Bio-Search Computing: Integration and global ranking of bioinformatics search results

Marco Masseroli\textsuperscript{1,2}, Giorgio Ghisalberti\textsuperscript{1}, Stefano Ceri\textsuperscript{1}

\textsuperscript{1} Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Summary

In the Life Sciences, numerous questions can be addressed only by comprehensively searching different types of data that are inherently ordered, or are associated with ranked confidence values. We previously proposed Bio-Search Computing to support the integration of the results of search engines with other data and computational resources. Here we present how well known bioinformatics resources can be described as search services in the search computing framework and integrated analyses over such services can be carried out. In particular, we described and registered in the search computing framework an initial set of bioinformatics services and created a bioinformatics search computing (Bio-SeCo) application using these services.

1 Introduction

In the life sciences, questions are often complex and simultaneously regard several different functional and structural aspects of an organism and its biomolecular entities. An example is: “Which drugs threat diseases that are likely to be associated with a given genetic mutation?” Such questions can be addressed only by exploring, comprehensively searching and globally evaluating the numerous available data and their relationships, which are of different types and often inherently ordered or associated with ranked confidence values. Access to these data is being increasingly provided by web services, which offer both generic and domain-specific \textit{search services}, i.e. bioinformatics services that provide results (often ranked) of user defined searches within data repositories. These services provide users with rapid and selective access to biomedical data from potentially huge repositories. However, individual search tools are often ineffective for use in applications in which the answer to a request involves combining results from more than one search engine.

Search Computing (http://www.search-computing.eu/) has been proposed to support the integration of search engine results from different domains with other data and computational resources. It can be used to describe well known bioinformatics and biomedical resources as search services, and carry out integrated analyses over such services in a way that takes account of the ranked results from the different services. Here we apply search computing in a bioinformatics use case by representing and integrating well known bioinformatics search services to be used to search for globally ranked answers to complex biomedical questions.

2 Results and Discussion

We described and registered an initial set of bioinformatics services in the developed search computing framework. A biomedical search computing (Bio-SeCo) application using these services was created and made available at http://www.search-computing.org/UIDemoBio/. It can answer the following multi-domain case study question: “Which genes encode proteins in

\textsuperscript{2} To whom correspondence should be addressed. E-mail: marco.masseroli@polimi.it
different organisms with high sequence similarity to a given protein and are significantly co-expressed in the same given tissue?" Such interesting and complex multi-domain query, which cannot be automatically answered by any other currently available system, can be decomposed into three single sub-queries: “Which proteins in different organisms have the highest sequence similarity to a given protein?”; “Which genes encode which proteins?”; and “Which genes are co-expressed in the same given tissue?”. Each of them can be respectively mapped to an available search service, i.e. a sequence similarity search program such as WU-BLAST, a query service in our Genomic and Proteomic Data Warehouse (GPDW), and a search engine over a repository of gene expression data such as ArrayExpress Gene Expression Atlas. In the search computing framework, each of these search services can be modelled with a service mart and one or more access patterns, which describe the service and its input (I), output (O) and ranked (R) attributes for the specific data accesses available. Examples of the access patterns and pair-wise coupling connection patterns useful for computing the answer to the considered case study question are as follows:

\[
WU-BLAST(SequenceAlignmentProgramI, SearchedDBI, QuerySequenceIDI, EmailI, FoundSequenceDBO, FoundSequenceIDO, FoundSequenceSymbolO, FoundSequenceDescriptionO, BestAlignment.ProbabilityR)
\]

\[
GPDW_Gene2Protein(ProteinDBI, ProteinIDI, GeneDBO, GeneIDO, GeneSymbolO, TaxonomyO)
\]

\[
\text{ExistsCodingGene}(WU-BLAST, GPDW_Gene2Protein):
\left[\text{\(WU-BLAST.\text{FoundSequenceDB} = GPDW_{\text{Gene2Protein}.\text{ProteinDB}}\)} \quad \text{AND} \quad \text{\(WU-BLAST.\text{FoundSequenceID} = GPDW_{\text{Gene2Protein}.\text{ProteinID}}\)}\right]\]

Through the WU-BLAST access pattern, the search computing platform calls one of the BLAST programs (e.g. SequenceAlignmentProgram = “BLASTP”) available in the WU-BLAST web service to search, in one of the protein sequence databases available to WU-BLAST (e.g. SearchedDB = “UniProtKB”), for protein sequences highly similar to the sequence of a user specified protein ID (e.g. QuerySequenceID = “uniprot:P26367”). Ranked IDs of the protein sequences found with the best alignment are obtained; then, based on the ExistsCodingGene connection pattern, they are passed as input to the GPDW_Gene2Protein query service. This query service is automatically invoked by the search computing platform, according to the defined query plan and its access pattern, in order to query the GPDW for genes encoding the proteins retrieved by the WU-BLAST web service. The symbol and organism name of the obtained genes are retrieved; then, based on the ExistsExpressedGene connection pattern and the ArrayExpress accession pattern, they and the user required expression type (e.g. Regulation = “up”) and tissue (e.g. Condition = “liver”) search constraints are sent as input to the ArrayExpress Gene Expression Atlas search engine. ArrayExpress ordered search results include those genes, among the input ones, that in the ArrayExpress Archive are reported in decreasing order of probability to be significantly over co-expressed in the liver. For the example input protein ID given above, the resulting genes are PAX6 in human, PAX6 in mouse and PAX6 in rat, respectively. They represent the ordered list of genes that encode proteins with the highest sequence similarity to the P26367 protein (human Paired box protein Pax-6) and are significantly over expressed in the liver. Hence, they constitute the global ranked answer to the considered multi-domain case study question that the search computing platform can build by integrating the partial ranked results.

The developed application demonstrates that search computing can be used to describe and compose search services, in order to automatically answer complex multi-domain biomedical questions. Such questions are numerous in life sciences; they can be addressed only by comprehensively searching different types of data that are inherently ordered, or associated with ranked confidence values. By providing direct support for ranking in data integration, search computing provides distinctive data integration features of relevance to biomedicine, where many ordered data types exist. In so doing, search computing can support curiosity driven exploratory searches of biomedical data that are difficult to be performed otherwise, thus enabling ambitious data driven biomedical knowledge discovery and verification.