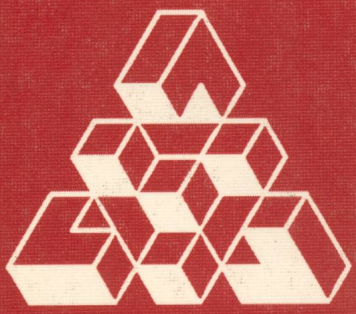
Jiří Fiala
Václav Koubek
Jan Kratochvíl (Eds.)

LNCS 3153

# Mathematical Foundations of Computer Science 2004

29th International Symposium, MFCS 2004
Prague, Czech Republic, August 2004
Proceedings



Springer

Jiří Fiala Václav Koubek
Jan Kratochvíl (Eds.)

# Mathematical Foundations of Computer Science 2004

29th International Symposium, MFCS 2004
Prague, Czech Republic, August 22-27, 2004
Proceedings



Springer

Volume Editors

Jiří Fiala
Jan Kratochvíl
Charles University, Department of Applied Mathematics
Faculty of Mathematics and Physics
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
E-mail: {fiala, honza}@kam.mff.cuni.cz

Václav Koubek
Charles University, Department of Theoretical Computer Science
and Mathematical Logic, Faculty of Mathematics and Physics
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
E-mail: koubek@ksi.mff.cuni.cz

Library of Congress Control Number: 2004109757

CR Subject Classification (1998): F.1, F.2, F.3, F.4, G.2, E.1

ISSN 0302-9743
ISBN 3-540-22823-3 Springer Berlin Heidelberg New York



Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 11310341 06/3142 5 4 3 2 1 0

# Preface

This volume contains the papers presented at the 29th Symposium on Mathematical Foundations of Computer Science, MFCS 2004, held in Prague, Czech Republic, August 22–27, 2004. The conference was organized by the Institute for Theoretical Computer Science (ITI) and the Department of Theoretical Computer Science and Mathematical Logic (KTIML) of the Faculty of Mathematics and Physics of Charles University in Prague. It was supported in part by the European Association for Theoretical Computer Science (EATCS) and the European Research Consortium for Informatics and Mathematics (ERCIM).

Traditionally, the MFCS symposia encourage high-quality research in all branches of theoretical computer science. Ranging in scope from automata, formal languages, data structures, algorithms and computational geometry to complexity theory, models of computation, and applications including computational biology, cryptography, security and artificial intelligence, the conference offers a unique opportunity to researchers from diverse areas to meet and present their results to a general audience.

The scientific program of this year's MFCS took place in the lecture halls of the recently reconstructed building of the Faculty of Mathematics and Physics in the historical center of Prague, with the famous Prague Castle and other celebrated historical monuments in sight. The view from the windows was a challenging competition for the speakers in the fight for the attention of the audience. But we did not fear the result: Due to the unusually tough competition for this year's MFCS, the admitted presentations certainly attracted considerable interest. The conference program (and the proceedings) consisted of 60 contributed papers selected by the Program Committee from a total of 167 submissions. These are accompanied in the proceedings by abstracts or full versions of the 10 invited talks. It has already become a tradition that EATCS offers a Best Student Paper Award for the best paper submitted to MFCS and authored solely by students. The winner of the award was announced during the conference.

As the editors of these proceedings, we would like to thank everyone who contributed to the success of the symposium and to its scientific merit. First of all the authors of the contributed papers for the record number of submissions, the invited speakers for accepting our invitation and sharing their knowledge and skills with us, the Program Committee members for their demanding and responsible work, their subreferees for careful reading of all the submissions, Springer-Verlag for excellent cooperation in the publication of this volume, and last but not least the Organizing Committee and Action-M Agency (our partner responsible for the local arrangements) for smooth running of the symposium. We hope the attendees all had a fruitful and enjoyable time in Prague.

August 2004

Jiří Fiala
Václav Koubek
Jan Kratochvíl

## Organizers

Faculty of Mathematics and Physics, Charles University, Prague

- Institute for Theoretical Computer Science
- Department of Theoretical Computer Science and Mathematical Logic

Action-M Agency (local arrangements)

## Supporters

European Association for Theoretical Computer Science (EATCS)
European Research Consortium for Informatics and Mathematics (ERCIM)
Institute for Theoretical Computer Science (ITI) — provided support for 11 students to participate at MFCS 2004

## Organizing Committee

Rostislav Caha
Ondřej Čepek
Jiří Fiala
Václav Koubek (Co-chair)
Antonín Kučera
František Mráz
Petr Štěpánek (Chair)

## Program Committee

| | |
|---|---|
| Manfred Broy | (Munich) |
| Bernard Chazelle | (Princeton) |
| Bruno Codenotti | (Pisa and Iowa City) |
| Josep Díaz | (Barcelona) |
| Peter van Emde Boas | (Amsterdam) |
| Michael Fellows | (Newcastle) |
| Irene Guessarian | (Paris) |
| Juhani Karhumaki | (Turku) |
| Václav Koubek | (Prague, Co-chair) |
| Dexter Kozen | (Cornell) |
| Dieter Kratsch | (Metz) |
| Jan Kratochvíl | (Prague, Chair) |
| Antonín Kučera | (Brno) |
| Janos Pach | (Budapest and New York) |
| Branislav Rovan | (Bratislava) |
| Georg Schnitger | (Frankfurt) |
| Paul Spirakis | (Patras) |
| Ondřej Sýkora | (Loughborough) |
| Jan Arne Telle | (Bergen) |
| Paul Vitanyi | (Amsterdam) |
| Peter Widmayer | (Zurich) |

## External Referees

Serge Abiteboul
Dimitris Achlioptas
Helmut Alt
Carme Àlvarez
Christoph Ambuhl
Luzi Anderegg
Albert Atserias
Jean-Michel Autebert
Volker Baier
Clemens Ballarin
Jiří Barnat
Eulalia Barriere
Andreas Bauer
Gertrud Bauer
Thomas Bayer
Marie-Pierre Béal
Daniéle Beauquier
Jean Berstel
Nicole Bidoit
Stephen L. Bloom
Nino Boccara
Ahmed Bouajjani
Gérard Boudol
Julian Bradfield
Andreas Brandstädt
Tomáš Brázdil
Herve Bronnimann
Wilfried Buchholz
Andrei Bulatov
Harry Buhrman
Olivier Carton
Didier Caucal
Patrick Cegielski
Ivana Černá
Chandra Chekuri
Christian Choffrut
Jan Chomicki
Anne Condon
Bruno Courcelle
Nadia Creignou
Eugen Czeizler
Artur Czumaj
Pierpaolo Degano
Jorg Derungs
Mariangiola Dezani-Ciancaglini
Volker Diekert
Rod Downey
Olivier Dubois
Pavol Duris
Jan Ernst
Panagiota Fatourou
Francesca Fiorenzi
Fedor Fomin
Lance Fortnow
Rusins Freivalds
Murdoch Gabbay
Paul Gastin
Viliam Geffert
Francoise Gire
Andreas Goerdt
Serge Grigorieff
Martin Grohe
Jozef Gruska
Peter Habermehl
Magnús Halldórsson
Hans-Dietrich Hecker
Jan Heering
Pinar Heggernes
Harald Hempel
Volker Heun
Daniel Hirschkoff
Ron Hirschorn
Hein van der Holst
Klaus Holzapfel
Markus Holzer
Mirko Horňák
Petr Jančar
Natasha Jonoska
Gabriel Juhas
Stasys Jukna
Marcin Jurdzinski
Jan Jürjens
Jarkko Kari
Hartmut Klauck
Jeroen Ketema
Astrid Kiehn
Lefteris Kirousis
Martin Klazar
Ondřej Klíma
Ton Kloks
Leonid Kof
Barbara König
Spyros Kontogiannis
Sven Kosub
Daniel Král'
Rastislav Kralovic
Evangelos Kranakis
Pavel Krčál
Mojmír Křetínský
Antonín Kučera
Gregory Kucherov
Armin Kuehnemann
Michal Kunc
Petr Kurka
Alexander Kurz
Jeff Lagarias
Klaus-Jörn Lange
Kim Larsen
Michel Latteux
Reinhold Letz
Leonid Libkin
Benedikt Löwe
Christos Makris
Maurice Margenstern
Jiří Matoušek
Richard Mayr
Catherine McCartin
Daniel Mölle
Kenichi Morita
Haiko Müller
Anca Muscholl
Rolf Niedermeier
Sotiris Nikoletseas
Tobias Nipkow
Marc Nunkesser
Jan Obdržálek
Vicky Papadopoulou
Daniël Paulusma
Leon Peeters
Radek Pelánek
Jean-Eric Pin

Wojciech Plandowski
Libor Polák
Wolfgang Prenninger
Andrzej Proskurowski
Rajiv Raman
Giovanni Resta
Stefan Richter
Zsuzsanna Roka
Jan Romberg
Peter Rossmanith
Joerg Rothe
Michel de Rougemont
Salvador Roura
Zdeněk Ryjáček
Kai Salomaa
Jacques Sakarovitch
Antonino Salibra
Robert Šámal
Massimo Santini
Bernhard Schaetz
Schmidt Schauss
Nicole Schweikardt
Thomas Schwentick
Alberto Segre
Géraud Senizergues
Olivier Serre
Jiří Sgall
Detlef Sieling
Daniel Singer
Anatol Slissenko
Jan Slovák
Maria Spichkova
Katharina Spies
Jiří Srba
Oldřich Stražovský
Gernot Stenz
Martin Strecker
Jan Strejček
Gabor Szabo
Dimitrios Thilikos
Wolfgang Thomas
Cesare Tinelli
Sophie Tison
Jacobo Torán
Daniele Varacca
Helmut Veith
Mahe Velauthapillai
Victor Vianu
Walter Vogler
Sergei Vorobyov
Imrich Vrt'o
Mirjam Wattenhofer
Birgitta Weber
Martin Wildmoser
Gerhard J. Woeginger
Ronald de Wolf
Christos Zaroliagis
Wieslaw Zielonka

## Previous MFCS

MFCS symposia have been organized in Czechia, Poland or Slovakia since 1972. The previous meetings took place in:

1972 Jablonna (Poland)
1973 Štrbské Pleso (Czechoslovakia)
1974 Jadwisin (Poland)
1975 Mariánské Lázně (Czechoslovakia)
1976 Gdańsk (Poland)
1977 Tatranská Lomnica (Czechoslovakia)
1978 Zakopane (Poland)
1979 Olomouc (Czechoslovakia)
1980 Rydzyna (Poland)
1981 Štrbské Pleso (Czechoslovakia)
1984 Praha (Czechoslovakia)
1986 Bratislava (Czechoslovakia)
1988 Karlovy Vary (Czechoslovakia)
1989 Porabka-Kozubnik (Poland)
1990 Banská Bystrica (Czechoslovakia)
1991 Kazimierz Dolny (Poland)
1992 Praha (Czechoslovakia)
1993 Gdańsk (Poland)
1994 Košice (Slovakia)
1995 Praha (Czech Republic)
1996 Kraków (Poland)
1997 Bratislava (Slovakia)
1998 Brno (Czech Republic)
1999 Szklarska Poreba (Poland)
2000 Bratislava (Slovakia)
2001 Mariánské Lázně (Czech Republic)
2002 Warsaw-Otwock (Poland)
2003 Bratislava (Slovakia)

# Table of Contents

## Invited Lectures



## Graph Algorithms

## Approximations



## Graphs and Complexity



## Circuits

## General Complexity



## Automata



## Parametrized and Kolmogorov Complexity

## Semantics



## Scheduling



## Algebraic Theory of Languages

## Games



## Languages



## Geometry



## Languages and Complexity

## Quantum Computing



## XML

# A Case Study of Genome Evolution: From Continuous to Discrete Time Model*

Jerzy Tiuryn[1], Ryszard Rudnicki[2], and Damian Wójtowicz[1]

[1] Institute of Informatics, Warsaw University
[2] Institute of Mathematics, Polish Academy of Sciences

**Abstract.** We introduce and analyse a simple model of genome evolution. It is based on two fundamental evolutionary events: gene loss and gene duplication. We are mainly interested in asymptotic distributions of gene families in a genome. This is motovated by previous work which consisted in fitting the available genomic data into, what is called *paralog distributions*. Two approaches are presented in this paper: continuous and discrete time models. A comparison of them is presented too – the asymptotic distribution for the continuous time model can be seen as a limit of the discrete time distributions, when probabilities of gene loss and gene duplication tend to zero. We view this paper as an intermediate step towards mathematically settling the problem of characterizing the shape of paralog distribution in bacterial genomes.

## 1 Introduction

Fitting data into various kinds of plots is a common practice of modern biology. A typical case is a study of genome organization and evolution, which can be viewed as a branch of a relatively new area of computational biology, called *comparative genomics* (see [9]). We can view a genome not simply as a set of genes, but rather as a dynamic collection of genes which changes in time. Various biochemical processes (e.g. point mutation, recombination, gene conversion, replication, DNA repair, translocation, horizontal transfer) constantly act on genomes and drive them to evolve dynamically. A problem which has been addressed in late 90's in this framework is an estimate of the distribution of paralogs in a genome. Two genes in a genome are said to be *paralogs* if they have evolved through duplication from a single ancestral gene. We do not discuss here the important issue of deciding which genes are paralogs. We assume that all genes have been clustered into groups of pairwise paralogous genes. The question which was asked in 1998 by P. Slonimski ([12,13]) and independently by M.A. Huynen and E. van Nimwegen [3] was about the distribution of the numbers of $i$-element clusters of paralogous genes (for consequtive $i$'s) in several microbial genomes which have been sequenced till then. The distribution was estimated by fitting

* This research was partially supported by the State Committee for Scientific Research (Poland) Grants No. 2 P03A 031 25, and 7 T11F 016 21 and by the EC programme *Centres of Excellence for States in phase of pre-accession*, No. ICA1-CT-2000-70024.

J. Fiala et al. (Eds.): MFCS 2004, LNCS 3153, pp. 1–24, 2004.

the available data. Since the method of deciding paralogy is only approximate and the size of the genomes was not large, as a consequence both authors came up with different answers: [12] claims that the distribution is logarithmic, while [3] claims that it follows the so called power law distribution. In 2001 Jordan et al. [4] have analysed 21 completely sequenced bacterial genomes and claimed that the logarithmic approximation fits the distribution slightly better than the power law approximation, although the difference between the two fits is not significant.

It should be obvious from the above description that it will be impossible to decide what actually is the observed distribution if we rely merely on the biological data. A decisive answer should come by adopting a certain mathematical model of genome evolution together with a rigorous analysis of the asymtotic distribution within this model. This is the main motivation for the present paper to build and analyse a simple model of genome evolution. The model we study is very simple indeed. It addresses only two evolutionary events: gene loss and gene duplication. Even though it is too simple to settle the problem of distribution of paralog families in genomes it can be used to study various subtleties of the model. We treat this paper as an intermediate step towards analysis of a more complicated model, which we postpone for future publication.

There is a short history of mathematical modeling of genome evolution. In 2000 Yanai et al. [16] designed a simple model of genome evolution based on random gene duplication and point mutations. The paper did not analyse the model. The main result consisted in showing that it is possible for each of the 20 microbal genomes to tune the parameters of the model so that the obtained distribution matches closely the paralog distribution of the genome. Recently Koonin's group has published in a series of papers [8,5,6,7] a simple model (called BDIM) of genome evolution which resembles our continuous time model. However, there are two important differences between the two models. In BDIM model in addition to gene loss and duplication there is an external source of new genes, called *invention*. This source is used to stabilize the asymptotic behavior of the model, i.e. to make sure that the supply of genes does not vanish at some point of evolution. On the contrary, we are interested in asymptotic distributions for the two extreme situtations: genome collapse and genome explosion. The reason for this is that if we assume that the two events: gene loss and gene duplication are independent of each other, it follows that we have to assume that their probabilities should not be equal. This leads the model to one of the two extreme situations. The second difference is more important. BDIM model sets an upper bound on the maximal size of gene family in the model. Technically speaking, this assumption implies that the system of differential equations is finite and the theory of finite dimentional matrices is applicable here. In the model which we investigate in this paper we do not impose any bound on the maximal size of a gene family and we end up with an infinite system of differential equations, for which existence of stationary solutions needs a special justification (see Theorem 1).

Continuous time model represents an ideal situation: in one unit of time two or more events can happen to a single gene, even though the probability of this is very low. In discrete time model we assume that in one unit of time every gene of the genome is subject to exactly one of the following events: removal, duplication, idle; each with a fixed probability. So, discrete time model is an approximation (and simplification) of the continuous model. Discrete model is much more suitable for computer simulations. Also, as we will see, the asymptotic distributions for both models are always different. The analysis of the discrete model is apparently more complicated, presumably due to lack of strong analytical tools. Moreover, as it follows from one of our results (see Theorem 7), the distribution for a continuous model can be seen as a limit of the discrete time distribution, when the probabilities of gene loss and duplication tend to zero. Another noteworthy property of the discrete model is a very nice isomorphism (see Theorem 8) between the situation of genome collapse (i.e. when Prob(gene duplication) $<$ Prob(gene loss)) and a genome explosion (when Prob(gene duplication) $>$ Prob(gene loss)). This allows us to reduce the latter situtation to the former. It appears that direct analysis of genome explosion is very difficult since the distribution looks more like a uniform distribution on an infinite set. The discrete model presented in this paper is in the same spirit as the model of DNA evolution presented in [14,15].

The paper is organized as follows. Section 2 contains a description of results for the continuous time model, together with asymptotic distributions for genome collapse (Theorem 2) and explosion (Theorem 3). Section 3 is devoted to discrete time. In particular we give a charcterization of a generating function for the asymptotic distribution for collapse (Theorem 6). All longer proofs are moved to the Appendix.

## 2 Continuous Time Model

Before we start a description of the genome evolution, let us introduce all entities used in our model: genes, gene families, class of gene families and genomes. *Genes* are atomic units, i.e. we do not assume any internal structure of these objects. A *genome* is a finite set of all genes. A *gene family* in a genome is a set of genes of that genome which are paralogs. We group families according to their size. *Classes* are sets of gene families which have the same number of elements, i.e. $S$ is a class $i$ if every family in $S$ has $i$ elements[1]. One gene duplication in a family belonging to class $i$ results in relocation of this family from class $i$ to class $i+1$. Conversely, one gene removal relocates the corresponding family to class $i-1$ or eliminates this family if $i = 1$. In this section, we consider time to be continous.

We assume that the probability of a gene duplication to happen during time interval of length $\Delta t$ is $d \cdot \Delta t + o(\Delta t)$. Similarly, probability of gene removal in time interval $\Delta t$ is $r \cdot \Delta t + o(\Delta t)$. It is assumed that $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$. Moreover, we assume that all elementary events (gene duplication and removal) are independent of each other.

[1] Obviously, a class may include families that are completely unreleted biologically.

Let $C_i(t)$ be the number of $i$-element families in our model at the time $t$. It follows from the description of our model that we have the following equations which describe an increment $\Delta C_i(t) = C_i(t+\Delta t) - C_i(t)$ during time interval $\Delta t$:

$$\Delta C_1(t) = -(d+r)C_1(t)\Delta t + 2rC_2(t)\Delta t + o(\Delta t)$$

and

$$\Delta C_i(t) = d(i-1)C_{i-1}(t)\Delta t - (d+r)iC_i(t)\Delta t + r(i+1)C_{i+1}(t)\Delta t + o(\Delta t),$$

for $i \geq 2$.

Hence, dividing both sides of the above equations by $\Delta t$ and passing with $\Delta t$ to 0, we obtain the following infinite system of differential equations:

$$C_i'(t) = d(i-1)C_{i-1}(t) - (d+r)iC_i(t) + r(i+1)C_{i+1}(t), \tag{1}$$

where $i \geq 1$. The above equation for $i = 1$ reduces to $C_1'(t) = -(d+r)C_1(t) + 2rC_2(t)$, independently of the value of $C_0(t)$. We assume that the latter is just $C_0(t) = 0$. Let us also observe that $C_i(t) = 0$, for $i \geq 1$ and $t \in \mathbb{R}$ is a (trivial) solution of (1).

**Theorem 1.** *If $r > 0$ and $d > 0$, then for each non-zero and non-negative absolutely summable sequence $(C_i(0))_{i\geq 1}$ equation (1) has a unique solution such that $C_i(t) > 0$ for all $t > 0$ and all positive integers $i$.*

The remainder of this section is devoted to the asymptotic behavior of a solution of (1), as $t \to \infty$. It turns out that the behavior of the system is quite different, depending on whether $r > d$, or $r < d$. In the former case all genes are eventually removed, while in the latter case we have an exponential explosion of the number of genes in the genome. We consider each case separately.

### 2.1 Collapse of the Genome: $r > d$

The next result characterizes an asymptotic behavior of solutions of (1) when $r > d$.

**Theorem 2.** *Let $(C_i(t))_{i\geq 1}$ be non-negative and non-zero solution of (1). If $r > d > 0$, then there exists a constant $c > 0$ such that for all $i \geq 1$,*

$$\lim_{t\to\infty} e^{(r-d)t} \cdot C_i(t) = c \cdot (\frac{d}{r})^i.$$

Hence for sufficiently large $t$ the number $C_i(t)$ of $i$ element gene families has the following asymptotics

$$C_i(t) \sim c \cdot e^{-(r-d)t} \cdot (\frac{d}{r})^i,$$

for all $i \geq 1$.