Chapter XV
Identification of Sequence Variants of Genes from Biomedical Literature:
The OSIRIS Approach

Laura I. Furlong
Research Unit on Biomedical Informatics (GRIB), IMIM-Hospital del Mar, Universitat Pompeu Fabra, Spain

Ferran Sanz
Research Unit on Biomedical Informatics (GRIB), IMIM-Hospital del Mar, Universitat Pompeu Fabra, Spain

ABSTRACT

SNPs constitute key elements in genetic epidemiology and pharmacogenomics. While data about genetic variation is found at sequence databases, functional and phenotypic information on consequences of the variations resides in literature. Literature mining is mainly hampered by the terminology problem. Thus, automatic systems for the identification of citations of allelic variants of genes in biomedical texts are required. We have reported the development of OSIRIS, aimed at retrieving literature about allelic variants of genes, a system that evolved towards a new version incorporating a new entity recognition module. The new version is based on a terminology of variations and a pattern-based search algorithm for the identification of variation terms and their disambiguation to dbSNP identifiers. OSIRISv1.2 can be used to link literature references to dbSNP database entries with high accuracy, and is suitable for collecting current knowledge on gene sequence variations for supporting the functional annotation of variation databases.

INTRODUCTION

In the last years the focus of biological research has shifted from individual genes and proteins towards the study of entire biological systems. The advent of high-throughput experimentation has led to the generation of large data sets, which is reflected in the constant growth of dedicated

repositories such as sequence databases and literature collections. Currently, MEDLINE indexes more than 17 million articles in the biomedical sciences, and it's increasing at a rate of more than 10 % each year (Ananiadou et al., 2006). In this scenario, text mining tools are becoming essential for biomedical researchers to manage the literature collection, and to extract, integrate and exploit the knowledge stored therein. Mining textual data can aid in formulating novel hypothesis by combining information from multiple articles and from biological databases, such as genome sequence databases, microarray expression studies, and protein-protein interaction databases (Jensen et al., 2006) (Ananiadou & McNaught, 2006). These kind of approaches are being applied in different scenarios: the prediction of the function of novel genes, functional annotation of molecules, discovering protein-protein interactions, interpreting microarray experiments and association of genes and phenotypes (for a review see (Ananiadou et al., 2006; Jensen et al., 2006)).

The basis of any text mining system is the proper identification of the entities mentioned in the text, also known as Named Entity Recognition (NER). Genes, proteins, drugs, diseases, tissues and biological functions are examples of entities of interest in the biomedical domain. It has been recognised that naming of these biological entities is inconsistent and imprecise, and in consequence tools that automatically extract the terms that refer to the entities are required to obtain an unambiguous identification of such entities (Park & Kim, 2006). In addition to the identification of a term that refer to, for instance, a protein in a text, it is very advantageous to map this term to its corresponding entry in biological databases. This process, also known as normalization, is very relevant from a biomedical perspective, because it provides the correct biological context to the term identified in the text.

NER has been an intense subject of research in the last years in the biology domain, specially for the identification of terms pertaining to genes and proteins (Jensen et al., 2006). Contrasting, few initiatives have been directed to the task of identification of Single Nucleotide Polymorphisms (SNPs) from the literature. Among other types of small sequence variants, SNPs represent the most frequent type of variation between individuals (0.1 % of variation in a diploid genome (Levy et al., 2007)). This observation, in addition to their widespread distribution in the genome and their low mutation rate, have positioned the SNPs as the most used genetic markers. SNPs are currently being used in candidate gene association studies, genome wide association studies and in pharmacogenomics. In this context they represent promising tools for finding the genetic determinants of complex diseases and for explaining the inter individual variability of drug responses.

From the point of view of NER, SNPs and other types of sequence variants represent a challenging task. Figure 1 illustrates the terminology problem for sequence variants of genes. The text fragments show different terms that can be used to refer to SNPs. It is important to note that the examples shown refer to a single SNP, for which different expressions are used, even in the same abstract. The same SNP can be referred to by its nucleotide representation as well as by its amino acid representation (in such cases where the SNP produces a change at the level of protein sequence). Even in each of these cases, different expressions can be used, and although there is a nomenclature standard for sequence variations (den Dunnen & Antonarakis, 2001) it is not widely used by the authors. The first approach related with NER for sequence variants was MuteXt, which was focused on collecting single point mutations for two protein families: nuclear hormone receptors and G-protein coupled receptors (Horn et al., 2004). A related approach has been implemented in MEMA (Rebholz-Schuhmann et al., 2004), in which regular expressions were used to extract variation-gene pairs from MEDLINE abstracts. The Vtag tagger (McDonald et al., 2004), based on Conditional Random Fields, was developed