An Extensible Light-Weight XML-Based Monitoring System for Sequence Databases

Dieter Van de Craen*, Frank Neven, and Kerstin Koch

Hasselt University and Transnational University of Limburg
School for Information Technology
firstname.lastname@uhasselt.be

Abstract. Life science researchers want biological information in their interest to become available to them as soon as possible. A monitoring system is a solution that relieves biologists from periodic exploration of databases. In particular, it allows them to express their interest in certain data by means of queries/constraints; they are then notified when new data arrives satisfying these queries/constraints. We describe a sequence monitoring system XSeqM where users can combine metadata queries on sequence records with constraints on an alignment against a given source sequence. The system is an XML-based solution where constraints are specified through search fields in a user-friendly web interface and which are then translated to corresponding XPath-expressions. The system is easily extensible as addition of new databases to the system then only amounts to the specification of new mappings from search fields to XPath-expressions. To protect private source sequences obtained in labs, it is imperative that researchers do not have to upload their sequences to a general untrusted system, but that they can run XSeqM locally. To keep the system light-weight, we therefore introduce an optimization technique based on query containment to reduce the number of XPath-evaluations which constitutes the bottleneck of the system. We experimentally validate this technique and show that it can drastically improve the running time.

1 Introduction

Motivation. Due to the increase in the speed of sequencing of genes and proteins, sequence databases, such as Genbank, double in size every two years [26]. This rapid expansion of data motivates researchers to repeat search queries over time. Indeed, a BLAST-search [13] that does not produce any useful result today might do so tomorrow. In this paper, we therefore propose a user-friendly sequence monitoring system XSeqM (*eXtensible Sequence Monitor*) that relieves researchers from repeating such searches over time.

We provide two motivating examples:

1. Researchers in a lab have obtained one or a few sequences of genes or proteins for which a BLAST-search only gives similarities for small regions of

* Corresponding author.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 280–296, 2006.

the sequence. No highly similar, annotated sequences are available in any database which might give hints for the function of the gene or protein. Therefore, the researchers regularly repeat BLAST-searches against several databases to find genes or proteins with a higher similarity.
2. A researcher has obtained a gene g expressed in the central nervous system (CNS) of the rainbow trout and is interested to learn about genes similar to g which are expressed in the peripheral nervous system (PNS) in any fish organism or mammal. She therefore repeats a BLAST-search with the gene g on a weekly basis.

The two tasks described above are tedious and time consuming when executed manually: not only the BLAST-searches themselves, but also the post-processing of the BLAST-reports (if any) to sort out relevant matches from irrelevant ones. Indeed, in situation (1), a match could be irrelevant as the matched part of the sequence is too small or the likelihood of the match expressed by the E-value is too large. In situation (2), all BLAST-hits from non-fish and non-mammal species should be discarded together with those that are not mRNA and that do not refer to the PNS.

A Solution: The XSeqM-System. In the XSeqM-system users can register BLAST-requests combined with constraints on the metadata of a sequence record. All requests are checked locally by the system after retrieval of the daily updates from the respective databases and users are informed, for instance through email, when relevant results are found. Figure 2 shows part of the monitor request related to situation (2). In brief, every such request specifies the following information:

- a database of interest (e.g., Genbank, SwissProt, ...),
- a sequence of interest (e.g., the gene g),
- constraints on the metadata (e.g., classification should contain the string 'fish' and molecular type should equal 'mRNA')
- an alignment program and its parameters (e.g., BLAST with word size 11 and matrix PAM30)
- relevance constraints (e.g., size of match should be greater than 20 and E-value should be smaller than e^{-10}).

The XSeqM-system has the following characteristics:

1. XSeqM is light-weight. It can be installed locally in a lab on a computer with average system requirements. This is important, as, referring to situation (1) above, research labs can be hesitant to upload their newly found sequences in a public system as some of them might be candidates for a patent application.
2. XSeqM is user-friendly as it hides all use of XML: users interact with the system through a Web-interface where search fields can be combined using the logical operators, much like other query and monitoring systems such as SRS and PubCrawler [22].
3. XSeqM is a flexible XML-based solution to which any sequence database can be added that makes updates available and whose format can be transformed