

25 Years of the Burrows-Wheeler Transform

Edited by

Travis Gagie¹, Giovanni Manzini², Gonzalo Navarro³, and
Jens Stoye⁴

¹ Dalhousie University – Halifax, CA, travis.gagie@dal.ca

² University of Eastern Piedmont – Alessandria, IT, giovanni.manzini@uniupo.it

³ University of Chile – Santiago de Chile, CL, gnavarro@dcc.uchile.cl

⁴ Universität Bielefeld, DE, jens.stoye@uni-bielefeld.de

Abstract

Dagstuhl Seminar 19241 (“25 Years of the Burrows-Wheeler Transform”) took place from June 10th to 14th, 2019, and was attended by 45 people from 13 countries and the three fields of Algorithms and Data Structures, Bioinformatics, and Combinatorics on Words. There were four talks and a panel session for each field. Feedback was generally positive and we are confident the seminar fostered interdisciplinary connections and will eventually result in noteworthy joint publications.

Seminar June 10–14, 2019 – <http://www.dagstuhl.de/19241>

2012 ACM Subject Classification Applied computing → Bioinformatics, Applied computing → Computational genomics, Applied computing → Molecular sequence analysis, Applied computing → Genomics, Mathematics of computing → Combinatorics on words, Theory of computation → Data compression, Theory of computation → Pattern matching, Theory of computation → Sorting and searching

Keywords and phrases Bioinformatics, Burrows-Wheeler Transform, Combinatorics on Words, Data Compression, Data Structures, Indexing, Sequence Alignment

Digital Object Identifier 10.4230/DagRep.9.6.55

1 Executive Summary

Travis Gagie

Giovanni Manzini

Gonzalo Navarro

Jens Stoye

License © Creative Commons BY 3.0 Unported license

© Travis Gagie, Giovanni Manzini, Gonzalo Navarro, and Jens Stoye

Dagstuhl Seminar 19241 marked the 25th anniversary of the publication of the Burrows-Wheeler Transform (BWT), which has had a huge impact on the fields of data compression, combinatorics on words, compact data structures, and bioinformatics. The 10th anniversary in 2004 was marked by a workshop at the DIMACS Center at Rutgers (<http://archive.dimacs.rutgers.edu/Workshops/BWT>) organized by Paolo Ferragina, Giovanni and S. Muthukrishnan, and it is exciting to see how far we have come. In the past 15 years, interest in the BWT has shifted from data compression to compact data structures and bioinformatics, particularly indexing for DNA read alignment, but seven of the 33 participants of that workshop (including Giovanni) also attended this seminar. Unfortunately, Professor Gørtz fell ill at the last minute and emailed us on June 11th to say she couldn’t attend, but everyone else on the



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

25 Years of the Burrows-Wheeler Transform, *Dagstuhl Reports*, Vol. 9, Issue 6, pp. 55–68

Editors: Travis Gagie, Giovanni Manzini, Gonzalo Navarro, and Jens Stoye

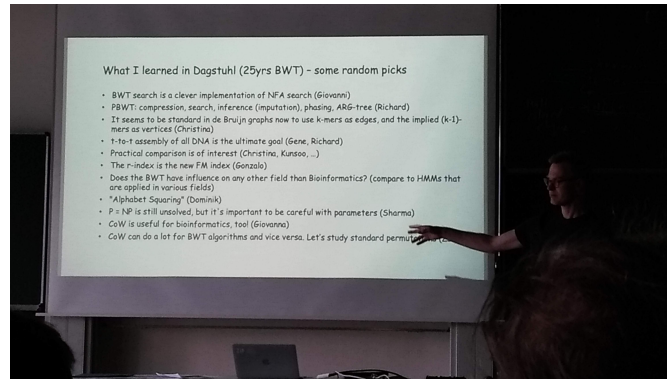


Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** Gonzalo with his birthday cake (featuring a BWT).



■ **Figure 2** Jens reviewing some points raised during the seminar.

final list of invitees was present for at least some of the seminar (although not everyone made it into the photo). In total there were 45 people (listed at the end of this report) from 13 countries, including ten women, six junior researchers and two researchers from industry. By happy coincidence, the seminar started the day after Gonzalo's 50th birthday, so we were able to celebrate that as well. We thank Professor Sadakane for the photos shown in Figures 1 and 2.

The schedule, shown in Figure 3, featured an introduction, 12 talks, three panel sessions and a closing. The talks were all timely and reflected the active and dynamic research being carried out on the BWT:

- Giovanni's introduction was a more in-depth version of his invited talk from DCC '19;
- Veli Mäkinen surveyed pan-genomic indexing, including work published in *BMC Genomics* last year;
- Richard Durbin surveyed results based on the positional BWT, published in *Bioinformatics* in 2014;
- Jouni Sirén presented work included in a *Nature Biotechnology* article last year;
- Christina Boucher surveyed compact data structures for de Bruijn graphs, including work from an ISMB/ECCB 2019 paper;
- Gonzalo Navarro reviewed BWT-based indexes, including work from a SODA '18 paper;
- Sandip Sinha presented work from a STOC '19 paper;
- Dominik Kempa presented work from another STOC '19 paper;
- Sharma Thankachan presented work from an ESA '19 paper;
- Nicola Prezza presented work from a STOC '18 paper;
- Marinella Sciortino gave a version of her invited lecture for IWOCA '19 a month later;
- Giovanna Rosone presented results about two extensions of the BWT, including work from a WABI '18 paper, now published in *Algorithms for Molecular Biology*;
- Dominik Köppl presented work from a CPM '19 paper.

We later received all the abstracts but one.

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	
07:30		BREAKFAST	BREAKFAST	BREAKFAST	BREAKFAST	
09:00		INTRO	ALG TALK 1	CoW TALK 1	WORK...	
09:45		BIO TALK 1	ALG TALK 2	CoW TALK 2		
10:30		BIO TALK 2	ALG TALK 3	CoW TALK 3		
11:15		BIO TALK 3	ALG TALK 4	CoW TALK 4		
12:15		LUNCH	LUNCH	LUNCH	LUNCH	
13:45		BIO TALK 4		CoW PANEL		
14:00		BIO PANEL	ALG PANEL			
14:30			WORK!	CLOSING		
15:00						
15:30	CAKE	CAKE	CAKE	CAKE		
16:00	WORK?	WORK	WORK!!	WORK!!!		
18:00	DINNER (buffet)	DINNER	DINNER	DINNER		
20:00	CHEESE?	CHEESE	CHEESE	CHEESE		
INTRO	Giovanni			BIO PANEL	ALG PANEL	CoW PANEL
BIO TALK 1	Veli	(Pan-genomic) alignment		Ben	Ian	Gabriele
BIO TALK 2	Richard	PBWT		Gene	Inge (chair)	Hideo
BIO TALK 3	Jouni	GBWT		Knut	Johannes	Jackie
BIO TALK 4	Christina	de Bruijn graphs		Kunsoo	Rahul	Pawel
ALG TALK 1	Gonzalo	r-index		Paola	Roberto	Sabrina (chair)
ALG TALK 2	Sandip	Local decodability		Richard	Simon G	Tomasz
ALG TALK 3	Dominik	BWT construction		Tony (chair)		Zsuzsa
ALG TALK 4	Sharma	Wheeler graphs				
CoW TALK 1	Nicola	String attractors		Jens chairs BIO talks		
CoW TALK 2	Marinella	Combinatorial properties		Giovanni chairs ALG talks		
CoW TALK 3	Giovanna	eBWT / BWT similarity		Travis chairs CoW talks		
CoW TALK 4	Dominik	Bijjective BWT				
CLOSING	Jens					

■ **Figure 3** The original seminar schedule. Inge Li Gørtz was unable to attend and so Tatiana Starikovskaya chaired the *Algorithms and Data Structures* panel. The talks and panel on *Bioinformatics* were held on the first day and those on *Algorithms and Data Structures* on the second day to accommodate participants' schedules.

2 Table of Contents

Executive Summary

Travis Gagie, Giovanni Manzini, Gonzalo Navarro, and Jens Stoye 55

Overview of Talks

25 Years of Burrows-Wheeler Transform: A review <i>Giovanni Manzini</i>	59
Scaling pan-genomic alignment using founders <i>Veli Mäkinen</i>	60
Genome Graphs and BWT-based Data Structures <i>Jouni Sirén</i>	60
BWT Meets the de Bruijn Graph: Results and Challenges <i>Christina Boucher</i>	61
Text Indexing with the BWT <i>Gonzalo Navarro</i>	61
Local Decodability of the Burrows-Wheeler Transform <i>Sandip Sinha</i>	61
BWT Construction: History, Techniques, State of the Art, Open Problems <i>Dominik Kempa</i>	62
On the Hardness and Inapproximability of Recognizing Wheeler Graphs <i>Sharma V. Thankachan</i>	63
String Attractors <i>Nicola Prezza, Travis Gagie, Dominik Kempa, and Gonzalo Navarro</i>	64
Combinatorial Properties of BWT <i>Marinella Sciortino</i>	64
BWT / eBWT similarity <i>Giovanna Rosone</i>	65
Searching Patterns in the Bijective BWT <i>Dominik Köppl</i>	65
Motivation (from proposal)	66
Feedback	66
Open Problems	67
Participants	68

3 Overview of Talks

Apart from the talks below, there were impromptu presentations by Jackie Daykin on “Order-based Burrows-Wheeler Transforms”, by Enno Ohlebusch on “An improved encoding of genetic variation in a Burrows-Wheeler transform”, by Kunsoo Park on “Comparing Pan-Genomic Indexes”. Slides from all of these and introductory slides from the Combinatorics on Words panel (submitted by Sabrina Mantaci) are available on the materials page.

3.1 25 Years of Burrows-Wheeler Transform: A review

Giovanni Manzini (University of Eastern Piedmont – Alessandria, IT)

License © Creative Commons BY 3.0 Unported license
 © Giovanni Manzini
Joint work of Giovanni Manzini, Paolo Ferragina, Travis Gagie, Raffaele Giancarlo, Marinella Sciortino, Jouni Sirén
Main reference Travis Gagie, Giovanni Manzini, Jouni Sirén: “Wheeler graphs: A framework for BWT-based data structures”, *Theor. Comput. Sci.*, Vol. 698, pp. 67–78, 2017.
URL <https://doi.org/10.1016/j.tcs.2017.06.016>

To establish a common ground, in this introductory talk we review the main properties of the BWT with respect to data compression and text indexing.

The “Block sorting data compression algorithm” by Mike Burrows and David Wheeler [1] was based on a data transformation, now called the BWT, designed “to make redundancy in the input more accessible”. While this is obvious at the intuitive level, it took ten years to formalize this notion in terms of empirical entropy. In [2] it was proven that the Burrows-Wheeler transform can be seen as a “compression booster”, that is a tool for transforming any order-0 encoder into a much more effective order- k encoder, and that this result holds simultaneously for every $k > 0$.

Starting from the year 2000, several researchers [3, 5, 6] observed that, because of the relationship between the BWT and the Suffix Array, the former can be used as a sort of full text index, possibly compressed if one takes advantage of the “boosting” properties of the BWT. Over the years, these ideas have been extended to design compressed data structures to index other discrete structures such as trees, graphs, automata, alignments, and so on.

We introduced the notion of Wheeler Graph that generalizes the BWT and provides a unified view of many of these extensions [4]. We show that pattern matching problems inside many of these discrete structures can be modeled using Nondeterministic Finite Automata which have the additional property of being Wheeler Graphs. We also show that we can compactly represent and navigate Wheeler Graphs using the well-known and highly optimized rank and select operations on linear arrays. Although not every BWT-related data structure fits in our framework, we believe our unifying view can help researchers develop new BWT variants and new indexing data structures.

References

- 1 M. Burrows, D. Wheeler, A block-sorting lossless data compression algorithm, Tech. Rep. 124, Digital Equipment Corporation (1994).
- 2 P. Ferragina, R. Giancarlo, G. Manzini, M. Sciortino, Boosting textual compression in optimal linear time, *J. ACM* 52 (2005) 688–713.
- 3 P. Ferragina, G. Manzini, Opportunistic data structures with applications, in: *Proc. 41st IEEE Symp. on Found. of Computer Science*, 2000, pp. 390–398.

- 4 T. Gagie, G. Manzini, J. Sirén, Wheeler graphs: A framework for BWT-based data structures, *Theor. Comput. Sci.* 698 (2017) 67–78.
- 5 R. Grossi, J. S. Vitter, Compressed suffix arrays and suffix trees with applications to text indexing and string matching, in: *Proc. of the 32nd ACM Symposium on Theory of Computing*, 2000, pp. 397–406.
- 6 V. Mäkinen, Compact suffix array, in: *Proc. of the 11th Symposium on Combinatorial Pattern Matching*, Springer-Verlag LNCS n. 1848, 2000, pp. 305–319.

3.2 Scaling pan-genomic alignment using founders

Veli Mäkinen (*University of Helsinki, FI*)

License  Creative Commons BY 3.0 Unported license
© Veli Mäkinen


The talk covers the PanVC framework for pan-genomic variant calling through multiple reference indexing for read alignment [1], as well as a lossy compression of multiple references into founders [2], which makes the whole framework scalable.

References

- 1 D. Valenzuela, Tuukka Norri, Niko Välimäki, Esa Pitkänen and Veli Mäkinen. Towards pan-genome read alignment to improve variation calling. *BMC Genomics*, Vol. 19, No. 87, 2018.
- 2 T. Norri, B. Cazaux, D. Kosolobov and V. Mäkinen. Linear time minimum segmentation enables scalable founder reconstruction. *Algorithms for Molecular Biology*, Vol. 14, No. 12, 2019.

3.3 Genome Graphs and BWT-based Data Structures

Jouni Sirén (*University of California – Santa Cruz, US*)

License  Creative Commons BY 3.0 Unported license
© Jouni Sirén

Joint work of Jouni Sirén, Erik Garrison, Adam M. Novak, Benedict Paten, Richard Durbin
Main reference Jouni Sirén, Erik Garrison, Adam M. Novak, Benedict Paten, Richard Durbin: “Haplotype-aware graph indexes”, in *Proc. of the 18th International Workshop on Algorithms in Bioinformatics, WABI 2018*, August 20–22, 2018, Helsinki, Finland, LIPIcs, Vol. 113, pp. 4:1–4:13, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2018.
URL <http://dx.doi.org/10.4230/LIPIcs.WABI.2018.4>

A reference sequence may represent a new dataset poorly if the sequenced individual diverges substantially at some location. Mapping reads to such a reference can introduce reference bias into the subsequent analysis. Genome graphs help to avoid the bias by including genetic variation in the reference. Although each path in the graph is a potential haplotype, most paths are unlikely recombinations of true haplotypes.

In this talk, I will show how we can use BWT-based methods to index genome graphs. We transform the graph into an equivalent Wheeler graph, or approximate it with Wheeler graphs when the equivalent Wheeler graph is too large or does not exist. I will also introduce the GBWT as a fast and space-efficient way of storing large collections of haplotypes as paths over the genome graph.

3.4 BWT Meets the de Bruijn Graph: Results and Challenges

Christina Boucher (University of Florida – Gainesville, US)

License © Creative Commons BY 3.0 Unported license
© Christina Boucher

The money and time needed to sequence a genome have decreased remarkably in the past decade. With this decrease has come an increase in the number and rate at which sequence data is collected for public sequencing projects. This led to the existence of GenomeTrakr, which is a large public effort to use genome sequencing for surveillance and detection of outbreaks of foodborne illnesses. This effort includes over 50,000 samples, spanning several species available through this initiative, a number that continues to rise as datasets are continually added. Unfortunately, analysis of this dataset has been limited due to its size. In this talk, I will describe our method for constructing the colored de Bruijn graph for large datasets that is based on partitioning the data into smaller datasets, building the colored de Bruijn graph using a FM-index based representation, and succinctly merging these representations to build a single graph. Finally, I will show its capability of building a colored de Bruijn graph for 16,000 strains from GenomeTrakr in a manner that allows it to be updated. Lastly, I conclude by outlining some opportunities for further study in this area.

3.5 Text Indexing with the BWT

Gonzalo Navarro (University of Chile – Santiago de Chile, CL)

License © Creative Commons BY 3.0 Unported license
© Gonzalo Navarro

Main reference Paolo Ferragina, Giovanni Manzini: “Opportunistic Data Structures with Applications”, in Proc. of the 41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA, pp. 390–398, IEEE Computer Society, 2000.

URL <https://doi.org/10.1109/SFCS.2000.892127>

The talk covers the history and functioning of the FM-index, since its first version (Ferragina and Manzini, 2000) to the latest one (Gagie, Navarro, and Prezza, 2018) aimed at repetitive datasets.

3.6 Local Decodability of the Burrows-Wheeler Transform

Sandip Sinha (Columbia University – New York, US)

License © Creative Commons BY 3.0 Unported license
© Sandip Sinha

Joint work of Sandip Sinha, Omri Weinstein

Main reference Sandip Sinha, Omri Weinstein: “Local decodability of the Burrows-Wheeler transform”, in Proc. of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019., pp. 744–755, ACM, 2019.

URL <https://doi.org/10.1145/3313276.3316317>


The Burrows-Wheeler Transform (BWT) is among the most influential discoveries in text compression and DNA storage. It is a reversible preprocessing step that rearranges an n -letter string into runs of identical characters (by exploiting context regularities), resulting in highly compressible strings, and is the basis of the **bzip** compression program. Alas, the decoding process of BWT is inherently sequential and requires $\Omega(n)$ time even to retrieve a *single* character.

We study the succinct data structure problem of locally decoding short substrings of a given text under its *compressed* BWT, i.e., with small additive redundancy r over the *Move-To-Front* (**bzip**) compression. The celebrated BWT-based FM-index (FOCS '00), as well as other related literature, yield a trade-off of $r = \tilde{O}(n/\sqrt{t})$ bits, when a single character is to be decoded in $O(t)$ time. We give a near-quadratic improvement $r = \tilde{O}(n \lg(t)/t)$. As a by-product, we obtain an *exponential* (in t) improvement on the redundancy of the FM-index for counting pattern-matches on compressed text. In the interesting regime where the text compresses to $o(n)$ (say, $n/\text{poly}(\lg(n))$) bits, these results provide an $\exp(t)$ *overall* space reduction. For the local decoding problem of BWT, we also prove an $\Omega(n/t^2)$ cell-probe lower bound for “symmetric” data structures.

We achieve our main result by designing a compressed partial-sums (Rank) data structure over BWT. The key component is a *locally-decodable* Move-to-Front (MTF) code: with only $O(1)$ extra bits per block of length $n^{\Omega(1)}$, the decoding time of a single character can be decreased from $\Omega(n)$ to $O(\lg n)$. This result is of independent interest in algorithmic information theory.

3.7 BWT Construction: History, Techniques, State of the Art, Open Problems

Dominik Kempa (University of Warwick – Coventry, GB)

License  Creative Commons BY 3.0 Unported license
© Dominik Kempa

Joint work of Dominik Kempa, Tomasz Kociumaka

Main reference Dominik Kempa, Tomasz Kociumaka: “String synchronizing sets: sublinear-time BWT construction and optimal LCE data structure”, in Proc. of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019., pp. 756–767, ACM, 2019.

URL <https://doi.org/10.1145/3313276.3316368>

Burrows-Wheeler transform (BWT) is an invertible text transformation that, given a text T of length n , permutes its symbols according to the lexicographic order of suffixes of T . BWT is one of the most heavily studied algorithms in data compression with numerous applications in indexing, sequence analysis, and bioinformatics. Its construction is a bottleneck in many scenarios, and settling the complexity of this task is one of the most important unsolved problems in sequence analysis that has remained open for 25 years. In this talk, I will review the recent progress made for the problem of BWT construction [1, 2] as well as summarize the existing algorithms and outline the main challenges lying ahead.

References

- 1 Dominik Kempa and Tomasz Kociumaka. *String Synchronizing Sets: Sublinear-Time BWT Construction and Optimal LCE Data Structure*. In Proceedings of the 51st Annual ACM SIGACT Symposium on the Theory of Computing (STOC), 2019, ACM.
- 2 Dominik Kempa. *Optimal Construction of Compressed Indexes for Highly Repetitive Texts*. In Proceedings of the 30th Annual ACM SIAM Symposium on Discrete Algorithms (SODA), 2019, SIAM.

3.8 On the Hardness and Inapproximability of Recognizing Wheeler Graphs

Sharma V. Thankachan (University of Central Florida – Orlando, US)

License © Creative Commons BY 3.0 Unported license

© Sharma V. Thankachan

Joint work of Daniel Gibney, Sharma V. Thankachan

Main reference Daniel Gibney, Sharma V. Thankachan: “On the Hardness and Inapproximability of Recognizing Wheeler Graphs”, in Proc. of the 27th Annual European Symposium on Algorithms, ESA 2019, September 9–11, 2019, Munich/Garching, Germany., pp. 51:1–51:16, 2019.

URL <http://dx.doi.org/10.4230/LIPIcs.ESA.2019.51>

In recent years several compressed indexes based on variants of the Burrows-Wheeler transformation have been introduced. Some of these are used to index structures far more complex than a single string, as was originally done with the FM-index [Ferragina and Manzini, J. ACM 2005]. As such, there has been an increasing effort to better understand under which conditions such an indexing scheme is possible. This has led to the introduction of Wheeler graphs [Gagie *et al.*, Theor. Comput. Sci., 2017]. Gagie *et al.* showed that de Bruijn graphs, generalized compressed suffix arrays, and several other BWT related structures can be represented as Wheeler graphs, and that Wheeler graphs can be indexed in a way which is space efficient. Hence, being able to recognize whether a given graph is a Wheeler graph, or being able to approximate a given graph by a Wheeler graph, could have numerous applications in indexing. Here we resolve the open question of whether there exists an efficient algorithm for recognizing if a given graph is a Wheeler graph. We present:

The problem of recognizing whether a given graph $G = (V, E)$ is a Wheeler graph is NP-complete for any edge label alphabet of size $\sigma \geq 2$, even when G is a DAG. This holds even on a restricted, subset of graphs called d -NFA’s for $d \geq 5$. This is in contrast to recent results demonstrating the problem can be solved in polynomial time for d -NFA’s where $d \leq 2$. We also show the recognition problem can be solved in linear time for $\sigma = 1$;

There exists an $2^{e \log \sigma + O(n+e)}$ time exact algorithm where $n = |V|$ and $e = |E|$. This algorithm relies on graph isomorphism being computable in strictly sub-exponential time;

We define an optimization variant of the problem called Wheeler Graph Violation, abbreviated WGV, where the aim is to remove the minimum number of edges in order to obtain a Wheeler graph. We show WGV is APX-hard, even when G is a DAG, implying there exists a constant $C \geq 1$ for which there is no C -approximation algorithm (unless $P = NP$). Also, conditioned on the Unique Games Conjecture, for all $C \geq 1$, it is NP-hard to find a C -approximation;

We define the Wheeler Subgraph problem, abbreviated WS, where the aim is to find the largest subgraph which is a Wheeler Graph (the dual of the WGV). In contrast to WGV, we prove that the WS problem is in APX for $\sigma = O(1)$;

The above findings suggest that most problems under this theme are computationally difficult. However, we identify a class of graphs for which the recognition problem is polynomial time solvable, raising the open question of which parameters determine this problem’s difficulty.

3.9 String Attractors

Nicola Prezza (*University of Pisa, IT*), Travis Gagie (*Universidad Diego Portales, CL*), Dominik Kempa (*University of Warwick – Coventry, GB*), and Gonzalo Navarro (*University of Chile – Santiago de Chile, CL*)

License © Creative Commons BY 3.0 Unported license

© Nicola Prezza, Travis Gagie, Dominik Kempa, and Gonzalo Navarro

Main reference Dominik Kempa, Nicola Prezza: “At the roots of dictionary compression: string attractors”, in Proc. of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018, pp. 827–840, ACM, 2018.

URL <https://doi.org/10.1145/3188745.3188814>

In this talk I show connections between the Burrows-Wheeler transform and popular compressors such as LZ77 and grammars. The first connection is that these compressors can be interpreted as approximation algorithms for computing a new combinatorial object: the string attractor [1]. A string attractor is a subset of the text’s positions such that each distinct text substring has at least one occurrence crossing at least one element in the set. It turns out that most dictionary compressors induce string attractors whose cardinalities are bounded by their outputs’ sizes, and that one can build a dictionary-compressed representation from a string attractor. It follows that these new objects allow one to prove new relations between the sizes of dictionary compressors, and to design universal compressed data structures. The second connection is through bidirectional parsings. A bidirectional parse is a generalization of LZ77 where phrases’ sources are not forced to precede their destination. I will show that the BWT induces a bidirectional parse with r phrases, where r is the number of equal-letter runs in the BWT. Unlike LZ77, this parse enjoys new fascinating properties that allow one to build an optimal-time index on top of it [2].

References

- 1 Dominik Kempa and Nicola Prezza *At the Roots of Dictionary Compression: String Attractors*. Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, Los Angeles, CA, USA, 2018.
- 2 Travis Gagie, Gonzalo Navarro, and Nicola Prezza *Optimal-time Text Indexing in BWT-runs Bounded Space*. Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, Louisiana, USA, 2018.

3.10 Combinatorial Properties of BWT

Marinella Sciortino (*University of Palermo, IT*)

License © Creative Commons BY 3.0 Unported license

© Marinella Sciortino

Although the BWT has been introduced in Data Compression, over the years it has found many applications in several different contexts. The outstanding versatility and efficacy of the BWT is based on some mathematical and combinatorial properties, i.e. its efficient reversibility and the “clustering effect” on the output. In this talk such properties are explored, highlighting the connections with well-known objects in Combinatorics of words, such as Lyndon words and Sturmian words. Furthermore, studying how the number of equal-letter runs varies after the BWT is applied, allows us to characterize infinite families of words based on the clustering effect produced by BWT. In such cases these characterizations are connected to still-open mathematical conjectures. Finally, some variants of the BWT

are described. The first one, denoted eBWT, is defined on multisets of strings and allows to establish a bijection between the multiset of conjugacy classes of strings and all the strings on a given alphabet, with interesting theoretical and applicative implications. The second variant, denoted ABWT, uses a different order (called alternating order) to sort the cyclic rotations of a string. It is interesting to note that the ABWT preserves many combinatorial and mathematical properties of the BWT and it can be used as a compressed index in the same way as the BWT.

3.11 BWT / eBWT similarity

Giovanna Rosone (University of Pisa, IT)

License © Creative Commons BY 3.0 Unported license
© Giovanna Rosone

Joint work of Giovanna Rosone, Veronica Guerrini, Sabrina Mantaci, Marinella Sciortino, Antonio Restivo

Main reference Sabrina Mantaci, Antonio Restivo, Giovanna Rosone, Marinella Sciortino: “A New Combinatorial Approach to Sequence Comparison”, *Theory Comput. Syst.*, Vol. 42(3), pp. 411–429, 2008.

URL <https://doi.org/10.1007/s00224-007-9078-6>

Sequence comparison has become a very essential tool in modern molecular biology. In fact, in biomolecular sequences high similarity usually implies significant functional or structural similarity. Traditional approaches use techniques that are based on sequence alignment able to measure character level differences. Here, we describe some similarity measures, alignment-free, based on the Burrows-Wheeler transform with several application in bioinformatics, such as the metagenomic problem.

3.12 Searching Patterns in the Bijective BWT

Dominik Köppl (Kyushu University – Fukuoka, JP)

License © Creative Commons BY 3.0 Unported license
© Dominik Köppl

Joint work of Hideo Bannai, Juha Kärkkäinen, Dominik Köppl, Marcin Piatkowski

Main reference Hideo Bannai, Juha Kärkkäinen, Dominik Köppl, Marcin Piatkowski: “Indexing the Bijective BWT”, in *Proc. of the 30th Annual Symposium on Combinatorial Pattern Matching, CPM 2019*, June 18-20, 2019, Pisa, Italy., *LIPICs*, Vol. 128, pp. 17:1–17:14, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2019.

URL <http://dx.doi.org/10.4230/LIPICs.CPM.2019.17>

We present an index data structure for the bijective Burrows-Wheeler transform [1]. The index data structure is based on the FM index [2]. Like the FM index, it reports the suffix array interval of all pattern occurrences by means of backward searches.

References

- 1 Joseph Yossi Gil and David Allen Scott. A bijective string sorting transform. *ArXiv 1201.3077*, 2012. [arXiv:1201.3077](https://arxiv.org/abs/1201.3077).
- 2 Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Proc. FOCS*, pages 390–398, 2000.

4 Motivation (from proposal)

When it was introduced in a technical report in May 1994, no one could have foreseen the impact the Burrows-Wheeler Transform (BWT) would have far beyond the field of data compression for which it was originally intended. Of course it first made a significant impact on compression, both in theory and in practice (e.g., as the basis for bzip2). New horizons opened up in 2000 with the introduction of the FM-index, a compressed suffix array based on the BWT. Among other applications, in the next decade FM-indexes became the heart of the DNA aligners such as Bowtie, BWA and SOAP 2 that helped pave the way for the genomics revolution.

Generalizations of BWT to labelled trees, de Bruijn graphs, automata, haplotype sequences and genomic reference graphs have kept the exchange of ideas lively between researchers in algorithms and data structures, bioinformatics, combinatorics on words and information retrieval. Burrows and Wheeler’s original technical report is still cited hundreds of times every year, subsequent papers are cited thousands of times, and new results about or using the BWT appear in the top conferences and journals.

By now, 25 years since its publication, probably no one person knows all the results that have been proven about the BWT, but we hope the expertise gathered together at this Dagstuhl Seminar will make progress on the following topics, among others:

FM-indexes for genomic databases: FM-indexes shine for indexing one or a few genomes, but they have not scaled well to indexing the genomic databases that have resulted from high-throughput sequencing technologies. An important problem has been that the suffix array samples used to locate occurrences of patterns must be fairly large or locating becomes very slow. Very recently, a way was discovered to greatly compress also the suffix array sample for repetitive texts, opening the door to indexing thousands of genomes. We expect this seminar will lead to a fuller understanding of this advance and how it can be applied in practice. Another challenge has been beating run-length compression of genomic databases’ BWTs, by identifying additional structure.

More generalizations of the BWT: Many BWT researchers have heard of the generalizations to trees and graphs mentioned above, but it seems few except specialists in algorithms and data structures know about its recent extension to indexed parameterized and order-preserving pattern matching, few except specialists in combinatorics on words know about the alternating BWT, and few except bioinformaticians know about the positional BWT – but each of these may have applications in the other areas. Also, a partially unifying framework has recently been proposed, but there are still many open problems.

New challenges in bioinformatics: Papers on the BWT are published in many venues and no single conference brings together all the experts from algorithms and data structures, combinatorics on words, and theoretical and applied bioinformatics. This disconnect between the areas hurts us all because it prevents knowledge being shared efficiently. The BWT has recently been applied to some surprising bioinformatics problems, such as building ancestral recombination graphs and optical read mapping, and we expect other possibilities will emerge from interdisciplinary discussions.

5 Feedback

All of the 19 respondents to the survey said they would definitely attend another Dagstuhl seminar (5, from 1 to 5) and the median rating of the scientific quality was 10 out of 11. One person was neutral about the seminar inspiring new ideas for work, research or teaching; four

agreed it did; and 13 agreed completely (and we do not know what the last person thought). Four people were neutral about it inspiring joint projects or publications; nine agreed it did; and six agreed completely. Two people disagreed that it led to insights from neighbouring fields; two were neutral; nine agreed; and five agreed completely. One person disagreed that it identified new research directions; four people were neutral; seven people agreed; and seven people agreed completely. The responses to the other questions were similar. The comments were generally positive, with people liking the mix of fields; the organization of the panels could have been improved, although they still offered some valuable insights and stimulated promising discussions.

6 Open Problems

Several general open problems were posed – e.g., generalizing the BWT to even more data types, merging BWTs of other data types (and, most generally, Wheeler graphs), improving the compression of BWT-based indexes for DNA readsets, applying current theory to practice – and some specific ones. Sharma Thankachan posed the problem of determining the degree of non-determinism that makes Wheeler graph recognition hard: it is currently known to take polytime when each node has at most two outgoing edges labelled with the same character, and it is NP-complete when nodes can have five outgoing edges labelled with the same character. Jackie Daykin asked about enhancing BWT performance with alphabet reordering (an issue about which Sharma now has preliminary results). Dominik Köppl posted a one-page description of an open problem (“Can we compute the Bijective BWT in linear time?”) to the materials page.

Participants

- Jarno Alanko
University of Helsinki, FI
- Hideo Bannai
Kyushu University –
Fukuoka, JP
- Paola Bonizzoni
University of Milan-Bicocca, IT
- Christina Boucher
University of Florida –
Gainesville, US
- Marilia Braga
Universität Bielefeld, DE
- Anthony J. Cox
Illumina – Saffron Walden, GB
- Fabio Cunial
MPI – Dresden, DE
- Jackie Daykin
Aberystwyth University, GB
- Richard Durbin
University of Cambridge, GB
- Gabriele Fici
University of Palermo, IT
- Johannes Fischer
TU Dortmund, DE
- Travis Gagie
Universidad Diego Portales, CL
- Pawel Gawrychowski
University of Wroclaw, PL
- Simon Gog
eBay Inc – San Jose, US
- Roberto Grossi
University of Pisa, IT
- Wing-Kai Hon
National Tsing Hua University –
Hsinchu, TW
- Tomohiro I
Kyushu Institute of Technology –
Fukuoka, JP
- Juha Kärkkäinen
University of Helsinki, FI
- Dominik Kempa
University of Warwick –
Coventry, GB
- Tomasz Kociumaka
Bar-Ilan University –
Ramat Gan, IL
- Dominik Köppl
Kyushu University –
Fukuoka, JP
- Ben Langmead
Johns Hopkins University –
Baltimore, US
- Zsuzsanna Liptak
University of Verona, IT
- Veli Mäkinen
University of Helsinki, FI
- Sabrina Mantaci
University of Palermo, IT
- Giovanni Manzini
University of Eastern Piedmont –
Alessandria, IT
- Ian Munro
University of Waterloo, CA
- Gene Myers
MPI – Dresden, DE
- Gonzalo Navarro
University of Chile –
Santiago de Chile, CL
- Yakov Nekrich
University of Waterloo, CA
- Enno Ohlebusch
Universität Ulm, DE
- Kunsoo Park
Seoul National University, KR
- Nicola Prezza
University of Pisa, IT
- Knut Reinert
FU Berlin, DE
- Giovanna Rosone
University of Pisa, IT
- Kunihiro Sadakane
University of Tokyo, JP
- Leena Salmela
University of Helsinki, FI
- Marinella Sciortino
University of Palermo, IT
- Rahul Shah
Louisiana State University –
Baton Rouge, US
- Sandip Sinha
Columbia University –
New York, US
- Jouni Sirén
University of California –
Santa Cruz, US
- Tatiana Starikovskaya
ENS – Paris, FR
- Jens Stoye
Universität Bielefeld, DE
- Sharma V. Thankachan
University of Central Florida –
Orlando, US
- Rossano Venturini
University of Pisa, IT

