Preface

A series of important applications of combinatorics on words has emerged with the development of computerized text and string processing, especially in biology and in linguistics. The aim of this volume is to present, in a unified treatment, some of the major fields of applications. The main topics that are covered in this book are

1. Algorithms for manipulating text, such as string searching, pattern matching, and testing a word for special properties.
2. Efficient data structures for retrieving information on large indexes, including suffix trees and suffix automata.
3. Combinatorial, probabilistic, and statistical properties of patterns in finite words, and more general pattern, under various assumptions on the sources of the text.
4. Inference of regular expressions.
5. Algorithms for repetitions in strings, such as maximal run or tandem repeats.
6. Linguistic text processing, especially analysis of the syntactic and semantic structure of natural language. Applications to language processing with large dictionaries.
7. Enumeration, generation, and sampling of complex combinatorial structures by their encodings in words.

This book is actually the third of a series of books on combinatorics on words. Lothaire's "Combinatorics on Words" appeared in its first printing in 1984 as Volume 17 of the Encyclopedia of Mathematics. It was based on the impulse of M. P. Schützenberger's scientific work. Since then, the theory developed to a large scientific domain. It was reprinted in 1997 in the Cambridge Mathematical Library. Lothaire is a *nom de plume* for a group of authors initially constituted of former students of Schützenberger. Along the years, it has enlarged to a broader community coordinated by the editors. A second volume of Lothaire's series, entitled "Algebraic Combinatorics on Words" appeared in 2002. It contains both complements and new developments that have emerged since the publication of the first volume.

The content of this volume is quite applied, in comparison with the two previous ones. However, we have tried to follow the same spirit, namely to present introductory expositions, with full descriptions and numerous examples. Refinements are frequently deferred to problems, or mentioned in Notes. There is presently no similar book that covers these topics in this way.

Although each chapter has a different author, the book is really a cooperative work. A set of common notation has been agreed upon. Algorithms are presented in a consistent way using transparent conventions. There is also a common general index, and a common list of bibliographic references.

This book is independent of Lothaire's other books, in the sense that no knowledge of the other volumes is assumed.

The book has been written with the objective of being readable by a large audience. The prerequisites are those of a general scientific education. Some chapters may require a more advanced preparation. A graduate student in science or engineering should have no difficulty in reading all the chapters. A student in linguistics should be able to read part of it with profit and interest.

Outline of contents.

The general organization is shown in Figure 0.1 and is described as follows.

The two first chapters are devoted to core algorithms. The first, "Algorithms on Words", is quite general, and is used in all other chapters. The second chapter, "Structures for Indexes", is fundamental for all advanced algorithmic treatment, and more technical.

Among the applications, a first domain is linguistics, represented by two chapters entitled "Symbolic Natural Language Processing" and "Statistical Natural Language Processing".

A second application is biology. This is covered by two chapters, entitled "Inference of Network Expressions", and "Statistics on Words with Applications to Biological Sequences".

The next block is composed of two chapters dealing with algorithmics, a subject which is of interest on its own in theoretical computer science, but is also related to biology and linguistics One chapter is entitled "Analytic Approach to Pattern Matching" and deals with generalized pattern matching algorithms. A chapter entitled "Periodic Structures in Words" describes algorithms used for discovering and enumerating repetitions in words.

A final block is devoted to applications to mathematics (and theoretical physics). It is represented by two chapters. The first chapter, entitled "Counting, Coding, and Sampling with Words" deals with the use of

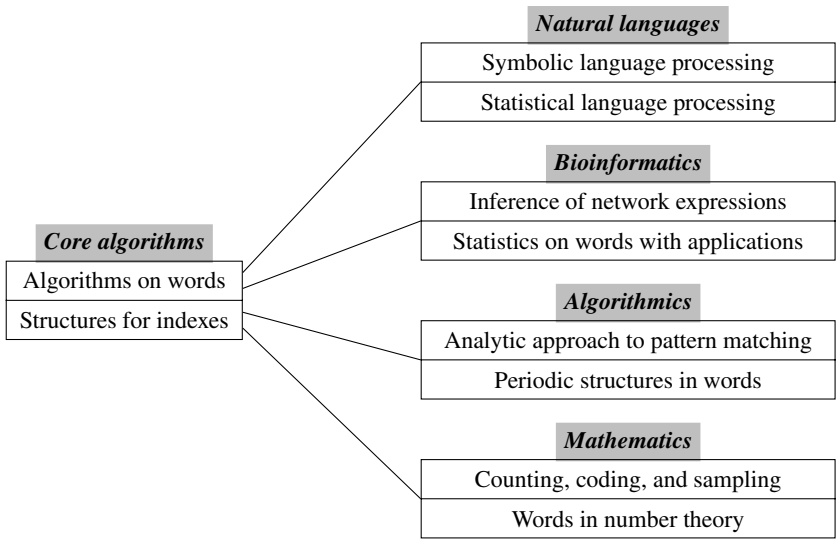

Figure 0.1. Overall structure of "Applied Combinatorics on Words".

words for coding combinatorial structures. Another chapter, entitled "Words in Number Theory" deals with transcendence, fractals, and dynamical systems.

Description of contents.

Basic algorithms, as needed later, and notation are given in Chapter 1 "Algorithms on Words", written by Jean Berstel and Dominique Perrin. This chapter also contains basic concepts on automata, grammars, and parsing. It ends with an exposition of probability distribution on words. The concepts and methods introduced are used in all the other chapters.

Chapter 2, entitled "Structures for Indexes" and written by Maxime Crochemore, presents data structures for the compact representation of the suffixes of a text. These are used in several subsequent chapters. Compact suffix trees are presented, and construction of these trees in linear time is carefully described. The theory and algorithmics for suffix automata are presented next. The main application, namely the construction of indexes, is described next. Many other applications are given, such as detection of repetitions or forbidden words in a text, use as a pattern matching machine, and search for conjugates.

The first domain of applications, linguistics, is represented by Chapters 3 and 4. Chapter 3, entitled "Symbolic Natural Language Processing" is

written by Eric Laporte. In language processing, a text or a discourse is a sequence of sentences; a sentence is a sequence of words; a word is a sequence of letters. The most universal levels are those of sentence, word, and letter (or phoneme), but intermediate levels exist, and can be crucial in some languages, between word and letter: a level of morphological elements (e.g. suffixes), and the level of syllables. The discovery of this piling up of levels, and in particular of word level and phoneme level, delighted structuralist linguists in the twentieth century. They termed this inherent, universal feature of human language as "double articulation".

This chapter is organized around the main levels of any language modelling: first, how words are made from letters; second, how sentences are made from words. It surveys the basic operations of interest for language processing, and for each type of operation, it examines the formal notions and tools involved. The main originality of this presentation is the systematic and consistent use of finite state automata at every level of the description. This point of view is reflected in some practical implementations of natural language processing systems.

Chapter 4, entitled "Statistical Natural Language Processing" is written by Mehryar Mohri. It presents the use of statistical methods to natural language processing. The main tool developed is the notion of weighted transducers. The weights are numbers in some semiring that can represent probabilities. Applications to speech processing are discussed.

The block of applications to biology is concerned with analysis of word occurences, pattern matching, and connections with genome analysis. It is covered by the next two chapters, and to some extent also by the algorithmics bloc.

Chapter 5, "Inference of Network Expressions", is written by Nadia Pisanti and Marie-France Sagot. This chapter introduces various mathematical models and algorithms for inferring regular expressions without Kleene star that appear repeated in a word or are common to a set of words. Inferring a network expression means to discover such expressions, which are initially unknown, from the word(s) where the repeated (or common) expressions will be sought. This is in contrast to the string searching problem considered in other chapters. It has many applications, notably in molecular biology, system security, text mining, etc. Because of the richness of the mathematical and algorithmical problems posed by molecular biology, we concentrate on applications in this area. Applications to biology motivate us also to consider network expressions that appear repeated not exactly but approximately.

Chapter 6 is written by Gesine Reinert, Sophie Schbath and Michael Waterman, and entitled "Statistics on Words with Applications to Biological

Sequences". Properties of words in sequences have been of considerable interest in many fields, such as coding theory and reliability theory, and most recently in the analysis of biological sequences. The latter will serve as the key example in this chapter.

Two main aspects of word occurrences in biological sequences are: where do they occur and how many times do they occur? An important problem, for instance, was to determine the statistical significance of a word frequency in a DNA sequence. The naive idea is the following: a word may be significantly rare in a DNA sequence because it disrupts replication or gene expression, (perhaps a negative selection factor), whereas a significantly frequent word may have a fundamental activity with regard to genome stability. Well-known examples of words with exceptional frequencies in DNA sequences are certain biological palindromes corresponding to restriction sites avoided for instance in *E. coli*, and the Cross-over Hotspot Instigator sites in several bacteria.

Statistical methods for studying the distribution of the word locations along a sequence and word frequencies have also been an active field of research; the goal of this chapter is to provide an overview of the state of this research.

Because DNA sequences are long, asymptotic distributions were proposed first. Exact distributions exist now, motivated by the analysis of genes and protein sequences. Unfortunately, exact results are not adapted in practice for long sequences because of heavy numerical calculation, but they allow the user to assess the quality of the stochastic approximations when no approximation error can be provided. For example, BLAST is probably the best-known algorithm for DNA matching, and it relies on a Poisson approximation. This is another motivation for the statistical analysis given in this chapter.

The algorithmics block is composed of two chapters. In Chapter 7, entitled "Analytic Approach to Pattern Matching", and written by Philippe Jacquet and Wojciech Szpankowski, pattern matching is considered for various types of patterns, and for various types of sources. Single patterns, sequences of patterns, and sequences of patterns with separation conditions are considered. The sources are Bernoulli and Markov, and also more general sources arising from dynamical systems. The derivation of the equations is heavily based on combinatorics on words and formal languages.

Chapter 9, written by Roman Kolpakov and Gregory Koucherov and entitled "Periodic Structures in Words", deals with the algorithmic problem of detecting, counting, and enumeration repetitions in a word. The interest for this is in text processing, compression, and genome analysis, where tandem repeats may have a particular significance. Linear time algorithms

exist for detecting tandem repeats, but since there may be quadratically many repetitions, maximal repetitions or "runs" are of importance, and are considered in this chapter.

A final block is concerned with applications to mathematics. Chapter 8, written by Dominique Poulalhon and Gilles Schaeffer, is entitled "Counting, Coding, and Sampling with Words". Its aim is to give typical descriptions of the interaction of combinatorics on words with the treatments of combinatorial structures. The chapter is focused on three aspects of enumeration: counting elements of a family according to their size, generating them uniformly at random, and coding them as compactly as possible by binary words. These aspects are respectively illustrated on examples taken from classical combinatorics (walks on lattices), from statistical physics (convex polyominoes and directed animals), and from graph algorithmics (planar maps). The rationale of the chapter is that nice enumerative properties are the visible traces of structural properties, and that making the latter explicit in terms of words of simple languages is a way of solving simultaneously and simply the above three problems.

Chapter 10 is written by Jean-Paul Allouche and Valérie Berthé. It is entitled "Words in Number Theory". This chapter is concerned with the interconnection between combinatorial properties of infinite words, such as repetitions, and transcendental numbers. A second part considers a famous infinite word, called the Tribonacci word, to investigate and illustrate connections between combinatorics on words and dynamical systems, quasicrystals, the Rauzy fractal, rotation on the torus, etc. Relations to the cut and project method are described, and an application to simultaneous approximation is given.

Acknowledgements.

Gesine Reinert, Sophie Schbath and Michael Waterman would like to thank Simon Tavaré for many helpful comments. Thanks go also to Xueying Xie for pointing out inconsistencies in a previous version concerning testing for the order of a Markov chain.

Their work was supported in part by Sandia National Laboratories, operated by Lockheed Martin for the U.S. Department of Energy under contract no. DE-AC04-94AL85000, and by the Mathematics, Information, and Computational Science Program of the Office of Science of the U.S. Department of Energy. Gesine Reinert was supported in part by EPSRC grant aGR/R52183/01. Michael Waterman was partially supported by Celera Genomics.

An earlier and shorter version of Chapter 6 appeared as "Probabilistic and statistical properties of words: an overview" in the *Journal of*

Computational Biology, Vol. 7 (2000), pp. 1–46. The authors thank Mary Ann Liebert, Inc. Publishers for permission to include that material here.

Philippe Jacquet and Wojciech Szpankowski thank J. Bourdon, P. Flajolet, M. Régnier and B. Vallée for collaborating on pattern matching problems, co-authoring papers, and commenting on this chapter. They also thank M. Drmota and J. Fayolle for reading the chapter and providing useful comments.

W. Szpankowski acknowledges NSF and NIH support through grants CCR-0208709 and R01 GM068959-01.

J.-P. Allouche and Valérie Berthé would like to express their gratitude to P. Arnoux, A. Rémondière, D. Jamet and A. Siegel for their careful reading and their numerous suggestions.

Jean Berstel
Dominique Perrin

Marne-la-Vallée, June 23, 2004