Gene Collector V1.1: a user-friendly bilingual gene data mining software

C. Pacheco1*, S.G. Lopes2, W.P. Silva2, J.L.C. Silva2, R.P. Vasconcelos2, A.A. Araujo-Neto3, S.M.S. Felipe3, M.M.D.C. Soares1, J. O. Alves1, Soares, P.M.1, D. P. Carvalho3, and V.M. Ceccatto1

1Instituto Superior de Ciências Biomédicas, Universidade Estadual do Ceará, Fortaleza, CE, Brazil.  
2Universidade Federal Rural do Semi Árido, Angicos, RN, Brazil.  
3Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro (IBCCF/UFRJ), Rio de Janeiro, RJ, Brazil.  
*christinaosvaldo@yahoo.com.br

Abstract: The amount of available biological information in public databases has grown, and in an exponential manner. In the light of this growth, scientists’ data analysis requirements have rapidly changed. We developed a computer software capable of capturing and summarizing information about a gene set from public databases. “Gene Collector v.1.1” is a bilingual data mining tool to aid in biomedical researches and also useful for educational uses. Using NCBI and Ensembl IDs as input, it is capable of gathering information on nomenclature (official symbol, aliases and full name), species, location (band, coordinates and strand), number of alternative transcripts and gene product type. The data set can be downloaded as a spreadsheet or as a local database. Some obtained parameters are summarized in a simple report. Using the program’s output, further analyses can be performed aiming to answer specific biological questions, making useful links with other bioinformatic tools.

Keywords: genetic data collection, bilingual data mining, biological database searches.

Paper submitted to BIOCOMP’14, the 2014 International Conference on Bioinformatics & Computational Biology.

1 Introduction

Wet-lab geneticists, molecular biologists, biomedical scientists and other life science researchers are producing and dealing with ever increasing volumes of genetic data. The amount of available biological information in public databases has grown, and in an exponential manner. In the light of this growth, scientists’ data analysis requirements have rapidly changed in the last decades. In the current “omics” era, scientists that study molecular biology, disease mechanisms, medical genetics, structural biology and population genetics, among other areas, have found the need to comprehend basic bioinformatics, focusing in resources and tools directed towards their needs [1].

Biological databases, such as NCBI gene [2] and Ensembl Genome Browser [3], offer access to a huge amount of information about genes, transcripts and their products, encompassing sequence, location, structural and functional data. With the advances in science, approaches that deal with large amounts of information, such as data mining and statistical methods, are now seen as complementary procedures [4]. Data mining can be defined as a process that handles large quantities of information, selecting, exploring and modeling data in search for regularities or relations in large data sets. The process envisages providing comprehensible and useful information for the user [5].

Information processing and data mining tools have, therefore, become key components in modern biomedical research, especially in genetics and protein biology [4, 6]. A great number of bioinformatics tools have been developed in the past few years, in order to address a wide variety of biological research problems. The current scenario for the use of tools for data mining includes the search and processing of biological information. This field includes the discovery and accumulation of information, such as IDs, aliases, official name, location, and genetic description parameters in general. "Gene Collector" is being presented as a new approach for collection and utilization of basic genetic data in a local database.

Even simple molecular biology projects, nowadays, tend to study a gene set rather than a single gene. As is well known that database searches can be time-consuming, especially when dealing with a large number of genes, our multi-disciplinary group aimed to develop a user-friendly tool for data mining basic genetic information on large gene sets. Our main objective was the development of a user friendly software capable of capturing and summarizing information about a gene set from the NCBI Gene database and Ensembl Genome Browser through a bilingual interface.
2 Methods

Parsing, a syntactic analysis of HTML pages which transforms text into a binary tree, was used in order to create the data structure using the API JSoup, a Java library for the extraction and manipulation of HTML (http://jsoup.org). Afterwards, it was possible to access the tags and save the information in the program’s variables.

Using this approach, the Gene Collector software was created in Java language, as shown in the pipeline in Figure 1. Gene Collector is capable of finding, extracting and classifying genetic information of interest from NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/) [2] and Ensembl Genome Browser (http://www.ensembl.org/) [3].

It is possible to input a list of genes for the software to download data, and after data collection it brings the option of saving the spreadsheet as an “.xIms” file. Another function that Gene Collector brings is to allow that the gathered data on the gene set is saved in a local database, which, in turn, allows for searches and manual alterations. After collecting all the information, modest data mining approaches were performed, creating the “Gene Set Summary” section, bringing a summary of the collected data.

In order to evaluate the efficacy of the searches, as a quality control measure, we used a comparative method. After performing searches on a gene set both manually and using the program, we compared the table produced by Gene Collector with a data set (with 1000 genes) previously manually characterized by our group (through exhaustive DB searches), revealing the percentage of correctly and incorrectly filled cells. Data mining procedures (the gene set summary) revealing aspects of the data set such as number of genes per gene type, division of gene set by organisms, as well as chromosome distribution were also evaluated by comparative methods.

Additionally, to characterize some biological aspects of gene sets, the program output data was further evaluated with other bioinformatic online tools, using 13 correlated genes for this purpose (cholesterol transporter activity - GO:0017127). Another set of 622 genes was used for evaluation of the output data with co-localization and gene synteny analyses.

Figure 1. Program development pipeline of the “Gene Collector” software.
3 Results and discussion

In the post-genomic era, the data explosion that is allowing for huge scientific progress is also threatening to drown scientists with ever increasing amounts of information [7]. Bioinformatics and Genomic approaches are helping in the understanding of biological systems with the development of computational tools that aim to solve biomedical problems [8]. Thousands of bioinformatics tools have already been developed, for a wide variety of biomedical research ends. However, there are still some questions that can only be solved through specific tools for the biomedical researcher.

Gene Collector is a bilingual data mining computer software capable of capturing and summarizing information about a gene set from the National Center for Biotechnology Information Gene database and Ensembl Genome Browser. Using NCBI and Ensembl IDs as input, the created algorithm is capable of gathering information on nomenclature (official symbol, aliases and full name), species, location (band, coordinates and strand), number of alternative transcripts and gene product type. The resulting data set can be downloaded as a spreadsheet or as a local database. Gene Collector’s user friendly data gathering interface can be seen in Figure 2. The program is capable of showing some parameters of gene nomenclature and identification (such as official symbol, gene IDs, aliases, official full name, description, name, synonyms).

One of the most common difficulties found by molecular biology researchers is the ambiguity in gene nomenclature (aliases). Scientific literature searches require text mining methods that are able to recognize the same gene in different databases, occurring when these aliases are not associated with the same subset of documents. These problems involve a) different names associated with the same gene, and b) one name associated with multiple genes or even with non-gene meanings [9].

The interface of the Gene Collector database is shown in Figure 3. This is a reasonably simple visualization of the results obtained by the search and allows to save the effect changes and generate a useful spreadsheet. The generated local database facilitates data maintenance and review.

The evaluation of the software involved a comparison with a manually filled spreadsheet produced from data collected from 77 papers. We used a set of 1000 genes obtained from the literature which showed genes expressed in human skeletal muscle. This evaluation of the current version of the software revealed that our newly developed software was reasonably efficient in the data gathering stage. The efficiency in this stage of the process must improve in latter versions of the program. The previously characterized gene set comprising 1000 genes was investigated using the newly developed tool. A significant proportion of genes (183 genes, 18.3% of the searched gene set) were not found by the program. Of the remaining 817 genes, we found, through comparisons between the manually filled table with the generated spreadsheet, that 96.0% of the automatically collected information matched data in the manually filled spreadsheet, with some columns being filled more accurately than others, as shown in figure 4. The number of transcripts and gene type columns, for example, were filled 99.9% correctly, whereas the aliases information was only identical for 74.7% of the genes.

Figure 2. Gene Collector’s user friendly interface.
The data summary report ("Gene Set Summary") was limited to a few parameters, such as number of genes per gene type, division of gene set by organisms, as well as chromosome distribution of the studied gene set. This is a simple but significant output which aims to summarize the collected data.

3.1. Further biological evaluations from Gene Collector v.1.1 output

One of the challenges of physiology nowadays arise from conducting large-scale analyses of the transcriptome of a tissue and the large scale of the data sets. Comparative studies of transcriptomes from different tissues or life stages from an organism allow for inferences of spatial or temporal regulation of gene expression, which may hold some additional clues to the complexity of biological systems.

Acquiring and understanding experimental data is the first step to turn large datasets useful and understandable in the search for biological patterns. The dry-lab must be accessible and comprehensible for small and large volumes of biological data. It is necessary to: a) define the questions to be made to the dataset; b) produce inferences concerning
biological responses and c) understand the inferences obtained. The biological question that led to the production of Gene Collector involved the need to obtain information on genes obtained by various proposals. Large datasets obtained by the use of microarray or RNA-Seq are now a reality for the biomedical researcher [11].

Some examples in biological, biomedical and physiological research are questions like: what are the metabolic pathways involved in the differentially expressed genes in a metabolic pathway? These differentially expressed genes are located in specific groups, jointly regulated? What are the connections and interactions between these genes? What is the location of these genes on chromosomes?

From the generated output it was possible to further evaluate the studied gene set and answer specific biological questions, as exemplified in figure 5. Two of the most relevant questions to physiology are a) ontology structure and b) chromosomal co-localization of genes and synteny. The bioinformatic tools used to perform in-depth gene set analyses were Gene ontology [12], Genemania [13], Enrichnet [14], Kerfule [15] and Idiographica [16]. The results were satisfactory taking into account the objectives of the software. The sorting and organizing of data for later analysis in a user friendly manner is one of the desires of wet-lab professionals. Future requirements may occur with the advent of new bioinformatics tools.

An set of 13 genes linked to cholesterol transporter activity (GO:0017127) was used to present the possible biological uses of Gene Collector’s output (figure 5). Cholesterol is the principal sterol of vertebrates and the precursor of many steroids, including bile acids and steroid hormones. The cholesterol transporter activity enables the directed movement of cholesterol into, out of or within a cell, or between cells [17]. The software produced a local database and a spreadsheet, in which the data organization enabled a comparative study of genes and a starting point for further assessments using other available tools, such as Gene Ontology, Genemania and Enrichnet [12, 13, 14].

Shared or conserved synteny describes preserved co-localization of genes on chromosomes which can reveal coregulation in genomic assembly or regulation clusters [18]. Using the software’s output table on a sample gene set it was possible to obtain both the chromosome location depicted in the graphic idiogram format [16] and information on gene co-localization, which may be connected to genetic co-regulation [15].

3.2. Educational uses of Gene Collector v.1.1

English is the universal language in all fields of science and technology. Scientific language for molecular genetics is an extensive vocabulary to learn [19]. For non-English speaking countries such as Brazil, this is a continuous challenge for undergraduate students in biological and technical areas. Using the language in which the students feel most comfortable could benefit, especially when teaching bioinformatic approaches.

Figure 5 – Some examples of user latter from the output of the Gene Collector v.1.1.
Gene Collector can be used as an educational tool for undergraduate students, with a bilingual interface, with the option to choose the parameters, as well as the tutorials, in English and Brazilian Portuguese. Since the searched databases are in English, the returned data will still be in English, although the column header will be shown in the chosen language, facilitating the comprehension. Considering curricular issues, courses in biological, biomedical, health and correlated sciences could be improved with bioinformatics practical classes [20]. Bypassing the language barrier with this new tool can ease this process.

Gene Collector is currently undergoing registration process and will be made freely available for download at Laboratory of Biochemistry and Genetic Expression (ISCB/UECE – Fortaleza/CE - Brazil) web page at http://www.uece.br/cmacf/index.php/salas-e-laboratorios/274.

4 Concluding remarks and future developments

As far as we know, Gene Collector v1.1 is relatively unique and partially solves some problems concerning the organization and visualization of data from important databases. Some improvements are still needed, especially in the efficiency of gene recognition. After inputting Entrez and Ensembl gene IDs, Gene Collector is capable of gathering and summarizing information about nomenclature, species, location, alternative transcripts and gene product type. These are questions concerning the various gene expression laboratories among others demands.

In future versions of the program we aim to improve the efficiency in gene recognition, as well as include more data to be collected. Our challenges include obtaining more relevant information and reducing error percentages. Future versions may also include links with exemplified bioinformatics tools.

5 Acknowledgements

We gratefully acknowledge the financial support from “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES), “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq) and “Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico” (FUNCAP).

6 References


