

Report from Dagstuhl Seminar 11081

Combinatorial and Algorithmic Aspects of Sequence Processing

Edited by

Maxime Crochemore¹, Lila Kari², Mehryar Mohri³, and
Dirk Nowotka⁴

1 King's College London, GB, maxime.crochemore@kcl.ac.uk

2 University of Western Ontario, London, CA, lila@csd.uwo.ca

3 New York University, US, mohri@cs.nyu.edu

4 University of Stuttgart, DE, nowotka@fmi.uni-stuttgart.de

Abstract

Sequences form the most basic and natural data structure. They occur whenever information is electronically transmitted (as bit streams), when natural language text is spoken or written down (as words over, for example, the latin alphabet), in the process of heredity transmission in living cells (as DNA sequence) or the protein synthesis (as sequence of amino acids), and in many more different contexts. Given this universal form of representing information, the need to process strings is apparent and actually a core purpose of computer use. Algorithms to efficiently search through, analyze, (de-)compress, match, learn, and encode/decode strings are therefore of chief interest. Combinatorial problems about strings lie at the core of such algorithmic questions. Many such combinatorial problems are common in the string processing efforts in the different fields of application.

Scientists working in the fields of *Combinatorics on Words*, *Computational Biology*, *Stringology*, *Natural Computing*, and *Machine Learning* were invited to consider the seminar's topic from a wide range of perspectives. This report documents the program and the outcomes of Dagstuhl Seminar 11081 "Combinatorial and Algorithmic Aspects of Sequence Processing".

Seminar 21.–25. February, 2011 – www.dagstuhl.de/11081

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, F.4.3 Formal Languages, G.2.1 Combinatorics, I.2.6 Learning, J.3 Life and Medical Sciences

Keywords and phrases Combinatorics on words, computational biology, stringology, natural computing, machine learning

Digital Object Identifier 10.4230/DagRep.1.2.47

1 Executive Summary

Maxime Crochemore

Lila Kari

Mehryar Mohri

Dirk Nowotka

License  Creative Commons BY-NC-ND 3.0 Unported license
© Maxime Crochemore, Lila Kari, Mehryar Mohri, and Dirk Nowotka

The object of concern of this seminar, *sequences*, implies a large degree of generality. It plays an essential rôle in many fields and constitutes a true cross section area. Hence, the seminar was designed to bring together researchers from different disciplines whose interest



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

Combinatorial and Algorithmic Aspects of Sequence Processing, *Dagstuhl Reports*, Vol. 1, Issue 2, pp. 47–66

Editors: Maxime Crochemore, Lila Kari, Mehryar Mohri, and Dirk Nowotka



Dagstuhl Reports

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

are string processing algorithms and related combinatorial problems on words. Scientists working in the following fields were invited to consider the seminar's topic from a wide range of perspectives:

- Combinatorics on Words,
- Computational Biology,
- Stringology,
- Natural Computing,
- Machine Learning.

This Dagstuhl seminar was attended by 40 researchers from 13 countries. Everyone of the five topics above was about equally represented. Given the extremely interdisciplinary approach of this meeting it was an obvious necessity to hold a tutorial on each one of the participating research areas. These tutorials were held over the first and the morning of the second seminar day (see the scientific schedule below). They provided a good introduction for the non-specialists and triggered the first scientific discussions and exchanges.

A second (and standard) element of this seminar were regular talks, of course. A total of 15 talks were presented. It has to be noted that one could experience a very productive atmosphere during the whole seminar. All talks were well-attended and accompanied with interesting comments. Plenty of time was reserved for questions and discussions which was actively used by the participants.

The third element of the seminar were open problem sessions which did yield a larger attention to a range of problems, only some of them are included in this report. These open problem sessions provided the ideal ground for the ignition of new research lines and cooperations. Just to mention one example, the paper "On the regularity of iterated hairpin completion of a single word" (arXiv:1104.2385v1) resulted from the collaboration of Steffen Kopecki and Shinnosuke Seki initiated at this Dagstuhl seminar. In the light of such developments, it can be safely claimed that this seminar was a success.

Given the quality of presentations on this seminar and the constructive intensity of discussions between and after the talks, it is self-evident that follow-ups will be attempted. After this initial meeting of different communities, where common problems were identified, personal contacts established and first cooperations initiated, further events can be sharpened in focus and more on particular cross section topics regarding combinatorial and algorithmic problems in sequence processing.

Finally, we would like to say that the organization of a meeting for researchers of so unusually diverse fields bears a certain risk. However, it can be said that the event turned out better than expected. It was more than worthwhile to have taken that risk. We are grateful to all participants for their contributions to this successful seminar as well as to the staff of Schloss Dagstuhl for their perfect service.

2 Table of Contents

Executive Summary

Maxime Crochemore, Lila Kari, Mehryar Mohri, and Dirk Nowotka 47

Overview of Tutorials

Data Structures for Text Indexing and String Algorithms
Roberto Grossi 51

Natural Computing Tutorial
Hendrik Jan Hoogeboom 51

Introduction to Sequence Learning
Mehryar Mohri 51

Combinatorics on Words: An Introduction
Jeffrey Shallit 52

Overview of Talks

Intelligent Strategies for Remote Homology Detection
Juliana Bernardes 52

Simple Real-Time Constant-Space String-Matching
Dany Breslauer 53

Combinatorial Measure of Co-evolving Blocks and their Evolutionary Pressure
Linda Dib 53

Non-Archimedean Words
Volker Diekert 54

Fixed Points of Nontrivial Morphisms
Štěpán Holub 54

Observations and Problems on k -abelian Avoidability
Juhani Karhumäki 55

Hairpin Completion versus Hairpin Lengthening
Steffen Kopecki 55

Sequence and Chromatin Signatures Predict Transcription Factor Binding in the Human Genome
Christina Leslie 56

Some algorithmic and combinatorial problems in the RNA and DNA world
Jan Mañuch 57

Exact ensemble properties in combinatorial dynamic programming schemes
Yann Ponty 58

On the Structure of Compacted Subword Graphs of Thue-Morse Words and Their Applications
Wojciech Rytter 58

Enumeration and Automatic Sequences
Jeffrey Shallit 59

Context Equivalence Problem	
<i>Arseny M. Shur</i>	60
Hashing for Strings	
<i>Alexander J. Smola</i>	60
Open Problems	
Word Equations with Loops	
<i>Štěpán Holub</i>	61
Is morphic primitivity hereditary?	
<i>Štěpán Holub</i>	62
Asymptotic Number of Long-Armed Palindromes in a Word	
<i>Gregory Kucherov</i>	62
The Separating Words Problem	
<i>Jeffrey Shallit</i>	63
Some open problems inspired by Dejean’s conjecture	
<i>Arseny M. Shur</i>	63
Scientific Schedule	65
Participants	66

3 Overview of Tutorials

3.1 Data Structures for Text Indexing and String Algorithms

Roberto Grossi (University of Pisa, IT)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Roberto Grossi

URL <http://www.dagstuhl.de/mat/Files/11/11081/11081.GrossiRoberto.Slides.pdf>

This is an introductory tutorial to the basic data structures employed in stringology: tries, compact tries, suffix trees, suffix arrays, and suffix automata. The tutorial considers also the case of large texts, discussing the external-memory model and the cache-oblivious model, with examples for suffix arrays, suffix trees, and string B-trees.

3.2 Natural Computing Tutorial

Hendrik Jan Hoogeboom (Leiden University, NL)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Hendrik Jan Hoogeboom

URL <http://www.dagstuhl.de/mat/Files/11/11081/11081.HoogeboomHendrikJan.Slides.pdf>

This is an overview of some concepts in the field of Molecular Computing (aka. DNA Computing): Adlemans experiment, TicTacToe computer (and beyond), evolutionary DNA computing, self assembly, bio-inspired formal models (splicing systems, new operations, membrane computing), nature as computer (gene assembly in ciliates).

3.3 Introduction to Sequence Learning

Mehryar Mohri (New York University, US)

License © ⓘ ⊖ Creative Commons BY-NC-ND 3.0 Unported license
© Mehryar Mohri


Joint work of Corinna Cortes; Mehryar Mohri

This tutorial presents an introduction to sequence learning. This includes a brief presentation of binary classification problems and solutions based on large-margin hyperplanes and kernel methods, and a detailed discussion of sequence kernels. In particular, we describe a general framework based on rational kernels, give the proof of the positive-definiteness of a general class of rational kernels, show how general families of count-based kernels can be defined using rational kernels, and give a variety of examples of PDS rational kernels relevant to computational biology and text and speech processing.

We also present a number of general open problems related to a faster computation of sequence kernels and to the characterization of the class of languages learnable with rational kernels. Finally, we discuss the problem of learning sequence kernels and that of determining more efficient optimization solutions using sequence kernels.

3.4 Combinatorics on Words: An Introduction

Jeffrey Shallit (University of Waterloo, CA)


License  Creative Commons BY-NC-ND 3.0 Unported license
 © Jeffrey Shallit
URL <http://www.cs.uwaterloo.ca/~shallit/Talks/introcw.pdf>

In this talk I surveyed some of the main themes in combinatorics on words: periodicity, patterns and pattern avoidance, equations in words, and infinite words and their properties. Among other things, I covered the Lyndon-Schutzenberger theorems, primitive words, conjugates, Lyndon words, fractional powers, unbordered words, Duval’s conjecture, the Fine-Wilf theorem, Sturmian words, the Thue-Morse sequence, construction of a square-free infinite word, Dejean’s conjecture, avoidance of abelian powers, Makanin’s algorithm, Plandowski’s PSPACE results, subword complexity, automatic sequences, and Christol’s theorem.

4 Overview of Talks

4.1 Intelligent Strategies for Remote Homology Detection

Juliana Bernardes (UPMC – Paris, FR)


License  Creative Commons BY-NC-ND 3.0 Unported license
 © Juliana Bernardes
Main reference Bernardes J.S, Carbone A. and Zaverucha Gerson. A discriminative method for family-based protein remote homology detection that combines inductive logic programming and propositional models. BMC Bioinformatics 2011, 12:83
URL <http://www.biomedcentral.com/1471-2105/12/83>

Remote homology detection is a hard computational problem. Most approaches have trained computational models by using either full protein sequences or multiple sequence alignments (MSA), including all positions. However, when we deal with proteins in the “twilight zone” we can observe that only some segments of sequences (motifs) are conserved. We introduce a novel logical representation that allows us to represent physico-chemical properties of sequences, conserved amino acid positions and conserved physico-chemical positions in the MSA. From this, Inductive Logic Programming (ILP) finds the most frequent patterns (motifs) and uses them to train propositional models, such as decision trees and support vector machines (SVM). Our results show that our methodology when using SVM performs significantly better than some of the state of the art methods, and comparable to other. However, our method provides a comprehensible set of logical rules that can help to understand what determines a protein function.

The strategy of selecting only the most frequent patterns is effective for the remote homology detection. This is possible through a suitable first-order logical representation of homologous properties, and through a set of frequent patterns, found by an ILP system, that summarizes essential features of protein functions.

4.2 Simple Real-Time Constant-Space String-Matching

Dany Breslauer (University of Haifa, IL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Dany Breslauer

Joint work of Dany Breslauer; Roberto Grossi; Filippo Mignosi

Main reference D. Breslauer R. Grossi, and F. Mignosi. Simple Real-Time Constant-Space String-Matching. 22nd Annual Symposium on Combinatorial Pattern Matching (CPM), 2011.

We use a simple observation about the locations of critical factorizations to derive a real-time variation of the Crochemore-Perrin constant-space string-matching algorithm. The real-time variation has a simple and efficient control structure.

4.3 Combinatorial Measure of Co-evolving Blocks and their Evolutionary Pressure


Linda Dib (UPMC – Paris, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Linda Dib

Co-evolution signals have been detected on a few divergent protein families while families of conserved protein sequences remain untractable by current methods. A large scale investigation of residue networks can only be made with the development of refined methods treating conserved sequences as well. We propose a new combinatorial approach to overcome this difficulty. Based on the observation that co-evolving positions are usually not isolated and that their co-evolving behaviour concerns adjacent positions as well, our combinatorial method, named Blocks In Sequences (BIS), studies co-evolution of blocks of contiguous positions in sequences, where a block is possibly constituted by a single position. BIS determines whether blocks of residues co-evolve or not and at which strength. BIS can be applied to sets of very conserved sequences, possibly made by a few sequences, and yet it is able to detect positional differences between these sequences and evaluate possible signals of co-evolution. BIS captures important information on folding processes. It gives no hint on the kinetics but rather on the actors (that is, residues, parts of secondary structures, 3D interactions) of the kinetics process. The level of importance of these actors is encoded on the strength of the co-evolution signal. This strength is measured by a symmetric signal coming from residue pairs and by the resemblance of residues in a network with their environment, but also by the combinatorics of the relationships possibly existing between networks that can be highlighted by the method. Network overlapping and connected components of the associated interval graph are used to bring up the intricate structure of co-evolving networks. Results obtained by BIS on the Protein A were compared to Φ -analysis and the outcomes are remarkably similar.

4.4 Non-Archimedean Words

Volker Diekert (University of Stuttgart, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Volker Diekert

Non-Archimedean words have been introduced as a new type of infinite words which can be investigated through classical methods in combinatorics on words due to a length function. The length function, however, takes values in the additive group of polynomials $\mathbb{Z}[t]$ (and not, as traditionally, in \mathbb{N}), which yields various new properties. Non-Archimedean words allow to solve a number of interesting algorithmic problems in geometric and algorithmic group theory. There is also a connection to logic and the first-order theory in free groups (Tarski Problems).


In my lecture I report on a joint work with Alexei Miasnikov. We provide a general method to use infinite words over a discretely ordered abelian group as a tool to investigate certain group extensions for an arbitrary group G . The central object is a group $E(A, G)$, which is defined in terms of a non-terminating, but confluent rewriting system. The group G as well as some natural HNN-extensions of G embed into $E(A, G)$ (and still “behave like” G), which makes it interesting to study its algorithmic properties.

The main result characterizes exactly when the Word Problem is decidable in all finitely generated subgroups of $E(A, G)$. We show that this property holds if and only if the Cyclic Membership Problem “ u in $\langle v \rangle$?” is decidable for all v in G .

The results combine methods from combinatorics on words, string rewriting and group theory.

4.5 Fixed Points of Nontrivial Morphisms

Štěpán Holub (Charles University – Prague, CZ)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Štěpán Holub

Main reference Holub, Štěpán, Polynomial algorithm for fixed points of nontrivial morphisms, *Discrete Mathematics* 309: 5069-5076 (2009)

URL <http://dx.doi.org/10.1016/j.disc.2009.03.019>

In this talk I presented an algorithm that decides whether given finite word is a fixed point of a nontrivial morphism. Such words are also called *morphically imprimitive*.

These words were characterized in [1] and [2]. In [3], the question was raised about the time complexity of the decision problem as well as about the possibility to find the corresponding morphism.

The algorithm works in a subquadratic time and outputs a morphism that is in a good sense unique minimal witness of the word being imprimitive.

I also mentioned a related open problem, known as The Conjecture of Billaud. Let w be a finite word, and let δ_x denote the morphism canceling the letter x and otherwise acting as the identity. The conjecture states that if $\delta_x(w)$ is imprimitive for all x from the alphabet of w , then also w is imprimitive. Some partial results regarding this conjecture can be found in [1].


References

- 1 T. Head. *Fixed languages and the adult languages of OL schemes*. *Int. J. Comput. Math.* 10(2) (1981) 103-107

- 2 D. Hamm and J. Shallit. *Characterization of finite and one-sided infinite fixed points of morphisms on free monoids*. Technical Report CS-99-17, University of Waterloo, July 1999
- 3 D. Reidenbach and J. C. Schneider. *Morphically primitive words*. *Theor. Comput. Sci.* 410 (2009) 2148-2161
- 4 F. Levé and G. Richomme. *On a conjecture about finite fixed points of morphisms*. *Theor. Comput. Sci.* 339(1)(2005) 103-128

4.6 Observations and Problems on k -abelian Avoidability

Juhani Karhumäki (University of Turku, FI)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Juhani Karhumäki

Joint work of Mari Huova; Juhani Karhumäki

Main reference M. Huova and J. Karhumäki, Observations and Problems on k -abelian Avoidability, Dagstuhl Preprint Archive

URL <http://arxiv.org/abs/1104.4273v1>


We introduce new avoidability problems for words by considering equivalence relations, k -abelian equivalences, which lie properly in between equality and commutative equality, i.e. abelian equality. For two k -abelian equivalent words the numbers of occurrences of different factors of length k coincide and the prefixies (resp. suffixies) of length $k - 1$ are equal as well.

The size of the smallest alphabet avoiding 2-repetitions of words, i.e. squares, is three and for abelian squares it is four. It follows that for 2-abelian squares this size has to be three or four. Similarly, the size of the smallest alphabet where 2-abelian cubes, i.e. 3-repetitions, can be avoided is two or three, because cubes (resp. abelian cubes) are avoidable in binary (resp. ternary) alphabet.

We show that for 2-abelian squares the required size is four, as in the case of abelian squares. The longest 2-abelian square-free ternary word is of length 537. The question for 2-abelian cubes is open. Though, we have computational evidence that the size would be two, since there exists 2-abelian cube-free binary word of length 100 000, meaning that the 2-abelian case would behave like that of words.

4.7 Hairpin Completion versus Hairpin Lengthening

Steffen Kopecki (University of Stuttgart, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Steffen Kopecki

URL <http://www.dagstuhl.de/mat/Files/11/11081/11081.KopeckiSteffen.Slides.pdf>

The hairpin completion and the hairpin lengthening are operations on formal languages that have been inspired by the hairpin formation in biochemistry. It is known that the hairpin completion (resp. hairpin lengthening) of a regular language is not in general regular but always linear context-free. As regularity of a (linear) context-free language is undecidable in general, we investigate the decidability problem whether the hairpin completion (resp. hairpin lengthening) of regular languages is regular again. For the hairpin completion we solved the problem positively in former papers. Even though both operations seem quite similar, we were not able to use the same approach for the hairpin lengthening. Here, we provide partial results on the decidability problem for the hairpin lengthening and discuss some differences between the two operations. To name one of them, the hairpin completion


of a regular language is an unambiguous (linear) context-free language, but the hairpin lengthening may lead to an inherent ambiguous (linear) context-free language.

References

- 1 D. Chepcea, C. Martin-Vide, and V. Mitrana. A new operation on words suggested by DNA biochemistry: Hairpincompletion. in *Proc. of Transgressive Computing*, 216–228 (2006)
- 2 F. Manea, C. Martin-Vide, and V. Mitrana. Hairpin lengthening. *Proc. of CiE*, LNCS 6158, 296–306 (2010)
- 3 V. Diekert and S. Kopecki. It is NL-complete to decide whether a hairpin completion of regular languages is regular. *CoRR*, abs/1101.4824, 2011.
- 4 S. Kopecki. On the iterated hairpin completion. *CoRR*, abs/1010.3640, 2011.

4.8 Sequence and Chromatin Signatures Predict Transcription Factor Binding in the Human Genome

Christina Leslie (Memorial Sloan-Kettering Cancer Center – New York, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christina Leslie

Gene regulatory programs are orchestrated by proteins called transcription factors (TFs), which coordinate expression of target genes both through direct binding to genomic DNA and through interaction with cofactors. Accurately modeling the DNA sequence preferences of TFs and predicting their genomic binding sites are key problems in regulatory genomics. These efforts have long been frustrated by the limited availability and accuracy of TF binding site motifs. Today, protein binding microarray (PBM) experiments and chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments are generating unprecedented high-resolution data on in vitro and in vivo TF binding. Moreover, genome-wide data on the cell-type specific chromatin state, including ChIP-seq experiments that profile histone modifications associated with active or inactive transcriptional states, provide additional information for predicting the genomic binding locations of TFs.


We will present a flexible new discriminative framework for representing and learning TF binding preferences using these massive data sets. We will first describe in vitro models of TF-DNA sequence affinities, where we train support vector regression (SVR) models with a novel string kernel on PBM data to learn the mapping from probe sequences to binding intensities. In a large data set of over 180 yeast and mouse TF binding experiments, our SVR models better predicted in vitro binding than popular motif discovery approaches or methods based on enrichment of k-mer patterns.

We will then show how to train kernel-based SVM models directly on TF ChIP-seq data to learn in vivo TF sequence models and present results from a large-scale evaluation on 184 TF ChIP-seq experiments from ENCODE. We confirmed that our discriminative sequence models significantly outperform existing motif discovery algorithms, and we found that ChIP-trained models greatly improved TF occupancy prediction over PBM-trained models, suggesting distinct in vivo sequence information (e.g. binding sites of cofactors). Finally, we trained discriminative chromatin models using histone modification ChIP-seq data and found that models combining sequence and chromatin signatures strongly outperformed using either one alone. We found that relatively few TFs in our study had pronounced cell-type specific binding patterns, but in those that did, we identified cell-type dependent sequence information. This work establishes effective new techniques for analyzing next generation

sequencing data sets to study the interplay of chromatin and sequence in TF binding in the human genome.

4.9 Some algorithmic and combinatorial problems in the RNA and DNA world

Jan Mañuch (Simon Fraser University – Burnaby, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Jan Mañuch

I have presented 3 problems in computational biology. The first problem, the energy barrier problem for pseudoknot-free RNA structures, asks whether what's the minimal energy barrier needed to transform from the initial to the final structure of an RNA molecule. We consider the simplest energy model in which only the number of basepairs is taken into account and transformation sequence removes and adds the basepairs in some order. We showed that the problem is NP-complete [1], but can be solved in polynomial time if the barrier is assumed to be constant [2]. I have also introduced a string displacement system which can model multi-strand scenario and can be modeled by a simple rewriting system with two types of strings (signals and templates).

The second and third problems arise in the DNA synthesis. DNA strand needs to be assembled from a shorter factors of the strand which avoid many types of collisions. The simplest type of collision is equality, which leads to the following word problem. Given a word on alphabet Σ and integer k , is it possible to partition it to distinct factors of length at most k . We show that this is NP-complete for alphabet size 4 [3]. We also consider other conditions which factors need to satisfy, e.g., prefix-freeness, factor-freeness, etc, and show that in all cases the problem is NP-complete even for the binary alphabet.


The third problem goes one step back. Starting from a protein sequence (sequence of amino acids), the task is to find the DNA sequence which is mapped that protein sequence and satisfies 2 constraints. One possible algorithm based on acyclic DFA leads to polynomial algorithm, however, its complexity is $O(n^{42})$. Is there a more efficient algorithm?

References

- 1 Mañuch, J., Thachuk, C., Stacho, L., Condon, A., NP-completeness of the energy barrier problem without pseudoknots and temporary arcs, *Nat. Comput.* **10**, No. 1, 391–405 (2011)
- 2 Thachuk, C., Mañuch, J., Rafiey, A., Mathieson, L-A., Stacho, L., Condon, A., An algorithm for the energy barrier problem without pseudoknots and temporary arcs, Proc. of *Pacific Symposium on Biocomputing* (PSB, Hawaii, USA, 2010), World Scientific Publishing, 108–119 (2010)
- 3 Condon, A., Mañuch, J., Thachuk, C., Complexity of a collision-aware string partition problem and its relation to oligo design for gene synthesis, Proc. of *Annual International Computing and Combinatorics Conference* (COCOON, Dalian, China, 2008), LNCS **5092**, 265–275 (2008).

4.10 Exact ensemble properties in combinatorial dynamic programming schemes

Yann Ponty (Ecole Polytechnique – Palaiseau, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Yann Ponty

Joint work of Yann Ponty; Cédric Saule


Main reference Y. Ponty and C. Saule. A combinatorial framework for the design of (pseudoknotted) RNA algorithms, *WABI*, 2011.

URL <http://www.lix.polytechnique.fr/~ponty/docs/EnsembleHypergraphsDP.pdf>

We extend an hypergraph representation, introduced by Finkelstein and Roytberg, to unify dynamic programming algorithms in the context of RNA folding with pseudoknots. Classic applications of RNA dynamic programming (Energy minimization, partition function, base-pair probabilities ...) are reformulated within this framework, giving rise to very simple algorithms. This reformulation allows one to conceptually detach the conformation space/energy model — captured by the hypergraph model — from the specific application, assuming unambiguity of the decomposition. To ensure the latter property, we propose a new combinatorial methodology based on generating functions. We extend the set of generic applications by proposing an exact algorithm for extracting generalized moments in weighted distribution, generalizing a prior contribution by Miklos and al. Finally, we illustrate our full-fledged programme on three exemplary conformation spaces (secondary structures, Akutsu's simple type pseudoknots and kissing hairpins). This readily gives sets of algorithms that are either novel or have complexity comparable to classic implementations for minimization and Boltzmann ensemble applications of dynamic programming.

4.11 On the Structure of Compacted Subword Graphs of Thue-Morse Words and Their Applications

Wojciech Rytter (University of Warsaw, PL)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Wojciech Rytter

Thue-Morse words (*TM* words, in short) form a famous family of words, due to many interesting properties related not only to text algorithms and combinatorics on words, but also to other disciplines, see [1]. In particular they do not contain factors of type $axaxa$, where a is a single letter (overlaps), consequently they do not contain cubes. A very good source for properties of these words is for example the book [3]. We rediscover/discover several known/unknown properties of *TM* words in a novel way: analyzing the subword graphs of finite and infinite *TM* words. This approach was already successfully applied by one of the authors to another well-known family of words, namely the Fibonacci words [8]. We also study how the cdawg of the infinite *TM* word is related to an infinite graph with 2-counting property and a numeration system, similar analysis for Fibonacci words and, in general, Sturmian words can be found in [7].

The structure of cdawg of a word w is closely related to right special factors of w (defined later on in the text). Such factors of *TM* words were already studied thoroughly in relation to the subword complexity function of the infinite *TM* word (i.e., the number of distinct factors of the word of a given length), see [4, 6, 9].

On the other hand, the vertices of cdawg of w can be seen as bispecial factors of w ; bispecial factors of the infinite TM word are characterized in [2, 5].


Using the special structure of cdawgs we present several unknown properties of Thue-Morse words as well as new (graph-based) proofs of some well-known properties. A slight modification of the compact dawg of the infinite Thue-Morse word yields an infinite graph with 2-counting property.

References

- 1 J.-P. Allouche and J. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. *Springer Ser. Discrete Math. Theor. Comput. Sci.*, pages 1–16, 1999.
- 2 L. Balkova, E. Pelantova, and W. Steiner. Return words in the Thue-Morse and other sequences. *arxiv:math/0608603v2*, 2006.
- 3 J. Berstel, A. Lauve, C. Reutenauer, and F. V. Saliola. *Combinatorics on Words: Christoffel Words and Repetitions in Words*. Amer. Mathematical Society, 2009.
- 4 S. Brlek. Enumeration of factors in the Thue-Morse word. *Discrete Applied Mathematics*, 24(1-3):83–96, 1989.
- 5 A. de Luca and L. Mione. On bispecial factors of the Thue-Morse word. *Inf. Process. Lett.*, 49(4):179–183, 1994.
- 6 A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theor. Comput. Sci.*, 63(3):333–348, 1989.
- 7 C. Epifanio, F. Mignosi, J. Shallit, and I. Venturini. On Sturmian graphs. *Discrete Applied Mathematics*, 155(8):1014–1030, 2007.
- 8 W. Rytter. The structure of subword graphs and suffix trees of Fibonacci words. *Theor. Comput. Sci.*, 363(2):211–223, 2006.
- 9 J. Tromp and J. Shallit. Subword complexity of a generalized Thue-Morse word. *Inf. Process. Lett.*, 54(6):313–316, 1995.

4.12 Enumeration and Automatic Sequences

Jeffrey Shallit (University of Waterloo, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Jeffrey Shallit

Main reference E. Charlier, N. Rampersad, and J. Shallit, Enumeration and decidable properties of automatic sequences, preprint, available at <http://arxiv.org/abs/1102.3698>.

URL <http://arxiv.org/abs/1102.3698>

In this talk, I discussed some new results, obtained with E. Charlier and N. Rampersad, on the decidability of properties of automatic sequences.

Here is a brief summary of our results:

1. Given a k -automatic sequence $\mathbf{x} = a(0)a(1)a(2) \cdots$, the sequence $\mathbf{b} = b(0)b(1)b(2) \cdots$ defined by $b(n) = 1$ if \mathbf{x} has an unbordered factor of length n and 0 otherwise, is also k -automatic.

2. The following questions are decidable:

- (a) given a k -automatic sequence, does it contain powers of arbitrarily large exponent?
- (b) given a k -automatic sequence, does it contain arbitrarily large unbordered factors?
- (c) given a k -automatic sequence, is it recurrent? linearly recurrent?

3. Many sequences counting properties of k -automatic sequences are k -regular, and constructively so. These include

- (a) the number of distinct factors of length n ;
- (b) the number of palindromic factors of length n ;
- (c) the number of unbordered factors of length n ;


and many other examples.

References

- 1 E. Charlier, N. Rampersad, and J. Shallit, Enumeration and decidable properties of automatic sequences, preprint, available at <http://arxiv.org/abs/1102.3698>.

4.13 Context Equivalence Problem

Arseny M. Shur (Ural State Univ. – Ekatarinenburg, RU)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Arseny M. Shur

The problem we study is a natural decision problem on words w.r.t. an arbitrary fixed language L . The instance is a pair of words; the problem is to decide whether these two words are “equally placed” in L in the sense that they have exactly the same contexts in L . By a context of a word v w.r.t. L we mean a pair (x, z) of words such that the word xvz belongs to L . From an algebraic point of view, the context equivalence problem is exactly the word problem in the syntactic monoid of the language L . It is little known about decidability and complexity of this interesting problem. Solving it for a given language L , we can get a lot of information about the internal structure of L .


We briefly explain some cases when this problem can be easily solved either because the language L is simple (the case of regular languages) or because the solution is trivial (the case of uniformly recurrent languages). Then we present a sophisticated, but linear-time solution of the context equivalence problem for the language of binary overlap-free words. Finally, we shortly discuss this problem for other power-free languages.

References

- 1 A. V. Klepinin. *On syntactic congruences of uniformly recurrent languages*, Proc. Ural State Univ. Ser. Computer Science. 2006. Vol. 1 (43). P. 38–44. [Russian]
- 2 A. M. Shur. *Syntactic semigroups of avoidable languages*, Sibirskii Matematicheskii Zhurnal. 1998. Vol. 39(3). P. 683–702. [Russian; Engl. Transl. in Siberian Math. J. 1998. Vol. 39(3). P. 594–610.]
- 3 A. M. Shur. *Deciding context equivalence of binary overlap-free words in linear time*, Semigroup Forum. 2011. (Submitted)

4.14 Hashing for Strings

Alexander J. Smola (Yahoo! Research – Santa Clara, US)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Alexander J. Smola

In this talk I discuss applications of hashing to fast computation of string similarity measures. For this purpose I first give an overview over string kernels using Suffix Trees, then I will discuss how hashing can deal with the problem of an ever increasing memory footprint for

suffix trees, simply by allowing collisions between its vertices. Applications to personalized spam filtering and approximate matching are provided to show the feasibility of this approach in practice.


References

- 1 Q. Shi, J. Petterson, G. Dror, J. Langford, A. J. Smola, A. L. Strehl, and V. Vishwanathan. Hash kernels. *Journal of Machine Learning Research - Proceedings Track*, 5:496–503, 2009.
- 2 S. V. N. Vishwanathan and Alexander J. Smola. Fast kernels for string and tree matching. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, pages 569–576. MIT Press, 2002.
- 3 K. Q. Weinberger, A. Dasgupta, J. Attenberg, J. Langford, and A. J. Smola. Feature hashing for large scale multitask learning. *CoRR*, abs/0902.2206, 2009.

5 Open Problems

5.1 Word Equations with Loops

Štěpán Holub (Charles University – Prague, CZ)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Štěpán Holub

In this talk I gave a short introduction into the problem of independent equivalent subsystems of word equations.

It is known (see [1] and [2]) that each infinite system of equations over a finite set of unknowns has a finite equivalent subsystem (where “equivalent” means having the same set of solutions).

On the other hand, little is known about possible size of such equivalent finite subsystems. It is even not known whether the size is bounded by a function in the number of unknowns.

I listed several examples where the bounds are known, including equations with particular looping properties, see [3] or [4].

In this framework I presented an open problem asking whether the following system of equations has a nontrivial solution for some number n of unknowns:

$$\begin{aligned}(x_1 \cdot x_2 \cdots x_n)^2 &= x_1^2 \cdot x_2^2 \cdots x_n^2 \\ (x_1 \cdot x_2 \cdots x_n)^3 &= x_1^3 \cdot x_2^3 \cdots x_n^3\end{aligned}$$

It is known that the answer is negative if, in addition to previous two equalities, also

$$(u_1 \cdot u_2 \cdots u_n)^4 = u_1^4 \cdot u_2^4 \cdots u_n^4$$


is required. For more details and bibliography see [5].

References

- 1 V. S. Guba. *The equivalence of infinite systems of equations in free groups and semigroups with finite subsystems*. *Mat. Zametki*, 40 (1986) 321-324
- 2 M. H. Albert and J. Lawrence. *A proof of Ehrenfeucht conjecture*. *Theoret. Comput. Sci.*, 41 (1985) 121–123
- 3 Š. Holub, *Local and global cyclicity in free semigroups*. *Theoret. Comput. Sci.*, 262 (2001) 25-36
- 4 Š. Holub and J. Kortelainen. *On systems of word equations with simple loop sets*. *Theoret. Comput. Sci.*, 380 (2007) 363-372
- 5 <http://www.karlin.mff.cuni.cz/~holub/soubory/prizeproblem.pdf>

5.2 Is morphic primitivity hereditary?

Štěpán Holub (Charles University – Prague, CZ)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Štěpán Holub

A word w is said to be morphically imprimitive if there is a nontrivial morphism f such that $f(w) = w$. Let δ_x , where x is a letter occurring in w , denote the morphism canceling x and being the identity on all other letters.

Prove or disprove the following claim, known as The Conjecture of Billaud:

If $\delta_x(w)$ is morphically imprimitive for all x occurring in w , then also w is imprimitive.


For more information and some partial results see [1].

References

- 1 F. Levé and G. Richomme. *On a conjecture about finite fixed points of morphisms*. Theor. Comput. Sci. 339(1)(2005) 103-128

5.3 Asymptotic Number of Long-Armed Palindromes in a Word

Gregory Kucherov (Université de Marne-la-Vallée, FR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Gregory Kucherov

Such a palindrome is called *long-armed* if $|v| \geq |u|$. Given a word w of length n , we are interested in all subwords of w that are long-armed palindromes. In [1], an algorithm has been proposed for computing all long-armed palindromes in time $O(n + S)$, where S is the size of the output, i.e. the number of long-armed palindromes found.

However, it is not known whether this number is linearly-bounded in n .


Trivially, for any fixed gap size ($|u|$) there can be a linear number of corresponding palindromes, as every position (or every letter) of w can be the center of only one palindrome. In a private communication after the Dagstuhl seminar, Jeffrey Shallit and Michael Domaratzki provided an example of a word containing order $3n$ long-armed palindromes. On the other hand, it is very easy to see that the number of long-armed palindromes is $O(n \log n)$. Proving (or refuting) the linear bound remains an open problem.

References

- 1 R. Kolpakov and G. Kucherov. Searching for gapped palindromes. *Theoretical Computer Science*, 410(51):5299–5382, 28 November 2009.

5.4 The Separating Words Problem

Jeffrey Shallit (University of Waterloo, CA)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Jeffrey Shallit

Main reference E. D. Demaine, S. Eisenstat, J. Shallit, and D. A. Wilson, Remarks on Separating Words. *Preprint*.
URL <http://arxiv.org/abs/1103.4513>

In this talk I discussed the separating words problem, as introduced by Goralčík and Koubek in 1986.

In this problem we are given two words w and x of length $\leq n$, and we want a good bound for the size of the smallest DFA that accepts one of $\{w, x\}$ and rejects the other.

If $|w| \neq |x|$ then w and x can be separated by a DFA of $O(\log n)$ states, so the only interesting case is where $|w| = |x|$.


I mentioned two new results:

(1) If the Hamming distance between w and x is $< d$, then w and x can be separated using $O(d \log n)$ states.

(2) There exists a sequence of words w, x such that nondeterministic separation is arbitrarily better than deterministic separation.

5.5 Some open problems inspired by Dejean's conjecture

Arseny M. Shur (Ural State Univ. – Ekatarinenburg, RU)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Arseny M. Shur

Joint work of Irina A. Gorbunova; Alexei V. Samsonov; Arseny M. Shur

Among the repetitions in words, fractional powers constitute one of the most natural classes. Any word, in which some prefix coincides with some suffix, is a fractional power. Each such power is numerically expressed by its exponent: $\exp(w) = |w|/\pi(w)$, where $|w|$ is the length of the word w , and $\pi(w)$ is the smallest period of w . A word v is β -free if no one of its factors has the exponent $\geq \beta$, and is β^+ -free if no one of its factors has the exponent $> \beta$. The repetition threshold on k letters, $RT(k)$, is the rational number γ_k such that the number of γ_k^+ -free k -ary words is infinite, while the number of γ_k -free k -ary words is only finite. The values of $RT(k)$ were conjectured by Dejean in 1972:

$$RT(3) = 7/4, \quad RT(4) = 7/5, \quad RT(k) = k/(k-1) \text{ for } k = 2 \text{ and } k \geq 5.$$

The conjecture is now confirmed in all cases, due to Dejean, Pansiot, Moulen-Ollagnier, Currie, Mohammad-Noori, Carpi, Rampersad, and Rao (1972 to 2009). The proof stimulated further research in several directions.

We mention three such directions and an open problem in each direction.

However, we are not pretending to present an exhaustive list of Dejean-like problems.

1. Strengthening original conjecture.

Problem: estimate the growth of the language of the k -ary $RT(k)^+$ -free language.

Exponential conjecture (folklore): all these languages have exponential growth. Confirmed by Ochem for $k = 3, 4$.

Growth rate conjecture (first stated by the author at JM 2008; see [2]): exponential growth rates of these languages tend to the limit $\alpha \approx 1.242$ as k approaches infinity.

2. Different notion of words.

Problem: find the analog(s) of repetition threshold for circular words.

There are three possible definitions of circular repetition threshold $CRT(k)$: weak (there are infinitely many circular $CRT(k)^+$ -free words), intermediate (there are circular $CRT(k)^+$ -free words of all but finitely many lengths), and strong (there are circular $CRT(k)^+$ -free words of any length). For $k = 2$, these bounds are 2, $7/3$, and $5/2$, respectively (Aberkane, Currie).

Conjecture for $k \geq 3$: weak and intermediate thresholds both coincide with $RT(k)$ (we have a proof for $k = 3$). The strong threshold is strictly bigger (for $k \geq 9$, it is at least $(k-3)/(k-4)$, as follows from the results of [3]).

3. Different notion of powers.

Problem: find the analog(s) of repetition threshold for Abelian powers.

We mention only one of several possible definitions of Abelian fractional powers. This definition was first given by Cassaigne and Currie and suits well for the powers less than 2. According to it, a word is Abelian β -free ($\beta < 2$) if it has no factors of the form xyz such that x and z are Abelian equivalent and $|xyz|/|xy| \geq \beta$.

Conjecture (first stated at JM 2010; see [1]): the Abelian repetition threshold (for the above definition of Abelian β -freeness) equals $9/5$ for $k = 4$ and $(k-2)/(k-3)$ for $k \geq 5$.

References

- 1 A. V. Samsonov, A. M. Shur. *On Abelian repetition threshold*, Proc. 13th Mons Days of Theoretical Computer Science. Univ. de Picardie Jules Verne, Amiens, 2010. P. 1–11.
- 2 A. M. Shur, I. A. Gorbunova. *On the growth rates of complexity of threshold languages*, RAIRO Inform. Theor. Appl. 2010. Vol. 44. P. 175–192.
- 3 A. M. Shur. *On the existence of minimal β -powers*, Proc. 14th Int. Conf. on Developments in Language Theory. Berlin: Springer, 2010. P. 411–422. (LNCS Vol. 6224).

6 Scientific Schedule

Monday

- 09:00–10:15 Tutorial: Combinatorics on Words – *Jeffrey Shallit*
 10:30–12:00 Tutorial: Machine Learning – *Mehryar Mohri*
 14:00–14:50 Hashing for Strings – *Alexander Smola*
 16:00–17:15 Tutorial: Natural Computing – *Hendrik Jan Hoogeboom*

Tuesday

- 09:00–10:30 Tutorial: Bioinformatics – *Rolf Backofen*
 10:50–11:45 Tutorial: Stringology (part 1) – *Roberto Grossi*
 11:45–12:30 Tutorial: Stringology (part 2) – *Alessandra Carbone*
 14:00–14:45 Non-Archimedean Words – *Volker Diekert*
 15:00–15:15 Open Problem – *Gad Landau*
 15:45–16:30 Sequence and Chromatin Signatures Predict Transcription Factor Binding in the Human Genome – *Christina Leslie*
 16:30–17:00 Combinatorial Measure of Co-evolving Blocks and their Evolutionary Pressure – *Linda Dib*
 19:30–20:00 Open Problems – *Volker Diekert, Štěpán Holub, Dirk Nowotka*

Wednesday

- 09:30–10:00 Real-Time, Constant Space String Matching – *Dani Breslauer*
 10:15–11:00 K-abelian Equivalence – *Juhani Karhumäki*
 11:15–12:00 Hairpin Completion versus Hairpin Lengthening – *Steffen Kopecki*
 12:00–12:15 Open Problems – *Štěpán Holub, Alexander Smola*

Thursday

- 09:30 - 10:15 Words and Permutations – *Antonio Restivo*
 10:30 - 11:15 Context Equivalence Problem – *Arseny Shur*
 11:30 - 12:15 Intelligent Strategies for Remote Homology Detection – *Juliana Bernardes*
 14:00 - 14:45 The Structure of Graphs Representing All Subwords of Thue-Morse Sequences – *Wojciech Rytter*
 14:45 - 15:30 Open Problems – *Gregory Kucherov, Jeffrey Shallit, Arseny Shur*
 15:45 - 16:30 Exact Ensemble Properties in Combinatorial Dynamic Programming Schemes – *Yann Ponty*

Friday

- 09:30 - 10:15 Energy Barrier Problem without Pseudo Knots – *Jan Mañuch*
 10:30 - 11:15 Polynomial Algorithm for Fixed Points of Nontrivial Morphisms – *Štěpán Holub*
 11:15 - 12:00 Some Decidable Properties of Automatic Sequences – *Jeffrey Shallit*

Participants

- Cyril Allauzen
Google – New York, US
- Rolf Backofen
University Freiburg, DE
- Marie-Pierre Beal
Univ. de Marne-la-Vallée, FR
- Juliana Bernardes
UPMC – Paris, FR
- Dany Breslauer
University of Haifa, IL
- Alessandra Carbone
UPMC – Paris, FR
- Corinna Cortes
Google – New York, US
- James Currie
University of Winnipeg, CA
- Alessandro De Luca
University of Napoli, IT
- Linda Dib
UPMC – Paris, FR
- Volker Diekert
University of Stuttgart, DE
- Mike Domaratzki
University of Manitoba, CA
- Roberto Grossi
University of Pisa, IT
- Stepan Holub
Charles University – Prague, CZ
- Hendrik Jan Hoogeboom
Leiden University, NL
- Costas S. Iliopoulos
King's College – London, GB
- Juhani Karhumäki
University of Turku, FI
- Juha Karkkainen
University of Helsinki, FI
- Steffen Kopecki
University of Stuttgart, DE
- Gregory Kucherov
Univ. de Marne-la-Vallée, FR
- Gad Landau
University of Haifa, IL
- Thierry Lecroq
Université de Rouen –
Mont-Saint-Aignan Cedex, FR
- Christina Leslie
Memorial Sloan-Kettering Cancer
Center – New York, US
- Jan Manuch
Simon Fraser University –
Burnaby, CA
- Mehryar Mohri
New York University, US
- Dirk Nowotka
University of Stuttgart, DE
- Enno Ohlebusch
Universität Ulm, DE
- Yann Ponty
Ecole Polytech. – Palaiseau, FR
- Svetlana Puzynina
University of Turku, FI
- Gunnar Rätsch
MPI für biologische Kybernetik –
Tübingen, DE
- Narad Rampersad
University of Liège, BE
- Antonio Restivo
Università di Palermo, IT
- Wojciech Rytter
University of Warsaw, PL
- Shinnosuke Seki
Univ. of Western Ontario, CA
- Jeffrey Shallit
University of Waterloo, CA
- Arseny M. Shur
Ural State Univ. –
Ekatarinenburg, RU
- Alexander J. Smola
Yahoo! Res. – Santa Clara, US
- Sören Sonnenburg
TU Berlin, DE
- German Tischler
Universität Würzburg, DE
- Chris J. Watkins
RHUL – London, GB

