Earmark graph approach to de novo genome assembly

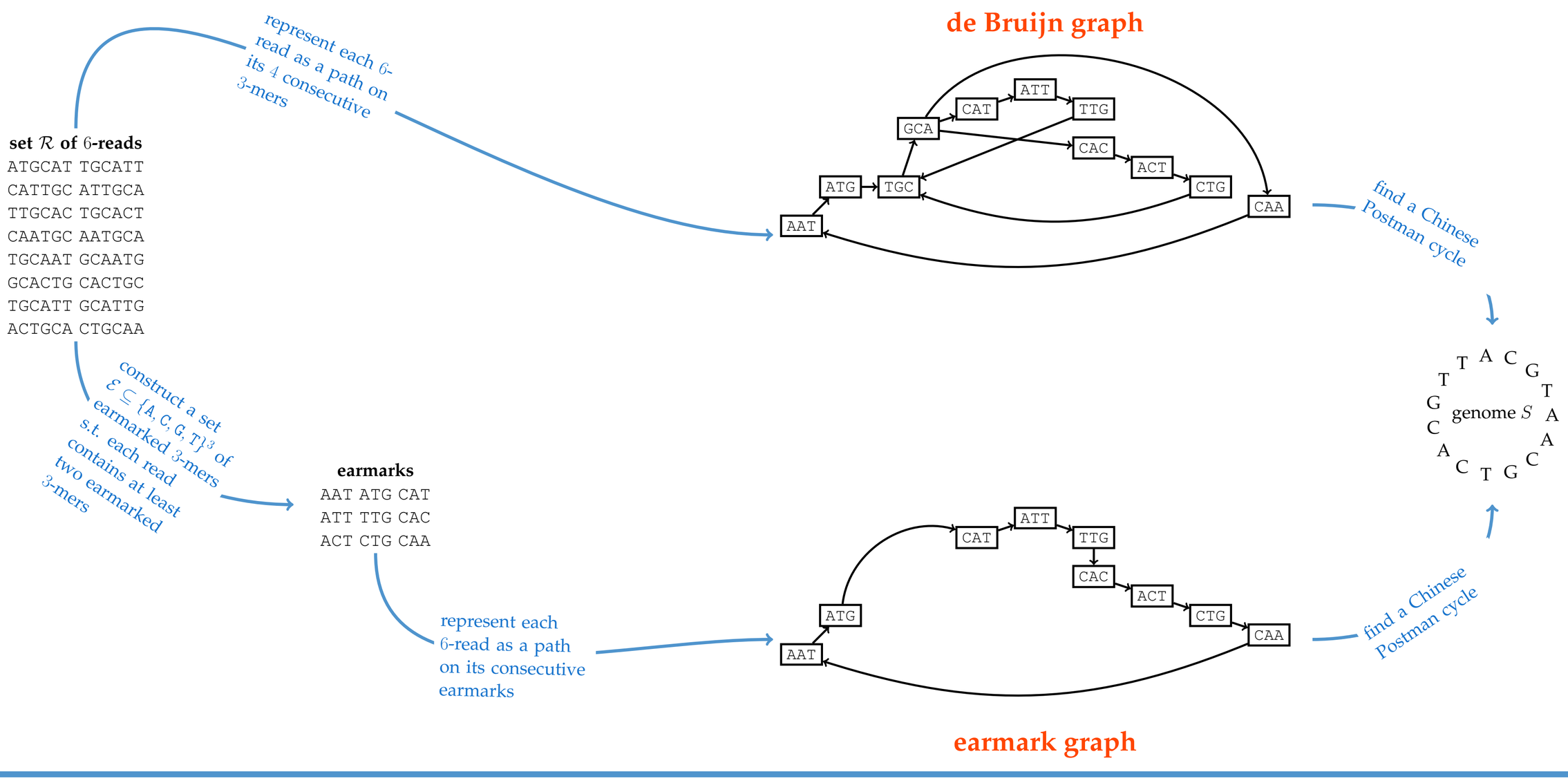
Mikhail E. Dvorkin
Algorithmic Biology Lab, Academic University, RAS
http://bioinf.spbau.ru/members/mikhail-dvorkin/

Alexander S. Kulikov
Algorithmic Biology Lab, Academic University, RAS and
Steklov Mathematical Institute at St. Petersburg, RAS
http://logic.pdmi.ras.ru/~kulikov/

Typical genome assembly setting

Given a set of reads $\mathcal{R} \subseteq \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^r$ of a genome $S \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^*$, find the genome S.

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 $\to$ GCA has two incoming edges and two outgoing edges. While spelling a genome through this part of the graph it is not clear which incoming edge corresponds to which outgoing edge. However in the earmarked graph this repeat may be already resolved if the 3-mers TGC and GCA are not earmarked, see Fig. **??**.

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

Practical results: to be rewritten

		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