

BIOINFORMATICS FOR BEGINNERS

BIOINFORMATICS FOR BEGINNERS

Genes, Genomes, Molecular
Evolution, Databases
and Analytical Tools

SUPRATIM CHOUDHURI

*With contribution from Dr Michael Kotewicz
on the Optical Mapping of DNA*

*Center for Food Safety and Applied Nutrition, FDA,
College Park, Maryland*



AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

2014 Published by Elsevier Inc.

The book was prepared by U.S. government employees in connection with their official duties, and therefore copyright protection is not available in the United States pursuant to 17 U.S.C. Section 105.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively, visit the Science and Technology Books website at www.elsevierdirect.com/rights for further information

Notice

The publisher and the author make no representations or warranties with respect to the accuracy and completeness of the contents of this work. No responsibility is assumed by the publisher and the author for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-410471-6

For information on all Academic Press publications
visit our website at elsevierdirect.com

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1



To my Family

Preface

As the title of the book suggests, this book is indeed for “beginners.” It is not intended for advanced students of bioinformatics or practicing bioinformaticians. This book has been written from the perspective of an end-user who wants to use the freely available web-based databases and tools for bioinformatic analysis. The audience of this book could include any scientist or student who has a background in basic molecular biology but has not used web-based databases and tools for sequence analysis, or has not done bioinformatic analysis on a regular basis. The total number of chapters is only nine. This is because related sections have been combined into one chapter for coherence and understanding. These sections could have been easily split into separate stand-alone chapters to increase the number of chapters.

More than a decade into the first human genome sequencing, the use of bioinformatic analysis has been steadily increasing. There are more web-based freely available databases and analytical tools than ever before. Modern biology has pervaded even the social sciences. For example, sociologists and psychologists are now probing how the epigenomic effects of environmental factors (including social factors) might shape the personality and behavior of the offspring postnatally. The National Center for Biotechnology Information has established an epigenomics database, which will be immensely useful to scientists in the near future. Thus, bioinformatics has been slowly but steadily pervading all branches of biology and beyond. In keeping with this, more and more bioinformatics books are being written for experts, which do not necessarily cater to the needs of the non-experts.

Because this book is about bioinformatic analysis using web-based databases and tools, the emphasis is on sequence analysis. Global gene-expression profiling has not been emphasized other than a short discussion. The makers of gene-expression analysis platforms provide necessary software for analysis. Lastly, it is not possible to show every type of analysis in a book with a defined word count; nor is it possible to discuss all the links and all the functions associated with a database or analysis. Therefore, this book should serve as an initial guide, and it is expected that the reader will take it upon himself/herself to explore further using the databases and tools. Terms such as program, tool, algorithm, and web server have been used interchangeably throughout the book. These terms essentially mean the same thing in the context of this book. However, the term web server could be used to mean both the hardware and the software.

Because the principal audience of the book is supposed to be non-specialists, it was felt necessary to introduce the science and some core concepts of genomics as well as some important genomic techniques before embarking on the bioinformatic analysis. By the same token, some fundamental aspects of molecular evolution have been discussed in this book because the goal of many applications of bioinformatics is to trace the signatures of molecular evolution, as well as study the relatedness of taxa. In order to minimize the number of references in the text, reviews are cited wherever possible.

Supratim Choudhuri

Acknowledgment

The author would like to acknowledge the invaluable contributions of all scientists and engineers who developed databases and online tools for analysis, and made them freely available. The author would also like to acknowledge the contributions of the groups/institutions/organizations for hosting and maintaining these resources on web servers. A number of links for freely available databases and web-based tools for analysis have been provided throughout the book. Wherever possible, the latest relevant publications (which usually include the previous publications as well) describing these resources have been cited to acknowledge the contribution. The scientific community is truly grateful to the developers of these

tools and databases and for making them freely available to facilitate bioinformatic analysis and learning.

The author would like to thank Dr Steve Gendel for his careful reading of the allergenicity prediction section in Chapter 8, and providing helpful suggestions.

The author would also like to thank many colleagues for their encouragement, enthusiasm, and support for the project.

Last but not the least, the author is grateful to Mr Graham Nisbet and Ms Catherine Mullane of Elsevier for making this project a reality, helping to bring it to successful completion, and being available whenever help and advice were needed.