CAMSA: A Tool For Comparative Analysis And Merging Of Scaffold Assemblies*

Sergey Aganezov and Max A. Alekseyev

The George Washington University, Washington, DC

Background

Despite the recent progress in genome sequencing and assembly, many of the currently available assembled genomes come in a draft form. Such draft genomes consist of a large number of genomic fragments (*scaffolds*), whose positions and orientations along the chromosomes are unknown. The *scaffold assembly* problem asks for reconstruction of chromosomes from a set of scaffolds by identifying pairs of scaffolds extremities (*assembly points*) to be glued together. While there exists a number of methods for solving the scaffold assembly (using various computational and wet-lab techniques), they often can produce only partial error-prone assemblies.

Depending on the utilized information and the underlying techniques, different scaffold assembly methods may produce results that differ from each other. Moreover, some scaffold assemblers can produce only non-oriented assemblies, where the relative orientation of (some) scaffolds in assembly points is yet to be determined. It therefore becomes important to compare and merge scaffold assemblies produced by different methods, thus combining their advantages and highlighting potential conflicts for further investigation. These tasks may be labor intensive if performed manually.

We present CAMSA, a tool for comparative analysis and merging of scaffold assemblies. CAMSA takes as an input two or more assemblies of the same set of scaffolds and generates a comprehensive comparative report for them. The report not only contains multiple numerical metrics for the input assemblies, but also provides an interactive framework for their visual comparison and analysis. CAMSA is available for download from `https://cblab.org/camsa/`.

Methods

CAMSA interprets the input assemblies as sets of assembly points, and further analyzes and classifies individual assembly points by a numbers of characteristics (e.g., uniqueness, orientation, conflictedness, etc). Results of this analysis are then reported at the levels of whole assemblies and individual assembly points.

For the purpose of comparative analysis and visualization of the input scaffold assemblies, CAMSA utilizes the *multiple breakpoint graph* (MBG) data structure traditionally used for analysis of gene orders across multiple species [4]. The MBG in CAMSA is formed by directed *scaffold edges* and undirected *assembly edges* of different colors representing the different input assemblies (Fig. 1). While conventional MBG is constructed for sequences of *oriented* genes (where orientation is defined by the strand), in CAMSA we extend it to support sequences of non-oriented scaffolds.

In addition to generating a comprehensive comparison report, CAMSA also produces a *merged assembly* that is most consistent with all input assemblies. CAMSA can take into account the level of confidence of each assembly point in each input assembly, which can be specified as the *confidence weight* on the scale from 0 to 1 (with 1 being the default value). These confidence weights contribute to the weights of assembly (multi-)edges in the MBG, which are then used to construct the *merged assembly* as the maximal matching on assembly edges (shown as bold colored edges in Fig. 1). We further use the constructed merged assembly to identify orientation for some non-oriented assembly points that is most consistent across the input assemblies (e.g., in Fig. 1 for the blue non-oriented assembly point $(\overrightarrow{F}, \overleftarrow{G})$ it suggests orientation $(\overrightarrow{F}, \overrightarrow{G})$), as well as to resolve issues of varying resolution across different assemblies, i.e., when a scaffold is missing in one assembly but is present in another (e.g., in Fig. 1 the scaffold D is missing in the red assembly, but is present in the blue assembly as well as in the merged assembly).

Results

The results of scaffold assembly analysis in CAMSA are presented in the form of an interactive report for the set of input assemblies, and an interactive visualization of the input and merged assemblies. Extensive interactive filtering options

*The work is supported by the National Science Foundation under the grant No. IIS-1462107.

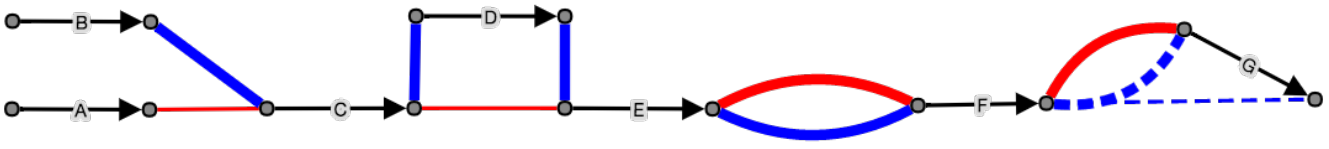

Figure 1: The MBG of "red" and "blue" assemblies of the same set of scaffolds $\{A, B, C, D, E, F, G\}$, where directed black edges correspond to scaffolds, red and blue edges correspond to assembly points, dashed edges represent alternative orientations for non-oriented assembly points, and bold edges indicate assembly points that participate in the merged assembly.

allow researchers to identify and work with groups of assembly points that are of most interest.

First section of the report produced by CAMSA focuses on comparison of each assembly to the others and presents several characteristics such as:

(i) number of *unique* assembly points (i.e., present only in one assembly);
(ii) percentage of *non-oriented* assembly points;
(iii) number of assembly points *shared* with other assemblies, with specification of particular subset of such assemblies (e.g., in Fig. 1 the assembly point $(\overrightarrow{E}, \overrightarrow{F})$ is shared by red and blue assemblies);
(iv) number of *conflicting* assembly points, i.e., scaffolds' extremities participating in different assembly points in other assemblies (e.g, in Fig. 1 the red assembly point $(\overrightarrow{A}, \overrightarrow{C})$ conflicts with the blue assembly point $(\overrightarrow{B}, \overrightarrow{C})$);
(v) proportion of assembly points that participate in the merged assembly.

Second section of the report addresses individual assembly points in the context of all input assemblies. For each assembly point P, CAMSA reports several characteristics such as:

(i) a set of *source assemblies* that contain P;
(ii) a flag specifying if P is oriented;
(iii) a set of non-source assemblies *conflicting* with P;
(iv) a subset of source assemblies that are uncertain about P (e.g., suggest alternative assembly points conflicting with P);
(v) a flag specifying if P is present in the merged assembly.

The interactive visualization of the input and merged assemblies is represented in the form of their MBG. This representation is dynamic with respect to the graph layout as well as the filtration of graph components.

Conclusions

CAMSA addresses the current deficiency of automated comparison and merging of multiple assemblies of the same scaffolds. Due to existence of various methods and techniques for scaffold assembly, identifying similarities and dissimilarities across different assemblies is beneficial both for developers of scaffold assembly algorithms and researchers improving genome assembly of specific organisms.

We remark that an alpha version of CAMSA is currently utilized in the study of Anopheles mosquito genomes, where multiple research laboratories (including ours) work on improving the existing assemblies for a number of mosquito species [5]. This project utilizes several scaffolding techniques [3, 1, 2], ranging from PacBio-based to homology-based assembly methods. CAMSA provides an automated framework for interactive comparison, analysis, and integration of constantly improving scaffold assemblies, thus helping the researchers to refine the resulting genome assemblies.

References

[1] Sergey Aganezov, Nadia Sydtnikova, AGC Consortium, and Max A. Alekseyev. Scaffold assembly based on genome rearrangement analysis. *Computational Biology and Chemistry*, 57:46–53, 2015.

[2] Yoann Anselmetti, Vincent Berry, Cedric Chauve, Annie Chateau, Eric Tannier, and Sèverine Bérard. Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*, 16(10):1–13, 2015.

[3] Lauren Assour and Scott Emrich. Multi-genome Synteny for Assembly Improvement. *Proceedings of 7th International Conference on Bioinformatics and Computational Biology*, pages 193–199, 2015.

[4] Pavel Avdeyev, Shuai Jiang, Sergey Aganezov, Fei Hu, and Max A. Alekseyev. Reconstruction of ancestral genomes in presence of gene gain and loss. *Journal of Computational Biology*, 23(3):1–15, 2016.

[5] D. E. Neafsey, R. M. Waterhouse, M. R. Abai, S. S. Aganezov, M. A. Alekseyev, et al. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522, 2015.