Knowledge Networks of Biological and Medical Data: An Exhaustive and Flexible Solution to Model Life Science Domains

(Systems Paper)

Sascha Losko, Karsten Wenger, Wenzel Kalus, Andrea Ramge, Jens Wiehler, and Klaus Heumann

Biomax Informatics AG, Lochhamer Str. 9, 82152 Martinsried, Germany

Abstract. The huge amount of unstructured information generated by academic and industrial research groups must be easily available to facilitate scientific projects. In particular, information that is conveyed by unstructured or semi-structured text represents a vast resource for the scientific community. Systems capable of mining these textual data sets are the only option to unveil the information hidden in free text on a large scale. The BioLT Literature Mining Tool allows exhaustive extraction of information from text resources. Using advanced tagger/parser mechanisms and topic-specific dictionaries, the BioLT tool delivers structured relationships. Beyond information hidden in free text, other resources in biological and medical research are relevant, including experimental data from "-omics" platforms, phenotype information and clinical data. The BioXM Knowledge Management Environment efficiently models such complex research environments. This platform enables scientists to create knowledge networks with flexible workflows for handling experimental information and metadata, including annotation or ontologies. Information from public databases can be incorporated using the embedded BioRS Integration and Retrieval System. Users can navigate and modify the information networks. Thus, research projects can be modeled and extended dynamically.

1 Introduction

Today, the life sciences generate an ever-increasing amount of information. This is mainly driven by two factors. First, the life sciences are highly complex fields of research. There are millions of enzymes, genes, chemical compounds, diseases, species, cell types and organs that interact and are related in many different ways. Second, new experimental methods are continuously developed; as their throughput increases, the amount of raw data generated increases with overwhelming speed.

For information technologies, the challenge remains to support scientists in the identification of relevant information, the integration of this information in specific "knowledge bases" and the formalization of this knowledge across multiple scientific domains to facilitate hypothesis generation and validation (and, therefore, the generation of new knowledge). Information technology (IT) solutions are needed to

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 232 – 239, 2006.

support the knowledge generation cycle [1, 2] to ultimately gain an adequate understanding of whole biological systems. Systems Biology is a new field of research that has an intrinsic hierarchical nature, presenting a multiplicity of applicative fields that must be interconnected to give a complete description of the fundamental biological system (E-cell, virtual organs).

1.1 From Information to Knowledge

The most important source to collect a comprehensive set of relevant information available to the scientific community is the text body of published papers. Although this body of information is mostly unstructured, text-mining techniques have been developed to analyze text syntax and semantics. Text mining may be the most important answer to the mass production of scientific literature. It is, however, confronted with the same phenomena as Natural Language Processing (NLP): complexity and ambiguity. Natural language is diverse and can express the same thought in many syntactically different ways. Ambiguity arises on all levels of natural language: lexical ambiguities such as "bank", syntactical ambiguities such as "the man watches the girl with the telescope", and semantical ambiguities such as "every man loves a woman". Complexity and ambiguity often come together and make the resolution of the underlying meaning almost impossible, even for the human mind.

With respect to database integration, solutions to make all information accessible exist. A common approach is based on flat-file indexing, which emerged due to the flat-file origins of most biological databases. SRS [3] and the BioRS Integration and Retrieval System (http://www.biomax.com) are prominent examples of such technical solutions. Relational database management systems (RDBMS) are also widely used, with Oracle being one of the most popular RDBMS in the life science domain.

One issue in the integration of multiple databases is mapping the data semantics. A simple example is a case where a protein identifier is designated "`prot_id`" in one database, but is designated "`id`" in another. This problem is rather easy to solve. Both identifiers designate semantically identical entities (the protein) by semantically identical attributes. Common data access systems implement mechanisms to provide a "unified" search semantic across databases using this simple mapping technique. However, this technique is insufficient to describe, for example, the relationship between a protein and a protein complex in which the protein is likely to participate. Here, the semantic of the relationship has to be explicitly described: "Protein A *participates_in_complex* Complex B". In this way, diverse information, such as molecular processes, disease phenotypes or clinical information about patients, can be modeled as *complex semantic networks.*

The above Protein-Complex relationship example illustrates a simple approach to formalized knowledge. Though the actual definition of "knowledge" is indistinct, knowledge can be seen as the awareness of a validated interconnection of details, which, in isolation, are of lesser value. That "Protein A *participates_in_complex* Complex B" should therefore be supplemented by evidence *why* Protein A participates in Complex B. That evidence is annotation of the relationship. If it is possible to provide evidence for a defined relationship from different, independent sources (e.g., multiple scientific experiments based on various methods), the validity of the relationship is strengthened. For both Protein A and Complex B, further