

# SpringerBriefs in Computer Science

## *Series editors*

Stan Zdonik

Peng Ning

Shashi Shekhar

Jonathan Katz

Xindong Wu

Lakhmi C. Jain

David Padua

Xuemin Shen

Borko Furht

V. S. Subrahmanian

Martial Hebert

Katsushi Ikeuchi

Bruno Siciliano

For further volumes:

<http://www.springer.com/series/10028>

Wyatt Travis Clark

# Information-Theoretic Evaluation for Computational Biomedical Ontologies

 Springer

Wyatt Travis Clark  
Department of Molecular Biophysics  
and Biochemistry  
Yale University  
New Haven, CT  
USA

ISSN 2191-5768                      ISSN 2191-5776 (electronic)  
ISBN 978-3-319-04137-7            ISBN 978-3-319-04138-4 (eBook)  
DOI 10.1007/978-3-319-04138-4  
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013957360

© The Author(s) 2014

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](http://www.springer.com))

# Preface

The development of effective methods for the prediction of ontological annotations is an important goal in computational biology, with protein function prediction and disease gene prioritization gaining wide recognition. While various algorithms have been proposed for these tasks, evaluating their performance is difficult due to problems caused both by the structure of biomedical ontologies and biased or incomplete experimental annotations of genes and gene products. In this work, we propose an information-theoretic framework to evaluate the performance of computational protein function prediction. We use a Bayesian network, structured according to the underlying ontology, to model the prior probability of a protein's function. We then define two concepts, misinformation and remaining uncertainty, that can be seen as information-theoretic analogs of precision and recall. Finally, we propose a single statistic, referred to as semantic distance, that can be used to rank classification models. We evaluate our approach by analyzing the performance of three protein function predictors of Gene Ontology terms and provide evidence that we address several weaknesses of currently used metrics. We believe this framework provides valuable and useful insights into the performance of protein function prediction tools.

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Background	4
1.2	Protein Function Prediction Scenarios	6
1.3	State of the Art Methods	7
	References	7
<b>2</b>	<b>Methods</b>	13
2.1	Calculating the Joint Probability of a Graph	13
2.1.1	Calculating the Information Content of a Graph	16
2.1.2	Comparing Two Annotation Graphs	17
2.1.3	Measuring the Quality of Function Prediction	18
2.1.4	Weighted Metrics	20
2.1.5	Semantic Distance	20
2.1.6	Precision and Recall	21
2.1.7	Supplementary Evaluation Metrics	22
2.1.8	Additional Topological Metrics	25
2.2	Confusion Matrix Interpretation of $ru$ and $mi$	25
2.3	Annotation Models	26
2.3.1	The Naïve Model	26
2.3.2	The BLAST Model	27
2.3.3	The GOtcha Model	27
	References	27
<b>3</b>	<b>Experiments and Results</b>	29
3.1	Average Information Content of a Protein	29
3.2	Comparative Examples of Calculating Information Content	30
3.3	Two-Dimensional Plots	33
3.4	Comparisons of Single Statistics	35
	References	40
<b>4</b>	<b>Discussion</b>	43
	References	44
	<b>Index</b>	45