Safe and complete genome assembly via omnitigs

Alexandru Tomescu
Department of Computer Science
University of Helsinki, Finland

1st Summer School on Bioinformatics Data Structures
August 9, 2016

This lecture was part of the 1st Summer School on Bioinformatics Data Structures, funded by BIRDS project (www.birdsproject.eu)
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690941

Central dogma of biology

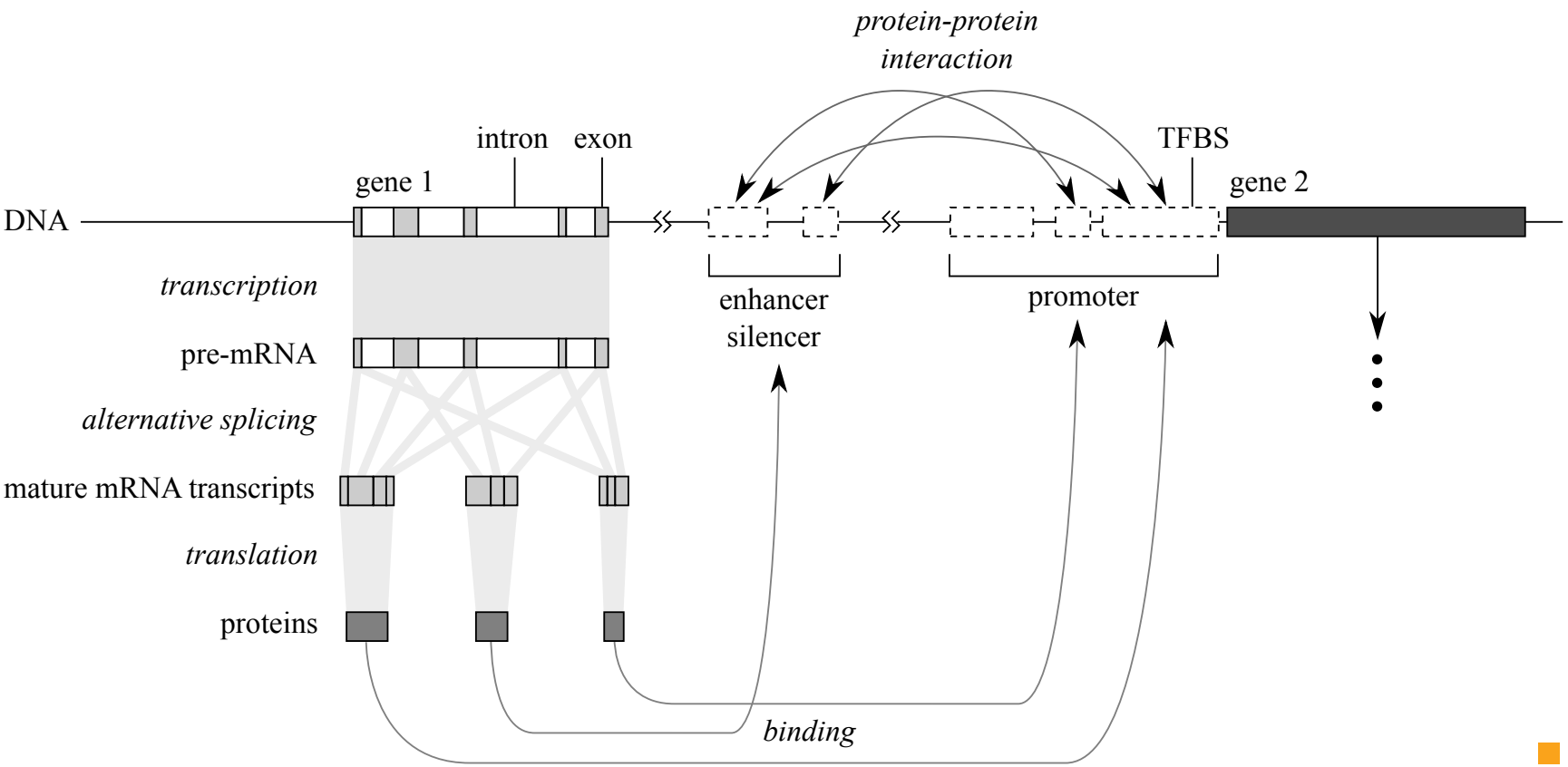


Image taken from
Genome-Scale Algorithm Design, Cambridge University Press, 2015

Sequencing atlas

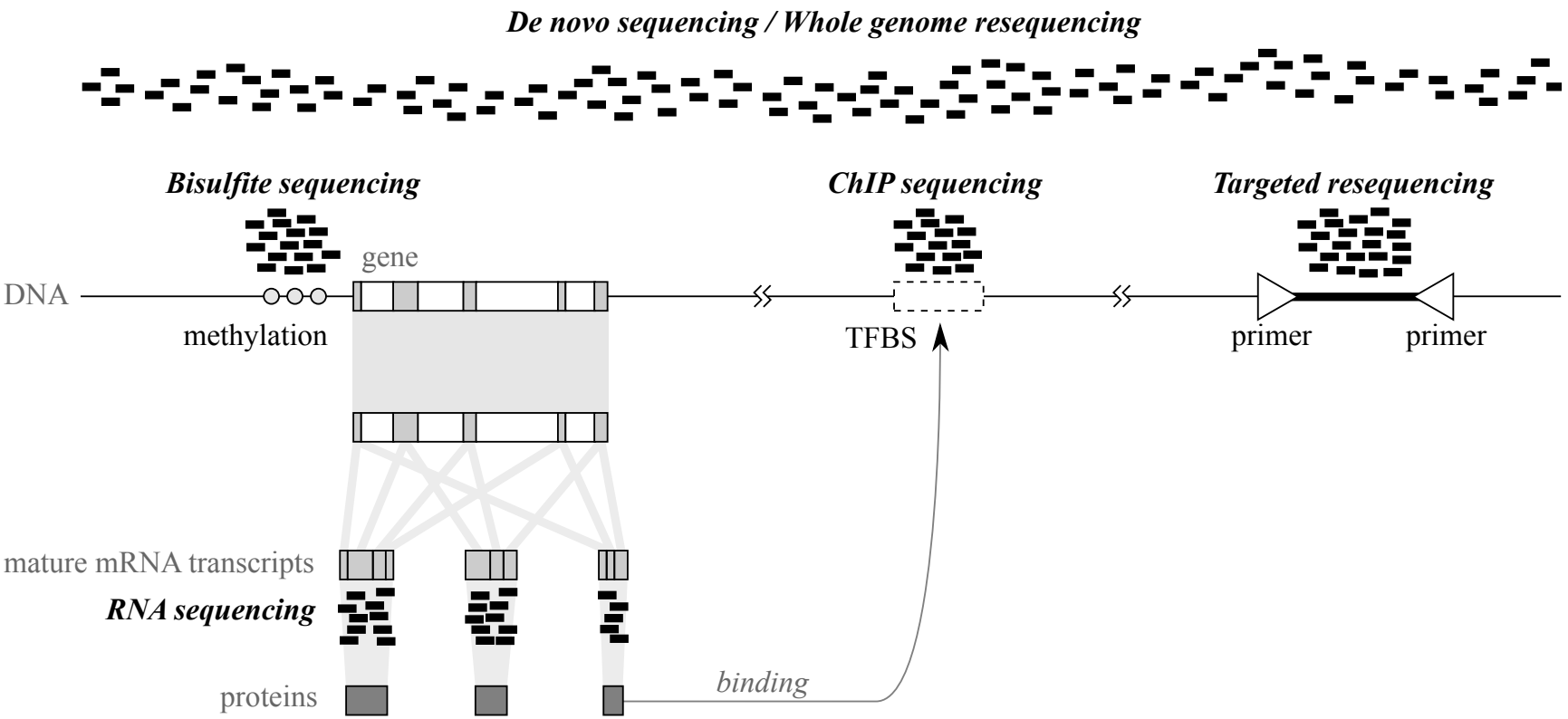


Image taken from
Genome-Scale Algorithm Design, Cambridge University Press, 2015

BLACKBOARD

The "safe and complete" framework is described in:

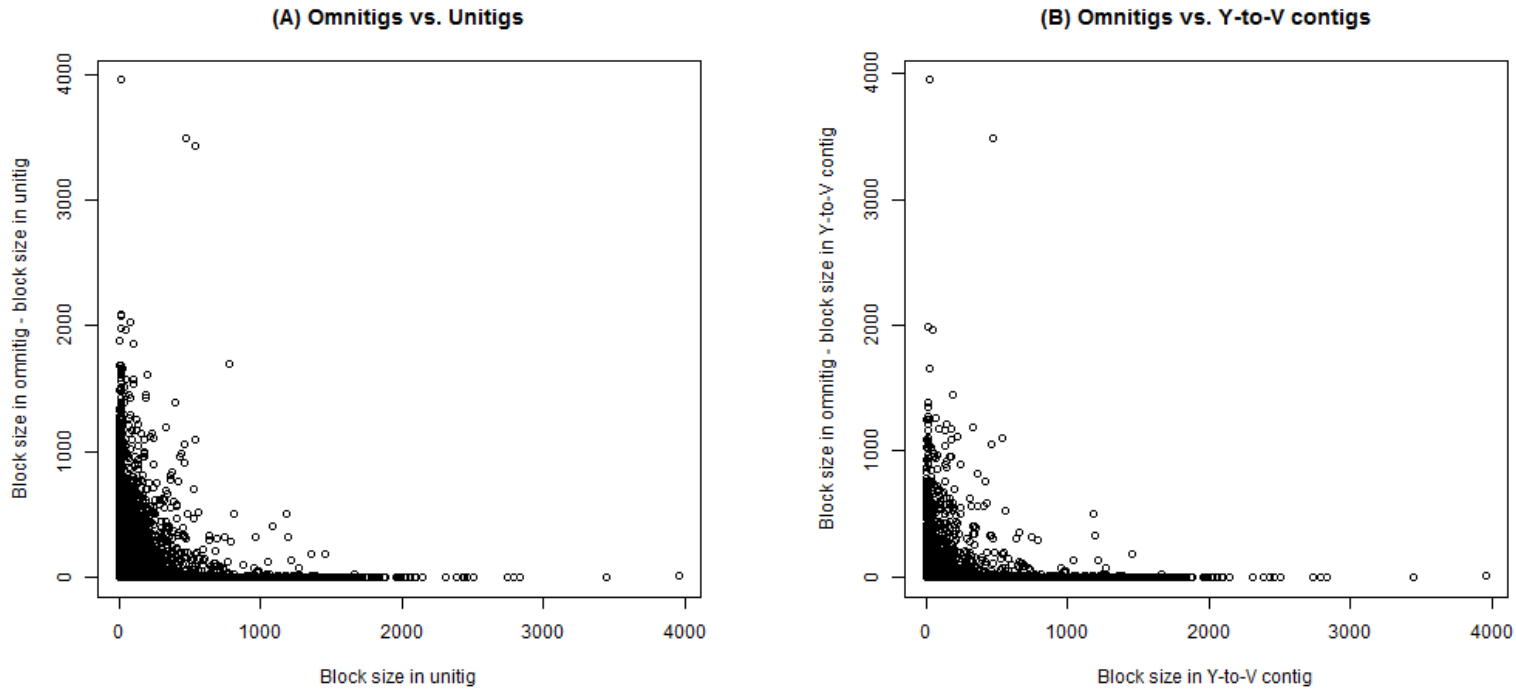
- A. I. Tomescu, P. Medvedev, *Safe and complete contig assembly via omnitigs* RECOMB 2016 - 20th Annual International Conference on Research in Computational Molecular Biology, LNCS 9649, 152-163, 2016.
  Extended version available at `http://arxiv.org/abs/1601.02932`

Experimental results

Table: Results for $DB^k_{ec}(R)$, where R is the set of all $(k+1)$-mers of the genome.

	E.coli ($k = 31$)				chr10 ($k = 55$)			
	# strings	avg len	E-size	time (s)	# strings	avg len	E-size	time (s)
unitigs	1,743	2,654	33,309	< 1	259,845	546	8,344	1
Y-to-V	1,004	4,682	33,632	< 1	159,101	878	8,376	2
omnitigs	983	4,832	34,557	< 1	158,236	887	8,401	1,046

Experimental results



Compared to unitigs, #SNPs whose block size

- increased: ~ 1.7 million (out of ~ 5.9 million)
- increased by more than 10: $\sim 137,000$

Compared to Y-to-V contigs, #SNPs whose block size

- increased: $\sim 266,000$ (out of ~ 5.9 million)