

# Earmark graph approach to *de novo* genome assembly

Mikhail E. Dvorkin

Alexander S. Kulikov

Max A. Alekseyev

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

The presentation of this work was made possible by a Travel Fellowship awarded by ISCB with grant funds obtained from the Department of Energy Office of Science, the National Science Foundation Bio-Directorate, and the NIH National Institute of General Medical Sciences.

## 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)  $\mathcal{R} \subseteq \{A, C, G, T\}^r$  of an unknown circular string  $S \in \{A, C, G, T\}^*$  (called genome).

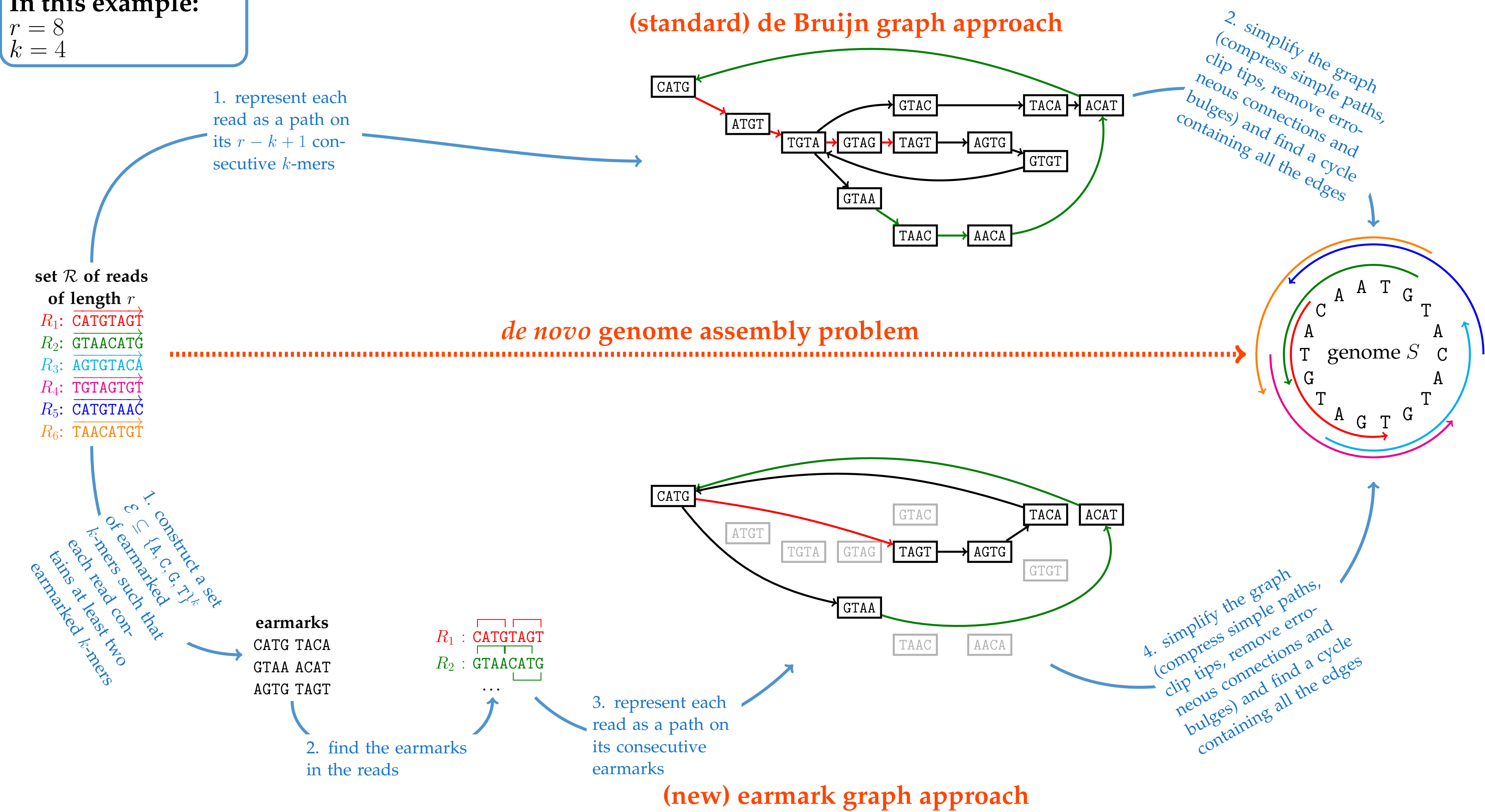
**Output:** the genome  $S$ .

**Complications:** reads may contain errors.

## Standard and new approaches: an example

In this example:

$r = 8$   
 $k = 4$



## Advantages of earmark graph over de Bruijn graph

**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.

**Simpler structure.** In the example above, the de Bruijn graph contains two vertices  $v$  with  $\text{indegree}(v) \times \text{outdegree}(v) \geq 2$ , while the earmark graph contains only one such vertex. This corresponds to a simpler representation of repeats in the genome and simplifies the problem of finding a cycle in the graph.

**Using trusted  $k$ -mers.** By restricting earmarks to  $k$ -mers present in many reads ("trusted"  $k$ -mers) one can reduce the number of erroneous edges resulting from sequencing errors in the reads.

## Practical results

| TO BE UPDATED                  |            |               |           |
|--------------------------------|------------|---------------|-----------|
|                                |            | de Bruijn     | earmarked |
| E.coli genome                  | # vertices | appr. 8097300 | 694592    |
|                                | # edges    | appr. 8103540 | 698170    |
| + compression                  | # vertices | appr. 50200   | 7862      |
|                                | # edges    | appr. 56440   | 11440     |
| + tips clipping                | # vertices | appr. 13100   | 4142      |
|                                | # edges    | appr. 19340   | 7720      |
| + bulge removal                | # vertices | appr. 9640    | 3030      |
|                                | # edges    | appr. 14100   | 5862      |
| + erroneous connection removal | # vertices | appr. 3980    | 1350      |
|                                | # edges    | appr. 5700    | 2514      |
| + tips clipping                | # vertices | appr. 3600    | 1304      |
|                                | # edges    | appr. 5320    | 2468      |
| + bulge removal                | # vertices | appr. 2420    | 1090      |
|                                | # edges    | appr. 3540    | 2080      |