ECRONICON OPEN ACCESS

EC MICROBIOLOGY
Research Article

Biological Databases Integration: A Data Warehouse Perspective Applied to Intergenic Regions

Daniel Luis Notari[1], Jovani Dalzochio[1], Camila RT Andrade[1], Jórdan R Rosa[1] and Scheila de Ávila e Silva[2]*

[1]Área do Conhecimento de Ciências Exatas e Engenharias, Universidade de Caxias do Sul, Brazil

[2]Instituto de Biotecnologia Universidade de Caxias do Sul, Brazil

***Corresponding Author:** Scheila de Ávila e Silva, Instituto de Biotecnologia, Universidade de Caxias do Sul, Brazil.

Received: June 07, 2019; **Published:** July 09, 2019

Abstract

There are several biological databases devoted to make available regulatory sequences for their analysis. However, there are not databases specialized on intergenic sequences of bacteria with information from each associated gene. Thus, the IntergenicDB portal is a public repository that is available on the web. IntergenicDB allows researchers to query information of intergenic regions and associated biological functions through a user-friendly interface. This article aims to describe the data retrieve, clean, association and load of the IntergenicDB Portal dataset, and present applications for this data

Keywords: *Intergenic Region; Biological Database; Data Warehouse*

Abbreviations

DDBJ: DNA Data Bank of Japan; DNA: Deoxyribonucleic Acid; DW: Data Warehouse; EBI: European Bioinformatics Institute; FTP: File Transfer Protocol; NCBI: National Center for Biotechnology Information; RNA: Ribonucleic Acid; SQL: Structured Query Language; SRA: Sequence Read Archive

Introduction

The biological phenomena are very complex and entail the integration of numerous areas of knowledge to validate or discredit hypotheses. The oldest (perhaps best known) interdisciplinary interface between Biology and Exact Sciences is Biostatistics. Gradually, Biology has used the tools provided by computing and mathematics to unravel complications in the most diverse fields: from Genetics to Ecology [1-3].

In this context, computational databases allow not only to store, organize and make accessible biological information, but also bring new views and discoveries to several applications [3]. A well-designed database is an efficient mechanism for large data sets [4,5]. It was not possible to find some estimative information in the literature about how many biological databases are available up to now. However, Biology inferences usually require the combination of data located in different or heterogeneous sources [3]. To address this issue, it is necessary the application of database integration approaches, such as data warehouse (DW), that can be described as an information repository where the data from different sources is stored in a standard design [5]. The process of the insertion of data in a DW comprises three main steps: (i) extraction of the data from heterogenous sources; (ii) transformation of the data in order to cleaning and improve the data accuracy; (iii) load the data originated from the second step into an integrated multidimensional schema.

Taking this in consideration, the IntergenicDB was created as a public database developed for the study of intergenic sequences [6] and was modeled to store relevant information on the structure of the intergenic sequences of Gram-negative bacteria. The portal has a restricted access administrative area and a public consultation area. The administration are enables the management of user and group registrations of the portal, and the group to which the user is associated will provide limited, or not, access. It is also possible to consult the biological data inserted in the database individually, not allowing insertion or maintenance of the data. The administrative area still provides the administrators with access to the users' access data and the administration of the texts of the portal's Home and Help pages.

IntergenicDB is justified since one of the utmost challenges of the post-genomic era is the determination of when, where, and how genes are expressed. The variances between two species is much more interrelated to the transcription of their genes than to the structure of these genes themselves. Thus, the study of gene regulation contributes to the construction of knowledge regarding the functionality of genes in distinct species, the cell differentiation in multicellular organisms, cellular response to environmental changes, among other issues [7].

In this context, regulatory essentials of gene transcription are in sequences known as intergenic, which consist of a non-transcribed element that comprises the sequences responsible for the process of regulating the onset and termination of gene expression. By creating an analogy, downstream elements (such as genes) represent the memory of a computer and the upstream elements (such as the promoters) the programs that act on that memory. Thus, the study of upstream elements can provide models about the constitution of the "program" and how it operates [8]. In prokaryotic organisms, such as bacteria and other single-celled organisms, the intergenic sequences are related to one or more genes [7]. Thus, biological information linked to a given intergenic sequence, as a gene associated with it, biological role of the gene and other information, contribute to the expansion of biological knowledge related to regulatory elements.

However, finding a biological database that has a gene connected to a specific intergenic region is not trivial. Some efforts in the obtention of intergenic sequences are the tools IntergenicS [9] and Junker [10]. Allied to this, computational knowledge is not homogeneous among Bioinformatics researchers who need to perform analysis of different data sources. To do so, this paper describes the data warehouse methodology used to improve the query in the IntergenicDB database. The article is organized with the sections of Materials and Methods, Results, Applications with Related Works and Conclusions.

Materials and Methods

This section presents the data sources used, the conceptual and logical model of the dataset, and the software description for generating the dataset.

Data from the biological databases GenBank [11] and Kegg [12] were used to collect the information of organisms, genes, family, kingdom and intergenic region the files of organisms were downloaded using GenBank FTP. The only information that does not originate from GenBank is the biological role. For such, a search was performed on the Keeg Brite database using the tool available on its website. The full description of IntergenicBD can be found in [6]. In order to insert the sequences in the database modeled as describe in figure 1, a tool named MMDB Import Tool was implemented. The figure 1 shows the Entity-Relationship diagram of the IntergenicDB database using Heuser notation [13]. The data stored in this database are: (i) each promoter region is linked to an organism, which has a name, a kingdom, a family, a type of molecule and a biological role; (ii) each gene has a name, a symbol, a start and end identification number, a function and a percentage of CG; (iii) each intergenic region has a number for the initial and final position in the sequence, a size, its nucleotide sequence and the type of tape to which it belongs.

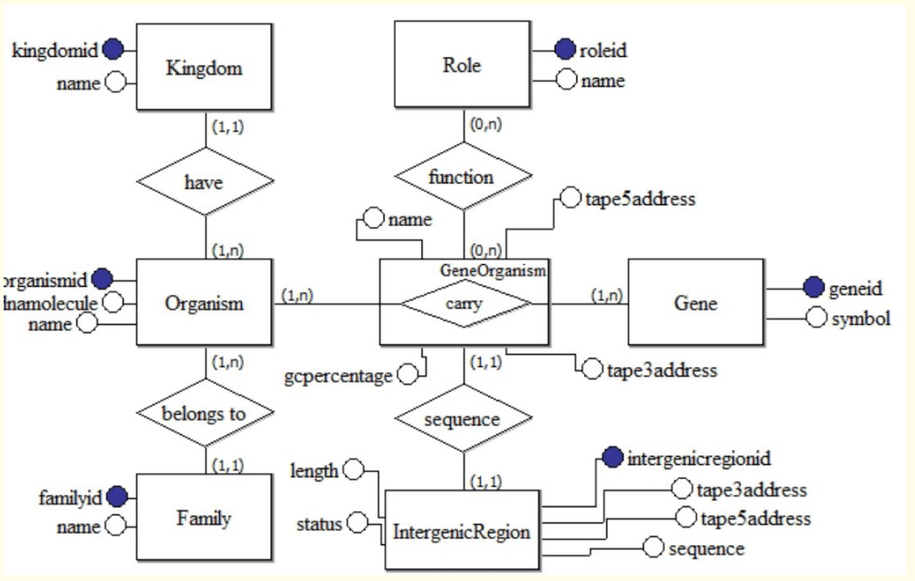

Figure 1: *Entity-Relationship diagram of the IntergenicDB.*

Aiming at optimizing database query performance, due to the various relationships and resulting volume, a data warehouse was implemented using the star model. The fact table (Figure 2) is formed only by the identifiers of the dimension tables. The fact table is static, that is, new data is not added to the table at the time of insertion. For this, it is necessary to run a routine that destroys the fact table and recreates.

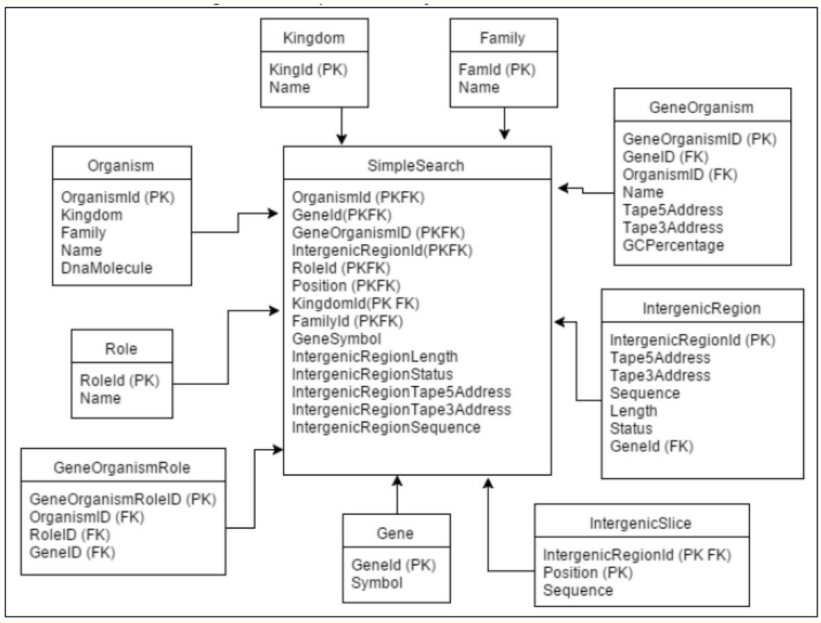

Figure 2: *Fact table and its tables dimensions.*

The query, available in the IntergenicDB portal, has an interface based on query builders of portals such as "PubMed Advanced Search Builder" of NCBI[1], "Free text search" of EBI[2] and "ARSA" of DDBJ[3]. From the parameters selected by the user, the query engine assembles an SQL query. Once the parameters are entered, the engine generates the query by looking for the identifiers of the parameters in the dimension tables and, after obtaining them, the query follows in the fact table. With identifiers filtered from the fact table, the query looks up their names in the dimension tables. The query only occurs within the range of user-determined pagination rows, that is, if the user parameterizes in the query that wants to view 50 rows per page, the query will only occur within that range to minimize the amount of data to be manipulated.

[1]https://www.ncbi.nlm.nih.gov/pubmed/advanced

[1]http://www.ebi.ac.uk/ena/data/warehouse/search

[1]http://ddbj.nig.ac.jp/arsa/

Results and Discussion

The execution of the MMDB Import Tool generated the data saved in the IntergenicDB. The structure of the dataset generated has the following structure: (i) organism: name of the organism; (ii) gene: gene symbol; (iii) kingdom: kingdom of the organism; (iv) family: family of the organism; (v) role: gene's biological role; (vi) tape5address: position of the sequence in the sense forward; (vii) tape3address: position of the sequence in the direction reverse; (viii) sequence: transgenic region data sequence.

The data set has 75252 lines entered with the data combined. In addition, it has data from 88 organisms, 15095 genes, 12565 biological functions and 55635 DNA sequences from different intergenic regions. Imported data are limited to gram negative bacteria. The present work highlights the analysis of DNA sequences for different organisms. IntergenicDB was created to store sequences of gene-based intergenic regions of gram-negative bacteria allowing future in silico analyzes using bioinformatics tools.

Conclusion

There is an assortment of tools employed in the analysis of regulatory elements of gene expression. Simultaneously, this widens the space of hypotheses generated, the range of standards (or lack thereof) of implementation becomes a restrictive factor in obtaining and contrasting results. The advance of a database to integrate the data of the intergenic regions into a single repository is central in helping the researcher to correlate this information with other online tools for the analysis of gene transcription regulatory elements. The intergenic dataset described in this paper was also presented in the Brazilian Data Base Symposium, in October, 2017 [14].

In addition, the relationship between the sequences deposited in this database with other online regulatory analysis tools is being developed. Considering this question, this project aims to amalgamate intergenic regions stored in IntergenicDB with other tools accessible on the internet for data analysis and/or prediction of sequences such as: promoters (e.g. BacPP [15], NNPP [16], BTSSFinder [17], PromPredict [18]), consensual motifs (e.g. ClustalO [19], WebLogo [20]), terminators (e.g. WebGeSTerDB [21], Arnold [22]) among other tools.

Acknowledgements

The authors would acknowledgement the University of Caxias do Sul for the financial support.

Conflict of Interest

The authors declare that there is no financial interest or any conflict of interest.

Bibliography

1. Barrera Junior., *et al*. "An Environment for Knowledge Discovery in Biology". *Computers in Biology and Medicine* 34.5 (2004): 427-447.
2. Lesk Arthur M. "Introduction to Bioinformatics". Oxford University Press (2013).
3. Marx Vivien. "Biology: The Big Challenges of Big Data". *Nature* 498.7453 (2013): 255-260.
4. Brown Justin R and Valentin Dinu. "High Performance Computing Methods for the Integration and Analysis Biomedical Data Using SAS". *Computer Methods and Programs in Biomedicine* 112.3 (2013): 553-562.
5. Elmasri Ramez and Shamkant B Navathe. "Fundamentals of Database Systems". 6th edition, Pearson Addison Wesley (2011).
6. Notari Daniel Luis., *et al*. "IntergenicDB: A Database for Intergenic Sequences". *Bioinformation* 10.6 (2014): 381-383.
7. Lehninger Albert L., *et al*. "Principles of Biochemistry". W.H. Freeman (2013).

8. Howard Daniel and Karl Benson. "Evolutionary Computation Method for Pattern Recognition of Cis-Acting Sites". *BioSystems* 72.1-2 (2003): 19-27.

9. Kurup Kavitha., *et al*. "Intergenics: A Tool for Extraction of Intergenicregions". *Bioinformation* 5.3 (2010): 83-84.

10. Sridhar Jayavel., *et al*. "Junker: An Intergenic Explorer for Bacterial Genomes". *Genomics, Proteomics and Bioinformatics* 9.4-5 (2011): 179-182.

11. Clark Karen., *et al*. "GenBank". *Nucleic Acids Research* 44 (2016): D67-D72.

12. Kanehisa Minoru., *et al*. "KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs". *Nucleic Acids Research* 45.D1 (2017): D353-D361.

13. Heuser Carlos Alberto. "Projeto De Banco De Dados". Volume 4, Bookman (2009).

14. Notari Daniel Luis., *et al*. "IntergenicDB: Banco de dados de regiões intergênicas de Bactérias Gram-Negativas". In: Brazilian Symposium on databases (SBBD) - Proceedings of the satellite events. Org: Carmem Satie Hara., *et al*. Uberlândia: SBC (2017): 234-244.

15. De Avila e Silva Scheila., *et al*. "BacPP: Bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria". *Journal of Theoretical Biology* 287 (2011): 92-99.

16. Burden S., *et al*. "Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences". *Bioinformatics* 21.5 (2005): 601-607.

17. Shahmuradov Ilham Ayub., *et al*. "bTSSfinder: A Novel Tool for the Prediction of Promoters in Cyanobacteria and Escherichia Coli". *Bioinformatics* 33.3 (2017): 334-340.

18. Rangannan Vetriselvi and Manju Bansal. "PromBase: A Web Resource for Various Genomic Features and Predicted Promoters in Prokaryotic Genomes". *BMC Research Notes* 4 (2011): 257.

19. Chojnacki Szymon., *et al*. "Programmatic Access to Bioinformatics Tools from EMBL-EBI Update: 2017". *Nucleic Acids Research* 45.W1 (2017): W550-W553.

20. Crooks Gavin E., *et al*. "WebLogo: A Sequence Logo Generator". *Genome Research* 14.6 (2004): 1188-1190.

21. Mitra Anirban., *et al*. "WebGeSTer DB-a Transcription Terminator Database". *Nucleic Acids Research* 39 (2011): D129-D135.

22. Naville Magali., *et al*. "ARNold: A web tool for the prediction of Rho-independent transcription terminators". *RNA Biology* 8.1 (2011): 11-13.

Volume 15 Issue 8 August 2019