UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

SOME ISSUES ON BIOLOGICAL DBS: DATA INTEGRATION AND MINING

Carlos Bilich
www.carlosbilich.com.ar

July 12th , 2005

Technical Report # DIT-06-060

Some Issues on Biological DBs: Data Integration and Mining

Carlos Gustavo Bilich

University of Trento, Faculty of Science, Department of Information and Communication Technology, Via Sommarive 14, 38050 Povo (Trento), Italy
carlos.bilich@dit.unitn.it

Abstract.

The usage of database technology has become essential in almost any research task in biology. There are plenty of examples of this usage, among one of the most important is the complete decoding of the human genome achieved in recent years, a task that could not have been completed in such a short time without the help of computers, and in particular, database technology. Its usage is nowadays so diffuse in many areas of biology, that databanks containing biological data are now commonly termed "biological databases". Among many of the issues that biological databanks face today, this article concentrates particularly on two: the multiplicity of sources and mining useful data from them. A brief description of each of the problems and possible solutions are provided.

1 Introduction

Contemporaneously with the widely spread usage of databases in many fields, biologists also started to use them to store the inputs and the results of their experiments. Later on, with the birth of bioinformatics, these databases evolved tailoring their features to match the challenges of the biological research field, and they started to be called: "biological databases" or "biological databanks". As a consequence, many biological databanks emerged during these recent years, for example [1] mentions: SWISS-PROT, EMBL, Genbank and ENZYME just to name a few; in fact there are more than 500 them, which in most of the cases presents their own particularities, i.e., interfaces, query language, data representation, etc.

Apart from data integration, there is another task that has always been an issue since the introduction of biological databases, and which up to some extend, motivated the birth of bioinformatics: The problem of how to effectively search for information in such huge repositories. Being this information biological in nature, it presents peculiarities specific of this area that cannot be addressed with solutions developed for other fields, and therefore requires specific approaches. Some of these tasks often involve some kind of iteration, something that is especially suitable for computers. Most of them involve some sort of pattern searching or matching for some predefined structure. Even more, since biological literature is increasingly growing at a fast pace, curators and researches have a hard time indexing and searching for rele-

vant information in such a large corpora, therefore not only database searching is of paramount importance but also literature classification is becoming crucial for efficient research on the field.

This article reviews five references on the field that clearly state the aforementioned issues and propose some solutions. Fig. 1 briefly describes the stated problems and list the approaches along with the reference that proposes a solution.

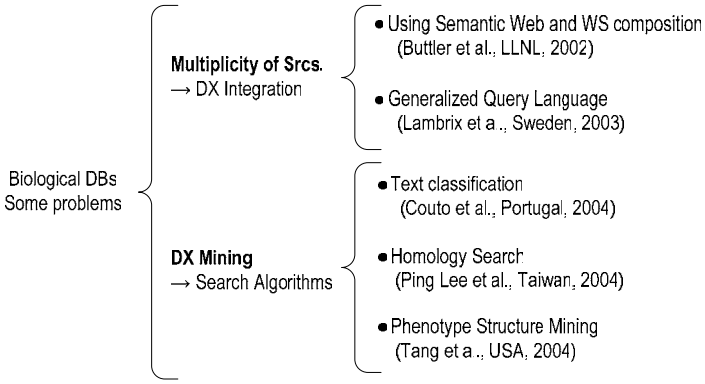

Fig. 1. Some of the problems that are still under development along with some proposed solutions and references.

The rest of the article is organized as follows. Section 2 states the problems that call for data integration and describes two approaches to the solution: one based on the utilization of semantic web concepts and the other through the design of a middleware architecture. Section 3 reviews algorithms aimed to solve three particular problems in data and text mining. Finally, section 4 contains the conclusions based on the reviewed material.

2 Data Integration

Among many of the problems that involve the integration of multiple data sources one can group them in four categories: Multiplicity; availability; accessibility and reusability.

Multiplicity is an important issue because today there are many databanks with similar but slightly different purposes in biological research. Reference [5] mentions there are more than a hundred different tools available online to process data, where the vast majority present different user interfaces which are not standardized. Moreover, in many cases those tools present different query languages each one tailored to render more efficient the specific service they intend to offer.

Availability is the second important issue and it refers to the time it takes to have the latest research results available online. It is well known that current research move at a fast pace and that there are usually several research groups that work on the same topic at the same time in different parts of the world, therefore it is desirable to have the results of the research available online as soon as a new result has been discovered. Nowadays, although the propagation of information is considerable fast, it can be further increased avoiding the delays incurred in moving data from the laboratory to large repositories, and one way to do that is by making data directly available from the laboratory instead of waiting until they are moved to the central databanks.

Accessibility is a third desirable feature in data integration and it refers to the possibility of data being accessed automatically without human intervention. As of today many of the system interfaces require at some point the intervention of the researcher to organize the input or to gather the results, and this fact is most noticeable when the output data from one tool becomes the input of another in what constitutes a workflow as it will be explained later. Because of multiplicity, many tools take part on the workflow and it is customary that at some point human intervention is needed to rearrange the data in its transition from one tool to the other. It is obvious that these interventions reduce the throughput of an experiment that can be otherwise totally automatic, with the consequent lost of productivity and performance.

Finally, reusability starts to become increasable important due to the fact that queries are getting more complex, and researchers would like to reuse the queries formulated by others maybe with only minor changes, or adding some special customization to obtain a slightly different result. This feature can also speed up high throughput experiments as well as build up research over past discoveries in a more efficient way.

One way to solve the aforementioned issues is through automation. According to the reviewed references, today's technology provides two approaches to integrate and automate data sources: one is through the construction of a middleware system which can be considered almost a classical approach, and the other one is through the utilization of the tools provided by the semantic web paradigm, something quite challenging since this technology is not yet completely developed.

2.1 The Semantic Web Contribution

To understand better what are the issues that semantic web[1] tools could solve, let's briefly enumerate the steps involved in gene research along with their problems and possible solutions [5].

First the process start with an analysis of the microarray[2] to measure and cluster the expression changes of the genes under study. Today there are many tools available to perform this task, therefore the researcher face the challenge of selecting the most suitable, running the data through it and gathering the result of the analysis. If one considers that a tool can be associated with a service, then in the semantic web para-

[1] According to the W3C: *"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* ─ Berners-Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web, Scientific American, May 2001.

[2] Piece of glass or plastic on which fragments of DNA have been affixed in a microscopy array for easy manipulation and study. Also known as DNA chip.

digm different tools can be viewed as different services where its features can constitute, for example, different attributes and therefore the automation features promised by semantic web services can help automate the selection and activation of these tools.

The second step constitutes the retrieval of the full sequence of the gene IDs previously clustered. This information is generally replicated in many sites and here the problem is that popular sites become sometimes overloaded while others are idle. The increase in the so called "cost of retrieval" affects the performance of the experiment, and here the solution should provide the ability to automatically discover new services, let them be mirrors or other services belonging to the same domain.

The third step involves the process of performing gene similarity matching aimed to identify homologs and promoter sequences. At this step the researcher is again in a situation where s/he has to choose among multiple services that provide different content, capabilities and load. Extracting data from these services involves dealing with multiple custom query interfaces which require some degree of human intervention. Again the semantic web paradigm arises as a possible solution by the utilization of features like multi-service integration, automated data extraction and semantic integration.

The fourth step deal with the analysis and identification of the promoter sequence in order to determine the regulatory profile. To perform this task the researchers again have many tools to choose from, and data from these tools has to be converted to a common well known format such as XML and then post-processed to select only the relevant parts required to feed the next step.

The fifth and sixth steps involve the analysis of the regulatory profile to identify the promoter sequence and the generation of the promoter model respectively. Once the model is ready it is followed by a search on multiple databases to find other possible candidate genes relevant to the study. All these last three tasks face similar problems and possible solutions such as those motioned in the third step.

The tasks described before show complex dependencies among them so that they can be conceived as a workflow with the particularity that it is a scientific workflow which is mainly discovery driven. This workflow has unique properties like flexibility, adaptability and construction that can be perfectly addressed by semantic web tools through its standard mechanisms for service description.

2.2 The Middleware Approach

Aside from the problems of multiplicity of sources, query languages and user interfaces that recurrently appeared at each of the steps previously described, there are other issues that inherently affect databanks. For example, due to the complexity and high variability on the representation of biological data, the schemas of biological databases tend to change at a rapid pace. Related with this, is the fact that data representation depends on the researcher since it is difficult that two different biologists store exactly the same information about the same data at all times in the same way. Moreover, most of the time it is common to find different names for the same entity in different databases. This, along with the problem that there are many databanks that are poorly designed, result in the tedious task of manually merge the queries submitted to different sources.

To cope with these issues, Lambrix et al. [1] propose a generalized query language together with an architecture to support it. It follows a brief description of each of these two components of the solution.

The query language aims to provide a tool to query multiple sources and it was inspired through interviews made to biologists and through studies of the problematic and deficiencies of current systems. It is based on an object model and has a SQL-like syntax, therefore it uses types and return complex objects, managing at the same time path expressions and path variables. It has boolean operators and there is also the possibility to perform complex operations on types like string search and string alignment, very useful and common nice-to-have features in gene research as it was explained in the workflow described in section 2.1. Table 1 provides some examples taken from [1]:

Table 1. Some examples that show the properties of the language. The complete syntax and semantics is given in an appendix of [1]

Query	Result
select * **where** *Signal-transducer*	Returns all objects of type "signal transducer" (e.g. if receptors belong to this type they will be returned)
select * **where** (**fills** *Source-organism Homo-Sapiens* **and some** *Related-clone* (**select** * **where** (**fills** *Vector-type Phage*)))	Finds all objects that have a source organism classified as "Homo-Sapiens" and have a possible related clone for which the vector type is a phage
select * **where function string-search** 'E. Coli'	Finds all documents containing the string 'E.Coli'
select * **where function align** 'MDHQDPYSVQ ...' **wrt** BLAST	Finds alignments for the sequence 'MDHQDPYSVQ...' using a well known alignment program called BLAST

The authors also mention that the language has enough room to be extended in several ways, for example, by adding constructions like "using-db" or "using-ontology" that can be used to specify preferred ontologies.

The architecture presented in [1] to support this language is reproduced in Fig. 2. There is no need to describe each component in detail because the architecture is almost self-explanatory and the description of each component clearly states its function. It can be seen how this architecture is thought as an overlay to operate above the databanks acting as an integrator of the different sources, that is the reason why here it is presented as a "middleware" [3] although it is not explicitly named in this way in [1].

[3] According to The American Heritage Dictionary of the English Language, middleware is a *"Software that serves as an intermediary between systems software and an application"*; and the Computer Desktop Encyclopedia defines it as *"Software that functions as a conversion or translation layer. It is also a consolidator and integrator"*

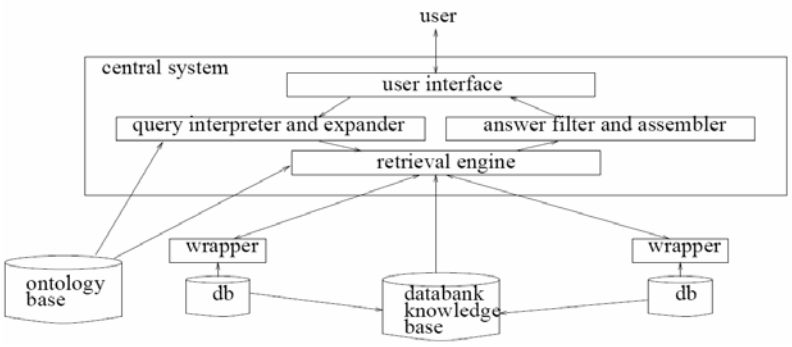

Fig. 2. Architecture reproduced from [1]. It contains a central system consisting of a user interface, a query interpreter and expander, an answer filter and assembler, and a retrieval engine. It also assumes the existence of an ontology base, a databank knowledge base as well as the use of wrappers that encapsulate the source databanks.

3 Search Algorithms for Data and Text Mining

3.1 Text Classification

The objective of this technique is that given a specific topic one wants to select a set of relevant articles form a large collection of literature, which in this particular case, is biomedical literature. Until now the classical approach is based strongly on domain knowledge that is built and maintained manually. The problem with this approach is that is not normally reusable for other domains and this render the approach expensive and limited. The reviewed solution build the knowledge by extracting information from databases commonly found on the web. This knowledge base can then be integrated with common statistical text classification methods.

The method assumes the authors of recently published biomedical articles also submit their results to public biological databases more or less at the same time of publication. The whole process can be briefly described using the example proposed in [2], where an article available in PubMed with the identifier 12803610, contains the following sentence: “The sequence of the nramp cDNA was filed at the EMBL/GenBank/DDBJ Databases under the accession number AJ514946."

The algorithm receives articles as inputs, each of them made of its content and the metadata that describes it; plus a set of biological databases. Then for each article it identifies the accession number of all databases (DB), in the example this number is AJ514946; retrieves the content and identify all the distinct terms that appears in the DB entry. Then, it computes the occurrence of each term in the article and builds a statistical representation of the article based on these occurrences. Finally, this statistical representation is ready to be used as input in classical classification methods based on statistical representation of documents.

According to the tests published in [2] the method achieved high precision, which is desirable, but low recall[4] which is not wanted. One possible solution to this problem is to increase the number of databanks in the input set, so that more descriptive terms can be found in order to build a more accurate representation, the drawback is the time it takes to search increases with the number of sources and this impacts the efficiency of the algorithm

3.2 Homology Search

For a given DNA sequence it is required to find similar sequences in the database, i.e. homologues. In this context "significant similarity" is user defined by basically two parameters: error ratio and seriate coverage. Finding similar sequences within the DNA of different species is useful because based on that it can be concluded that one protein of one species has a similar biological function on the other.

Current state of the art approaches do not consider seriation[5] and therefore produce a high number of false positives, and classical exhaustive all-against-all searches are considerable slow. In order to wade into these problems Ping Lee et al. have put together already known techniques in a smart algorithm that increases both the speed and the precision of the search. The idea described in [3] is based in the following key points: First, preprocess and index the genomic sequence so as to quickly locate substrings of certain length and decrease the cost of searching. Then, the trick is to transform the task into a variation of the well known "Longest Increasing Subsequence" (LIS) problem, by intelligently converting the user's search criteria, expressed by means of an error and a seriate coverage, into thresholds of interest. After that, use the minimum comparison strategy to generate a hit list that will be search by means of another efficient algorithm especially developed to extract the homology candidates.

All these new enhancements produce an astonishing improvement on execution efficiency, attaining high levels of precision because of seriation, achieving a filtration efficiency around 75% and a speed up of around 600 times compared with state of the art approaches and more than 10000 times when compared with classical solutions.

3.3 Phenotype Structure Mining

The last reference reviewed describes a novel algorithm to mine phenotype structures among a set of biological samples. In their paper Tang et al. [4], describe how by using heuristics it is possible to approximate a solution of a problem that is otherwise NP-hard. The phenotype displays the characteristics of a biological structure derived from the interactions between its genotype and environmental influences. Therefore different phenotypes can have the same genotype and the ability to detect a unique set of genes responsible for several different phenotype structures helps understanding the nature of diseases as well as the development of new drugs. Tang et al. define

---

[4] precision = tp/(tp+fp); recall = tp/(tp+fn); where tp: relevant retrieved articles; fp: Not relevant retrieved articles; fn: Not retrieved relevant articles.

[5] Seriation refers to the relative position of genes in the DNA sequence. For example, given the sequences: S1: TACTGTTC and S2: TATCTT; they share 4 sequences of length 2 (TA, CT, TT, TC), but only 3 of them also come one after the other in both S1 and S2 (TA, CT, TT).

novel phenotype quality metrics, called intra-consistency and inter-divergence to measure the quality of detection at each iterative adjustment, and stops the search only when this quality cannot be ameliorated during one whole iteration. Moreover, the algorithm has been conceived in a way that can mine both empirical[6] and hidden[7] phenotypes structures at the same time, something that was not previously done in other algorithms. Finally, according to the tests realized in [4], the procedure shows better efficiency and effectiveness than previous proposals.

4 Conclusions

Regarding data integration, this article tried to show that it is still an open issue but there exists several approaches to address it and two where exemplified. Building a middleware that can act as an overlay, although challenging, can be qualified as a rather conservative approach. Instead, using semantic web tools with automatic service discovery and composition offer better scalability, but to insure the success of this approach, it is necessary that the institutions that host data and applications participate in the effort by providing the information necessary to enable the appropriate semantic web tool.

Regarding data mining, the algorithms presented are not at all optimal therefore there is still room for improvement. Nevertheless, some intractability issues could prevent the development of an optimal solution, therefore the quest to find and optimal approximation continues to be a challenging problem.

References

1. Lambrix, P., Jakoniene, V.: Towards transparent access to multiple biological databanks. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003 - Volume 19. January 2003.
2. Couto, F.M., Martins, B., Silva, M.J.: Classifying biological articles using web resources. Proceedings of the 2004 ACM symposium on applied computing. March 2004.
3. Hsiao Ping Lee, Yin Te Tsai, Chuan Yi Tang: A seriate coverage filtration approach for homology search. Proceedings of the 2004 ACM symposium on Applied computing. March 2004.
4. Tang, C., Zhang, A.: Mining multiple phenotype structures underlying gene expression profiles. Proceedings of the twelfth international conference on Information and knowledge management. November 2003.
5. Buttler, D., Coleman, M., Critchlow, T., Fileto, R., Han, W., Pu, C., Rocco, D., Xiong, L.: Querying multiple bioinformatics information sources: can semantic web research help?. ACM SIGMOD Record, Volume 31 Issue 4. December 2002

[6] Empirical phenotypes are those controlled by the experiment, i.e. appositely chosen and known by the researcher.

[7] Hidden phenotypes are those that cannot be explicitly recognized beforehand and can only be discovered only when the algorithm finds a common set of genes that manifest them.