Interactive Visual Support for Metagenomic Contig Binning

Bertjan Broeksema* Fintan McGee† Magdalena Calusinska‡ Mohammad Ghoniem§

Centre de Recherche Public Gabriel Lippmann

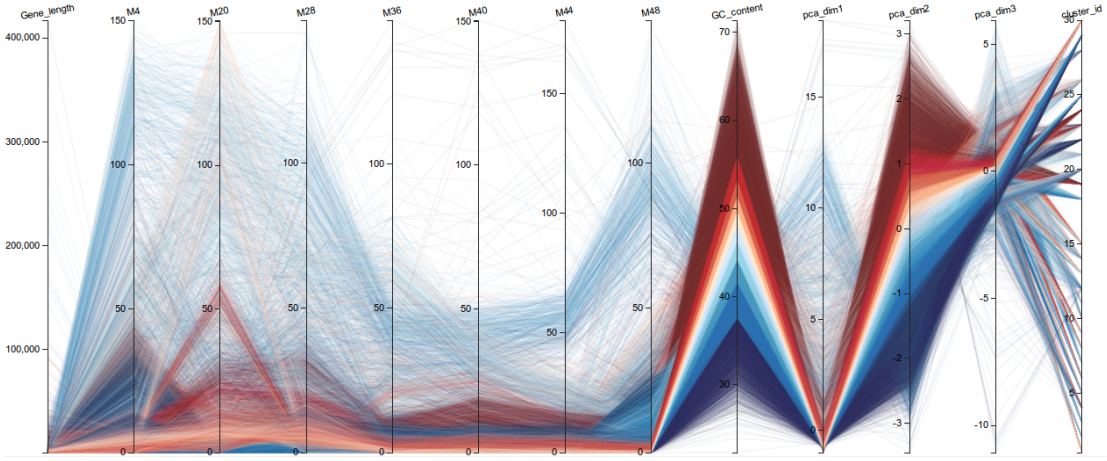

Figure 1: Interactive web front-end to visually inspect results and adapt analysis parameters.

Abstract

Within metagenomics, "Contig Binning" is an important step in the process of reconstructing genomes of species in mixed cultures and environmental samples. We present an interactive visual environment which enables a biologist to statistically analyze the multiple dimensions of data that are typically used during binning, and integrate and compare the results of various binning methods. Our system features a web-based parallel coordinate visualization at the front end and a R server back end for analysis and semi-supervised clustering of contig data.

1 Introduction

The process of species DNA reconstruction starts with the sequencing of biological material, such as samples of nucleic acid material from the same pool at different points in time, resulting in small fragments of deoxyribonucleic acid (DNA). As a first preprocessing step, fragments of insufficient quality are filtered out. The remaining segments are assembled to obtain longer genomic contigs, using specialized software. In order to analyze longer and more complex genetic elements, further assembly is required. To steer the assembly process in a meaningful way, contigs are organized into groups which might represent an individual genome or genomes from closely related organisms. This process is referred to as binning. Finally, contigs in a single bin are re-assembled to obtain even longer fragments of genomic data. The resulting reconstructed genetic material can then be used for further analysis such as species identification or functional analysis.

*e-mail:broeksem@lippmann.lu
†e-mail:mcgee@lippmann.lu
‡e-mail:calusins@lippmann.lu
§e-mail:ghoniem@lippmann.lu

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

We present an interactive visual environment which enables biologists to combine the results of various binning methods and relate them to the data that is used during binning. Our goal is to enhance automated binning methods by using multi-dimensional data visualization that allows for interactive curation by expert users. We are closely collaborating in an iterative process with biologists, who investigate bacteria populations involved in biogas production.

Many approaches to contig analysis and binning are taxonomy dependent. These are often similarity based, and directly consider the similarity of the contig to a reference data set of known origin. Non-taxonomic approaches, utilizing techniques such as Dimensionality Reduction (DR) [4] or Self Organizing Maps (SOMs) [2],are often necessary, as there is no reference genome for the vast majority of species. Compositional approaches, utilizing DNA characteristics such as *GC-content* or Tetranucleotide frequency (TNF) [3, 6], are an alternative to similarity based approaches. Recent work has also focused on examining different DNA sequences from the same source[1, 5]. [3] uses a combination of TNF and contig abundance data from multiple samples and clusters them using pre-trained probabilistic models.

Visualization is not a prominent feature of many contemporary contig binning approaches. The described approaches are highly automated. User control is usually limited to parametrization of algorithms, or in the case of [1], manually curating the output, using existing graph visualization applications. [4] utilizes a form of DR not seen previously in meta-genomic analysis, but the final visualization is a simple 2D scatter plot. The concepts of user interaction at all stages of the process and exploratory data analysis are not considered, within the existing metagenomics literature on the contig binning problem.

2 Current Prototype

Our approach is a non taxonomy-dependent compositional approach to binning, utilizing a sequential time series of samples. We visualize the data interactively using parallel coordinates, to explore multiple dimensions of the input data. Our prototype currently consists of three main components: an R back-end, a node-js middle-

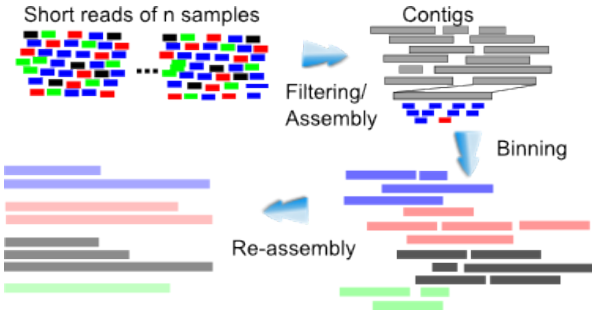

Figure 2: Simplified view of the assembly process.

ware server and a web front-end.

The R back-end reads the data source and provides various scripts to analyze the data. The output of analysis (clustering and Principal Component Analysis) can be seen as axes in figure 1. The flexible design of our system allows for utilization of publicly available R scripts for alternative approaches, such as [1]. This allows for comparing the effectiveness of existing approaches, as well as enhancing our interactive exploratory approach by including results of established contemporary techniques in the analysis.

The node-js middleware server provides a REST API for executing various R-scripts with different parameter settings. Together with the web front-end this lowers the entry barrier for less technical users. They can start exploring the data visually right away, and slowly move to adapting existing or writing new R scripts.

The web front-end enables interactive exploration of the data and analysis of results. A parallel coordinates view (Fig. 1) was chosen because it allows for easy detection of observations that have similar trends. To further enhance this detection we provide various coloring methods, such as coloring per decile for a given attribute or based on cluster id. Filtering is provided to allow the user to get rid of outliers or to zoom in on a particular selection of interest. Additionally, we keep track of the filtering history so that the user can track back the steps that led to a particular selection of contigs. At any given point a selection of contigs can be exported in a format suitable for re-assembly. Finally, we automatically generate parts of the UI based on a special comment line in the R-scripts to enable parameter settings for each of the scripts in the web front-end.

2.1 Analytical process

We support the contig binning as follows. Different types of clustering are calculated by the backend and the resulting cluster data is added as a variable to the data set. This data set is next visualized in the parallel coordinates view. Multiple coloring options are provided for visual inspection of patterns in the data. For example, the user can color based on one of the clustering results to inspect what makes clustered contigs similar. Axes can be brushed to highlights part of the data. Upon brushing data can be filtered, by either keeping the brushed subset or by removing it. This helps in filtering out outliers or for inspecting a subset of particular interest. Finally, we perform Correspondence Analysis on the TNF data for the contigs in the parallel coordinates. Results of this analysis are shown in a separate scatter plot which is linked with the parallel coordinates. It serves as a means to find out structural properties of a selection of contigs. Once the biologist has identified a subset of contigs that seem to belong to one species, these contigs can be exported for reassembly.

3 Current Open Challenges

Many of the visualization specific challenges in the DNA reconstruction process revolve around uncertainty. The multi-step DNA reconstruction process adds uncertainty at each step due to the differing quality of the sequence reads. Typically, thresholding is applied after which all remaining reads are treated equally, even though there is still variation in the quality. An interesting challenge would be to clearly convey the trade-off that is being made when discarding part of the reads and the quality of the remaining ones. After the initial assembly a similar thresholding can be applied to leave out contigs for which the confidence is too low. Again, there will be still variability in the confidence for the remaining contigs. Additionally, some forms of binning such as [3] and [4] require contigs with lengths of at least 1000 nucleotides. As a result, more data is discarded in the reconstruction process. Conveying uncertainty of quality will allow users to make an informed decision about which contigs should be used in analysis.

Another challenge is that the data is a mixture of time-series (abundance level of a contig for each sample), contig properties (e.g. length) and summary data (e.g. gc-content). Additionally, we deal with both numerical data (e.g. contig abundance level) and categorical data (e.g. TNF). Analyzing and presenting this data in ways that speed up the overall analysis of the user and properly treats each data-type (time-series, categorical, numerical) remains an open challenge.

4 Future Work

Parallel coordinates excel at displaying trends across multiple dimensions. An enhanced brushing approach, allowing selection of contigs based on correlation with a selected contig or a user defined pattern, may be useful for exploring relationships between contigs. This would allow users to explore potentially related contigs as well as search for potential symbiotic relationships between species.

The large number of contigs in a sample can lead to a significant amount of overdrawing in a parallel coordinates visualization. While this currently is mitigated by brushing and filtering, the use of bundling and aggregation techniques may further enhance the clarity of the visualization, as well as providing another technique to convey groupings of contigs.

Our access to Subject Matter Experts, who are target end users, will allow us to provide an expert evaluation of the system, as well as provide realistic comparisons of different approaches to contig binning. One potentially very interesting avenue in future work is a comparative evaluation of the effectiveness of the TNF and contig abundance data sources. Exported contig clusters are assembled externally and rated based on the quality of the resulting genomes. This could provide an evaluation metric for different binning techniques. Additionally, the integration of environmental parameters, known for shaping the microorganism community structure, within the parallel coordinates visualization, is perceived as potent information to facilitate full genome assembly and to unravel microbe-environment interactions.

References

[1] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538, May 2013.

[2] G. J. Dick, A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield. Community-wide analysis of microbial genome sequence signatures. *Genome biology*, 10(8):R85+, 2009.

[3] D. D. Kang, J. Froula, R. Egan, and Z. Wang. Metabat: Metagenome binning based in abundance and tetranucleotide frequency, 2014. Poster presented at the JGIUM.

[4] C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Scientific reports*, 4:4516, Jan. 2014.

[5] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120, 2013.

[6] M. Strous, B. Kraft, R. Bisdorf, and H. Tegetmeyer. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Frontiers in Microbiology*, 3(410), 2012.