Extraction and Prediction of Biomedical Database Identifier Using Neural Networks towards Data Network Construction

Hendrik Mehlhorn
Institute of Plant Genetics and Crop Plant Research, Germany

Uwe Scholz
Institute of Plant Genetics and Crop Plant Reseearch, Germany

Matthias Lange
Institute of Plant Genetics and Crop Plant Research, Germany

Falk Schreiber
Institue of Plant Genetics and Crop Plant Research, Germany & Martin Luther University, Germany

EXECUTIVE SUMMARY

Knowledge found in biomedical databases is a major bioinformatics resource. In general, this biological knowledge is represented worldwide in a network of thousands of databases, which overlap in content, but differ substantially with respect to content detail, interface, formats, and data structure. To support a functional annotation of lab data, such as protein sequences, metabolites, or DNA sequences, as well as a semi-automated data exploration in information retrieval environments, an integrated view to databases is essential. A prerequisite of supporting the concept of an integrated data view is to acquire insights into cross-references among database entities.

In this work, we investigate to what extent an automated construction of an integrated data network is possible. We propose a method that predicts and extracts cross-references from multiple life science databases and possible referenced data targets. We study the retrieval quality of our method and report on first, promising results.

1. INTRODUCTION

Bioinformatics is the field of science in which biology, computer science, and in particular information retrieval merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights. The first step in this direction is already done. High throughput biotechnologies, like next generation sequencing, proteomics and metabolomics techniques produce a massive amount of data (Galperin & Fernandez-Suarez, 2012). But the data gathered in biology or medicine is as manifold as the biological research areas itself. If we will narrow down in this chapter the complex areas of biomedical research to molecular biology, bioinformatics attempts to model and interprets this data pathway: *genome*, *gene sequence*, *protein sequence*, *protein structure*, *protein function*, *cellular pathways & networks*, and *biomedical literature*. The first consequence of this revolution is the explosion of available data that biomolecular researchers have to harness and exploit (Roos, 2001) (e.g., as of March 2012, Genbank provides access to 150,000,000 DNA sequences[1] and in PubMed there are 2,400,000 research articles listed). The number of public available databases passed currently the number of high water mark of 1,200 (Galperin & Fernandez-Suarez, 2012).

The big players in this context are on the one hand companies like pharmaceutical or plant breeders on the other hand public or private financed research institute. Their role is either a data consumer or a data producer. In consequence there is a raising need for find, extract, merge, and synthesize information from multiple, disparate sources. Convergence of biology, computer science, and information technology will accelerate this multidisciplinary endeavor. The basic needs are formulated in Lacroix & Critchlow, 2003:

1. On demand access and retrieval of the most up-to-date biological data and the ability to perform complex queries across multiple heterogeneous databases to find the most relevant information.
2. Access to the best-of-breaded analytical tools and algorithms for extraction of useful information from the massive volume and diversity of biological data.
3. A robust information integration infrastructure that connects various computational steps involving database queries, computational algorithms, and application software.