like to model the evidence that has led to these instances. This leads to a clash between our intention of enriching a biological model, and representing the artifacts of a text mining procedure such as 'term', 'interaction assertion', or 'term collocation'. For these, we have concrete instance but that have no direct meaning in the biological domain. Within our OWL representation, we purposefully kept five distinct OWL models in order to avoid the conflation of knowledge from the different stages of our knowledge extraction process. Our models represent:

- Biological knowledge for our hypothesis (Protein, Association)
- Documents (Terms, PubMed Identifiers)
- Knowledge extraction process (Workflows, Processes)
- Mined results (Extracted terms, extracted relationships)
- Mapping model to integrate the above through references.

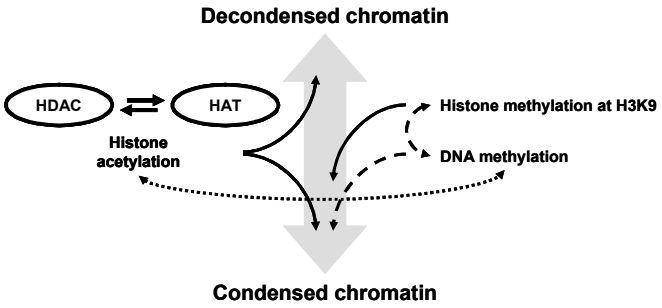

Fig. 2 – Example biological model: cartoon representation of a hypothesis for a chromatin (de)condensation mechanism. HDAC and HAT refer to enzymes with histone deacetylase activity and histone acetylase activity, respectively. For more details see figure 3 in [12] on which this figure is based.

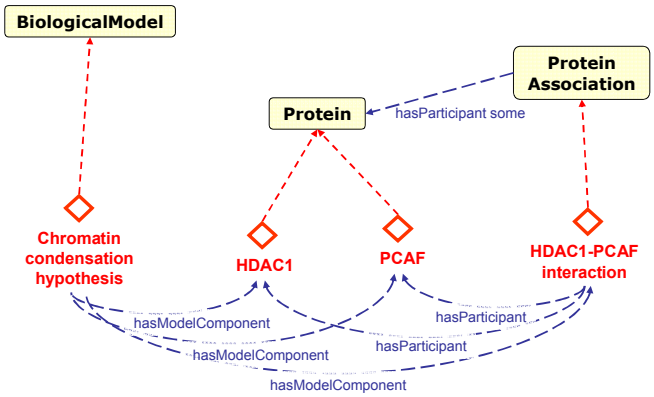

Fig. 3 - Biological domain model for hypothesis support with example instances. HDAC1[1] and PCAF[2] are examples of proteins implied in chromatin (de)condensation and known to interact. In this and following figures, diamonds represent instances, dashed arrows connected from diamonds instance-of relationships. The other dashed arrows represent properties between classes or instances. For clarity inverse relationships are not shown.

3.1.1 Biological model

In the context of our example hypothesis (Fig. 2) we start with a minimal set of classes for a biological model with proteins and protein-protein associations (Fig. 3). We cannot directly inspect concrete instances of proteins or their interactions. We regard instances in the biological model as interpretations of certain observations, in our case, of text mining results. We also do not consider such instances as biological facts; they are restricted to a hypothetical model. The evidence for the interpretation is important, but it is not within the scope of this model. In the case of text mining, evidence is modeled by the document and text mining models.

3.1.2 Document model

A model of the structure of documents and statements therein is less ambiguous than the biological model, because we can directly inspect concrete instances such as (references to) documents or pieces of text (Fig. 4). We can be sure of the scope of the model and we can be clear about the distinction between classes and instances because we computationally process the documents. For our knowledge extraction experiment, we have created classes for documents, protein or gene terms, and mentions of associations between proteins or genes. Unfortunately, we cannot make a distinction between proteins and genes at this stage due to the limits of biological text mining.

[1] http://www.uniprot.org/uniprot/Q13547
[2] http://www.uniprot.org/uniprot/Q92831

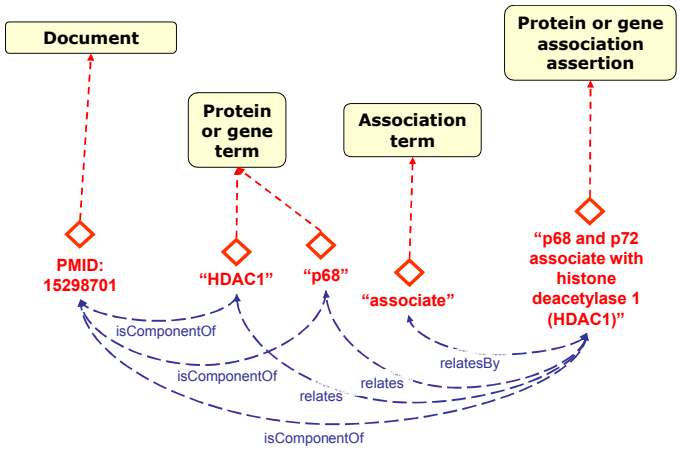

Fig. 4 – Basic ontological model that represents the relationship between documents and terms and statements used in the text.

3.1.3 Text mining model

Next, we want to structure what we know of the knowledge extraction process that may serve as evidence for the population of our biological model (**Fig. 5**). The aim of this step is to create assertions about instances of text mining processes, which

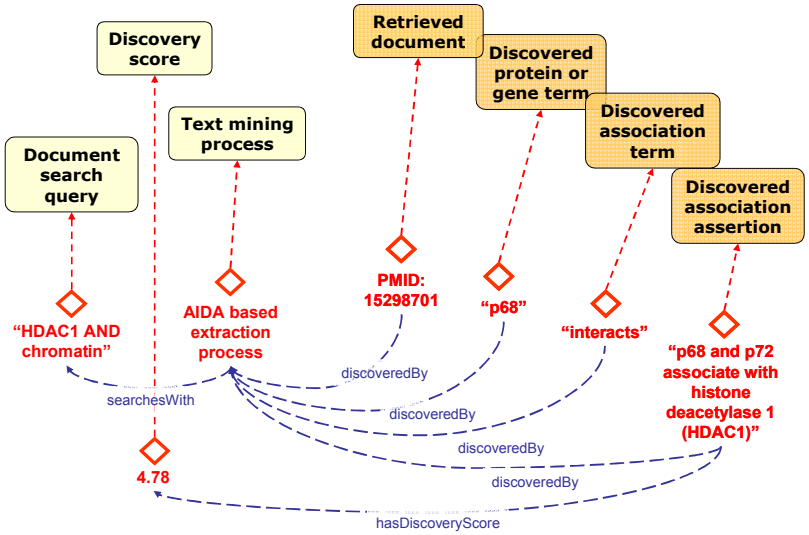

Fig. 5 – The knowledge extraction model with defined classes that classify instances from the text model as text mining discoveries. For clarity, property restrictions between classes and model components of the text mining process are not shown.

process instances of documents that contain instances of terms. In addition, in this model we represent information about the likelihood of terms and relationships being found in the literature. We also gain valuable knowledge provenance that can be used to track down any conflicting statements later on. This allows us to qualify the uncertainty of the text mining procedure. For more complete knowledge provenance, we have also created a semantic model representing the implementation of the text mining process as a workflow of (AIDA) Web Services (not shown).

3.1.4 Mapping model

At this point, we have a clear framework for the description of our biological domain and the documents and the text mining results as instances in our document and process ontologies. The next step is to relate the mined information to the biological domain model. Our strategy is to initially keep the domain model simple at the class and object property level, and to map sets of instances from our results to the domain model. For this, we created an additional mapping model that defines reference properties between the models (Fig. 6). We can now see that an interaction between the proteins labeled 'p68' and 'HDAC1' in our hypothetical model is referred to by a mention of an association between the terms 'p68' and 'HDAC1', with a likelihood score for finding this combination in literature.

The difficulty of distinguishing between genes and proteins during text mining also presents a problem for mapping to the biological model. When the number of proteins is small enough we may choose to initially map the text mining results to proteins, or we could create a perhaps more factual 'gene or protein' class in the biological model.

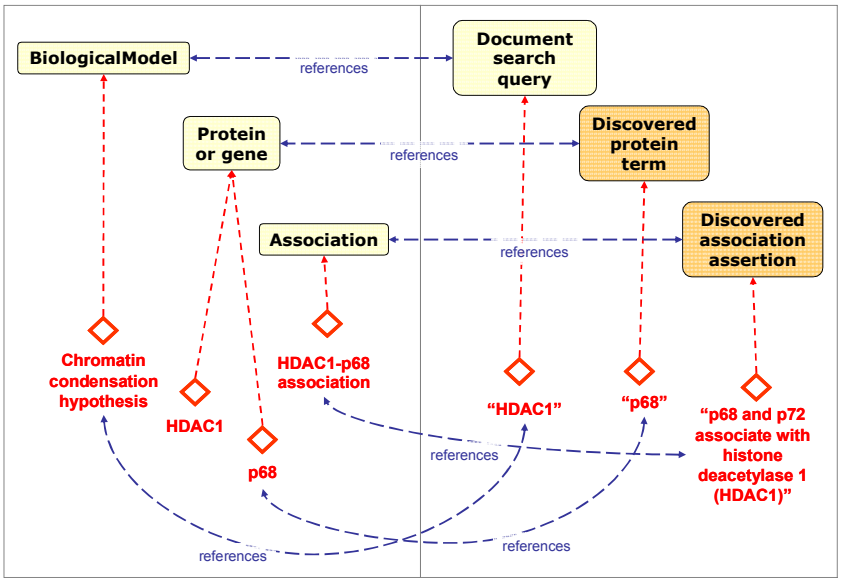

Fig. 6 – Mined knowledge mapping strategy. Instances from the results set (right) refer to instances in the domain model (left).

4 Preliminary results

The final result of the knowledge extraction workflow is a knowledge base extended with text mining results captured in OWL. We performed an example experiment starting with the query 'HDAC1 AND chromatin'. As a result we could query our knowledge base to find an instance of our biological hypothesis model and its partial representation by the input query and its expanded form (35 synonyms were added for document retrieval). We could further find 257 proteins linked to this model as putative components. We could also recover that these links were discovered through 489 protein terms found in 276 documents, and by what process, Web Service and workflow. The data is per individual: for each we stored its specific links to other individuals within a domain (e.g. the biological) and between domains. For instance, NF-KappaB is linked to our initial hypothesis and 'HDAC1' within the biological model, and to its associated term which was found in 10 abstracts. As our knowledge base grows with instances and different types of evidence we can perform increasingly interesting queries in search of novel relations with respect to our nascent hypothesis. A prototypical example is the protein referred to by the term 'p68' that was found to be collocated with the query term 'HDAC1' and also in a direct mention of this interaction in an abstract by Wilson *et al.* [13], suggesting p68 as a candidate for investigating its role in relation to HDAC1 and chromatin.

5 Conclusion

We have demonstrated first steps towards automating support for the processes involved in the formation of scientific hypotheses, particularly in studying biomolecular mechanisms. Text mining supports a researcher by inspecting more papers than an individual could and without human bias, while the use of an OWL-based knowledge base supports exploration of semantic relationships of one or many experiments. Our focus is on modeling information that is extracted during a computational experiment, rather than on improving a particular text mining procedure. The approach is not limited to the modeling of text mining results but could be applied to the results of other computational experiments. Our method shares some features with the general task of ontology learning from text [2, 9], and that of populating a predefined ontology with instances obtained from text mining [14]. However, our aim is to provide a method for improving and reusing a biological hypothesis. We do not aim to construct a comprehensive hierarchy for a domain, nor are we specifically interested in recall as long as the text mining is reasonably unbiased. Semantic Web standards and tools allow us to explicitly represent the biological knowledge, share it as a resource online, and make it interoperable with other knowledge resources. Models representing provenance add a layer of trust into the results because the biological assertions are verifiable. It will be interesting to see how much our approach can make use of the data provenance in future versions of Taverna [8]. The rich potential of Semantic Web technologies will support the future extension of the domain model to suit more complex knowledge; its exploration hopefully supported by increasingly user friendly query tools and DL-reasoners [11].

Acknowledgements

We thank Edgar Meij, Sophia Katrenko, Willem van Hage, and Martijn Schuemie for providing Web Services, and the myGrid team and OMII-UK for their support. This work was carried out for the Virtual Laboratory for e-Science project (http://www.vl-e.nl) and BioRange, supported by BSIK grants from the Dutch Ministry of Education, Culture and Science (OC&W). VL-e is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

References

1. Broekstra, J., Kampman, A. and van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. The Semantic Web - ISWC 2002: First International Semantic Web Conference, Vol. 2342/2002. Springer Berlin / Heidelberg, Sardinia, Italy (2002) 54
2. Gomez-Perez, A. and Manzano-Macho, D.: An overview of methods and tools for ontology learning from texts. Knowledge Engineering Review, 19(3):187-212, (2004)
3. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T.: Taverna: a tool for building and running workflows of services. Nucl. Acids Res., 34(Web Server issue):W729-W732, (2006)
4. Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G. and Kors, J.A.: Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome biology, 9(6):R96, (2008)
5. Katrenko, S. and Adriaans, P.W.: Using Semi-Supervised Techniques to Detect Gene Mentions. In Proc. Second BioCreative Challenge Workshop, Madrid, Spain (2007)
6. Krallinger, M. and Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome biology, 6(7):224, (2005)
7. Meij E J., IJzereef L H L., Azzopardi L A., Kamps J., de Rijke M., Voorhees E.M. and P., B.L.: Combining Thesauri-based Methods for Biomedical Retrieval. The Fourteenth Text REtrieval Conference (TREC 2005). National Institute of Standards and Technology. NIST Special Publication (2006)
8. Missier, P., Belhajjame, K., Zhao, J. and Goble, C.: Data lineage model for Taverna workflows with lightweight annotation requirements. IPAW'08, Salt Lake City, Utah (2008)
9. Missikoff, M., Velardi, P. and Fabriani, P.: Text mining techniques to automatically enrich a domain ontology. Applied Intelligence, 18(3):323-340, (2003)
10. Natarajan, J., Berrar, D., Hack, C.J. and Dubitzky, W.: Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. Critical reviews in biotechnology, 25(1-2):31-52, (2005)
11. Ruttenberg, A., Rees, J., Samwald, M. and Marshall, M.S.: Life Sciences on the Semantic Web: The Neurocommons and Beyond. Brief Bioinform, (invited paper accepted for publication in HCLS special issue), (2008)
12. Verschure, P.J.: Chromosome organization and gene control: it is difficult to see the picture when you are inside the frame. Journal of cellular biochemistry, 99(1):23-34, (2006)
13. Wilson, B.J., Bates, G.J., Nicol, S.M., Gregory, D.J., Perkins, N.D. and Fuller-Pace, F.V.: The p68 and p72 DEAD box RNA helicases interact with HDAC1 and repress transcription in a promoter-specific manner. BMC molecular biology, 5:11, (2004)
14. Witte, R., Kappler, T. and Baker, C.J.O.: Ontology Design for Biomedical Text Mining. In: Baker, C.J.O.andCheung, K.-H. (eds.): Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. Springer Science+Business Media, New York (2007) 281-313