SpringerBriefs in Electrical and Computer Engineering

For further volumes:
http://www.springer.com/series/10059

Jesse M. Lingeman · Dennis Shasha

Network Inference in Molecular Biology

A Hands-on Framework

Springer

Jesse M. Lingeman
New York University
4 Washington Place
New York, NY 10003
USA

Dennis Shasha
New York University
251 Mercer Street
New York, NY 10012
USA

ISSN 2191-8112 ISSN 2191-8120 (electronic)
ISBN 978-1-4614-3112-1 ISBN 978-1-4614-3113-8 (eBook)
DOI 10.1007/978-1-4614-3113-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012939634

© The Author(s) 2012
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

We would like to thank Alex Greenfield, Rich Bonneau, and Alex Rubinsteyn for their comments both technical and stylistic. We would also like to thank Juliana Freire for her help with workflow concepts and the Vistrails software.

This work has been partly supported by the US National Institutes of Health grant 2R01GM032877-25A1 and the US National Science Foundation grants MCB-0929339, IOS-0922738, and IIS - 1050388. That support is greatly appreciated.

Contents

1 Overview of Network Inference 1
1.1 Inferring Causality 1
1.1.1 Basics of Genomics 2
1.1.2 Observation, Intervention, and Inference 3
1.1.3 The Data 4
1.1.4 Data Sources 4
1.1.5 Basic Data Metrics 5
1.1.6 How to Read This Book 8
1.1.7 How to Use the Software 8

2 Step 1: Clustering Data 11
2.1 Introduction 11
2.1.1 Clustering 11
2.1.2 Biclustering 12
2.2 cMonkey 12
2.2.1 What it Does 12
2.2.2 The Data 12
2.2.3 The Strategy 12
2.2.4 Walkthrough Example on Toy Data 16
2.3 Factor Analysis for Bicluster Acquisition (FABIA) 18
2.3.1 What it Does 18
2.3.2 The Data 19
2.3.3 The Strategy 19
2.3.4 Walkthrough Example on Toy Data 21

3 Step 2: Use Steady State Data for Network Inference 23
3.1 Introduction 23
3.2 Median-Corrected Z-Scores (MCZ) 23
3.2.1 What it Does 23
3.2.2 The Data 24
3.2.3 The Strategy 24

3.2.4 Performance on Examples 25
3.3 Network Identification by Multiple Regression (NIR) 28
3.3.1 What it Does 28
3.3.2 The Data 28
3.3.3 The Strategy 29
3.3.4 Performance on Examples 30
3.4 Gene Network Inference with Ensemble of trees (GENIE3) 34
3.4.1 What it Does 34
3.4.2 The Data 34
3.4.3 The Strategy 34
3.4.4 Performance on Examples 37
3.5 Context Likelihood of Relatedness (CLR) 41
3.5.1 What it Does 41
3.5.2 The Data 41
3.5.3 The Strategy 41
3.5.4 Performance on Examples 42
3.6 Semidefinite Programming 46
3.6.1 What it does 46
3.6.2 The Data 46
3.6.3 The Strategy 46
3.6.4 Performance on Examples 48

4 Step 3: Using Time-Series Data 51
4.1 Introduction 51
4.2 Time-Delay ARACNE 51
4.2.1 What It Does 51
4.2.2 The Data 52
4.2.3 The Strategy 52
4.2.4 Performance on Examples 54
4.3 Time-Lagged Context Likelihood of Relatedness (tlCLR) 56
4.3.1 What it Does 56
4.3.2 The Data 56
4.3.3 The Strategy 56
4.3.4 Performance on Examples 58
4.4 Inferelator 62
4.4.1 What it Does 62
4.4.2 The Data 62
4.4.3 The Strategy 62
4.4.4 Performance on examples 64
4.5 Dynamic Factor Graphs (DFG) 65
4.5.1 What it does 65
4.5.2 The Data 65
4.5.3 The Strategy 65
4.5.4 Performance on examples 67
4.6 Bayesian Network Inference with Java Objects (BANJO) 72

4.6.1 What it does 72
4.6.2 The Data 72
4.6.3 The Strategy 73
4.6.4 Performance on Examples 74

5 Step 4: Pipelines 77
5.1 Consensus Step: Combining Results of Different Approaches 77
5.2 Creating Pipelines and Ensemble Networks 77
5.3 Pipeline 1: Steady State algorithm + Dynamic Factor Graphs 78
5.3.1 What it does 78
5.3.2 The Strategy 78
5.3.3 Performance on examples 78
5.4 Pipeline 2: Inferelator 2.0 82
5.4.1 What it does 82
5.4.2 The Strategy 82
5.4.3 Performance on examples 83
5.5 Ensemble 1: Voting 86
5.5.1 What it does 86
5.5.2 The Strategy 86
5.5.3 Performance on examples 87
5.6 Ensemble 2: Simulated Annealing 90
5.6.1 What it does 90
5.6.2 The Data 90
5.6.3 The Strategy 90
5.6.4 Performance on examples 91
5.7 Conclusions 95
References 97

Index 99