METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

The Gene Ontology Handbook

Edited by

Christophe Dessimoz

Department of Genetics, Evolution & Environment, University College London, London, UK; Swiss Institute of Bioinformatics, Lausanne, Switzerland; Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; Center of Integrative Genomics, University of Lausanne, Lausanne, Switzerland; Department of Computer Science, University College London, London, UK

Nives Škunca

Department of Computer Science, ETH Zurich, Zurich, Switzerland; SIB Swiss Institute of Bioinformatics, Zurich, Switzerland; University College London, London, UK

Humana Press

Springer Open

Editors
Christophe Dessimoz
Department of Genetics
Evolution and Environment
University College London
London, UK

Swiss Institute of Bioinformatics
Lausanne, Switzerland

Department of Ecology and Evolution
University of Lausanne
Lausanne, Switzerland

Center of Integrative Genomics
University of Lausanne
Lausanne, Switzerland

Department of Computer Science
University College London
London, UK

Nives Škunca
Department of Computer Science
ETH Zurich
Zurich, Switzerland

SIB Swiss Institute of Bioinformatics
Zurich, Switzerland

University College London
London, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-3741-7 ISBN 978-1-4939-3743-1 (eBook)
DOI 10.1007/978-1-4939-3743-1

Library of Congress Control Number: 2016943478

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC New York

Preface

The Gene Ontology (GO) is the leading project to organize biological knowledge on genes and their products in a formal and consistent way across genomic resources. This has had a profound impact at several levels. First, such standardization has made possible the integration of multiple resources and sources of knowledge, thereby increasing their discoverability and simplifying their usage. Second, it has greatly facilitated—some might say *exceedingly so*—data mining, aggregate analyses, and other forms of automated knowledge extraction. Third, it has led to an increase in the overall quality of the resources by enforcing minimum requirements across all of them.

Even considering these advantages, the rapid adoption of the GO in the community has been remarkable. In the 15 years since the publication of its introductory article [1], over 100,000 scientific articles containing the keyword "Gene Ontology" have been published and the rate is still increasing (Google Scholar).

However, despite this popularity and widespread use, many aspects of the Gene Ontology remain poorly understood [2], at times even by experts [3]. For instance, unbeknownst to most users, routine procedures such as GO term enrichment analyses remain subject to biases and simplifying assumptions that can lead to spurious conclusions [4].

The objective of this book is to provide a practical, self-contained overview of the GO for biologists and bioinformaticians. After reading this book, we would like the reader to be equipped with the essential knowledge to use the GO and correctly interpret results derived from it. In particular, the book will cover the state of the art of how GO annotations are made, how they are evaluated, and what sort of analyses can and cannot be done with the GO. In the spirit of the *Methods in Molecular Biology* book series in which it appears, there is an emphasis on providing practical guidance and troubleshooting advice.

The book is intended for a wide scientific audience and makes few assumptions about prior knowledge. While the primary target is the nonexpert, we also hope that seasoned GO users and contributors will find it informative and useful. Indeed, we are the first to admit that working with the GO occasionally brings to mind the aphorism "the more we know, the less we understand."

The book is structured in six main parts. Part I introduces the reader to the fundamental concepts underlying the Gene Ontology project, with primers on ontologies in general (Chapter 1), on gene function (Chapter 2), and on the Gene Ontology itself (Chapter 3).

To become proficient GO users, we need to know where the GO data comes from. Part II reviews how the GO annotations are made, be it via manual curation of the primary literature (Chapter 4), via computational methods of function inference (Chapter 5), via literature text mining (Chapter 6), or via crowdsourcing and other contributions from the community (Chapter 7).

But can we trust these annotations? In Part III, we consider the problem of evaluating GO annotations. We first provide an overview of the different approaches, the challenges associated with them, but also some successful initiatives (Chapter 8). We then focus on the more specific problem of evaluating enzyme function predictions (Chapter 9). Last, we

reflect on the achievements of the Critical Assessment of protein Function Annotation (CAFA) community experiment (Chapter 10).

Having made and validated GO annotations, we proceed in Part IV to use the GO resource. We consider the various ways of retrieving GO data (Chapter 11), how to quantify the functional similarity of GO terms and genes (Chapter 12), or perform GO enrichment analyses (Chapter 13)—all the while avoiding common biases and pitfalls (Chapter 14). The part ends with a chapter on visualizing GO data (Chapter 15) as well as a tutorial on GO analyses in the programming language Python (Chapter 16).

Part V covers two advanced topics: annotation extensions, which make it possible to express relationships involving multiple terms (Chapter 17), and the evidence code ontology, which provides a more precise and expressive specification of supporting evidence than the traditional GO annotation evidence codes (Chapter 18).

Part VI goes beyond the GO, by considering complementary sources of functional information such as KEGG and Enzyme Commission numbers (Chapter 19), and by considering the potential of integrating GO with controlled clinical nomenclatures (Chapter 20).

The final part concludes the book with a perspective by Suzi Lewis on the past, present, and future of the GO (Chapter 21).

London, UK *Christophe Dessimoz*
Zurich, Switzerland *Nives Škunca*

References

1. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
2. Thomas PD, Wood V, Mungall CJ et al (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. PLoS Comput Biol 8:e1002386
3. Dessimoz C, Škunca N, Thomas PD (2013) CAFA and the open world of protein function predictions. Trends Genet 29:609–610
4. Tipney H, Hunter L (2010) An introduction to effective use of enrichment analysis software. Hum Genomics 4:202–206

Acknowledgements

We thank all chapter authors for their contributions to the book. We are particularly indebted to Pascale Gaudet and Ruth Lovering for contributing multiple chapters and demonstrating unabated enthusiasm throughout the process. All chapters were reviewed by at least two independent peers, which represents a considerable effort. Peer reviews were mostly contributed by chapter authors, but also by the following people: Adrian Altenhoff, Natasha Glover, Debra Klopfenstein, Chris Mungall, Prudence Mutowo, Marc Robinson-Rechavi, Kimberly Van Auken, and Haibao Tang. Last but not least, we thank support by John Walker, our series editor, and Patrick Marton from Springer.

Funding for the Open Access charges was generously provided by the University College London Library, the Swiss Institute of Bioinformatics, the Agassiz Foundation, and the Foundation for the University of Lausanne.

Contents

Contributors

Eyal Akiva • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA*

Patricia C. Babbitt • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA*

Sebastian Bauer • *PRIVATE, Berlin, Germany*

Seth Carbon • *Berkeley Bioinformatics Open-Source Projects, Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*

Marcus C. Chibucos • *Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA*

Domenico Cozzetto • *Bioinformatics Group, Department of Computer Science, University College London, London, UK*

Rebecca Davidson • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA*

Spiros C. Denaxas • *Farr Institute of Health Informatics Research, University College London, London, UK; Institute of Health Informatics, University College London, London, UK*

Christophe Dessimoz • *Department of Genetics, Evolution & Environment, University College London, London, UK; Swiss Institute of Bioinformatics, Lausanne, Switzerland; Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; Center of Integrative Genomics, University of Lausanne, Lausanne, Switzerland; Department of Computer Science, University College London, London, UK*

Iddo Friedberg • *Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA*

Nicholas Furnham • *Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK*

Pascale Gaudet • *CALIPHO Group, Swiss Institute of Bioinformatics, Geneva, Switzerland; Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland*

Michelle Giglio • *Department of Medicine, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA*

Janna Hastings • *Cheminformatics and Metabolism, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridgeshire, UK*

Gemma L. Holliday • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA*

James C. Hu • *Department of Biochemistry and Biophysics, Texas A&M University and Texas AgriLife Research, College Station, TX, USA*

Rachael P. Huntley • *Functional Gene Annotation Initiative, Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, University College London, London, UK*

David T. Jones • *Bioinformatics Group, Department of Computer Science, University College London, London, UK*

Suzanna E. Lewis • *Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
Ruth C. Lovering • *Functional Gene Annotation Initiative, Centre for Cardiovascular Genetics, Institute of Cardiovascular Science, University College London, London, UK*
Monica Munoz-Torres • *Berkeley Bioinformatics Open-Source Projects, Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
Catia Pesquita • *LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*
Sylvain Poux • *Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland*
Predrag Radivojac • *Department of Computer Science and Informatics, Indiana University, Bloomington, IN, USA*
Richard J. Roberts • *New England Biolabs, Ipswich, MA, USA*
Patrick Ruch • *SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland; BiTeM Group, HES-SO\HEG Genève, Carouge, Switzerland*
Deborah A. Siegele • *Department of Biology, Texas A&M University, College Station, TX, USA*
Nives Škunca • *Department of Computer Science, ETH Zurich, Zurich, Switzerland; SIB Swiss Institute of Bioinformatics, Zurich, Switzerland; University College London, London, UK*
Martin Steffen • *Department of Biomedical Engineering, Boston University, Boston, MA, USA; Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA, USA*
Fran Supek • *Division of Electronics, Ruder Boskovic Institute, Zagreb, Croatia; EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain*
Paul D. Thomas • *Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA*
Alex Warwick Vesztrocy • *Department of Genetics, Evolution and Environment, University College London, London, UK; Swiss Institute of Bioinformatics, Lausanne, Switzerland*