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

Meili Vanegas-Hernandez*, Fabio Andres Lopez-Corredor*, Tiberio Hernandez*, Alejandro Reyes* and John Alexis Guerra-Gomez**

**Universidad de los Andes, Colombia **Northeastern University, Silicon Valley, USA.*

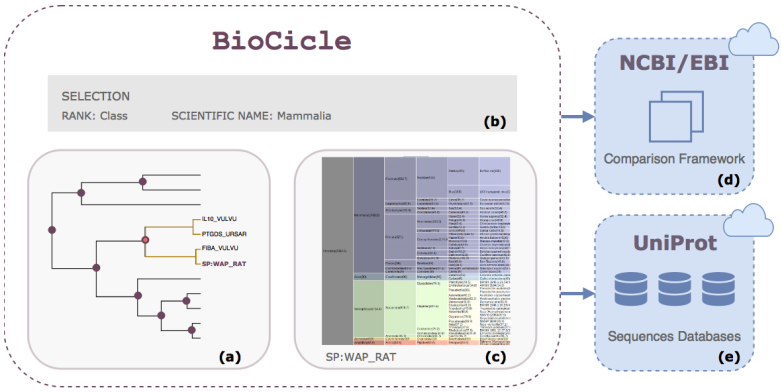

Figure 1: `BioCicle` key features. We present `BioCicle`, an open source and web-based application for summarizing and comparing taxonomic profiles for biological sequence alignments. `BioCicle` supports several input formats, as well as direct sequence comparisons using the NCBI/EBI's (d) and UniProt's (e) APIs. Once data has been loaded, `BioCicle` presents an overview of sequence comparisons results (a) and allows the user to filter regions of interest (b) to specifically analyze the most representative species that match to the each query according to the score of similarity (c).

Keywords: Information Visualization, Bioinformatics, Biological Sequence Comparison,

1. Introduction

A common practice in metagenomics consists in collecting genetic material from environmental samples, for later classifying them by comparing them to existing sequences in large known biological databases. BLAST and HMM are among the most common tools used for achieving this goal. The result of the

comparison consists of a large output that includes information about the similarity of the sampled sequences against the closest matches in the database. There is a lack of effective tools for summarizing and comparing such output, which leads overwhelmed analyst to keep only the first most probable result, ignoring all the remaining ones. This can lead to miss-classification, as more meaningful matches could be hidden in the remaining results. For example, think of a result that ranks a frog as the first result with a score of 70%, but the remaining results are bacteria with scores around 60%. An analysts could miss-classify the sequence as a frog, when there are high chances of being a bacteria.

To address this problem, this paper presents BioCicle, an interactive visual analytics system that summarizes all the results in a sequence comparison output, highlighting the corresponding scores, and allowing analysts to make decisions based on the whole input. Moreover, BioCicle also supports the summarization of multiple outputs, to identify patterns in groups of samples. We validated BioCicle by working closely with domain experts and by means of a case study presented in the paper.

BioCicle was designed following Munzner's visual analytics framework [1], identifying the main analysis tasks that analysts usually performed when doing sequence comparisons. From our close work with domain experts, we identified that sequence comparisons usually output three type of results:

- (RS1) Sequence alignments
- (RS2) Taxonomic reports
- (RS3) Aligned sequences' descriptions

Building on this types of results, and considering that domain experts can run single or multiple queries depending on the number of samples they collected, this paper contributes an analysis task taxonomy illustrated in 2.

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 six tasks were either *summarizing* or *comparing* for the three different outputs of interest: sequence alignments (AT1), taxonomic reports (AT2) and sequences' descriptions (AT3). `BioCicle` focuses on AT2 allowing taxonomic profiles summarization (AT2a) and

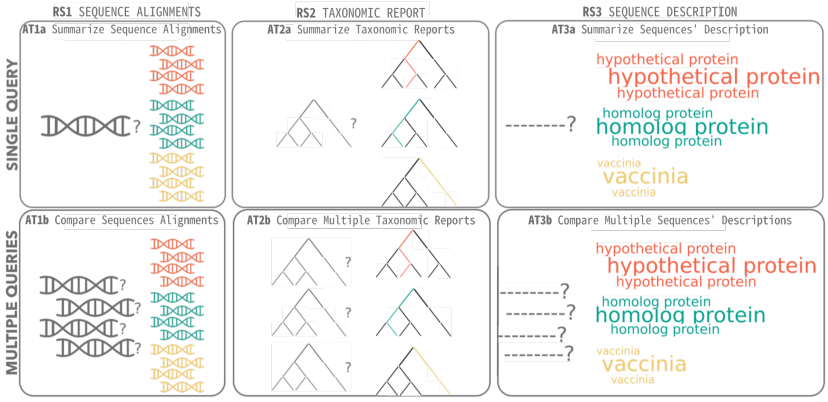

Figure 2: Taxonomy of Analysis tasks for biological sequence alignments.

	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		

Figure 3: Taxonomy of state-of-the-art implementations supporting the analysis tasks defined in Figure 2.

comparisons (AT2b) for single and multiple queries using visual analytics design principles [1] and withstanding multiple input formats.

We made a revision of 17 tools presenting sequence alignment results and classified each tool considering which analysis task they addressed. As a result, we present a taxonomy (Table 3).

The vast majority of tools are based in single-sequence alignment summarization and comparison (AT1). Nonetheless, some approaches tackle taxonomic reports summarization [2], [3], [4], [5] for single-query alignments (AT2a) and comparison [6], [7], [8] for multiple comparisons results (AT2b).

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.

3. Implementation

3.1. AT2a: Summarize Taxonomic Reports for a Single Sequence Comparison

The visualization consisted in three different components: an icicle map, a collapsible tree and a set of small multiples, as shown in Figure **??**.

1. Icicle Tree: 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.

2. Collapsible Tree: 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 species 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.

3. Small Multiples: We presented the total amount of icicles as small multiples for each of the unclassified sequences. indicating over which subgroup it was being iterating.

Our implementation allows to dynamically explore taxonomic reports out of multiple comparisons comparisons and compare general characteristics out of them, supporting a considerable amount of unclassified species.

4. Results

In order to validate the framework, this case study considered a set of 179 sequences. These sequences were obtained by sequencing a genetic marker over a single sample. The gene was the 16S ribosomal rRNA, which is a molecular genetic marker and it is used to detect variations in DNA. In this case, sequencing such gene is meant to identify the diversity of the sample. The original dataset

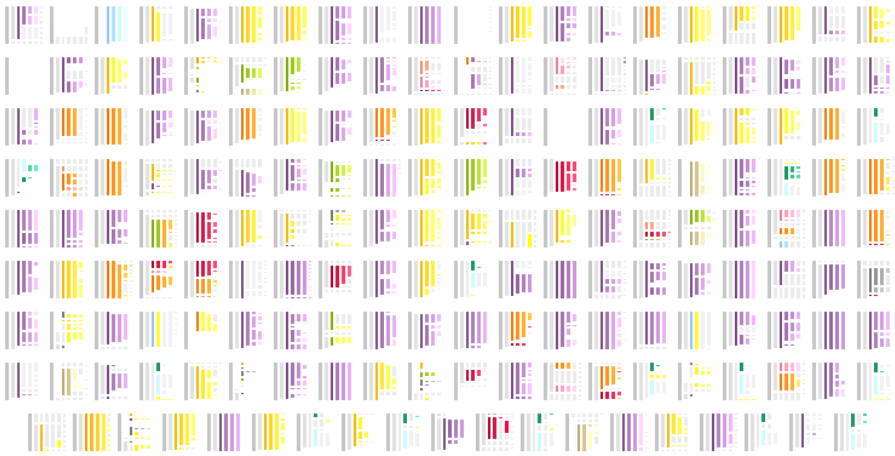

Figure 4: Overview of the taxonomic profiles for the entire 16S dataset using *small multiples*. This visualization shows an overview of the trees generated for each of the compared sequence. Colors and spatial regions are preserved for nodes all over the entire for ease of understanding. Prevailing tones are purple, yellow, and orange.

consisted in 23.000 sequences but, considering time limitations, only 179 were employed for this study. The results are shown in Figure4.

The small multiples allow the user to have a general overview of the results. In Figure 4 it is possible to see there is a diverse distribution of ranks over the sample, however, yellow, purple, and orange tones prevail in the general sight.

After analyzing the results for this specific sample, it was possible to quickly describe the dataset as sequences belonging mostly to class *Proteobacteria*, with a high amount (49,16%) having high similarities to sequences belonging to the order *Gammaproteobacteria*. Additional sequences had high similarities with sequences belonging to classes *Alphaproteobacteria* (29,05%), *Deltaproteobacteria* (7,82%) and *Flavobacteria* (12,29%).

5. Validation

To test the contribution of `BioCicle` in the performing of the identified tasks by bioinformaticians, potential users were asked to use the tool and validate the easier approach to common tasks. A recorded session with two computational biology students and a researcher in bioinformatics was done, in which we could evaluate both the usability and usefulness of the platform. They were also prompted to identify the potential tasks that `BioCicle` allows them to do. A standard usability scale was used to evaluate how well each task is performed by the web platform.

From the recordings of testing sessions we got the proof from actual users that `BioCicle` is able to help in a quick processing of the results from a BLASTp execution. The findings from the tests, aside from web development usability, allowed the team to define the respective upcoming developments around two fronts: Specific tasks and new functionalities.

5.1. Specific tasks

The users determined which tasks, in terms of their domain vocabulary, are able to be executed with the tool's assist. For each task they graded `BioCicle` in a difficulty scale from 1 to 7, where 1 was a task hard to perform with the web page, and 7 a task easily executed with the help of the platform.

The description in the terms of the platform of these tasks, with their user-assigned grades, is well explained in Table 5. In first instance, we recognized most identified tasks as similar to the original analysis tasks identified (AT2a and AT2b from Figure 2). User-defined tasks 1, 3, 4 and 6 make reference to sequence alignments, including both single and multiple queries. Task 2 is an analysis work that could be done after performing the identified analysis task, while task 5 is an additional function of filtering to give more control to users of the information entered to the application.

User Defined Task	Grade (1-7)	Description in terms of BioCicle	Related Analysis Tasks
UD1	5	Visualization of taxonomy from entered comparison file from sequences of virus proteins	**AT2a** & **AT2b**
UD2	6	Explore nodes in collapsible tree and Icicle and determine which nodes are from a different nature from the expected from the initially uploaded sequences	Analysis done after results from **AT2b**
UD3	2	For an Icicle leave, understand how it ranks among other leaves according to their score.	**AT2a**
UD4	6	Identify the nodes from the collapsible tree or Icicle with the highest scores	**AT2b**
UD5	6	Select a node from the collapsible tree, and download the filtered xml file.	Filtering results from **AT2b**
UD6	7	For a comparison file from a single sequence, visualize the taxonomy in the Icicle	**AT2a**

Figure 5: Comparison between User-defined tasks and Analysis tasks identified in Figure 2.

The grades for the user-defined tasks turned out to be favorable with the exception of user-defined task 3. In the session the user who identified and graded this task, a student mainly interested in assembling and classification of virus, stated that the results were displaying both the taxonomy for the virus and the bacteria where the virus is found at. As it is unclear if the taxonomy is calculated from the sequence from the virus or the bacteria containing it, it is hard for the user to understand if the sampled virus actually corresponds to the retrieved genome.

6. Conclusions and Future Work

Biological sequence comparisons are a widely used methodology for sequence identification and characterization. Such methodology assists detection of regions

of similarity between DNA, RNA or protein sequences, which may imply evolutionary relationships between species. Sequence's comparison outputs have often a considerable amount of information, thus, fast extraction of relevant information is a very costly process. Besides, sequence 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 misclassifications, as such sequence will be part of the comparison set in future sequence alignments.

This project introduces BioCicle, a tool to summarize and compare either single or multiple results displays for taxonomic profiles in sequence alignments (AT2). It was developed following the visual analytics principles. In addition, the application was directly connected to the NCBI/EBI and UniProt API's, allowing custom comparison generation on demand. BioCicle was constantly tested and evaluated along with a group of bioinformaticians and presented as a proof of concept.

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 multiple comparisons displays (AT3). Ideally, BioCicle could be extended as a decision support system framework for sequence alignment analysis for taxonomic reports and sequence's description. Methods such as text-analysis, feature selection and data mining could ease sequence's description analysis and decrease incorrect insertions rates in biological databases.

7. References

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

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

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

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

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

[6] Y. Zhai, J. Tchieu, 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 (2002) 69–70.

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

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