BioCicle: A Tool for Summarizing and Comparing Taxonomic Profiles out of Biological Sequence Alignments

Meili Vanegas-Hernandez, Tiberio Hernandez, Alejandro Reyes, and John Alexis Guerra-Gomez

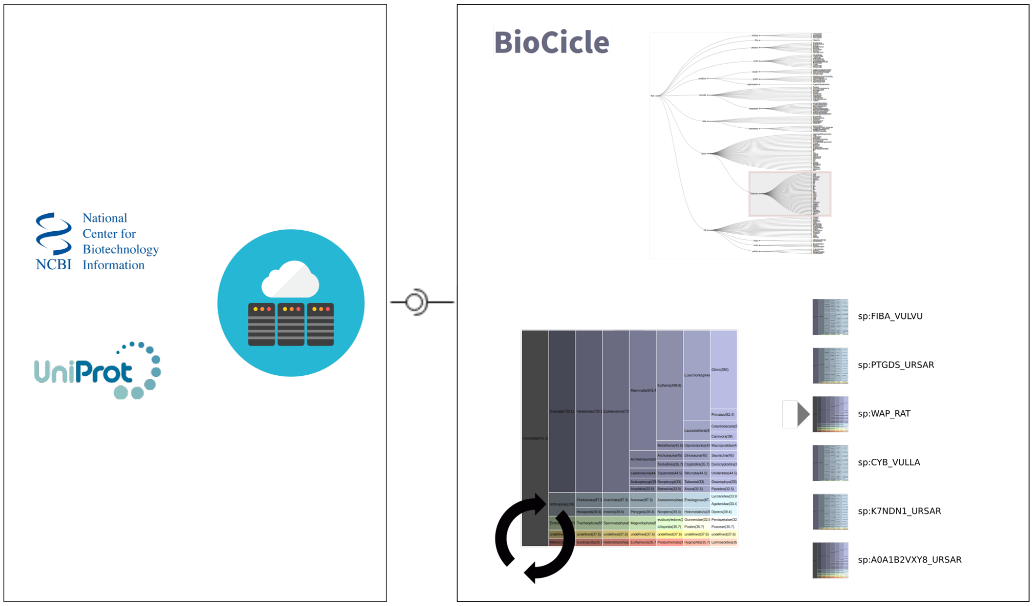

Fig. 1: BioCicle key features. We present BioCicle, an open source and web-based application for summarizing and comparing taxonomic profiles for single and multi-query biological sequence alignments. BioCicle supports several input formats, as well as direct sequence comparisons using the NCBI/EBI's APIs and UniProt.

Abstract—Biological sequence comparison is a crucial step towards the process of analyzing and cataloging new species. To achieve this, bioinformaticians must compare a new candidate to the universe of known species. This comparison produces a myriad of results, from where extracting useful information is highly cost-intensive given the lack of tools providing an overview of the results. Moreover, it is possible to mistakenly catalog new species due to poor analysis over comparison's outputs. To address this, we present BioCicle, a web-based and open-source system to summarize and compare single and multiple taxonomic reports out of biological sequence comparisons using straightforward visualizations. BioCicle is the outcome of a close collaboration with domain experts and a thorough study of the state of the art, from which we identified six analysis tasks commonly performed by biologists. Each task consists either in summarize (for single query results) or compare (for multi-query results): sequence alignments (AT1), taxonomic reports (AT2), and sequences' descriptions (AT3). Out of the revision, we concluded AT1 is very well covered by the state of the art, AT2 still has space for improvement and AT3 has been mostly unattended. BioCicle focuses in AT2 but, contrary to previous work, applies visual analytics principles from its inception. Furthermore, instead of requiring restrictive formats, our tool supports the most common outputs of the standard de-facto sequencing tools, or even allows to generate comparisons with the algorithm of choice directly using the NCBI/EBI's and UniProt APIs. BioCicle is released as an open source project and web-based tool available for usage by the community.

Index Terms—Information Visualization, Bioinformatics, Biological Sequence Comparison

1 Introduction

Currently, we are witnessing the acceleration of the accumulation of biological data that needs to be organized, classified and parsed into useful information. Somewhere between 15,000 and 18,000 new species are discovered yearly [11]. When a new organism is discovered, it must be studied and classified in the phylogenetic tree. Such classification process is done by acquiring a DNA, RNA or protein sequence of the unclassified organism and comparing it with sequences of previously classified organisms.

Algorithms such as Basic Local Alignment Search Tool (BLAST) and Hidden Markov Models (HMM) are used to compare sequences to one or multiple others to achieve this goal. The methodology performed in both algorithms is to arrange the sequences of interest to assist the detection of regions of similarity that can lead to evolutionary relationships between organisms [13]. Results in biological sequence analysis are generating large datasets, and fast retrieving of useful

- *Meili Vanegas-Hernandez is with Universidad de los Andes. E-mail: m.vanegas10@uniandes.edu.co.*
- *Tiberio Hernandez is with Universidad de los Andes. E-mail: jhernand@uniandes.edu.co.*
- *Alejandro Reyes is with Universidad de los Andes. E-mail: a.reyes@uniandes.edu.co.*
- *John Alexis Guerra-Gomez is with Universidad de los Andes, and UC Berkeley. E-mail: john.guerra@gmail.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

information is highly cost-intensive without an overview of the results.

The lack of manageable tools supporting results' summarization and comparison may lead to misclassification. Therefore, when a organism is misclassified, incorrect information is recorded in the database and future comparisons may consider inaccurate information when comparing new sequences. Rapid and widespread misclassification can be easily be attained out of few erroneous records. To address this issue, tools that enable fast retrieving of useful information must be reinforced and strengthen.

Nowadays, there are multiple platforms for downloading, submitting and analyzing biological data, such as the National Center for Biotechnology Information (NCBI) [22]. Although these platforms allow several interactions with the data, there are some limitations in the analysis tools for such output. The output for one sequence comparison consists in three different sets of results:

RS1 Sequence Alignments The sequence comparisons indicating which regions of the query sequence have significant similarities with the sequences in the database.

RS2 Taxonomic Report The linked-detailed information of taxonomic profiling for each sequence producing significant results.

RS3 Aligned Sequences' Description The list of sequences' descriptions with significant similarities, along with the calculated score of similarity.

A list of analysis tasks encouraged by the result sets provided by the NCBI platform is presented below. In Figure 2 we describe the different scenarios: two for each result set, in which we consider either single or multiple query displays.

AT1a Summarize sequence alignments for a single query comparison.

AT1b Compare sequence alignments for multiple query comparisons.

AT2a Summarize taxonomic reports for a single query comparison.

AT2b Compare taxonomic reports for multiple query comparisons.

AT3a Summarize sequence's descriptions for a single query comparison.

AT3b Compare sequence's descriptions for multiple query comparisons.

Out of this analysis tasks identification, we proposed a taxonomy of state-of-the-art implementations. Most of the tools currently available cover the analysis tasks concerning single query alignments for RS1 and RS2 (AT1a and AT2a Summarizing Sequence Alignments and Taxonomic Reports). Some of them support as well multiple query displays. However, the limited number of tools that support such task have restrictions either in the number of comparisons, in the portability of the tool or in the flexibility in input file formats.

This paper introduces `BioCicle`: a web-based and open source application to (1) summarize similarities in organisms' taxonomies out of biological sequences alignments and (2) compare multiple-query displays. Our main contributions are listed below.

- We present a taxonomy of state-of-the-art research projects and commercial tools supporting summarization and comparison of sequences alignments, classified by type of results, number of queries, and flexibility.

- We propose a visualization that summarizes the results of taxonomic profiles (**AT2a**) and follows good practices of visualization design.

- We introduce a visualization that supports exploration and comparison of multiple taxonomic profiles' results (**AT2b**), and follows good practices of visualization design.

- Our prototype is a web-based and open source application directly connected with both: the NCBI and the UniProt API, to instantly compute comparisons and visualize its results.

2 Related work

After a task-driven analysis developed along with a group of bioinformaticians, we proposed a taxonomy of the state of the art. The classification criteria consisted in whether the tools supported the analysis tasks or not. The six tasks were either *summarizing* or *comparing* for the three different outputs of interest: sequence alignments (AT1), taxonomic reports (AT2) and sequences' descriptions (AT3). This section presents how AT1 is widely covered by previous work, while AT3 is mostly unattended. Although AT2a is supported by some implementations, most of them have restrictive formats and do not assume some fundamentals of visual analytics design. `BioCicle` focuses on AT2 allowing taxonomic profiles summarization (AT2a) and comparisons (AT2b) for single and multiple queries using visual analytics design principles and withstanding multiple input formats.

BLAST and HMM are the most well-known methods for biological sequences' comparisons. The methodology performed in both algorithms is to arrange the sequences of interest to detect regions of similarity that can lead to evolutionary relationships between organisms [13]. Despite the differences between BLAST and HMM, both methods have similar outputs that need to be analyzed and processed in order to summarize and understand the results.

The NCBI website is a robust platform that allows bioinformaticians to upload, download and analyze sequences online. The analysis feature supports multiple comparison algorithms, such as BLAST, for which the user is asked to upload a sequence, choose a database and a program.

NCBI is one of the most used platforms to generate sequence comparisons, however visualization of the results is not one of their priorities. For that reason, there are multiple tools that provide an interface for sequence alignment results or comparable datasets. We made a revision of 17 of them and classified each tool considering which analysis task they addressed. As a result, we present a taxonomy (Table 1), which was developed focusing in single and multi-query analysis for sequence alignments and taxonomic reports (AT 1, 2), since non of the reviewed implementations supported sequence description analysis (AT3).

The vast majority of tools are based in single-sequence alignment summarization (AT1a) [8], [2], [4], [6], [19], [9] or multiple sequences alignment comparisons (AT1b) [21], [5], [3], [18]. Nonetheless, some approaches tackle taxonomic reports summarization [10], [17], [7], [14] for single-query alignments (AT2a) and comparison [23], [20], [12] for multi-query results (AT2b).

An additional criteria was considered regarding if the tool was tied to a specific comparison algorithm (Restrictive) or not (Non-Restrictive). As a result, we identified several restrictive tools tied to either BLAST analysis [8], [2], [4], [19], [10], [7], [14], [23] or HMMsearch comparisons [21], [5], [3], [18], [20], [12].

In conclusion, most of the tools identified are focused in single-query displays. The main drawback is that each sequence alignment must be analyzed independently, which leads to a highly cost-intensive understanding of the results. Besides that, the limited number of tools that support multi-query comparisons have a restrictive input or comparison algorithm and overlooks some essential visual analytics principles. Such approximations are described in detail as follows.

As stated before, our main interest relies upon taxonomic reports result set (RS2). Hence, we focus our attention in tools supporting taxonomic reports summarization and comparison (AT2). In particular, MG-Rast [14], AmphoraVizu [12], MetaPhlAn [20] and Krona [17] support single-query comparisons (AT2a). Tools such as Megan [10], Metarep [7] and BlastGrabber [16] tackle not only single-query results summarization (AT2a) but also multi-query comparisons (AT2b). However, tools belonging to this last group are all stand-alone

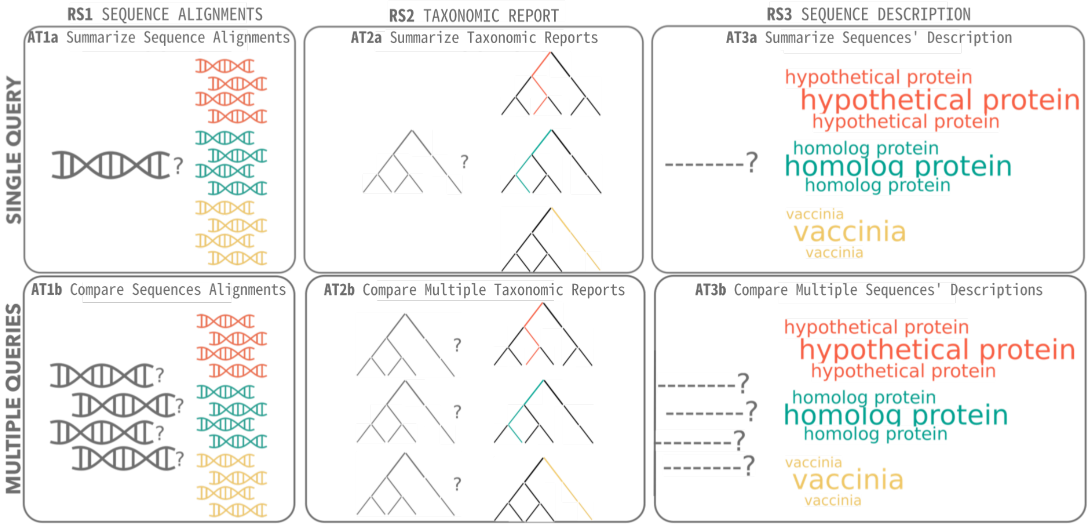
RS1 SEQUENCE ALIGNMENTS
RS2 TAXONOMIC REPORT
RS3 SEQUENCE DESCRIPTION
SINGLE QUERY
MULTIPLE QUERIES
AT1a Summarize Sequence Alignments
AT2a Summarize Taxonomic Reports
AT3a Summarize Sequences' Description
AT1b Compare Sequences Alignments
AT2b Compare Multiple Taxonomic Reports
AT3b Compare Multiple Sequences' Descriptions
hypothetical protein
homolog protein
vaccinia
?

Fig. 2: Analysis tasks identification for biological sequence alignments. We defined two types of tasks for each type of result set: summarization for single-query alignments and comparison for multi-query alignments. Subsequently, we propose 6 different analysis tasks consisting in summarize and compare: sequence alignments (AT1), taxonomic reports (AT2) and sequences' descriptions (AT3).

	RS1 Sequence Alignments		**RS2** Taxonomic Report		**RS3** Sequences' Descriptions	
	AT1a Summarize Sequence Alignments	**AT1b** Compare Multiple Sequences' Alignments	**AT2a** Summarize Taxonomic Reports	**AT2b** Compare Multiple Taxonomic Reports	**AT3a** Summarize Sequences' Descriptions	**AT3b** Compare Multiple Sequences' Descriptions
Restrictive	Bov Blast2Go Artemis HmmEditor	Circoleto ClustalW Hmmer Geneclusterviz Megan BlastGrabber	MG-Rast	Megan Metarep BlastGrabber	OUR ONGOING WORK	
Non-Restrictive	Clans GenoPlotR		AmphoraVizu MetaPhlAn Krona BioCicle	BioCicle		

Table 1: Taxonomy of state-of-the-art implementations supporting the analysis tasks defined in Figure 2. Neither of the reviewed implementations stand on sequences' descriptions (AT3). Most of the tools that support summarization and comparison of sequence alignments (AT1) and taxonomic profiles (AT2) have restrictive input formats or comparison algorithms. BioCicle introduces a non-restrictive platform for summarizing and comparing single and multi-query segment alignments taxonomic reports.

applications that run directly segment comparisons using a specific algorithm, thus, are classified as restrictive.

Regarding visual encodings, AmphoraVizu [12], MetaPhlAn [20] and Metarep [7] summarize aggregated results for a specific taxonomic rank either using heatmaps [20], [7], bar charts or pie charts [12]. Even though single-rank approaches facilitate score representation and readability by selecting a subset of the results, they instantly lose the advantages of an overview, relying in the user's memory to keep track of the storyline.

A more challenging approach considers the entire tree representation. Tools such as MG-Rast [14], Krona [17], Megan [10] and BlastGrabber [16] summarize entire taxonomic profiles either by concentric circles [17], node-link diagrams (layered [10] or radial [14]) or intended outlines [16]. Such representation allows the user to understand and identify the different classifications in the phylogenetic tree. However, as each sequence comparison result groups several taxonomic profiles, resulting trees are usually very big and hard to represent. One important disadvantage is that most of the tools do not express score values as part of the tree. This issue was gracefully addressed by Krona [17], which adopts a Radial Space-Filling (RSF) display engaging tree-like data with score values.

Krona is a metagenomic visualization tool, however, its RSF display can be used to present taxonomic reports' results for single-query alignments (AT2a), as shown in Figure 3. The entire implementation is presented as a web-based solution supporting several input formats. Representing trees using concentric circles allows an efficient use of the space, supporting big trees representation. In addition, the interactive design supports multiple tasks using dynamic visualizations. Despite its multiple strengths, Krona does not supports multiple query reports comparison (AT2b).

Although our proposal uses layering to represent multiple taxonomic profiles, same as Krona's implementation, we opted for an icicle visualization instead of a radial one. Despite Krona's implementation employs the available space more efficiently, our approximation eases label reading and score value interpretation. To tackle label reading, we aligned text to the horizontal axis. To interpret score values, we use bar's width rather than the area to represent the numerical value. Both decisions were taken considering expressiveness and effectiveness visualization principles [15].

After presenting a taxonomy of the current implementations supporting summarization and comparison of sequence alignments, we identify important analysis tasks that are not addressed by the state of the art. Following such discoveries, we present a visualization technique that supports taxonomic report summarization

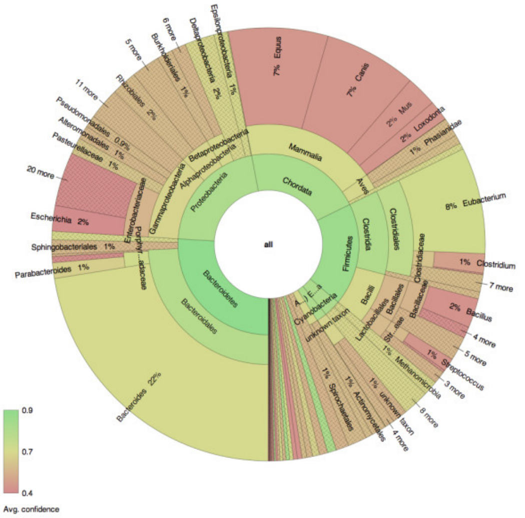

Fig. 3: The KRONA RSF Display. This visualization presents the results for a human gastrointestinal canal sample (MH0072) classification using PhymmBL taken from the MetaHIT project.

and comparison for single and multiple query results (AT2) following good practices of visualization design. In addition, we support multiple input formats, thus, comparisons can be either uploaded or generated through the web-based and open source application using preferred parametrization.

3 Implementation

BioCicle was conceived as a web-based application profiting from the rendering process, fast deployment, and easy access from any web browser. The visualization was developed using D3.js [1] and designed following the expressiveness and effectiveness principles [15]. This section will provide a quick look to the application's architecture and the visualization design process.

3.1 Architecture

BioCicle was built using Python in the server side and Javascript in the client. The server consists in two different modules: sequence comparison and taxonomic profiler. The first module can be activated or not, given the user's requirements. If the user uploads its own comparison (or comparisons) file, the comparison module is not called. Whereas if the user aims to directly perform the comparisons from the application, such is in charge to call the NCBI/EBI's API to apply the algorithm of choice.

When the comparisons are retrieved, the taxonomic profiler extracts the id for each of the compared sequences, consuming the UniProt's API. Later on, the taxonomic report is generated out of the NCBI's open source taxonomic database. Such report must be generated either if the user uploads an offline version of the results or generates it using the application. Afterwards, the client parses the resulting tree from multiple individuals and renders the icicle. The client side is developed using React Library and D3.js. A global presentation of BioCicle's architecture is presented in Figure 4.

3.2 Visual Design

The visualization design process was conceived differently for each of the analysis tasks addressed. Consequently, each process will be explained independently.

3.2.1 AT2a: Summarize Taxonomic Reports for a Single Sequence Comparison

The retrieved output provided from the server for a given unclassified sequence consists in list as long as the amount of classified sequences with which it was compared. Each record provided information about

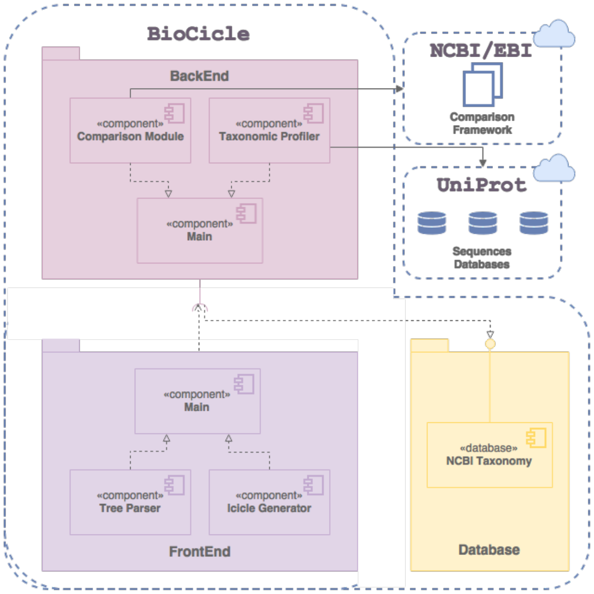

Fig. 4: BioCicle Architectural Diagram BioCicle's server was built entirely in Python, using the Python's APIs for the NCBI/EBI and UniProt connection. The server and client communicate by a REST service. The client was built with ReactJS and D3.

the taxonomic profile for the sequence and the similarity score resultant from the comparison. The bigger the score was, the more similar it was with the individual.

As the task we were tackling consisted in providing an overview of the result set, the representation of the entire tree was crucial for the process. The tree mapping was built based on the lowest common ancestor of the highest scoring alignments. Accordingly, we had two different variables to represent: the resulting tree out of grouping the entire taxonomies for every individual and a numeric score for each leaf of the tree.

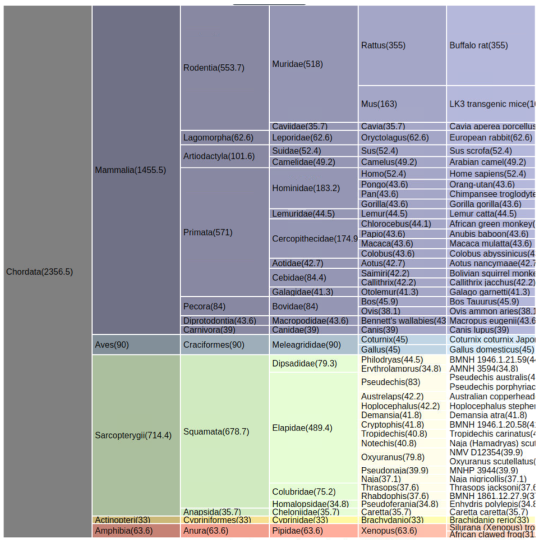

Fig. 5: Exploration prototype for taxonomic profiling summarization (AT2a).

Considering the comparisons result dataset we designed an icicle tree, as shown in Figure 5. Each dimension of the taxonomy (i.e. class, genus, order, etc.) was located in a different column of the icicle. Also, each dimension was threated as a nominal variable and represented using the spatial region in the limited column. The score value, as a numerical variable, was represented using the length of the columns divisions. Therefore, the height of each level was calculated with a linear scale having the score value as the domain and the number of pixels as the range. As an usual icicle tree, the child nodes score values/length contributed to the parent's value, meaning that an entire column considered the 100% of the sequences displayed.

Unlike Krona's implementation [17], BioCicle has more space-filling restrictions. However, we focused our design decisions in representing as clear as possible the score value for each compared sequence. Hence, our proposal ensures better score comparison representing the score variable using the rectangle's length rather than the polygon's area, as used in Krona's RDF display. In addition, BioCicle provides better taxonomy readability, locating labels parallel to the horizontal axis.

3.2.2 AT2b: Compare Taxonomic Reports for Multiple Sequences Comparisons

In the previous subsection, the result set considered was the comparison's results for a single unknown organism. Furthermore, the analysis task was to summarize such comparison's results and aimed to help the user to classify such organism in the phylogenetic tree. In this case, we did not considered one unknown organism, but multiple. With this, the users were willing to identify possible taxonomic similarities out of multiple unknown organisms.

With this in mind, the result set considered for this task was a grouping of multiple comparison results. Each comparison's result was represented as a list, as the one described in the previous subsection. As the explained before, each of these lists represented a tree, having as leaves the organisms with which the unknown was compared and the score of similarity. This resulting tree is called a taxonomic report. The grouping of those taxonomic reports could be interpreted as a conglomeration of trees.

Taking that into account, the new objective was visualizing several trees in a glance. Our proposal consisted in a constant iteration over comparison's results. Each result was shown in an icicle tree, as the visualization presented in Figure 5, and each icicle was presented for a customizable period of time (the default is 100 ms). The colors of each dimension were preserved for the entire iteration. For example, if 5 of the unknown individuals were compared to organisms in the Family *Elapidae*, the 5 icicles representing each of those results would preserve the color while rendering such family. This aimed to help the analyzer identify which were the most common ranks for the entire taxonomic reports.

Iterating over multiple trees provided an overview of the group of unknown organisms dataset. Yet, such implementation entirely relied in the user's memory and, for a considerable amount of organisms, analysis results were not profitable. In addition, if the . Therefore, we proposed to make clusters of unknown organisms, considering similarities in the taxonomic reports.

In order to group similar unclassified organisms, we had to assign them a possible taxonomy. Such possible taxonomy relied in a threshold set by the user. That means, if the user set a threshold of 80%, the summation of scores for each compared organism was calculated and the one that exceeded 80% of the score summation was assigned. It was possible that an organisms was not entirely classified for a demanding score. In those cases, the organism was assigned until the rank that fulfilled the given threshold.

Having that possible assignation for the entire unclassified organisms, they could be grouped in a new tree. This allowed us to present an overview of the entire group of unclassified organisms and with an interactive implementation, the user could select which node or rank was willing to analyze, focusing on a subgroup of the dataset. By doing so, we could present the visualization iterating over multiple icicles for a reduced subgroup of organisms. The iteration could be paused and resumed whenever the user desired to. Moreover, one sparkline for each of the unclassified organism was presented, indicating over which subgroup it was being iterating.

Megan [10] supports multiple comparison's results (AT2b) by grouping the entire taxonomic reports for each unclassified organism. However, they do not represent the score as part of the visualization. As a result, they just provide an overview of common ranks all the taxonomic reports.

Our implementation allows to dynamically explore taxonomic reports out of multi-query comparisons and compare general characteristics out of them, supporting a considerable amount of unclassified organisms.

4 Conclusions and Future Work

Biological sequence comparisons are a widely used methodology for organism classification in the phylogenetic tree. Such methodology assists detection of regions of similarity between DNA, RNA or protein sequences, which may imply evolutionary relationships between organisms. Sequence's comparison outputs have often a considerable amount of information, thus, fast extraction of relevant information is a very costly process; besides, organism misclassification can be easily achieved if the sequence comparison output is misread. Comparable mistakes affect not only the new individual classification, but also it ensures future misclassification, as such organism will be part of the comparison set in future sequence alignments.

Several visualization tools aim to address biological sequence alignment analysis. However, sequence comparisons results cover multiple result sets each one of them providing relevant information for different analysis. Of particular relevance there are three: sequence alignments (RS1), taxonomic reports (RS2) and sequence description (RS3).

Considering the different result sets of interest, we identified six different analysis tasks that should be covered by the state of the art. Two tasks were proposed for each of the result sets, one for a single-query sequence alignment and another for multi-query results. The resulting analysis tasks consist in *summarize* (for single queries) and *compare* (for multiple queries) either sequence alignments (AT1), taxonomic profiles (AT2) or sequence descriptions (AT3).

Once identified the different tasks concerning sequence alignment comparisons, we presented a taxonomy of state-of-the-art implementations. As an additional criteria, we evaluated if the tools had restrictive input formats or comparison algorithms. The classification yielded that all the reviewed tools focus their attention in sequence alignment analysis (AT1) and taxonomic reports (AT2). None of the implementations supported sequence description analysis (AT3). In addition, we discovered most of the tools considered single-query results (AT 1a and 2a), and the few supporting multi-query results have restrictive input formats or algorithms, which severely limits its functionality.

Consequently, we presented BioCicle, a tool built following the visual analytics principles that summarizes and compares either single or multi-query displays for taxonomic profiles in sequence alignments (AT2). BioCicle is an open source and web-based application that supports several input formats such as pregenerated comparisons result sets. In addition, the application is directly connected to the NBI/EBI and UniProt API's, allowing to generate a custom comparison on demand. BioCicle was constantly tested and evaluated along with a group of bioinformaticians.

Although BioCicle tackles an unaddressed problematic (AT2b) with non-restrictive characteristics, there is still an untapped potential in sequence description analysis for either single or multi-query displays (AT3). We are currently working in an extensive platform which, using text-analysis methods, feature selection and data mining, eases sequence description analysis for biological sequence alignments.

References

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[2] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. Gene Ontology Database Blast2GO:A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676, 2005.

[3] J. Dai and J. Cheng. HMMEditor: a visual editing tool for profile hidden Markov model. *BMC Genomics*, 9(Suppl 1):S8, 2008.

[4] N. Darzentas. Circoletto: Visualizing sequence similarity with Circos. *Bioinformatics*, 26(20):2620–2621, 2010.

[5] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(SUPPL. 2), 2011.

[6] T. Frickey and A. Lupas. CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, 20(18):3702–3704, 2004.

[7] B. A. Goll, Johannes; Rusch, Douglas B; Tanenbaum, David M; Thiagarajan, Mathangi; Li, Kelvin; Methé and S. Yooseph. METAREP: JCVI Metagenomics Reports - an open source tool for high- performance comparative metagenomics. *Bioinformatics*, 26:2631–2632, 2010.

[8] R. Gollapudi, K. Revanna, C. Hemmerich, S. Schaack, and Q. Dong. BOV âĂŞ a web-based BLAST output visualization tool. *BMC Genomics*, 9(1):414, 2008.

[9] L. Guy, J. R. Kultima, S. G. E. Andersson, and J. Quackenbush. GenoPlotR: comparative gene and genome visualization in R. In *Bioinformatics*, volume 27, pages 2334–2335, 2011.

[10] D. Huson. MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequening data. *PLos Computational Biology*, 12(6):e1004957, 2016.

[11] IISE. Retos SOS 2000 - 2009: A decade of Species Discovery in Review. Technical report, International Institute for Species Exploration, Tempe, AZ, 2011.

[12] C. Kerepesi, B. Szalkai, and V. Grolmusz. Visual Analysis of the Quantitative Composition of Metagenomic Communities: the AmphoraVizu Webserver. *Microbial Ecology*, 69(3):695–697, 2015.

[13] H. Li, X. Dai, and P. X. Zhao. A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. *Bioinformatics*, page 73401, 2008.

[14] F. Meyer, D. Paarmann, M. D'Souza, and Etal. The metagenomics RAST serverâĂŤa public resource for the automatic phylo- genetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.

[15] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[16] R. S. Neumann, S. Kumar, T. H. A. Haverkamp, and K. Shalchian-Tabrizi. Blastgrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive blast data. *BMC Bioinformatics*, 15(1):128, May 2014.

[17] B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385, 2011.

[18] V. R. Pejaver, J. An, S. Rhee, A. Bhan, J. H. Choi, B. Liu, H. Lee, P. J. Brown, D. Kysela, Y. V. Brun, and S. Kim. Geneclusterviz: A tool for conserved gene cluster visualization, exploration and analysis. *Bioinformatics*, 28(11):1527–1529, 2012.

[19] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000.

[20] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–4, 2012.

[21] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, mar 1994.

[22] D. Wheeler, T. Barrett, D. B. N. acids . . . , and undefined 2007. Database resources of the national center for biotechnology information. *academic.oup.com*.

[23] Y. Zhai, J. Tchieu, and M. H. Saier. JMMB Bioinformatics Corner A Web-Based T ree V iew (TV) Program for the Visualization of Phylogenetic Trees. *Journal of molecular microbiology and*, 4:69–70, 2002.