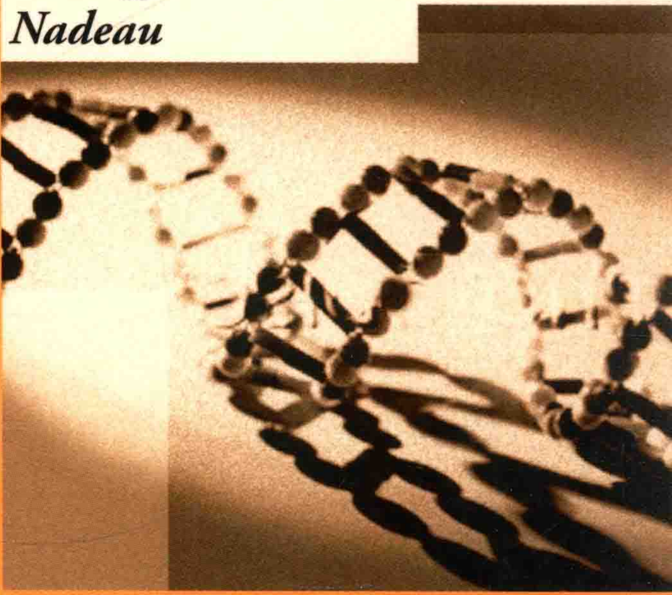
Computational
Biology Series
Volume 1

Edited by
David ***Sankoff***
Joseph H. ***Nadeau***



Comparative Genomics

Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families

Kluwer Academic Publishers

COMPARATIVE GENOMICS

EMPIRICAL AND ANALYTICAL APPROACHES TO GENE ORDER DYNAMICS, MAP ALIGNMENT AND THE EVOLUTION OF GENE FAMILIES

edited by

DAVID SANKOFF
Université de Montréal

and

JOSEPH H. NADEAU
Case Western Reserve University

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 0-7923-6583-6 (HB)
ISBN 0-7923-6584-4 (PB)

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AH Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Cover Design Boris Kessler
Digital Imagery@copyright 2000 Photodisc, Inc

Printed on acid-free paper

Printed in the Netherlands.

COMPARATIVE GENOMICS

Computational Biology

VOLUME 1

PREFACE

Genomic data may be analyzed in many ways. Most of these are extensions of methods previously applied to gene sequences or sequences of larger stretches of DNA. The global study of gene *order*, however, is meaningful only at the genomic level, and the advent of genomic sequencing has given a particular impetus to this approach. Though it has antecedents in long-standing traditions in genetics, the comparative analysis of gene order has seen rapid development over the last ten years, with the participation of scholars in a number of biological and mathematical sciences. In putting together this collection, we have been fortunate in recruiting key researchers, responsible for the most exciting current advances in the field, to contribute reports of their latest work or reviews and commentary on the newest trends. As a result, this volume is a compendium of the ongoing work on the processes that affect gene order, their consequences for evolution and analytical approaches to studying them.

These papers were discussed at the DCAF workshop (Gene Order Dynamics, Comparative Maps and Multigene Families) held at Le Chantecler in Sainte-Adèle, Québec on September 22–25, 2000. This was hosted and underwritten by the Centre de recherches mathématiques (CRM) of the Université de Montréal in the context of a theme year (2000-2001) on Mathematical Methods in Biology and Medicine.

This conference was the culmination of series of roundtables on the subject of genome rearrangements, including many of the same participants, the first organized by Pavel Pevzner and Mike Waterman in Los Angeles in March, 1994, and others at annual meetings (1995, 1998) of the Evolutionary Biology Program of the Canadian Institute for Advanced Research (CIAR). The role of the CIAR was fundamental to the DCAF workshop, not only because of a direct financial contribution, but because of the ongoing salary and interaction support that many of the participants receive from the Institute, fostering and facilitating research in this field for over ten years.

All the invited and submitted papers in this volume were read by the organizers or external referees; a few manuscripts were rejected and a good proportion underwent major revisions. We have not tried to standardize notation or even terminology, respecting the interdisciplinarity of the collection. Computer scientists use *reversal* to refer to the same process that biologists call *inversion*. *Translocation* usually refers to the exchange of genetic material between two chromosomes and *transposition* to the movement of such material within a single chromoosome, but these terms are some times used differently. *Synteny* refers sometimes to genes located on a common chromosome and sometimes to conserved order relations.

Random usually refers to the probabilistic component of any nondeterministic process but is sometimes taken to refer to a *uniform* distribution. *Homology* has many specific submeanings. These and other varying usages are disambiguated either explicitly or by the context in which they appear.

We have organized the papers into sections partly according to type of genome: organelles, prokarytes, higher eukaryotes. The first section, however, surveys mechanisms of genome rearrangement in various systems and at several levels of analysis, and most of the papers focusing on algorithms appear in a separate section. Papers on genome duplication and multigene families, a preoccupation relatively new within the rearrangements field but crucial to further development, are grouped together at the end of the collection.

Most of the papers could well have been placed in different sections; the mathematical papers touch on a number of specific biological questions, and some of the biologically-oriented papers contain new and significant analytical developments.

We have introduced each section with a discussion of related literature or comments on some of the issues which seem important for further research.

Thanks to Jacques Hurtubise and Martin Goldstein, Director and Deputy Director of the CRM, respectively, for incorporating the preparation of this manuscript into the CRM publication schedule, and to André Montpetit and his team of LaTeX experts Louise Letendre, Diane Brulé-DeFilippis, Diane Poulin and Fritz Pierre, for their professionalism and cooperation.

We mention with regret the loss of Mary Elizabeth Cosner (1994), Robert J. Cedergren (1998) and Susumu Ohno (2000), all of whose contributions were important to the early development of the field.

David Sankoff
Joseph H. Nadeau

FOREWORD

A reasonable complaint about genomics so far is that it has taught us a lot about genes, but very little about genomes. Publications describing genome sequences (really announcements in the form of scientific papers) usually read like telephone directories—but without the simple organizing principle which makes those at least useful. This is nobody's fault: the emphasis so far has been on data collection. Indeed, gene catalogs are an invaluable aid in understanding organismal biology, and gene by gene comparisons remain the best way to establish relationships among species. But still, there ought to be a science of whole genomes, genomology, perhaps. This science would address higher-order questions: what difference does genome organization make, how does it evolve, are there processes intrinsic to genomes which facilitate or retard phenotypic evolution, do these processes vary between groups (such as prokaryotes and eukaryotes) in ways which can explain differences in evolutionary trajectory, and so on.

Genomology, like all good biological sciences, will have to be comparative in its fundamental approach. Fortunately, I think, we have moved beyond the biopolitical research agenda which gave the genomic juggernaut its initial push—the human genome as holy grail, with other species genomes serving only a humble models, pilot projects along the way. So there will be more than enough data with which to establish the general rules and lineage-specific peculiarities of genome function and evolution at levels above that of the individual gene.

Will there be enough methods? This book, and the meeting which gave rise to it, give ample reason for confidence. Sankoff, Nadeau and their colleagues have recognized for several years that there is phylogenetic signal in gene order, if only we could measure it and reconstruct possible ancestral gene arrangements from comparative data. Much of this volume is devoted to showing that both pattern and process of genome rearrangements are indeed accessible from comparative data, and to addressing specific evolutionary questions with these new methods. Rearranging is of course not all that genes can do: polyploidization and the duplications of regions of genomes and genes are arguably what has made the evolution of complex life from simple life possible. A general theory is emerging here too, and several of this book's chapters articulate and apply it.

I am proud to say that the Canadian Institute for Advanced Research, through its support of a half-dozen of the scientists contributing to this volume (and through support of the meeting it reports) can claim some share of the credit for the existence and health of the new way of looking at genomes, indeed the new science of genomology, represented here.

W. Ford Doolittle
Fellow and Director
Canadian Institute for Advanced Research
Program in Evolutionary Biology

TABLE OF CONTENTS

Introduction

COMPARATIVE GENOMICS

David Sankoff
Joseph H. Nadeau

1. Toward a concerted approach to gene orders and comparative maps

The interdisciplinarity of the approach to comparative genomics exemplified in this volume is largely a matter of researchers from many disciplines studying the same biological phenomenon. It has not yet the case that techniques and concepts from two or more disciplines are being systematically combined to make discoveries about genomes or create new insights about rearrangement processes, although there are a few happy exceptions.

In particular, notions of genomic rearrangement, such as inversion and translocation, have been borrowed by theoretical computer scientists as the source of a series of new and challenging combinatorial problems. These problems and their solutions, partial and complete, well represented in Section 3 and elsewhere in this collection, form a whole new area within the field known as computational biology. Little of this has had any impact, however, on biological practice in studying rearrangements.

Even within empirical biology, the tradition of comparative mapping (Section 6) which has acquired its own statistical methodology over more than fifteen years (Section 5), does not always seem pertinent to researchers carrying out comparative studies of gene order (Section 4) on the complete genome sequences of prokaryotes now available.

The main goal of this collection is to marshall the range of expertise represented in this volume as a first step to a concerted attack on the central problem blocking our further understanding of gene order, its dynamics, and the evolving patterns seen in comparative maps. This problem is how conserved segments or gene clusters arise and/or disperse. In the short term, these can be explained in terms of the operations of a few rearrangement events, but to what extent are these events constrained by functional considerations? In the longer term, are some segments protected by functional relationships? Do some new clusters actually form over time in response to some selective forces? Biologists must realize

D. Sankoff and J.H. Nadeau (eds.), Comparative Genomics, 3–7.