A Knowledge Base Approach for Genomics Data Analysis

Leila Kefi-Khelif
(INRIA Sophia Antipolis, France
leila.khelif@inria.sophia.fr)

Michel Demarchez
(IMMUNOSEARCH, Grasse, France
mdemarchez@immunosearch.fr)

Martine Collard
(INRIA Sophia Antipolis, France
University of Nice-Sophia Antipolis, France
martine.collard@inria.sophia.fr)

Abstract: Recent results in genomics and proteomics and new advanced tools for gene expression data analysis with microarrays have produced so huge amounts of heterogeneous data that biologists driving comparative genomic studies face a quite complex task for integrating all relevant information. We present a new framework based on a knowledge base and semantic web techniques in order to store and semantically query a consistent repository of experimental data.
Key Words: Genomics, Knowledge base, Ontology, Semantic web technology
Category: H.2, H.4, H.3.2

1 Introduction

Recent results in genomics and proteomics and new advanced tools for gene expression data analysis with microarrays have led to discovering gene profiles for specific biological processes. Data on gene profiles are now available for the entire scientific community from public databases such as the Gene Express Omnibus[1](GEO) or the ArrayExpress[2] repository. So it becomes conceivable for a biologist to take advantage of this whole set of responses in order to compare them and characterize the underlying biological mechanisms. Nevertheless, biologists that are interested in studying these data and finding novel knowledge from them face a very complex task. Navigating into huge amounts of data stored in these public repositories is such a tedious task that they lead restricted studies and make limited conclusions. Indeed, one can observe that publications dedicated to gene expression data analysis generally focus on the hundred first differentially expressed genes among thousands of a whole genome and they deeply discuss on

[1] http://www.ncbi.nlm.nih.gov/geo/.
[2] http://www.ebi.ac.uk/microarray-as/aer/.

ten of them only. In order to highlight similar and specific biological responses to a particular biological test, it seems promising to transversally analyze the largest set of related data. A meta-analysis on multiple independent microarray data sets may provide more comprehensive view for intervalidating previous results and comparing to novel analyses. In the past few years some attempts were made on meta-analyses and focused on differentially expressed genes or on co-expressed genes. Moreau and al. [Moreau et al., 2003] discussed different issues for an efficient integration of microarray data. [Hong and Breitling, 2008] evaluated three statistical methods for integrating different microarray data sets and concluded that meta-analyses may be powerful but have to be led carefully. Indeed a critical point is to combine directly data sets derived from different experimental processes. Our approach for a better insight into huge amounts of independent data sets, is quite different. Our proposition is to build a kind of warehouse for storing expression data at a more synthetic level. We have designed a specific framework organized on two main tools: *a knowledge base* that structures and stores refined information on experiments and *an intelligent search engine* for easy navigation into this knowledge. The knowledge base is expected to include correlated information on experiments such as refined expression data, descriptive data on scientific publications and background knowledge of biologists. In this paper, we present the overall approach of the *AMI (Analysis Memory for Immunosearch)* project which aims at providing the scientist user with semi-automatic tools facilitating navigation and comparative analyses into a whole set of comparable experiments on a particular biological process. This work is done in collaboration with the Immunosearch organization[3] whose projects focus on human biological responses to chemicals. The system should allow to confront novel analyses to previous comparable results available in public repositories in order to identify reliable gene signature of biological responses to a given product. In a first stage AMI is devoted to human skin biological reactions only. Technical solutions in the AMI knowledge base and its search engine take mainly advantage of semantic web techniques such as semantic annotation languages and underlying ontologies in order to integrate heterogeneous knowledge sources, and query them in an intelligent way. The following is organised in four sections: Section 2 gives a global overview of AMI, Section 3 is devoted to the AMI knowledge base and details how its semantic annotations and ontologies are exploited, in Section 4 we demonstrate the benefit of the semantic search through examples and we conclude in Section 5.

[3] http://www.immunosearch.fr.

2 AMI Overview

A central point in our solution is to build the knowledge base on semantic annotations. Each relevant source of available information on a genomic experiment is represented as a set of semantic annotations. A semantic search engine relying on semantic ontological links is a powerful tool which may retrieve interesting approximate answers to a query as well as inferred knowledge deduced from logical rule annotations. The AMI knowledge base consists on three underlying ontologies and four sets of semantic annotations. As presented in Figure 1, AMI provides the biologist with three main tools: *ANNOTATER, ADVANCED MINER* and *SEMANTIC SEARCH*. The system takes input data describing experiments either from public repositories or from new experiments driven specifically by the system user. Semantic annotations represent different kinds of information: (i) background knowledge of biologists which has to be explicitly stated through logical facts and rules, (ii) scientific publications selected by the biologist into public microarray data repositories like GEO, ArrayExpress or PUBMED[4], (iii) descriptive information on experiments (laboratory, microarray) and conditions (tissue, treatment), (iv) synthetic data obtained from numeric raw expression data by processing transformation, statistical and data mining tools. The *ANNOTATER* tool takes each kind of available information as inputs and generates semantic annotations. It produces annotations on textual sources as scientific publications by extracting them from the text. It annotates data resulting from statistical and mining operations on raw expression data provided by the *ADVANCED MINER*. Semantic annotations include expressions of the expert background knowledge that the biologist clarifies through dialogs and graphical interfaces. The *ADVANCED MINER* tool allow the users to process data transformations for further combined meta-analysis and to run statistical and data mining tasks such as differentially expressed gene analysis or co-expressed genes clustering on relevant subspaces of the data set. The *SEMANTIC SEARCH* tool is invoked to navigate into the knowledge base and retrieve experiments, conditions or genes according more or less complex criteria. This tool generates either exact answers and approximate answers extracted according similarity links in ontologies or deduced answers obtained by logic inference rules.

3 Knowledge base

Ontologies and annotations in AMI are expressed in RDFS and RDF[5] languages as recommended by the World Wide Web Consortium (W3C)[6], respectively to represent light ontologies and to describe web resources using ontology-based

[4] http://www.ncbi.nlm.nih.gov/PubMed/.
[5] http://www.w3.org/RDF/.
[6] http://www.w3.org/.

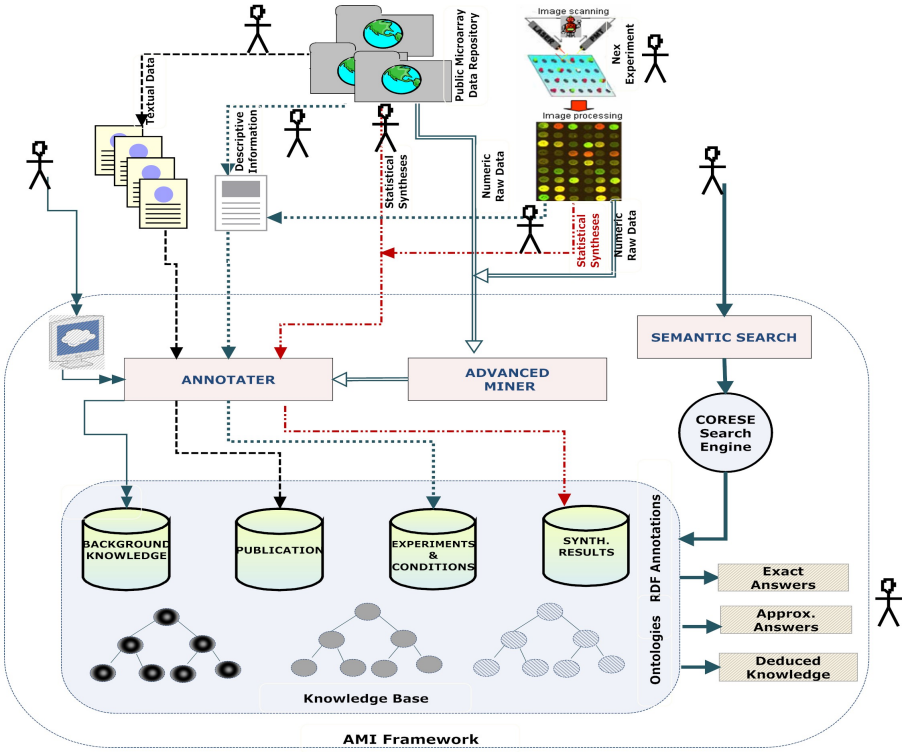

Figure 1: Global Overview of AMI framework

semantic annotations. This choice enables to use the semantic search engine CORESE [Corby et al., 2006] as explained in Section 4.

3.1 Ontologies

Ontologies provide an organizational framework of concepts and a system of hierarchical and associative relationships of the domain. In addition to the possibility of reuse and sharing allowed by ontologies, the formal structure coupled with hierarchies of concepts and relations between concepts offers the opportunity to draw complex inferences and reasoning. In AMI, we chose to reuse the existing ontology MeatOnto [Khelif et al., 2007] in order to annotate biomedical literature resources, and to develop two new ontologies: *GEOnto* for experiments and conditions, and *GMineOnto* to annotate statistical and mining results on numeric experimental data. Both ontologies will be implemented in RDF and RDFS loaded into CORESE. *MeatOnto* is based on two sub-ontologies: UMLS(Unified Medical Language System) semantic network (which integrates

the Gene Ontology[7]) enriched by more specific relations to describe the biomedical domain, and DocOnto to describe metadata about scientific publications and to link documents to UMLS concepts. *GEOnto (Gene Experiment Ontology)* is devoted to concepts related to an overall microarray expression experiment (contributors, pubmedId, keywords, general description...) and its experimental conditions (sample, treatment, subject...). While ontologies describing experiments are already available (MGED Ontology) [Stoeckert et al., 2002] and OBI (Ontology for Biomedical Investigations) [Smith et al., 2007], we choose to propose a dedicated ontology which integrates original concepts specifically relevant in our context. In fact, OBI and MGED Ontology provides models for the design of an investigation (protocols, instrumentation, material, data generated, analysis type) while GEOnto allows the description of its experimental conditions. Some of the MGED ontology concepts are included in GEOnto but they are differently structured in order to support the annotation of the experiments in our context. Some concepts in GEOnto cover general biology fields (in vivo, inductor, subject, sample, etc.) and others are specific to a particular field. In a first step, as presented in 2, we limit it to dermatology (skin, eczema, contact dermatitis, etc.) but GEOnto can be extended towards other biologic fields. To build GEOnto, we rely on (i) a corpora of experiment descriptions used to pick out candidate terms, (ii) biologists who help us to structure the concepts and validate the proposed ontology and (iii) existing ontologies (UMLS and OntoDerm[8]) to extract specific concepts (for example, UMLS to enrich the concept "cell of the epidermis" and OntoDerm to enrich the concept "disease of the skin"). *GMineOnto* provides concepts for the description of basic statistical analysis and more complex mining processes on expression data. Gene expression data are stored in two different modes: (i) refined gene expression value in a given condition, (ii) gene regulation (up, down or none) behaviour in a given condition compared to another condition. Figure 2 and Figure 3 give respectively fragments of GEOnto and GMineOnto ontologies.

3.2 Annotations

Annotations on a resource attach the most relevant descriptive information to it. In this section, we focus on the AMI approach for annotating experiments. We consider here two types of experiments: so-called "public" experiments selected by biologists from the public repositories and so-called "local" experiments led directly by the biologist. For instance, if we consider a public experiment selected from the public repository GEO, we annotate the MINiML formatted family file which is an XML document relying on the MIAME formalism [Brazma et al., 2001]. MINiML assumes only basic relations between objects:

[7] http://www.geneontology.org/.
[8] http://gulfdoctor.net/ontoderm/.

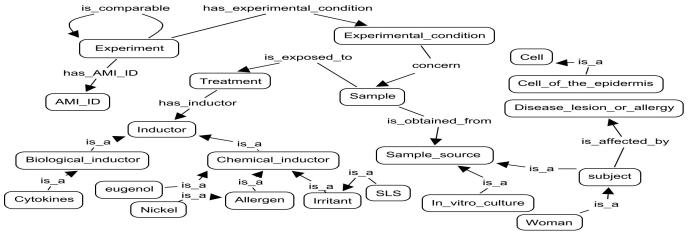

Figure 2: Fragment of GEOnto

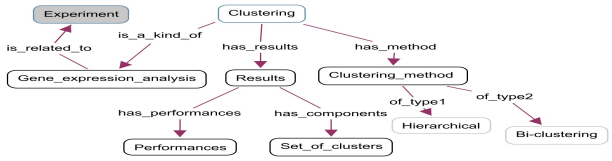

Figure 3: Fragment of GMineOnto

Platform, Sample (e.g., hybridization), and Series (experiment). The annotation process is semi-automatic. Instances of GEOnto concepts are detected in the document, some instances are directly used to generate the annotation describing the experiment (exp. contributors, pubmedID, keywords, condition titles), and others are proposed to the biologist who selects the more relevant instance for each condition (exp. time point, treatment, subject). For local experiments, the biologist has to give a structured description of the experiment and its conditions. In both cases, he uses an interactive interface. The background knowledge of biologists may be embedded into annotations on experiments too. For instance, information about *comparable* experiments can be stated by biologists solely. The RDF code below provides partly an example: The experiment annotated has the accession number GSE6281 in the GEO repository. The pubmedID 17597826 references the published article describing this experiment: "Gene expression time-course in the human skin during elicitation of allergic contact dermatitis". The experiment is declared to be comparable to experiment AMI_2008_125. It concerns a patch test with 5% nickel sulfate taken from a nickel allergic woman (age range 33-49). The patch test was exposed for 48h immediately followed by a skin biopsy.

```
<geo:Experiment rdf:about="GSE6281">
 <geo:has_AMI_ID>AMI_2008_41</go:has_AMI_ID>
 <geo:has_PMID>17597826</go:has_PMID>
 <geo:has_title>Gene Expression time_course in the human skin ...
```

```
 </geo:has_title>
 <geo:is_comparable rdf:resource="#AMI_2008_125"/> ...
 <geo:has_experimental_condition rdf:resource="#GSM144432"/>...
 <geo:Experimental_condition rdf:ID="GSM144432">
 <geo:concern><rdf:Description rdf:about="#GSM144332_BioSample">
 <geo:has_type rdf:resource="#Skin"/>
 <geo:is_analysed_at rdf:resource="#0h"/>
 <geo:is_obtained_from
                rdf:resource="#1_nickel-allergic_Woman_33-49"/>
 <geo:is_exposed_to rdf:resource="#Nickel5_48h_patch"/>
 </rdf:Description></geo:concern>...</geo:Experimental_condition>
 ...
 <geo:Treatment rdf:ID="Nickel5_48h_patch">
 <geo:has_dose>5%</geo:has_dose>
 <geo:is_exposed_for>48h</geo:is_exposed_for>
 <geo:has_delivery_method rdf:resource="#Patch_test"/>
</geo:Treatment>... </geo:Experiment>
```

4 Semantic search

AMI SEMANTIC SEARCH tool uses the semantic search engine CORESE [Corby et al., 2006] which supports navigation and reasoning on a whole base of annotations taking into account concept and relation hierarchies defined into ontologies. In addition, CORESE allows defining logic rules which extend basic annotations. The benefit for AMI is to provide search capacities on its knowledge base built from different heterogeneous sources (publications, gene expression data analyses, domain knowledge). CORESE interprets SPARQL[9] queries as sets of RDF triples with variables. Let us consider the SPARQL query presented below to retrieve all experimental conditions where the sample was exposed to a nickel patch and where the genes IL1β and TNFa are highly expressed.

```
SELECT MORE  ?c    WHERE {
?c   rdf:type     geo: Experimental_condition
?c   geo:concern   ?s
?s  is_exposed_to  ?treat
?treat  geo:has_inductor geo:Nickel
?treat  geo:has_delivery_method geo: Patch_test
?g1  rdf:type     umls:Gene_or_Genome
?g1  go:name  ?n1 filter(regex(?n1, '^ IL1 '))
?g2  rdf:type     umls:Gene_or_Genome
```

[9] http://www.w3.org/TR/rdf-sparql-query/.

```
?g2  m:name  ?n2 filter(regex(?n2, '^ TNFa '))
?g1  gmo:is_highly_expressed_in ?c
?g2  gmo:is_ highly_expressed_in ?c}
```

The *MORE* keyword in the query *SELECT* clause enables to ask for an approximate answer. An approximate search for *a sample exposed to Nickel* can retrieve *a sample exposed to eugenol* since *eugenol* is defined as a very closed concept in the GEOnto ontology. A similar approximate search to retrieve genes involved in the same cluster as IL1β and obtained by hierarchical clustering method on comparable experiments would produce results derived from bi-clustering method too since hierarchical clustering and bi-clustering are very close concepts in the GMineOnto ontology. CORESE rule language provides an inference mechanism to deduce new facts from declared annotations. Thus inferring rules on the annotation base reduces silence in the information retrieval (IR) phase. In AMI, rules are a good mean to reflect background knowledge. For instance the following rule: *If the sample studied in an experimental condition, is taken from a subject affected by psoriasis, then we can consider this condition as using the IL22 inductor* may be coded by the following lines:

```
IF         ?c  rdf:type    geo: Experimental_condition
             ?c  geo:concern  ?s
             ?s is_obtained_from ?subj
             ? subj is_affected_by geo:psoriasis
THEN   ?s geo:is_exposed_to ?t
            ?t geo:has_inductor ?geo:IL22
```

In AMI, the rule inference mechanism will provide the system with much more abilities to assist the biologist in exploring the huge information space. For instance, if the previous rule is inferred, a query asking for *all experimental conditions where this condition is using the IL22 inductor will automatically suggest extended answers with subject affected by psoriasis* avoiding a tedious manual search on well known topics closed to IL22 inductor.

5 Conclusions and Future Work

In this paper we have introduced the AMI designed to offer an easy-to-use and customized environment for assisting the biologist in comparative genomic studies. The main originality is to offer the ability to take advantage of most public available information about genomics experiments through automatic and semi-automatic tools. We have highlighted AMI originality relying on semantic web techniques such as ontologies, RDF annotations and semantic search engine. AMI is in its preliminary development phase which focused on the ANNOTATER tool. Further works will consist partly on solutions devoted to collect all heterogeneous data in order to drive real scale tests on the system.

References

[Brazma et al., 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–71.

[Corby et al., 2006] Corby, O., Dieng-Kuntz, R., Faron-Zucker, C., and Gandon, F. (2006). Searching the semantic web: Approximate query processing based on ontologies. *IEEE Intelligent Systems*, 21(1):20–27.

[Hong and Breitling, 2008] Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382.

[Khelif et al., 2007] Khelif, K., Dieng-Kuntz, R. ., and Barbry, B. (2007). An ontology-based approach to support text mining and information retrieval in the biological domain. *J. UCS*, 13(12):1881–1907.

[Moreau et al., 2003] Moreau, Y., Aerts, S., Moor1, B., Strooper, B., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.*, 19:570–7.

[Smith et al., 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The obo foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–5.

[Stoeckert et al., 2002] Stoeckert, C., Causton, H., and Ball, C. (2002). Microarray databases: standards and ontologies. *Nature Genetics*, 32:469 – 473.