Bioinformatics

Sequence and Genome Analysis

David W. Mount

University of Arizona, Tucson

COLD SPRING HARBOR LABORATORY PRESS
Cold Spring Harbor, New York
Bioinformatics

Sequence and Genome Analysis

All rights reserved
Printed in the United States of America

Developmental Editor Judy Cuddihy
Project Coordinator Joan Ebert
Production Editor Patricia Barker

Interior Designer Denise Weiss
Cover Designer Ed Atkeson, Berg Design

Library of Congress Cataloging-in-Publication Data

Mount, David W.

Bioinformatics : sequence and genome analysis / David W. Mount.
p. cm.
Includes bibliographical references and index.
QH441.2.M68 2000
572.8'633—dc21
00-060252

10 9 8 7 6 5 4 3 2 1

Front cover: Illustration inspired by the relationship of the 6000 genes in yeast to each other (Lisa Mount and Adam Sherman).

Students and researchers using the procedures in this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials.

Laser radiation, visible or invisible, can cause severe damage to the eyes and skin. Take proper precautions to prevent exposure to direct and reflected beams. Always follow manufacturers’ safety guidelines and consult your local safety office.

Procedures for the humane treatment of animals must be observed at all times. Check with the local animal facility for guidelines.

All WorldWideWeb addresses are accurate to the best of our knowledge at the time of printing.

Certain experimental procedures in this manual may be the subject of national or local legislation or agency restrictions. Users of this manual are responsible for obtaining the relevant permissions, certificates, or licenses in these cases. Neither the authors of this manual nor Cold Spring Harbor Laboratory assumes any responsibility for failure of a user to do so.

The polymerase chain reaction process is covered by certain patent and proprietary rights. Users of this manual are responsible for obtaining any licenses necessary to practice PCR or to commercialize the results of such use. COLD SPRING HARBOR LABORATORY MAKES NO REPRESENTATION THAT USE OF THE INFORMATION IN THIS MANUAL WILL NOT INFRINGE ANY PATENT OR OTHER PROPRIETARY RIGHT.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press, provided that the appropriate fee is paid directly to the Copyright Clearance Center (CCC). Write or call CCC at 222 Rosewood Drive, Danvers, MA 01923 (508-750-8400) for information about fees and regulations. Prior to photocopying items for educational classroom use, contact CCC at the above address. Additional information on CCC can be obtained at CCC Online at http://www.copyright.com/

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803-2500. Phone: 1-800-843-4388 in Continental U.S. and Canada. All other locations: (516) 349-1930. FAX: (516) 349-1946. E-mail: cshpress@cshl.org. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site http://www.cshl.org/
Contents

Preface xi

1 Historical Introduction and Overview 1

2 Collecting and Storing Sequences in the Laboratory 19

3 Alignment of Pairs of Sequences 51

4 Multiple Sequence Alignment 139

5 Prediction of RNA Secondary Structure 205

6 Phylogenetic Prediction 237

7 Database Searching for Similar Sequences 281

8 Gene Prediction 337

9 Protein Classification and Structure Prediction 381

10 Genome Analysis 479

Glossary 533

Index 547