A Study of Challenges and Obstacles in Biomedical Data Search

Faris Abomelha
Institute for Research in Applicable Computing
University of Bedfordshire
Luton, UK
faris.abomelha@beds.ac.uk

Ingo Frommholz
Institute for Research in Applicable Computing
University of Bedfordshire
Luton, UK
ifrommholz@acm.org

Paul Sant
Institute for Research in Applicable Computing
University of Bedfordshire
Luton, UK
paul.sant@beds.ac.uk

ABSTRACT

The fields of Bioinformatics and Biomedicine are knowledge intensive disciplines where information changes over time due to either the availability of new data or its re-analysis and refinement (e.g., removing errors, adding new annotations),and biological data is being produced at a phenomenal rate. This study has investigated the challenges and obstacles scientists and Bioinformaticians in the Medical Biology field face when working with a variety of databases as well as having to combine heterogeneous structured and unstructured data. A qualitative research methodology (interviews, observation) has been used to examine the main challenges that face information seekers in Biomedicine. Based on the discussion of the results, an information retrieval framework is outlined with the aim of supporting information seekers in Biomedicine in their daily tasks.

CCS Concepts

•Information systems → Information retrieval; •Applied computing → Digital libraries and archives;

Keywords

Information Retrieval; Bioinformatics; Biomedical Data

1. INTRODUCTION

Biomedical data has traditionally been stored on local databases and shared as flat files with a small group of scientists. Nowadays, however, databases are available online and in vast numbers for everyone to access. Bioinformatics revolves around data; scientists and bioinformaticians are facing additional issues due to the analysis of large-scale data sets, which is increasing at an exponential rate [6]. For example, as of February 2016, the GenBank repository of nucleic acid sequences contained 190,250,235 entries compared to 171,744,486 in April 2014 which is an increase of 10.8% within two years [4]. Within the field of Bioinformatics, "Translational Bioinformatics" has been defined by the American Medical Information Association as: "The development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, preventative, and participatory health" [1]. Some of the main goals of bioinformatic text analyses are to supply effective access to unstructured information by improving searches; providing automatically generated summaries; relating publications to structured resources; visualising content for a better understanding, and guiding scientists to formulate novel hypotheses and to discover knowledge [2]. The widely held common obstacles to analysing bioinformatics data include their heterogeneous nature (e.g. unstructured text documents and structured data items together may be relevant); information overloading; personalised information selection, and incomplete information (a user has to search different data sources to obtain the complete information). [5]

One way of overcoming these obstacles is by designing and developing a framework that will retrieve related data from disparate resources and consolidate most of the databases and tools into a single featured application. Thus allowing the scientist to be able to search and retrieve information that is best related to their requirements. To this end, it is necessary to learn first about the obstacles and challenges scientists face from current systems, which is the target of this study. From the insights gained here, it has been possible to derive the suitable requirements for an IR framework that supports scientists' information needs.

The remainder of this paper is structured as follows: in the next section the methodology of the study is introduced along with some descriptive statistics, before discussing the results and findings in Section 3. In Section 4 the findings, and the requirements derived from them are discussed. Based on the findings a proposed information retrieval framework is outlined in Section 5. Finally, the conclusions are presented in Section 6.

2. METHODOLOGY

A qualitative methodology has been implemented to conduct the research and analysis. The use of interviews and observation of daily tasks was selected to collect the desired information. Convenience sampling was used in the selection of the participants, and the criteria are that they had to be either a genetic scientist or a Bioinformatician working with biomedical data and using a number of the available databases. Three laboratories and five scientists were

MedIR 2016 in Pisa, Italy

DOI: 10.475/123_4

selected from the Genetics Department at King Faisal Specialist Hospital and Research Centre (KFSHRC), Riyadh, Saudi Arabia; interviews were conducted with two scientists at the European Bioinformatics Institute (EBI), four from University College London (Centre for Cardiovascular Genetic), and four from the NIHR Biomedical Research Centre at Guy's and St. Thomas' Hospital / King's College London, and all participants were chosen at random. The total number of participants is fifteen as presented in Figure 1. Oral consent was obtained from all participants before the interviews began, and a brief description of the research was also presented. No personal or sensitive data has been collected or stored, and the use of the data collected was explained to them. The daily activities of the scientists working in the following laboratories at KFSHRC were observed: the National Laboratory for Newborn Screening (NLNBS), Genotyping Core and Saudi Diagnostics Laboratory (SDL), and interviews were conducted with five random participants. The questions that were asked during the interviews to collect the data regarding the current challenges that they are facing are as follows:

1. What are the databases that you search?
2. What analysis tools do you use?
3. What kind of equipment do you use?
4. Do you know how to program in any programming language (if so which one)?
5. What are the difficulties that you face from retrieving data?

These questions were chosen after the observation of the daily routine of the scientists at KFSHRC, which will be described in Section 4. The other participants were only interviewed, and no observations of their routine were carried out. Notes were taken to collect the participant's answers. A thematic analysis of the data has been used to identify the common points that recurred and to identify the main themes that summarise all the views.

3. RESULTS

The participants were categorised into two groups: scientists with computer programming (IT) knowledge, and without computer programming (IT) knowledge. The total number of participants without such IT knowledge was ten. The common attributes among them are that the majority, at some point, needed training on the tools before they could use them, and they required some guidance as to which tools and databases are more relevant to their work. The total number of participants with relevant IT knowledge is five. The common attributes among them are that they have all created a tool at least once during their career, and they each have their own preferred programming language for creating their tools (Perl, Python, MATLAB, R). Fig. 1 shows the distribution of the two categories at each research centre.

The main points that have arisen from the analysis of the data collected are:

1. There are too many databases to choose from (P1)
2. Duplication of data and missing data (P2)
3. Different formats of data (P3)
4. The need for training to be able to use the tools available (P4)
5. The lack of standards for naming, definition and format (P5)
6. The quality and currency of the data differs from one database to another (P6)
7. Potential of human error with the upload of data to the databases (P7)
8. Lack of documentation for locally developed tools (P8)
9. The constant improvement of current tools and databases is hard to keep up with (P9)

The participants' agreement on the points above is shown in Fig. 2. The majority of the participants have highlighted the point that there is a huge amount of data out there and some of it is duplicated. The different format of the data causes some issues for a number of them, especially for those with no relevant IT knowledge. The lack of standards creates another issue when collaborating with others; it requires a new step to convert the data according to the user's needs. If there is no standard that the research centre implements within all departments, then each department within a centre creates their own protocols with regards to which databases they use, and the platform, tools, and output format of their data. In some cases, the lead scientist or the head of each lab sets the protocol, databases, machine platform, and samples that are going to be used by all who work in the same lab. An important finding has been noted in that if the centre has a support system for the scientists and bioinformaticians, whether it is a person, group, or a full department that caters to their requirements, the scientists will know which databases they need to access, what types of formats to use, and they will have access to training and support for the tools that they use. Research centres have started requesting that the in-house support team must include at least one bioinformatician with knowledge of the research that the centre is undertaking. This position has also been created in different universities in the US, and is called the Bioinformatics Librarian. This highlights the growth in demand for the need for support in the Bioinformatics field, which is growing at an exceptional rate [7].

4. DISCUSSION

After the analysis of the data gathered from the observations and interviews, it has become clear that, in general, a genetic scientist compares the data that they produce in the lab to other research data and publications available through different databases and tools. The categorisation of the participants into two groups was carried out according to their answers; importantly, some said that if they do not receive proper training or support from a bioinformatician they will not be able to continue their work. The advancement in the field of sequencing, especially with regards to next-generation-sequencing has seen the introduction of high throughput platforms that require training and some knowledge in IT to use them. The participants with IT knowledge have the expertise to build their own tool to retrieve information from different databases; also, they have the skills

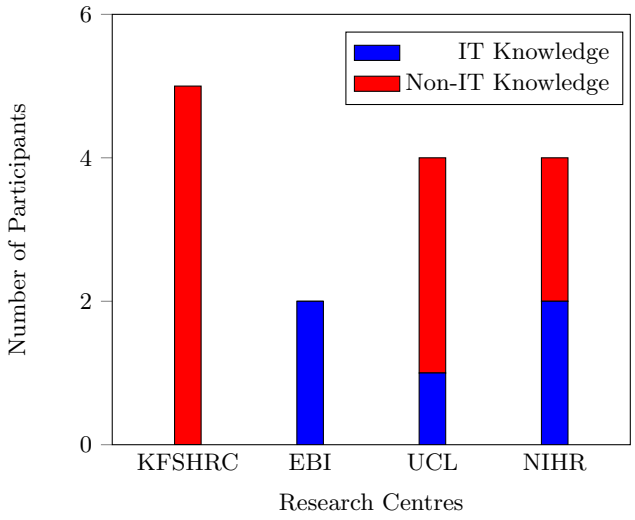

Figure 1: Number of participants

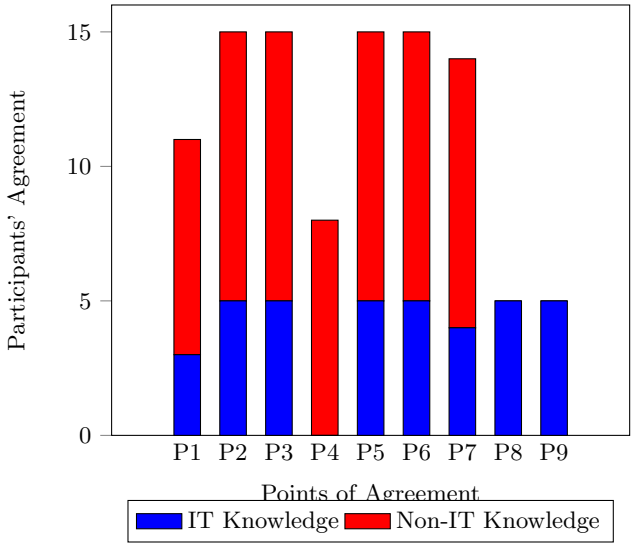

Figure 2: Participants' agreement with the points raised

to annotate information in some databases that allow user input. To understand the problems users in Biomedicine face, the basic workflow for DNA sequencing is outlined as follows [3]:

- DNA sample preparation (evaluate the quality of the sample it's integrity and purity)
- Library construction and validation. This stage has four steps: fragmentation of DNA, end repair of the fragment, ligation of adapter sequences, and optional library amplification; These steps vary depending on which platform will be used.
- Massive parallel clonal amplification of library molecules (to generate sufficient copies of the sequencing template).
- Sequencing (sequence the sample into the chosen platform).

The workflow highlights three areas where the scientists do their research; i) the lab where the samples are prepared and amplified, ii) the use of sequencing platforms to produce data and iii) the use of a computer to seek relevant information. Regarding information seeking, the needs of the scientists differ according to how they answer the three questions below:

- What kind of data is required?
- What kind of equipment do you use to produce your data?
- What level of computer programming skills do you possess?

These three questions determine the category of the databases that are relevant to the scientist, what kind of data is produced, its format, and whether they are able to use existing tools or have to create their own. Once all the questions have been answered, the scientists can narrow down the number of databases that they can actually utilise to search in. They will then be able to know if they need to convert the format of the data that is retrieved from the databases to compare it with the local data they produce, and determine if they require the help of a bioinformatician. However, there is a gap between the user needs and how the user can achieve their requirements. The majority of the tools that are being developed require training, computer knowledge, and most of their interfaces are designed with specific users in mind. The lack of documentation also makes the update or refinement of these tools an almost impossible task to complete [2]. Fig. 2 shows the agreement of the participants on the nine points derived from this study (briefly introduced in Section 3). The yearly addition of databases has produced a flood of data and databases (P1), which has led to the troublesome task of searching for the required data. Often, the same data is replicated over different databases (P2); whereas others lack crucial information which then means the user must search for it in other locations or databases. The different platforms available and the lack of standards (P3) have caused manufacturers to use different proprietary file formats thus forcing users to follow the manufacturer's format. The majority of scientists will require training to use these tools, which are produced at a fast rate, whether in-house or commercially (P4). The combination of the lack of naming standards (P5) (which creates a vocabulary mismatch) and some missing information in databases (P2) poses another challenge to the scientists –, as they have to search for the missing data elsewhere and need to figure out which other terms have been used to make sure their search has not missed any information. The degree of human error and the quality of the data (P7) is highly influenced by the scientists' own criteria. Some will have low standards concerning how they produce their data compared to others, or they may upload incomplete or wrong data due to human error. Therefore, this requires the scientist to choose the data they will compare carefully and assess it's quality (P6). This includes checking fields like author, quality score, laboratory, and the year that are associated with the data. Scientists with programming knowledge usually do not have the time to document the development of their own tools,

which makes it hard to update, edit and use them, in particular for other users who might be interested in these tools (P8). Finally the rapid update of databases (P9) has created a gap for scientists due to two points: Firstly, a major change to an underlying database requires additional training. Secondly, in-house tools built on top of existing databases will need to be altered or even scrapped to build new ones, due to the fact that the underlying databases may have a new build, interface, or changes to their access protocols.

5. TOWARDS A BIOMEDICAL IR FRAMEWORK

This study regarding the obstacles and challenges that scientists face, has motivated the development of an IR framework that a non-IT individual can use to search for biomedical data and retrieve it from disparate sources in different formats that can be manipulated either by editing them, updating, or adding information without the need to have computer programming knowledge. The proposed framework is divided into four layers: i) data retrieval (retrieving data from disparate databases), ii) data preparation (filters, ranks, and indexes data), iii) data presentation and analysis (e.g, a user interface which incorporates multiple tabs for different databases and tools), iv) local database or personal library (user data will be stored, preferences, and exclusion criteria). The main features of the proposed IR framework are:

- Retrieve biomedical data from disparate databases (heterogeneous, textual and non-textual data)
- Incorporate a focused ranking system for the data (Phred Quality Scoring[1])
- Provide the ability for the user to take full control of the records retrieved (edit, update, add)
- Create a reject list for the user to exclude specific results from appearing again
- Incorporate a single featured application that will include most of the required databases and tools as tabs
- Create a straightforward interface that is easy to navigate

The development of the proposed framework will be carried out in two phases: Phase one is to create a first prototype system that will be validated by installing it at KFSHRC, allowing feedback to be gained from the users to refine and tweak the system. Phase two comprises the creation of a fully fledged system with the help of the Scientific Computing department. The system will be installed at the Research Centre in Saudi Arabia with continuous development to accommodate the constant progress of the users' needs and their research.A rough sketch of the user interface is seen in Fig. 3. The prototype will be called Bioinformatics Information Retrieval System (BIRS). The aim is to make it as simple as possible with clear sections.Further refinement of the interface will be made once the feedback has been gathered.

[1]This is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing.

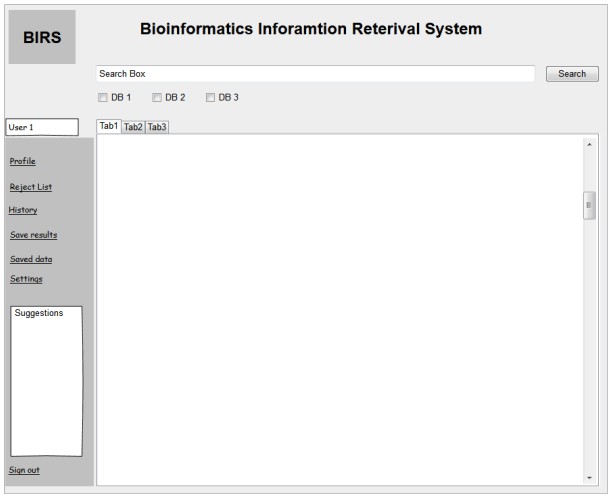

Figure 3: BIRS User Interface

6. CONCLUSION

Information seekers in Biomedicine are faced with a range of different databases and information sources, having to cope with and combine heterogeneous structured and unstructured (textual) data from different sources. This requires constant review and reassessment to be able to keep up with recent developments. In this study the main challenges faced by information seekers in Biomedicine when working with a variety of databases and tools have been identified. The different points users agree are the main obstacles in their daily work have been discussed, along with differentiating between users knowledgeable in programming (thus able to create their own tools) and those who are not. From the results and discussion, some requirements for an IR framework that supports information seekers in Biomedicine have been derived, along with discussing future developments.

7. REFERENCES

[1] AMIA. Translational Bioinformatics – The American Medical Information Association, 2014.

[2] A Manconi, E Vargiu, G Armano, and L Milanesi. Literature retrieval and mining in bioinformatics: state of the art and challenges. *Advances in Bioinformatics*, 2012:573846, 2012.

[3] Elaine R Mardis. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9:387–402, January 2008.

[4] NCBI. GenBank and WGS Statistics, 2016.

[5] L Stein. Creating a bioinformatics nation. *Nature*, 417(6885):119–120, May 2002.

[6] Ronald C Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics*, 11(Suppl 12)(S1), January 2010.

[7] Michele R. Tennant. Bioinformatics librarian. *Reference Services Review*, 33(1):12–19, March 2005.