Critique: "Reference-assisted chromosome assembly"

Jaebum Kim, Denis M. Larkin, Qingle Cai, Asan, Yongfen Zhang,
Ri-Li Ge, Loretta Auvil, Boris Capitanu, Guojie Zhang, Harris A. Lewin
and Jian Ma

1 Abstract

One of the most difficult problems in modern genomics is the assembly of full-length chromosomes using next generation sequencing (NGS) data. Since the emergence of NGS technology, several groups have developed de novo assemblers based on NGS data, but the limitation of NGS read length makes it extremely difficult to assemble the reads into chromosomes for large genomes. This problem becomes even more challenging in the case of mammalian genomes, which contain a high fraction of repetitive elements.
To address this problem, the authors of the paper developed "reference-assisted chromosome assembly" (RACA), an algorithm to reliably order and orient sequence scaffolds generated by NGS assemblers into longer chromosomal fragments using comparative genome information and paired-end reads.

2 Critique

The algorithm described in this paper is novel and interesting. The article gives concise and clear description of the assembly process of NGS data and identifies problems of full-genome (full-chromosome) assembly. Then it describes a novel way of concatenating contigs, using information of special blocks - syntenic fragments (SF). Key idea of the proposed method is to derive SF for the genome being assembled from lots of already assembled genomes of related organisms. Using the probability model of their adjacency we can to infer the ordering of contings required for full chromosome or genome assembly. The paper includes a comprehensive evaluation section, which presentes a lot of information about results on synthetic and real data. Thus, a biologist reading this paper should have a sure sense that he gets a great tool for solving all of his problems. But, when he tries to understand how do this instrument works, he is going to be disappointed.

The first problem meets us in algorithm overview: the text includes key terms, the meaning of which is not explained. A good example of this is the notion of synteny, which must be defined clearly in order for the following sections (including evaluation) to make sense. Also the overview is very general and the algorithm itself is separated into some parts. Some mathematical explanation of the algorithm can be found only in the very end of the paper. So, if you really want to understand the heart of the matter, you should go to supplementary materials part.

Supporting information represents a 46 pages of additional material not included in the article, but necessary for understanding of many subtle details in it. We are given the formula for probability of a SF adjacency, still without the definition of syntenic fragments. There we can also see the origin of this formula and much more mathematical explanation of all the algorithm

parts. An important factor is the presence of pseudocode of SF ordering algorithm. It would seem that the most information we get from the source codes of the algorithm, distributed under a free license, but they are written on Perl and have not any documentation. Also, the algorithm‘s input must have the special format, different from SAM, so we have to convert all our data to work with it.

However, despite criticism, a great job on the analysis of the correctness of the algorithm should be noted. The represented materials impress by inclusive approach and attention to details. This gives a hope, that the method is really promising in the task of assembling full chromosomes and genomes.

Pavel Yakovlev
SPbAU RAS
27.04.2013