Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 899–904

Procedia
Computer Science

www.elsevier.com/locate/procedia

The 5th International Workshop on Big Data and Networks Technologies (BDNT)
April 6-9, 2020, Warsaw, Poland

Enrichment of Metabolic Routes through Big Data

Amelec Viloria[a]*, Marisela Torres[b], Jesus Vargas[c], Omar Bonerge Pineda[d]

[a,c] *Universidad de la Costa, Barranquilla, Colombia*
[b]*Universidad Simon Bolivar, Barranquilla, Colombia*
[d]*Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

The Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway is a database that contains a graphical representation of cellular processes. Cellular processes are basic systems involving biochemical reactions at the cellular level such as transport, catabolism, metabolism, growth and cell death. The KEGG Pathway information is shown through the use of graphs, in which the molecular interactions between genes, processes and chemical compounds are represented. This paper proposes to perform Data Analytics using the Big Data Analytics Life Cycle methodology to enrich the metabolic pathways of the KEGG Pathway database by applying the Target Fishing technique.

Keywords: Chemical-Biological, Chemical Compound, Data Analytics, Metabolic Pathways, Target Fishing.

1. Introduction

Data analytics is a discipline that includes data lifecycle management and encompasses their data collection, cleansing, organization, storage, analysis, and governance. Data analytics uses predictive analysis based on patterns, trends, and relationships found in current or historical data [1]. The processing of these results can lead to the

* Corresponding author. Tel.: +57-3046238313
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2020.03.113

discovery of information and benefits in different branches of knowledge, as in the case of Computational Biology. For example, the analysis of information from chemical-biological databases in order to obtain predictions based on experimentally proven data [2], [3].

This research proposes the enrichment of information contained in the KEGG Pathway metabolic pathways. In this case, the enrichment of the routes consists of proposing alternatives of chemical compounds and targets through the integration of information from other databases such as UniProt, ChEMBL and ChEBI, based on experimental data, as well as the use of Target Fishing technique. However, localization of information has become a challenge for researchers because it is stored in different external or internal sources, and also in different formats, making it difficult to integrate [4] [5], [6]. A relatively large amount of data, with information in different formats from different sources, will be used in this paper. For these reasons, the Big Data Analytics Life Cycle method proposed by [1] is chosen.

The background check and theoretical aspects on which this work is based are found in:[7][8][9][10][11][12][13][14][15].

2. Method

The procedure used to obtain and analyze the data is framed in the stages of the Big Data Analytics Life Cycle proposed in [1] and is detailed as follows: Data Identification, Data Acquisition and Filtering, Data Extraction, Data Validation and Cleaning, Data Aggregation and Representation, Data Analysis, Data Visualization and, finally, Use of Analysis Results.

1. Data identification. Identifies the data sets and resources that are required for project analysis, which may come from external or internal sources to be collected.
2. Data acquisition and filtering. Collects the information established in the previous step, which is filtered depending on the type of source and provider, in order to eliminate unnecessary or corrupt data that will have no value for the purpose of the analysis.
3. Data extraction. It extracts data from different formats and transforms them into an understandable format for data processing and analysis.
4. Data validation and cleaning. It establishes data validation rules, removes any invalid data, and determines whether redundancy allows a set of interconnected data to be explored to complete the information with missing valid data.
5. Data aggregation and representation. It is dedicated to integrating multiple datasets to arrive at a single unified view. To do this, data structure and semantics must be considered.
6. Data analysis. Information analysis is performed using exploratory analysis, which is characterized by an inductive approach that facilitates the discovery of patterns based on methods or techniques.
7. Data visualization. It is dedicated to using data visualization techniques and tools to graphically communicate analysis results.
8. Use of analysis results. It determines how and where the processed data from the analysis can be used in the case study.

The Big Data Analytics Life Cycle will be applied in the present research based on the chemical-biological databases mentioned below:

Kyoto Encyclopedia of Genes and Genomes (KEGG)

An integrated database resource for understanding high-level functions and usefulness of the biological system such as cell, organism and ecosystem, from molecular-level information generated through genome sequencing and high-performance experimental technologies [13].

2.1 Universal Protein Resources (UniProt)

A database collection containing sequence information and annotations associated with proteins. Annotations consist of the analysis, comparison and fusion of all available sequences for a given protein, as well as a critical review of experimental data and associated prognoses [14]. To do this, biological information is extracted from the literature and numerous computational analyses are performed to provide known relevant information about a particular protein. It describes, in a single register, the different protein products derived from a given gene of a specific species. UniProt is structured in two sections: Swiss-Prot and TrEMBL. The first one is characterized by a review of experimental results, computational characteristics and scientific conclusions; on the other hand, TrEMBL includes protein annotations, but without a review of the data [15].

2.2 *ChEMBL*

An open database containing information about bioactive compounds, i.e. compounds that have biological activity within the organism. In addition, information from scientific publications is entered manually. This information is clean and standardized, which is a criterion that maximizes its quality and usefulness in solving biological research problems and in drug discovery. The information is published and can be accessed via a web interface [16].

23.3 ChEBI

ChEBI is a database of chemical entities of biological interest [21] that stores information on natural products or synthetic products used to intervene in the processes of living organisms [17]. The ChEBI database combines chemical nomenclature, structures, synonyms and chemical information related to chemical compounds to provide a wide range of related data such as formulas, links to other databases and an ontological classification, whereby parent-child relationships between molecular entities are specified [18].

3. Results

The enrichment of metabolic pathways information is made possible by the relationship between chemical compounds and targets. According to the principles of systems biology, a simple chemical compound that disturbs a network of targets within a metabolic pathway can trigger complex reactions and, in this way, it is possible to connect *in vitro* experimental data with data obtained by *in silica* analysis. These connections are based on the *Target Fishing* technique*,* since it allows predicting the targets of a chemical compound based on the calculation of similarity of the chemical structure and using information in chemical-biological databases.

In order to validate the results of the compound - metabolic path relationships, the hypergeometric distribution of probability with a *pvalue* less than or equal to 0.05 is established as a method. This is because the probability value obtained wants to validate that the results are not random, but that there is a statistically significant relationship between the data used in the analysis [19]. This is why it is considered to calculate the *pvalue* to show which metabolic pathways were enriched by the chemical compounds and their *KEGG targets.*

3.1 Chemical compounds

When selecting a chemical compound, information is displayed on the alternatives of *KEGG* targets and enriched metabolic pathways that met the established value for *p-value*, respectively; for the chemical compound C00002 *(Adenosine 5'-triphosphate).*

For example, for the metabolic route hsa04080 (*Neuroactive ligand-receptor interaction*), the following information is defined for the parameters mentioned in the method:

- $N = 2457$
- k is the number of *KEGG targets* that were obtained from *Target Fishing* for the chemical compound C00002, whose value is equivalent to 52 and the identifiers can be observed in Table 1.
- x is the total number of KEGG targets for the metabolic route hsa04082, which equals 232 and the identifiers are given in Table 2.
- m is the number of KEGG targets found in both k and x, which equals 16 for the compound C00002 and the metabolic route hsa04082, as shown in Table 3.

When replacing the function: hypergeom.sf (16,2457,52,219), it is obtained that the pvalue is equal to 0.00263, this being less than the established value of 0.05. Therefore, it can be said that there is an enrichment in the metabolic route hsa04080 by the compound C00002 and their respective KEGG targets.

Table 1. Identifiers of KEGG targets for k

144	141	1203	2298	3239	3985	5014	5074	6074	22975
162	192	1212	2547	3962	3879	504	5164	6747	60498
144	203	1341	2568	3780	5043	5699	5332	7223	64835
195	247	2674	2147	3686	5064	5890	5334	9224	65274
177	465	2253	3965	3474	5755	5781	5632	10780	84457

Table 2. Identifiers of KEGG targets for x

117	886	9002	2918	5732	45289	1909	2831	4886
134	887	9127	2925	5733	47852	1910	2837	4887
135	1128	9170	3061	5734	45697	2147	2846	4889
136	1129	9294	3062	5737	47524	2149	2847	4923
140	1131	9340	3269	5739	4236	2150	2859	4985

While two of the objectives described were to aggregate chemical information on metabolic pathway files and to develop a web interface to present enriched metabolic pathways, it became evident that a chemical cannot be said to have direct activity with a metabolic pathway protein. This is because there are certain biological limitations, such as that the cell membrane can prevent the chemical compound from entering, which would restrict any chemical compound from being part of a metabolic pathway, even if there is the potential for interaction between the compound and its target. Therefore, it was considered to show in the web page the chemical compounds that have activity on the metabolic pathways based on the calculation of the hypergeometric distribution of probability with p-value less than or equal to 0.05, considering the relationship of chemical compound and metabolic pathway.

This study managed the storage of information through files instead of using a database manager. This is due to the fact that files do not require a large amount of memory to store information, unlike databases, which, being complex programs, need disk space and memory to work efficiently [20]. In addition, performance is better through files since access to information is simple and light, unlike databases, where, if no indexes have been defined during the creation of the table, it could slow down the query time [21].

Table 3. Identifiers of KEGG targets of m

116	887	1919	2835	2928	4890	5892	9012	134758
133	888	1910	2838	2935	4897	5963	9128	339587
134	1129	2120	2847	3063	4890	5354	9173	100996698
137	1120	2121	2848	3065	4935	5967	9298	
142	1541	2122	2858	3268	4925	5589	9336	
148	1252	2386	2863	3272	4947	5695	9978	

Additionally, data migration is an easy process, since the files can be copied and pasted to a destination folder and provide write or read permissions; on the other hand, in the case of a database, the tables must be exported in the available formats, create the database and the table, and then import them into the other database, which means a time-consuming process for the user [22]. Another reason why files were used is due to the ease of reading for end users, since some of them are not computer technicians and it is not necessary to have specialized personnel to handle the information [23]. However, it is important to mention that if a database manager had been used to store the information, the programming time would have been optimized. This is due to the fact that the information would be accessed directly by specific queries in which the relationship conditions between the tables would be defined and navigation between directories and files would not be necessary to obtain data.

The reading and writing of data during the execution of the codes was done through RAM memory and not through disk memory, this is because access to information is instantaneous and concurrent, therefore, data should not be located, read and sent before processing, as in the case of disk memory [24] [25]. In addition, the performance for reading via RAM is better, since the information is accessed in the order of nanoseconds as opposed to disk memory, which takes milliseconds for reading [26]. However, there may be drawbacks such as loss of information in the event of a computer crash or power failure [27], as well as storage problems since it is not used to store information in long-term memory [28].

4. Conclusions

The research used Big Data Analytics Life Cycle method to organize activities and tasks related to data acquisition, processing and analysis, because it is characterized by its large data sets, stored in different sources and various formats. For this reason, data processing was structured on the basis of the stages proposed by this method. In addition, it was necessary to unify the following stages of the Big Data Analytics Life Cycle: Data Acquisition and Filtering, Data Extraction, Data Validation and Data Cleansing, in a single stage. This is because some datasets are used for the extraction and cleaning of other datasets. For example: KEGG Pathway information, after being extracted, filtered and verified, is used as base data for the acquisition, filtering, extraction and validation of UniProt, ChEMBL and ChEBI data.

The implementation of multiprocessing in the execution of the code in the data analysis stage of the Big Data Analytics Life Cycle allowed optimizing the execution time of Target Fishing, since it divided the processing according to the number of available nuclei and the amount of input chemical compounds.

The use of the hypergeometric probability distribution and the probability value equivalent to 0.05 made it possible to distinguish which metabolic pathways were enriched by the chemical compounds and their KEGG targets. This was done with the purpose of eliminating random results and determining which compounds have activity on the metabolic pathway.

In this sense, the research took the set of compounds contained in *KEGG Pathway* as the basis for *Target Fishing.* A future study could add more information to the metabolic pathways by performing *Compund Fishing*. For this purpose, in the Data Analysis stage of the Data Analytics Life Cycle, the chemical compounds of the ChEMBL database should be taken as base data. This will result in more chemical alternatives for the metabolic pathway.

Similarly, the present study implemented *Target Fishing* using a chemical compound from the *KEGG Pathway database*. Another future project would consist of carrying out *Target Fishing* on a set of compounds to be used in alternative mixtures such as natural products or in drug repositioning, which is the process of identifying new applications for existing, discontinued, or archived drugs that are currently under development for other medical conditions, in which the dose used may even vary [29] [30].

Finally, toxicological information on the side effects available in the *Side Effects Resource database* for *DRUG-type* chemicals could be added using the name stored in the *KEGG DATABASE. Side Effects Resource* contains information on side effects, frequency of effects, information on the drug at the commercial level and identifiers to other databases [31].

References

[1] T. Erl, W. Khattak y P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Indiana: Prentice Hall, 2016, p. 19.

[2] J. D. J. Durán, F. Astier y S. Banov, «Bases de Datos vs Sistemas de Archivos,» 22 enero 2014. [En línea]. Available: https://prezi.com/jgrydc9ncude/bases-de-datos-vs-sistema-de- archivos/. [Último acceso: 12 Noviembre 2018].

[3] A. Sulaiman, «File System vs. Database, » 27 Abril 2017. [En línea]. Available: https://dzone.com/articles/which-is-better-saving-files-in-database-or-in-fil. [Último acceso: 12 Noviembre 2018].

[4] Fundamentos de Bases de Datos, «1.4 Sistemas de bases de datos frente a los sistemas de archivos,» mayo 2010. [En línea]. Available: https://fundamentosdebasededatos.files.wordpress.com/2010/05/equipo2.pdf. [Último acceso: 12 noviembre 2018].

[5] International Multimedia Resource Center, «RAM vs. Hard Drive Memory, » 2018. [En línea]. Available: https://www.lehigh.edu/~inimr/computer-basics- tutorial/ramvsdiskspacehtm.htm. [Último acceso: 13 noviembre 2018].

[6] Kanehisa Laboratories, «KEGG: Kyoto Encyclopedia of Genes and Genome, » 2018. [En línea]. Available: https://www.genome.jp/kegg/. [Último acceso: 25 07 2018].

[7] United States Environmental Protection Agency, Appendix F. SMILES Notation Tutorial, Washington D.C., 2017.

[8] United States Environmental Protection Agency, «SMILES Tutorial,» 21 febrero 2016. [En línea]. Available: https://archive.epa.gov/med/med_archive_03/web/html/smiles.html. [Último acceso: 26 Julio 2018].

[9] Daylight Chemical Information Systems, «4. SMARTS - A Language for Describing Molecular Patterns, » 2008. [En línea]. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. [Último acceso: 26 Julio 2018].

[10] TOCRIS, «Cell Biology,» 2018. [En línea]. Available: https://www.tocris.com/cell-biology. [Último acceso: 16 octubre 2018].

[11] Kyoto Encyclopedia of Genes and Genomes, «KEGG PATHWAY Database, » 21 Agosto 2018. [En línea]. Available: https://www.genome.jp/kegg/pathway.html. [Último acceso: 16 octubre 2018].

[12] Bucci, N., Luna, M., Viloria, A., García, J. H., Parody, A., Varela, N., & López, L. A. B. (2018, June). Factor analysis of the psychosocial risk assessment instrument. In International Conference on Data Mining and Big Data (pp. 149-158). Springer, Cham.

[13] Gamero, W. M., Ramírez, M. C., Parody, A., Viloria, A., López, M. H. A., & Kamatkar, S. J. (2018, June). Concentrations and size distributions of fungal bioaerosols in a municipal landfill. In International Conference on Data Mining and Big Data (pp. 244-253). Springer, Cham.

[14] Kyoto Encyclopedia of Genes and Genomes, «KEGG release history, » 2018. [En línea]. Available: https://www.genome.jp/kegg/docs/upd_all.html. [Último acceso: 17 octubre 2018].

[15] M. Linderman, J. Sorenson, L. Lee y G. Nolan, «Computational solutions to large-scale data management and analysis, » Nature Reviews Genetics, vol. 11, pp. 647-657, 2010.

[16] L. Wang y X. Qung Xie, «Computational target fishing: what should chemogenomics researchers expect for the future of in silico drug design and discovery? » Future Med Chem, vol. 6, nº 3, pp. 247-249, 2014

[17] Viloria, A., Bucci, N., Luna, M., Lis-Gutiérrez, J. P., Parody, A., Bent, D. E. S., & López, L. A. B. (2018, June). Determination of dimensionality of the psychosocial risk assessment of internal, individual, double presence and external factors in work environments. In International Conference on Data Mining and Big Data (pp. 304-313). Springer, Cham.

[18] J. Swamidass† y P. Baldi, «Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval, » Journal of Chemical Information and Modeling, vol. 47, nº 1, pp. 952-964, 2006.

[19] S. Arif, J. Holliday y P. Willett, «Comparison of chemical similarity measures using different numbers of query structures, » Journal of Information Science, vol. 39, nº 1, pp. 1-8, 2013.

[20] G. Landrum, «RDKit Documentation,» 01 marzo 2018. [En línea]. Available: https://www.rdkit.org/RDKit_Docs.current.pdf. [Último acceso: 10 septiembre 2018].

[21] L. Sánchez, «Distribución hipergeométrica de probabilidad,» 29 octubre 2014. [En línea]. Available: https://estadisticayadministracion.wordpress.com/2014/10/29/distribucion- hipergeometrica-de-probabilidad-cero-complicada/. [Último acceso: 16 Noviembre 2018].

[22] X. Su, «Introduction to Big Data, » 29 Agosto 2017. [En línea]. Available: https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf. [Último acceso: 16 enero 2018].

[23] K. Minoru y G. Susumu, «KEGG: Kyoto Encyclopedia of Genes and Genomes, » Nucleic Acids Research, vol. 28, nº 1, pp. 27-30, 2000.

[24] The UniProt Consortium, «UniProt: the universal protein knowledgebase, » Nucleic Acids Research, vol. 45, nº 5, p. 2699, 2018.

[25] The UniProt Consortium, «UniProt: the Universal Protein, » [En línea]. Available: https://www.uniprot.org/docs/uniprot_flyer.pdf. [Último acceso: 29 Julio 2018].

[26] A. Gaulton, L. Bellis, P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani y J. Overington, «ChEMBL: a large-scale bioactivity database for drug discovery, » Nucleic Acids Research, vol. 40, nº 1, pp. 1100-1107, 2012.

[27] F. Haseltine, M. Huerta, Y. Liu, G. Downing y B. Seto, «NIH Working Definition of Bioinformatics and Computational Biology, » 17 Julio 2000. [En línea]. Available: http://www.bisti.nih.gov/docs/CompuBioDef.pdf. [Último acceso: 6 agosto 2018].

[28] M. Cruz Monteagudo, E. Tejera, Y. Pérez, J. Medina Fronco, A. Sánchez Rodríguez y F. Borges, «Systemic QSAR and phenotypic virtual screening: chasing butterflies in drug discovery, » Drug Discovery Today, vol. 22, nº 7, pp. 994-1007, 2017.

[29] N. Wale y G. Karypis, «Target Fishing for Chemical Compounds Using Target-Ligand Activity Data and Ranking Based Methods, » Journal of Chemical Information and Modeling, vol. 49, nº 10, p. 2190–2201, 2009.

[30] El Pasante, «Ventajas y desventajas de las bases de datos,» 17 junio 2015. [En línea]. Available: https://educacion.elpensante.com/ventajas-y-desventajas-de-las-bases-de- datos/. [Último acceso: 12 Noviembre 2018].

[31] Probability Formula, «Hypergeometric Distribution,» [En línea]. Available: http://www.probabilityformula.org/hypergeometric-distribution.html. [Último acceso: 16 noviembre 2018].