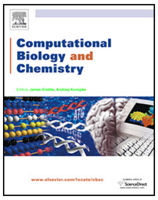
Computational Biology and Chemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Research Article

Scaffold assembly based on genome rearrangement analysis

Sergey Aganezov[b,*], Nadia Sitdykova[a], Max A. Alekseyev[b], AGC Consortium[1]

[a] Academic University, St Petersburg, Russia
[b] The George Washington University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 1 January 2015
Accepted 3 February 2015
Available online xxx

Keywords:
Scaffolding
Genome rearrangements
Breakpoint graph
MGRA
Genome assembly

ABSTRACT

Advances in DNA sequencing technology over the past decade have increased the volume of raw sequenced genomic data available for further assembly and analysis. While there exist many algorithms for assembly of sequenced genomic material, they often experience difficulties in constructing complete genomic sequences. Instead, they produce long genomic subsequences (*scaffolds*), which then become a subject to scaffold assembly aimed at reconstruction of their order along genome chromosomes. The balance between reliability and cost for scaffold assembly is not there just yet, which inspires one to seek for new approaches to address this problem. We present a new method for scaffold assembly based on the analysis of gene orders and genome rearrangements in multiple related genomes (some or even all of which may be fragmented). Evaluation of the proposed method on artificially fragmented mammalian genomes demonstrates its high reliability. We also apply our method for incomplete anophelinae genomes, which expose high fragmentation, and further validate the assembly results with referenced-based scaffolding. While the two methods demonstrate consistent results, the proposed method is able to identify more assembly points than the reference-based scaffolding.

1. Background

Genome sequencing technology has evolved over time, increasing availability of sequenced genomic data. Modern sequencers are able to identify only short subsequences (*reads*) in the supplied genomic material, which then become an input to genome assembly algorithms aimed at reconstruction of the complete genome. Such reconstruction is possible (but not guaranteed) only if each genomic region is covered by sufficiently many reads. Lack of comprehensive coverage (particularly severe in single-cell sequencing Chitsaz et al. (2011), Nikolenko et al. (2013)) and presence of long similar subsequences (*repeats*) in genomes pose major obstacles for existing assembly algorithms. They therefore often are able to reliably reconstruct only long subsequences of the genome (interspaced with low-coverage regions and repeats), called *scaffolds*.

The challenge of reconstructing a complete genomic sequence from scaffolds is known as the *scaffolds assembly* problem. It is often addressed technologically by generating so-called long-jump libraries Talkowski et al. (2012), Collins and Weissman (1984) or by using a related complete genome as a reference. Unfortunately, the technological solution may be expensive and inaccurate Hunt et al. (2014), while the reference-based approach is obfuscated with structural variations across the genomes Feuk et al. (2006).

In the current study, we assume that the constructed scaffolds are accurate and long enough to allow identification of orthologous genes. The scaffolds then can be represented as ordered sequences of genes and we pose the scaffolds assembly problem as the reconstruction of the global gene order (along genome chromosomes) from the gene sub-orders defined by the scaffolds. We view such gene sub-orders as the result of both evolutionary events and technological fragmentation in the genome. Evolutionary events that change gene orders are *genome rearrangements*, most common of which are *reversals*, *fusions*, *fissions*, and *translocations*. Technological fragmentation can be modeled by artificial “fissions” that break genomic chromosomes into scaffolds. Scaffold assembly can therefore be reduced to the search for “fusions” that revert technological “fissions” and glue scaffolds back into chromosomes. This observation inspires us to employ the genome rearrangement analysis techniques for scaffolding purposes.

Rearrangement analysis of multiple genomes relies on the concept of the breakpoint graph. While traditionally the breakpoint graph is constructed for complete genomes, it can also be constructed for fragmented genomes, where we treat scaffolds as “chromosomes”. We will demonstrate that the breakpoint graph of multiple genomes possesses an important property that its connected components are robust with respect to genome fragmentation. In other words, connected components of the

[*] Corresponding author. Tel.: +1 7039498047.
E-mail address: aganezov@gwu.edu (S. Aganezov).
[1] See Appendix A.

http://dx.doi.org/10.1016/j.compbiolchem.2015.02.005
1476-9271/© 2015 Elsevier Ltd. All rights reserved.

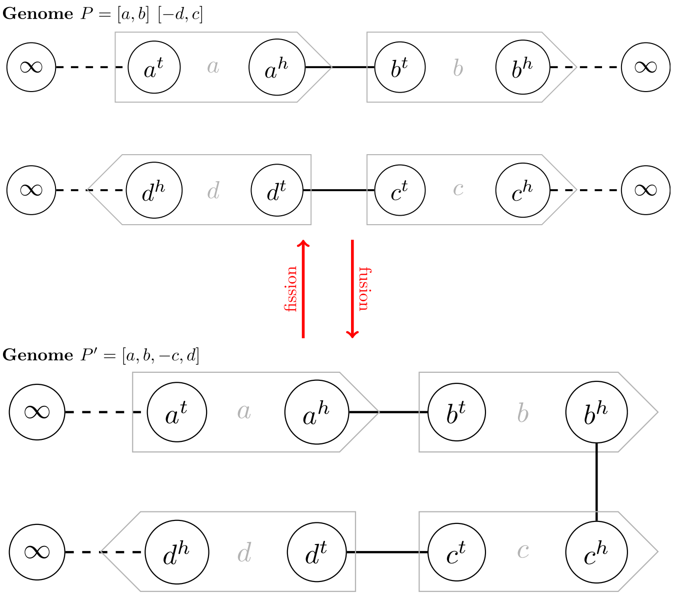


Fig. 1. Fusion/fission operations between the genome graphs of two-chromosomal genome $P = [a, b]\,[-d, c]$ and unichromosomal genome $P' = [a, b, -c, d]$, where regular and irregular edges are represented as solid and dashed, respectively. Grey boxes enclose pairs of vertices representing genes.

breakpoint graph mostly retain information about the complete genomes, even when the breakpoint graph is constructed on their scaffolds. We will show how to utilize connected components of the breakpoint graph for the scaffold assembly of fragmented genomes.

The paper is organized as follows. In Section 2, we provide background information about breakpoint graphs and genome rearrangements, discuss connected components of breakpoint graphs with respect to genome fragmentation, and describe our scaffold assembly algorithm. In Section 3, we evaluate our proposed algorithm on both simulated and real data. We summarize and discuss the paper results in Section 4.

2. Methods

2.1. Genome and breakpoint graphs

We start with defining a graph representation for a single genome, which may consist of multiple chromosomes and/or scaffolds commonly referred to as *fragments*. We represent each fragment with n genes as an undirected graph on $2 \cdot n$ *regular* vertices representing gene extremities and several *irregular* vertices, labeled by ∞, encoding fragment ends (telomeres, if a fragment is a chromosome). A gene a is represented by two regular vertices labeled as a^t and a^h denoting its *tail* and *head* extremities, respectively. Vertices corresponding to extremities of adjacent genes are connected by *regular* edges. Each vertex corresponding to a gene extremity at a fragment end is connected to an irregular vertex with an *irregular* edge. The *genome graph* of a genome is formed by the graph representing its fragments.

We remark that a single genome rearrangement affects the genome graph as follows: a pair of edges are removed and a new pair of edges on the same four vertices is created (Fig. 1). Genome rearrangements are therefore often modeled as DCJ Yancopoulos et al. (2005) or 2-break Alekseyev and Pevzner (2008) operations on graphs.

The *breakpoint graph* of k genomes composed of the same ℓ genes consists of $2 \cdot \ell$ regular vertices (representing gene extremities), a number of irregular vertices (representing fragment ends) and undirected edges of k colors (one color reserved for each of the genomes) encoding adjacencies between genes and/or fragment ends in the genomes. The breakpoint graph can be viewed as the superposition of k genome graphs of individual genomes (Fig. 2). All edges connecting a pair of vertices in the breakpoint graph form a *multiedge*, whose *multicolor* is the set of individual colors in the multiedge[3] (e.g., in Fig. 2 vertices a^h and b^t are connected by a multiedge of the red–black multicolor).

We remark that traditionally breakpoint graphs are constructed on synteny blocks whose endpoints represent *breakpoints* (thus the name) in the genomes. In contrast, we construct the breakpoint graph directly on genes, whose extremities may or may not form breakpoints. Such graph therefore can contain *trivial multiedges* formed by parallel edges of all colors, which correspond to gene adjacencies shared across all the genomes and would be hidden within synteny blocks. Clearly, each trivial multiedge in the breakpoint graph forms its own connected component, which we also call *trivial*.

2.2. Connected components and fragmentation

Under a *connected component* in the breakpoint graph we will understand any largest set of regular vertices such that any two of them are connected by a path consisting of regular edges of any colors. The connected components form a partition of the regular vertices. We will show that this partition is robust with respect to fragmentation of the genomes. Namely, we observe that the connected components of the breakpoint graph of multiple genomes are strongly connected and can be hardly broken by technological "fissions". To support this observation, we applied a number of random fissions[4] to six complete mammalian genomes and analyzed how such fissions affected the connected components of the breakpoint graph.

Using Ensembl BioMart tool Kasprzyk (2011), we obtained the following six complete mammalian genomes and pairwise orthologous gene mappings between them: *Homo sapiens* (GRCh38), *Mus musculus* (GRCm38.p2), *Rattus norvegicus* (Rnor_5.0), *Canis familiaris* (CanFam3.1), *Macaca mulatta* (MMUL_1.0), and *Pan troglodytes* (CHIMP2.1.4). From the orthologous gene mappings, we constructed gene families and filtered some of them so that each genome was represented as sequences of the same 11816 genes, each appearing in a single copy.

In order to determine how robustness of the connected components depends on the number of genomes, we analyzed different subsets of mammalian genomes of various sizes from 3 to all 6. For each subset of size ℓ, we considered all possible combinations of ℓ mammalian genomes and constructed their breakpoint graph. After the same number of random fissions[5] was applied to every genome, we computed the averaged number of nontrivial connected components in the breakpoint graphs.

The results in Fig. 3 demonstrate that as the number of genomes grows, random fissions are less likely to break connected components into smaller ones. In other words, most connected components in the breakpoint graph of fragmented genomes are likely not affected by random fissions and thus represent connected components also in the breakpoint graph of complete genomes. We will employ this robustness property of the connected components and use them to guide our scaffolding algorithm.

[3] Each regular vertex may be adjacent to at most one irregular vertex (e.g., in Fig. 2 the vertex a^t is connected to a single irregular vertex with red, black, and blue edges forming the red–black–blue mutiedge (a^t, ∞)).

[4] Unfortunately we do not have information about the actual mechanism of fragmentation of a genome into scaffolds.

[5] A random fission operation uniformly selects a regular edge in the genome graph performs a fission on this edge (Fig. 1).