HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA
MATEMATISK-NATURVETENSKAPLIGA FAKULTETEN
FACULTY OF SCIENCE

Indexing Finite Language Representation of Population Genotypes

Jouni Sirén, Niko Välimäki, Veli Mäkinen

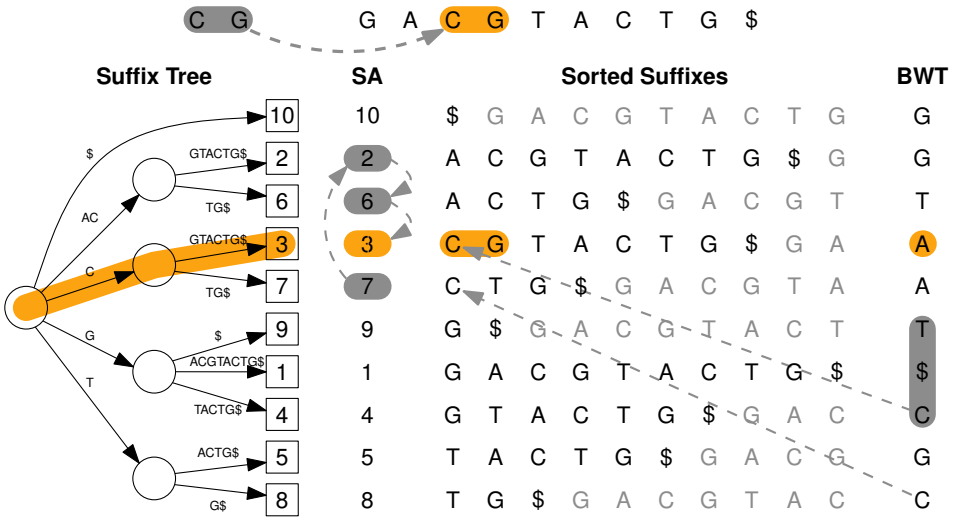
ABSTRACT

Compressed full-text indexes [6] based on the *Burrows-Wheeler transform (BWT)* are widely used in bioinformatics. Their most succesful application so far has been mapping short reads to a reference sequence (e.g. Bowtie [3], BWA [4], SOAP2 [5]). These indexes use the BWT to simulate the *suffix tree* or the *suffix array (SA)*, while using much less space than either of them. A simple generalization allows indexing a set of sequences.

We propose a biologically motivated generalization of the BWT to finite languages. Given a multiple alignment of sequences (e.g. individual genomes), we build a compressed index capable of simulating the suffix array over plausible recombinations of the sequences. Alternatively, we start from a reference sequence and a set of mutations, and build the index over sequences containing any subset of the mutations.

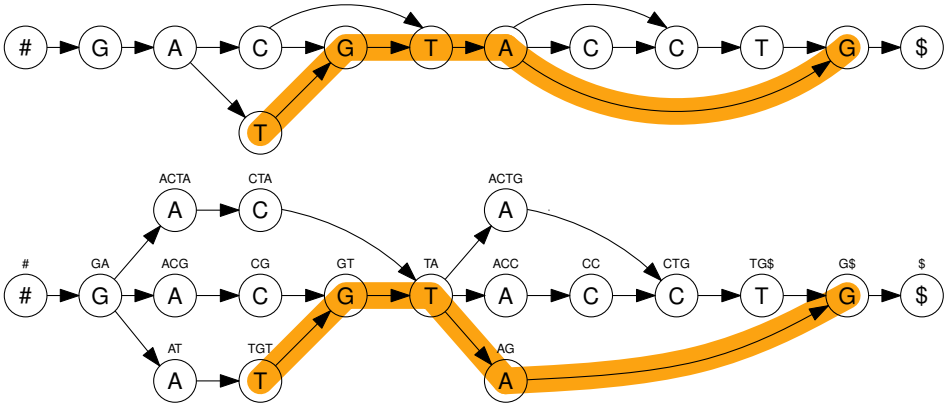
Our approach is based on finite automata. We start with an automaton recognizing the input language. This automaton is transformed into an equivalent automaton, where each state corresponds to a lexicographic range of suffixes of the language. A generalization of the XBW transform for labeled trees [2] is used to index the transformed automaton.

FULL-TEXT INDEXES FOR PATTERN MATCHING AND SEQUENCE ANALYSIS



A MATCH IN MULTIPLE ALIGNMENT

INITIAL AUTOMATON AND SORTED AUTOMATON



GENERALIZED COMPRESSED SUFFIX ARRAY

	$	ACC	ACG	ACTA	ACTG	AG	AT	CC	CG	CTA	CTG	G$	GA	GT	TA	TG$	TGT	#
BWT	G	T	G	G	T	T	G	A	A	A	AC	AT	#	CT	CG	C	A	$
Edges	1	1	1	1	1	1	1	1	1	1	1	1	100	1	100	1	1	1

Basic operations are about 2 times slower than in regular BWT-based indexes. For reasonable mutation frequencies f, the expected size of the sorted automaton is $n(1+f)^{O(\log n)}$, where n is the length of the reference sequence. For $1/f = \Omega(\log n)$, this becomes $O(n)$. In our experiments, an index built for the human reference genome and the genetic variation found in the Finnish population sample of the *1000 Genomes Project* took approximately 2.8 gigabytes.

FUTURE DIRECTIONS

- With our current algorithm, the construction of a genome-scale index requires 12 hours and 192 gigabytes of memory. We are currently investigating other algorithms, such as external memory construction and distributed construction in the MapReduce framework [1].
- In principle, our index can be used in any algorithm using a regular BWT-based index. What can be done efficiently in practice?
- We are currently investigating several ways to use the generalized index in read alignment. Are there other applications, where our index could be superior to the existing approaches?

REFERENCES

[1] J. Dean, S. Ghemawat: *Simplified Data Processing on Large Clusters*. OSDI 2004.

[2] P. Ferragina et al.: *Compressing and indexing labeled trees, with applications*. Journal of the ACM, 2009.

[3] B. Langmead et al.: *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology, 2009.

[4] H. Li, R. Durbin: *Fast and accurate short read alignment with Burrows-Wheeler Transform*. Bioinformatics, 2009.

[5] R. Li et al.: *SOAP2: an improved ultrafast tool for short read alignment*. Bioinformatics, 2009.

[6] G. Navarro, V. Mäkinen: *Compressed full-text indexes*. ACM Computing Surveys, 2007.