Alexander Bolshoy, Zeev (Vladimir) Volkovich, Valery Kirzhner, and Zeev Barzily

Genome Clustering

Studies in Computational Intelligence, Volume 286

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 265. Zbigniew W. Ras and Li-Shiang Tsay (Eds.)
Advances in Intelligent Information Systems, 2009
ISBN 978-3-642-05182-1

Vol. 266. Akitoshi Hanazawa, Tsutom Miki, and Keiichi Horio (Eds.)
Brain-Inspired Information Technology, 2009
ISBN 978-3-642-04024-5

Vol. 267. Ivan Zelinka, Sergej Celikovský, Hendrik Richter, and Guanrong Chen (Eds.)
Evolutionary Algorithms and Chaotic Systems, 2009
ISBN 978-3-642-10706-1

Vol. 268. Johann M.Ph. Schumann and Yan Liu (Eds.)
Applications of Neural Networks in High Assurance Systems, 2009
ISBN 978-3-642-10689-7

Vol. 269. Francisco Fernández de de Vega and Erick Cantú-Paz (Eds.)
Parallel and Distributed Computational Intelligence, 2009
ISBN 978-3-642-10674-3

Vol. 270. Zong Woo Geem
Recent Advances In Harmony Search Algorithm, 2009
ISBN 978-3-642-04316-1

Vol. 271. Janusz Kacprzyk, Frederick E. Petry, and Adnan Yazici (Eds.)
Uncertainty Approaches for Spatial Data Modeling and Processing, 2009
ISBN 978-3-642-10662-0

Vol. 272. Carlos A. Coello Coello, Clarisse Dhaenens, and Laetitia Jourdan (Eds.)
Advances in Multi-Objective Nature Inspired Computing, 2009
ISBN 978-3-642-11217-1

Vol. 273. Fatos Xhafa, Santi Caballé, Ajith Abraham, Thanasis Daradoumis, and Angel Alejandro Juan Perez (Eds.)
Computational Intelligence for Technology Enhanced Learning, 2010
ISBN 978-3-642-11223-2

Vol. 274. Zbigniew W. Raś and Alicja Wieczorkowska (Eds.)
Advances in Music Information Retrieval, 2010
ISBN 978-3-642-11673-5

Vol. 275. Dilip Kumar Pratihar and Lakhmi C. Jain (Eds.)
Intelligent Autonomous Systems, 2010
ISBN 978-3-642-11675-9

Vol. 276. Jacek Mańdziuk
Knowledge-Free and Learning-Based Methods in Intelligent Game Playing, 2010
ISBN 978-3-642-11677-3

Vol. 277. Filippo Spagnolo and Benedetto Di Paola (Eds.)
European and Chinese Cognitive Styles and their Impact on Teaching Mathematics, 2010
ISBN 978-3-642-11679-7

Vol. 278. Radomir S. Stankovic and Jaakko Astola
From Boolean Logic to Switching Circuits and Automata, 2010
ISBN 978-3-642-11681-0

Vol. 279. Manolis Wallace, Ioannis E. Anagnostopoulos, Phivos Mylonas, and Maria Bielikova (Eds.)
Semantics in Adaptive and Personalized Services, 2010
ISBN 978-3-642-11683-4

Vol. 280. Chang Wen Chen, Zhu Li, and Shiguo Lian (Eds.)
Intelligent Multimedia Communication: Techniques and Applications, 2010
ISBN 978-3-642-11685-8

Vol. 281. Robert Babuska and Frans C.A. Groen (Eds.)
Interactive Collaborative Information Systems, 2010
ISBN 978-3-642-11687-2

Vol. 282. Husrev Taha Sencar, Sergio Velastin, Nikolaos Nikolaidis, and Shiguo Lian (Eds.)
Intelligent Multimedia Analysis for Security Applications, 2010
ISBN 978-3-642-11754-1

Vol. 283. Ngoc Thanh Nguyen, Radoslaw Katarzyniak, and Shyi-Ming Chen (Eds.)
Advances in Intelligent Information and Database Systems, 2010
ISBN 978-3-642-12089-3

Vol. 284. Juan R. González, David Alejandro Pelta, Carlos Cruz, Germán Terrazas, and Natalio Krasnogor (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), 2010
ISBN 978-3-642-12537-9

Vol. 285. Roberto Cipolla, Sebastiano Battiato, and Giovanni Maria Farinella (Eds.)
Computer Vision, 2010
ISBN 978-3-642-12847-9

Vol. 286. Alexander Bolshoy, Zeev (Vladimir) Volkovich, Valery Kirzhner, and Zeev Barzily
Genome Clustering, 2010
ISBN 978-3-642-12951-3

Alexander Bolshoy, Zeev (Vladimir) Volkovich,
Valery Kirzhner, and Zeev Barzily

Genome Clustering

From Linguistic Models to Classification of Genetic Texts

Springer

Alexander Bolshoy
The Department of Evolutionary and Environmental Biology and the Institute of Evolution
University of Haifa, Haifa 39105
Israel
E-mail: bolshoy@research.haifa.ac.il

Valery Kirzhner
The Institute of Evolution
University of Haifa,
Haifa 39105
Israel
E-mail: valery@research.haifa.ac.il

Zeev (Vladimir) Volkovich
Software Engineering Department
ORT Braude College
P.O. Box: 78
Karmiel, 20101
Israel
E-mail: vlvolkov@ort.org.il

Zeev Barzily
Software Engineering Department
ORT Braude College
P.O. Box: 78
Karmiel, 20101
Israel
E-mail: zbarzily@ort.org.il

ISBN 978-3-642-12951-3 e-ISBN 978-3-642-12952-0

DOI 10.1007/978-3-642-12952-0

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2010926032

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Foreword

Knighting in sequence biology

Edward N. Trifonov

Genome classification, construction of phylogenetic trees, became today a major approach in studying evolutionary relatedness of various species in their vast diversity. Although the modern genome clustering delivers the trees which are very similar to those generated by classical means, and basic terminology is the same, the phenotypic traits and habitats are not anymore the playground for the classification. The sequence space is the playground now. The phenotypic traits are replaced by sequence characteristics, “words”, in particular. Matter-of-factually, the phenotype and genotype merged, to confusion of both classical and modern phylogeneticists.

Accordingly, a completely new vocabulary of stringology, information theory and applied mathematics took over. And a new brand of scientists emerged – those who do know the math and, simultaneously, (do?) know biology.

The book is written by the authors of this new brand. There is no way to test their literacy in biology, as no biologist by training would even try to enter into the elite circle of those who masters their almost occult language. But the army of informaticians, formal linguists, mathematicians humbly (or aggressively) longing to join modern biology, got an excellent introduction to the field of genome clustering, written by the team of their kin.

The analogy genomic sequences – texts is both an immediate simple thought, and an open door to the depths of genetic information and intricacies of its organization. The most fascinating and unique features of these texts are multiplicity, degeneracy and overlapping of various codes carried by the genetic sequences. In this respect mere transfer of techniques used for analysis of familiar “monocode” texts to the “polycode” sequences would be naïve. But no one would deny importance of such transfer, to begin with, to reveal, at least, the amazing specifics of the new reality. Another interesting aspect of the genomes is the uncertainty of the species’ formal definition. Already in classical genetics this was a stumbling block. The fertile progeny based definition of Dobzhansky[1], though broadly accepted, does not fit all diversity of species. In the genomics the matter becomes even more complicated, in particular, due to horizontal gene transfer. It appears that the species is not an elementary node of evolution. Rather, the gene, or (again uncertain) DNA segment in general, is the node.

Principally new techniques have to be introduced to cope with this very special language. The monograph is a rather comprehensive outline of the state of art in the field, introducing as well some original developments. The appreciation of the principal differences of the natural sequence language from all we knew before is an important merit of the book.

1. Dobzhansky, T.: Genetics and the Origin of Species. Columbia Univ. Press, New York (1937)

Preface

People like to compare and do it in a great variety of fields and with all kinds of objects. In particular, comparative biological studies of different species of living beings lead us to better understanding of known biological phenomena and even to novel discoveries in the biological science. In modern biology, species are often presented by their genomes. Thus, instead of comparing external organisms' features such as the length of the tail, the shape of the wings, etc., it is possible now to compare organisms' genomes, which are represented as long texts over the alphabet of four letters.

There exist different methods of analyzing texts which are written in human languages or composed of special symbols (e.g., computer programs). Although these methods had been developed long before the discovery of genetic texts, many of them are applicable to the genomic text analysis as well, and some are described in this book. However, there also exist methods which were not borrowed from the studies of natural languages, but were developed especially for the comparison of genomic texts.

This book deals with the methods of text comparison which are based on different techniques of converting the text into a distribution on a certain finite support, be it a genetic text or a text of some other type. Such distribution is usually referred to as "spectrum". The measure of dissimilarity of two texts is formally expressed as a certain "distance" between the spectra of these texts. Such definition implies that the similarity of the texts results from the similarity of the random processes generating the texts. It is obvious, thus, that the zero distance between two texts does not necessarily imply their letter-by-letter coincidence; for example, the texts may be just different implementations of the same process. The spectrum support usually represents a finite set of words. In a natural language, the latter may be the words of the language, while in a genetic text, particular patterns may be considered as words. The patterns range from the simplest signals to genes, which are parts of the genetic text. However, in the natural language analysis, formal, meaningless words, which are called *N*-grams, are also successfully employed. Since the repertoire of different patterns in genetic texts is relatively small, the use of *N*-grams for the genetic text analysis appears to be still more beneficial. In certain applications, the spectrum support may be a set of relative positions in the text, but in this case, too, the distribution value in each position is evaluated as some function of words which are connected, in a certain way, with each position. The fact that these are the words, whatever their definition may be, that are used as the basis for the spectrum evaluation, allows viewing the methods under consideration as a part of a more general field which may be called "DNA linguistics".

Genetic texts have certain features which are used for their analysis. The essential features, as well as some relevant information on the molecular biology of the cell, are presented in Chapter 1. Additionally, the reader can refer to several excellent introductory courses such as [8], [297] and [184].

Since this book is dedicated to the methods of genetic sequence comparison or, in other words, to a particular approach to genetic text classification, we review some classical general approaches to classification in Chapter 2. This chapter provides a brief introduction to the Linnaean classification system, to modern *taxonomy*, and to the field of molecular evolution called *phylogenetic systematics*. In the text of this book, we often compare the described results with the above classifications. The following books may be recommended for further reading on the topic of molecular evolution: [95], [204] and [200].

Chapter 3 provides a review of the main *data mining* models generating the text spectra which were developed for the analysis of texts written in natural languages. In particular, in the framework of some models, the coincidence (or similarity) of the spectra suggests the common author or the same topic of the two documents. The models which are based on the "letter-by-letter" comparison of texts are also described. They are further used in the book for constructing the spectra of genetic texts.

In Chapter 4, the questions are discussed as to the standpoint from which the DNA molecule can be viewed as a certain text and how this text can be evaluated in terms of formal grammar. Another essential question considered in the chapter concerns the process of creating genetic texts. While texts in natural languages are written by people, countless numbers of genetic texts (a unique text for each species and even, as it appears now, a unique text for each individual organism's genome) are "written" in the course of evolution. The models of special mechanisms which evolution uses for writing genetic texts are also described in the chapter. Obviously, the fact that *DNA texts* are, actually, the result of the evolution process should be employed for the comparison of these texts.

In Chapter 5, the particular case of *digrams* (N-gram with N=2) is described in detail, including the results of bacterial genome classification obtained by this method. Moreover, the concepts of *fuzzy N-grams* and of compositional spectra (CS) based on such N-grams are introduced. The evaluation of CS is a complicated computational problem; hence some plausible algorithms for its solution are also discussed in the chapter. Quite a few examples of genetic texts are employed to assess the properties and the biological appropriateness of different distance functions.

Chapter 6 elaborates on the *application of the CS* model to the genome classification; in particular, the optimal parameters of the model are obtained. Finally, two possible classifications of species "across life" are derived and their relevance to the standard classification is discussed.

In Chapter 7, a different *profile-based approach* to classification is presented. As a result of the suggested technique, the text is converted to a point in the

K-dimensional Euclidean space. The general description of the profile-construction method is followed by consideration of two important applications: in the first example, the *linguistic complexity* measure is employed, while in the second example, the measure based on *DNA curvature* is used.

In Chapter 8, the new approach to phylogenetics based on considering the whole-genome information is illustrated. This approach, called *phylogenomics*, is closely related to the main topic of the book since it also deals with embedding of the genome into a coordinate space. The sets of all the genes of particular prokaryotic genomes were used in the framework of the Information Bottleneck method adapted for genome clustering. The dendrogram of the genome classification obtained by this method represents, actually, a phylogenetic tree.

In Appendix A, the reader is introduced to the main ideas and techniques of the *clustering* approach to classification.

In Appendix B, a short review of three *sequence complexity* measure methods is compiled.

Appendix C is devoted to the introduction to the issue of *DNA curvature*.

The book is written by four co-authors, whose fields of expertise are close, but still represent different lines of research. Therefore, it would be virtually impossible to bring in harmony a great many details and maintain a uniform style of the text without the help of our persistent and careful scientific editor, Tanya Pyatigorskaya, PhD in molecular biophysics, to whom the authors express their deep gratitude.

Contents