CONAN: An Integrative System for Biomedical Literature Mining

Rainer Malik and Arno Siebes

Universiteit Utrecht, Institute for Information and Computing Sciences,
PO Box 80.089, 3508TB Utrecht, The Netherlands
rainer@cs.uu.nl

Abstract. The amount of information about the genome, transcriptome and proteome, forms a problem for the scientific community: how to find the right information in a reasonable amount of time. Most research aiming to solve this problem, however, concentrate on a certain organism or a very limited dataset. Complementary to those algorithms, we developed CONAN, a system which provides a full-scale approach, tailored to experimentalists, designed to combine several information extraction methods and connect the outcome of these methods to gather novel information. Its methods include tagging of gene/protein names, finding interaction and mutation data, tagging of biological concepts, linking to MeSH and Gene Ontology terms, which can all be found back by querying the system. We present a full-scale approach that will ultimately cover all of PubMed/MEDLINE. We show that this universality has no effect on quality: our system performs as well as existing systems.

1 Introduction

It is an often quoted fact that the number of articles in MEDLINE and PubMed is growing exponentially [1]. The problem for the scientist is that interesting and useful information, like interaction data and mutation data, could appear in papers they have not read. Therefore, important facts might get overlooked and the scientific work might be affected. To overcome these problems, many systems have been developed that search the literature automatically for the relevant information [2]. Most systems, however, focus only on a very specific aspect of literature, on a very limited dataset or on a certain organism.

Complementary to those systems, we want to address - as completely as possible - the problem of experimentalists to find certain information "hidden" in the abstracts of biomedical literature. We present here the first release of CONAN, a system which is as complete as possible, offering a wide range of information. This information is also combined to construct new information, e.g. the output of a protein name tagging method is used as input for a method which finds Protein-Protein-Interaction Data and as input to find protein synonyms. Our system can be regarded as the "right-hand" of a scientist: given a query, it hands the researcher back a set of essential results.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 248–259, 2005.

Our goal is not to find new algorithms, but to integrate interesting and important algorithms into one system. The system can be installed locally or accessed via a web-service.

The road map of this paper is as follows: In the next section, we describe the general architecture of CONAN and its components. In Section 3, we show the performance evaluation, and we discuss the results. In Section 4, we draw the conclusion and give future directions.

2 Approach

The general architecture is shown in Figure 1. It shows that MEDLINE XML Files, containing abstracts, serve as input for several processing steps, namely BLAST-searching and the tagging of Gene and Protein names. These Gene and Protein names serve as input for the detection of Protein-Protein-Interaction Data. Mutation Data is also extracted from the abstracts. MeSH- and Gene Ontology (GO)-terms serve as additional input, the data is combined and integrated in the Data Integrator-Step (see Section 2.4), before it gets stored in an XML-File. This XML-File can be queried directly by XPath-queries or via a Web-Interface, using pre-defined queries.

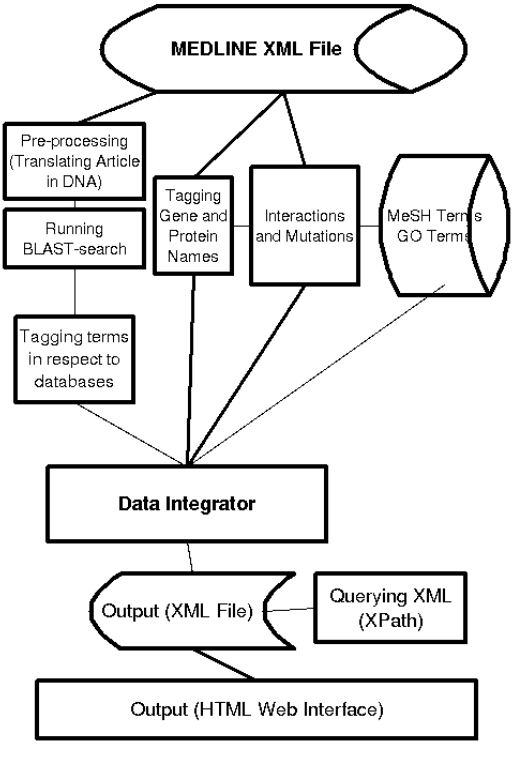

Fig. 1. Flow Diagram of CONAN