Earmark graph approach to de novo genome assembly

Max Alekseyev **Mikhail E. Dvorkin** **Alexander S. Kulikov**

Algorithmic Biology Lab at St. Petersburg Academic University of the Russian Academy of Sciences
http://bioinf.spbau.ru/en/

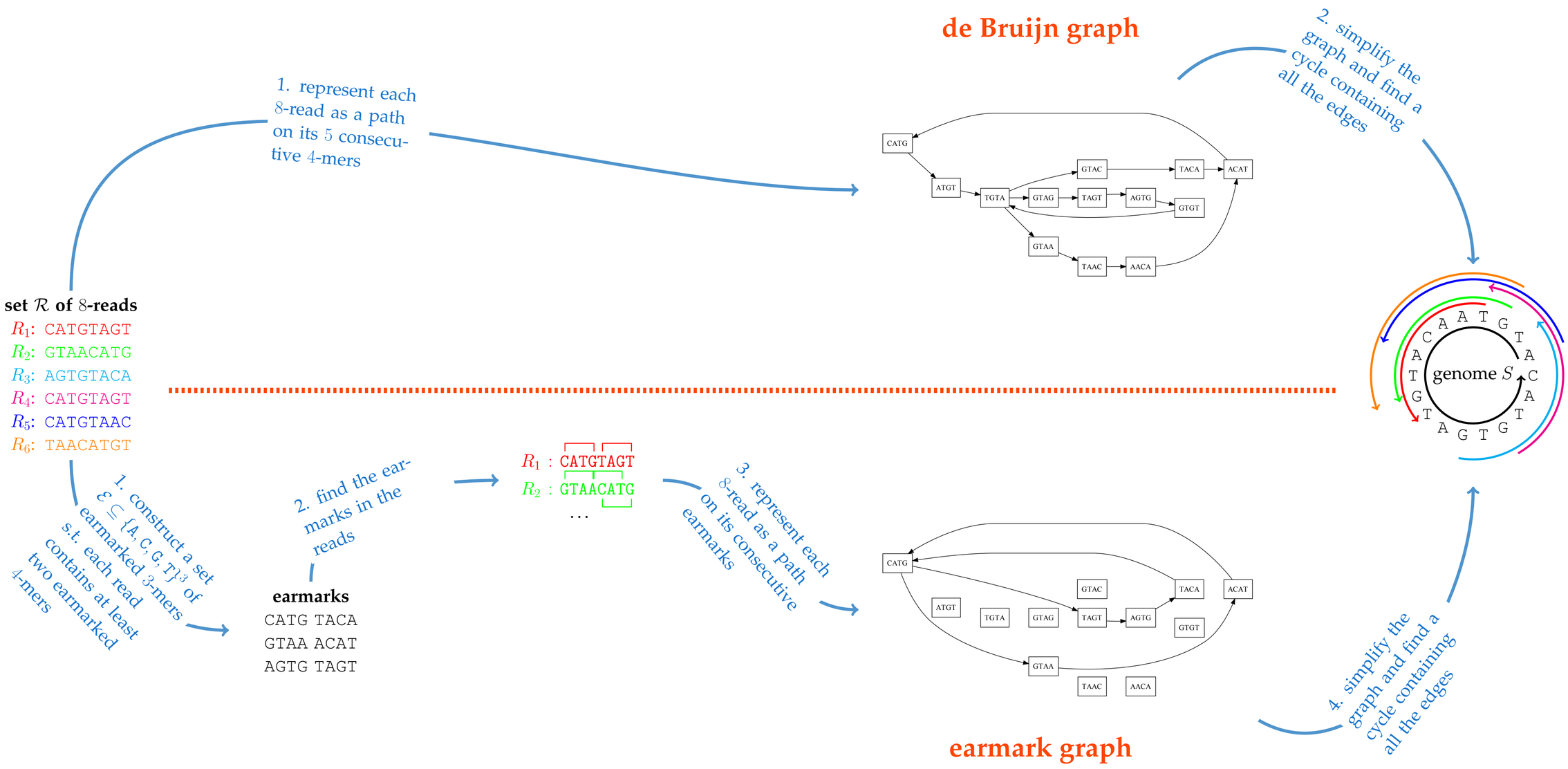
Abstract

A common approach to assembling a genome from short reads is constructing the de Bruijn graph on all k-mers from the given set of reads and finding a traversal of edges in this graph. We propose a new approach that allows to decrease the graph size without losing the essential information from the input data. Instead of using all the k-mers from a read we take only a few of them (and call them earmarked). Besides an obvious advantage of requiring less memory and time for constructing, the resulting earmark graph has several other advantages over the de Bruijn graph. We discuss them in the paper and also present some experimental results.

Typical genome assembly setting

Input: a set of substrings (called reads) $\mathcal{R} \subseteq \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^r$ of an unknown circular string $S \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^*$ (called genome).
Output: the genome S.
Complications: blah-blah-blah...

Standard and new approach: example



Advantages of earmark graph over de Bruijn graph: to be rewritten

Smaller size. The earmark graph requires less time and memory for construction. Note also that one can control the size of the earmark graph by varying the size of the set of earmarks (e.g., by varying the value of the parameter t of Algorithm **??**).

Some of short repeats are already resolved. To give an example, consider two reads TTGCAC and ATGCAT sharing a 4-mer TGCA. They are represented as two paths, shown in Fig. **??**, in the de Bruijn graph built on 3-mers (this is a part of the de Bruijn graph from Fig. **??**). This is a typical repeat. The edge TGC → GCA has two incoming edges and two outgoing edges. While spelling a genome lear which incoming edge corresponds to which outgoing edge. However in the earmarked ved if the 3-mers TGC and GCA are not earmarked, see Fig. **??**.

Practical results: to be rewritten

Less errors. Erroneous k-mers can be excluded from the graph already on the construction stage.

		de Bruijn	earmarked
first 10% of E.coli	# vertices	809730	62420
	# edges	810354	62900
+ compression	# vertices	5020	1376
	# edges	5644	1856
+ tips clipping	# vertices	1310	818
	# edges	1934	1298
+ bulge removal	# vertices	964	472
	# edges	1410	750
+ erroneous connection removal	# vertices	398	212
	# edges	570	314
+ tips clipping	# vertices	360	172
	# edges	532	274
+ bulge removal	# vertices	242	106
	# edges	354	166