Earmark graph approach to de novo genome assembly

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

Algorithmic Biology Lab at St. Petersburg Academic University of the Russian Academy of Sciences

http://bioinf.spbau.ru/en/

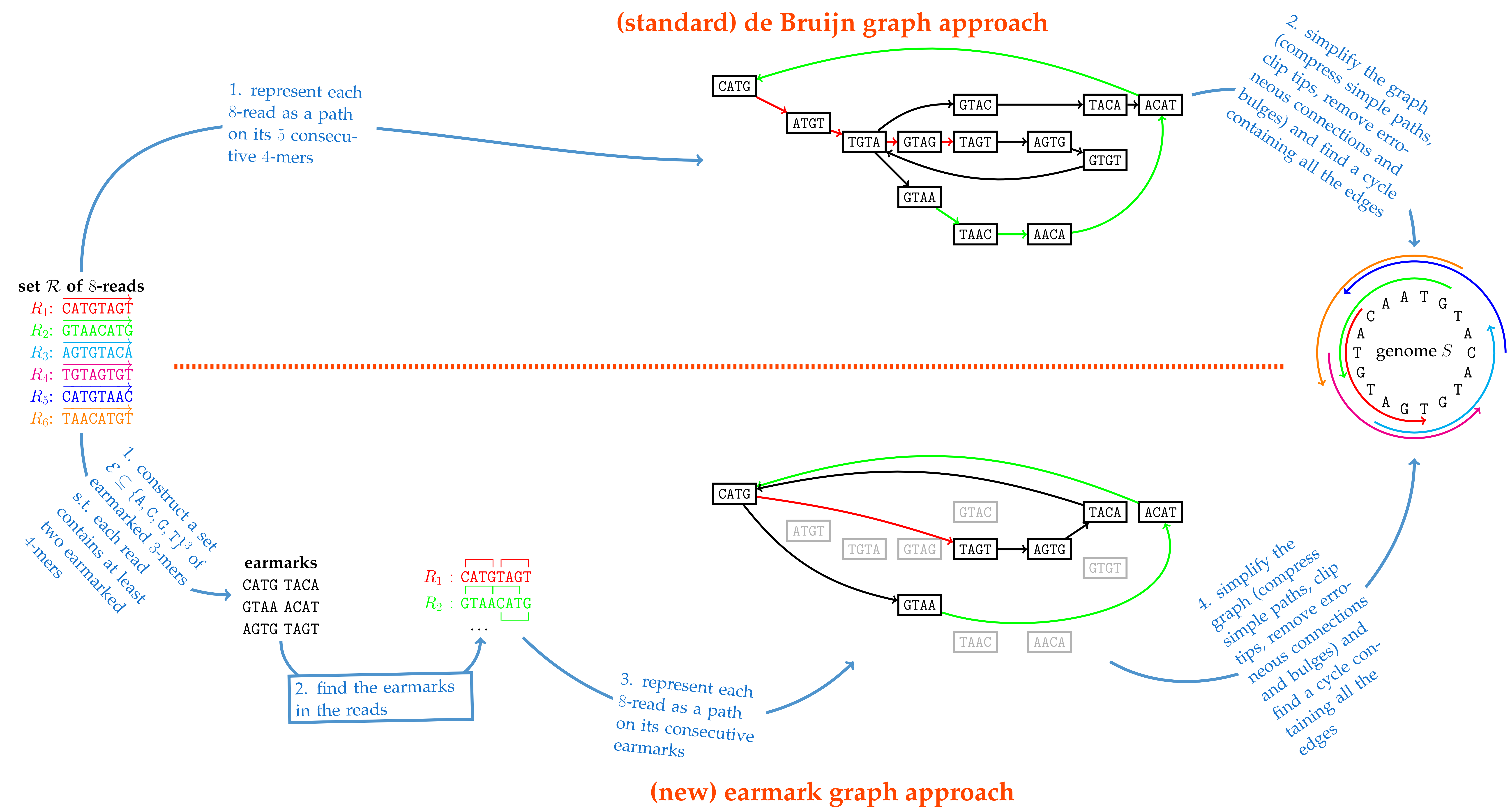
Abstract

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.

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: reads may contain errors.

Standard and new approach: example



(standard) de Bruijn graph approach

1. represent each 8-read as a path on its 5 consecutive 4-mers

2. simplify the graph (compress simple paths, clip tips, remove erroneous connections and bulges) and find a cycle containing all the edges

set $\mathcal{R}$ of 8-reads

R_1: $\overrightarrow{\mathtt{CATGTAGT}}$
R_2: $\overrightarrow{\mathtt{GTAACATG}}$
R_3: $\overrightarrow{\mathtt{AGTGTACA}}$
R_4: $\overrightarrow{\mathtt{TGTAGTGT}}$
R_5: $\overrightarrow{\mathtt{CATGTAAC}}$
R_6: $\overrightarrow{\mathtt{TAACATGT}}$

1. construct a set $\mathcal{E} \subseteq \{A, C, G, T\}^3$ of earmarked 3-mers s.t. each read contains at least two earmarked 4-mers

earmarks
CATG TACA
GTAA ACAT
AGTG TAGT

R_1 : CATGTAGT
R_2 : GTAACATG
…

2. find the earmarks in the reads

3. represent each 8-read as a path on its consecutive earmarks

4. simplify the graph (compress simple paths, clip tips, remove erroneous connections and bulges) and find a cycle containing all the edges

(new) earmark graph approach

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. Also, it is easy to see that in an extreme case when $\mathcal{E}$ is just the set of all the k-mers of the input reads, the earmark graph coincides with the de Bruijn graph.

Some of short repeats are already resolved. Blah-blah-blah…

Using trusted info. While constructing the set of earmarks $\mathcal{E}$, one can use an information about trusted k-mers (if available) to get a more accurate earmark graph.

Practical results

TO BE UPDATED

		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