BestMasters

Thorsten Will

Predicting Transcription Factor Complexes

A Novel Approach to Data Integration in Systems Biology

Springer Spektrum

BestMasters

Springer awards „BestMasters“ to the best master's theses which have been completed at renowned universities in Germany, Austria, and Switzerland.
The studies received highest marks and were recommended for publication by supervisors. They address current issues from various fields of research in natural sciences, psychology, technology, and economics.
The series addresses practitioners as well as scientists and, in particular, offers guidance for early stage researchers.

Thorsten Will

Predicting Transcription Factor Complexes

A Novel Approach to Data Integration in Systems Biology

Springer Spektrum

Thorsten Will
Saarbrücken, Germany

BestMasters
ISBN 978-3-658-08268-0 ISBN 978-3-658-08269-7 (eBook)
DOI 10.1007/978-3-658-08269-7

Library of Congress Control Number: 2014956553

Springer Spektrum
© Springer Fachmedien Wiesbaden 2015
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.
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. The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Spektrum is a brand of Springer Fachmedien Wiesbaden
Springer Fachmedien Wiesbaden is part of Springer Science+Business Media
(www.springer.com)

Geleitwort des Betreuers

Die genetische Information jeder biologischen Zelle ist bekanntlich in deren Erbsubstanz kodiert. Insofern gibt es beispielsweise keinerlei Unterschiede zwischen den etwa 200 verschiedenen Zelltypen eines Menschen (Nervenzellen, Muskelzellen, Knochen, Haut etc). Obwohl diese Zellen natürlich große morphologische Unterschiede aufweisen, enthalten alle Zellen im Prinzip dieselbe Information. Entscheidend für den jeweiligen Zustand einer menschlichen Zelle - d.h. deren Differenzierung in die einzelnen Gewebetypen - ist deshalb nicht, welche Informationen die Erbsubstanz prinzipiell enthält, sondern welche der etwa 22.000 Gene in ihr tatsächlich „abgelesen" werden und welche nicht. Man kann dies vereinfacht mit dem Lesen eines Buches vergleichen, bei dem auch jeweils eine Seite aufgeschlagen wird. Wer daher in der Zelle diesen Lese-Prozess reguliert, bestimmt quasi das Schicksal der Zelle. Die Entschlüsselung dieser Prinzipien zur Regulierung der Gentranskription ist somit ein sehr wichtiger Schlüssel zum Verständnis von Zellen.

Die wichtigste Rolle bei der Entscheidung, welche Gene abgelesen werden sollen, übernehmen Eiweißmoleküle, sogenannte Transkriptionsfaktoren, die gezielt an Bindungsstellen an der Erbsubstanz binden können. Wenn diese Transkriptionsfaktoren direkt „vor" dem Beginn einer Gensequenz an die Erbsubstanz binden, kann der Kopierapparat der Zelle, die RNA-Polymerase, gezielt an diese Region rekrutiert werden und beginnt, die sich anschließende Region der Erbsubstanz zu kopieren, d.h. abzulesen.

Man könnte sich nun vorstellen, dass zu jedem einzelnen Gen ein bestimmter Transkriptionsfaktor gehört, der den Ableseprozess dieses Gens reguliert. Jedoch müsste ja auch dieser Transkriptionsfaktor in der Zelle durch Ablesen eines anderen Gens produziert werden, so dass man dann zweimal so viele Gene bräuchte. Und so weiter. Man sieht leicht, dass solch eine Variante nicht funktionieren kann. Eine andere Möglichkeit wäre, dass ein einzelner Transkriptionsfaktor jeweils das Ablesen von vielen anderen Genen reguliert. Dann käme man mit einer wesentlich geringeren Anzahl an Transkriptionsfaktoren aus, könnte aber die Regulation der einzelnen Gene nicht mehr so feinkörnig steuern. Eine dritte Möglichkeit, die nun

tatsächlich in Zellen realisiert wird, ist dass jeweils mehrere Transkriptionsfaktoren gemeinsam den Ableseprozess von einzelnen Genen kontrollieren. Damit reichen wenige hunderte an Transkriptionsfaktoren aus um eine enorme Anzahl an kombinatorischen Varianten zu erzeugen.

Der Autor entwickelte in seiner Masterarbeit im Fachgebiet Bioinformatik einen neuen Ansatz um Eiweißkomplexe zu identifizieren, die aus mehreren Transkriptionsfaktoren sowie aus weiteren Proteinen bestehen. Um die Praktikabilität der Methode zu testen, wurde als Modellorganismus die Bäckerhefe (S. cerevisiae) ausgewählt, da hierfür besonders gute experimentelle Daten zu paarweisen Proteininteraktionen vorliegen. Der Algorithmus verwendet die in der Informatik oft eingesetzte Baumstruktur zur Aufzählung aller möglichen Komplexe, die an die Erbsubstanz binden und maximal 10 Eiweißmoleküle enthalten. Ein neuartiger Beitrag bestand darin, die grundlegende Datenstruktur für die Interaktionen nicht auf gesamten Proteinen aufzusetzen, sondern auf deren Domänenbausteinen. So konnten zum einen weitere Interaktionsdaten zwischen Proteindomänen eingebunden werden. Zum anderen ergibt sich eine feinere strukturelle Auflösung der in Konkurrenz miteinander stehenden Kontakte. Mit dem neuen Ansatz konnten für Hefe mehr als 10 mal so viele unterschiedliche Proteinkomplexe generiert werden wie mit anderen derzeit verfügbaren Methoden. Der Autor zeigte zudem, dass die Ergebnisse eine bessere Abdeckung der bisher experimentell charakterisierten Komplexe liefern als alle anderen Methoden und dass die vorhergesagten Komplexe eine hohe biologische Plausibilität besitzen.

Diese vielversprechenden Ergebnisse lassen es denkbar erscheinen, ähnliche Methoden auch für komplexere Lebewesen wie die Maus oder sogar den Mensch einzusetzen. Dies wäre ein wichtiger Schritt dabei, die Mechanismen der Genregulation besser zu verstehen, da deren Fehlfunktionen natürlich auch zur Entstehung vieler Krankheiten beitragen. Die Bioinformatik übernimmt bei solchen Projekten meist die wichtige Aufgabe der Datenintegration und ermöglicht es, die Existenz bestimmter Szenarien oder Mechanismen zu postulieren, deren Korrektheit dann im Experiment gezielt getestet werden kann.

Prof. Dr. Volkhard Helms

Institutsprofil

Das Zentrum für Bioinformatik an der Universität des Saarlandes[1] ist eine interfakultäre Einrichtung zwischen der Informatik, Medizin und den Lebenswissenschaften Biologie, Chemie und Pharmazie. Ein wichtiger Schwerpunkt der Saarbrücker Bioinformatik-Forschung ist die medizinische Bioinformatik, d.h. die Anwendung von Bioinformatik-Methoden für die Bearbeitung von biomedizinischen Daten. Man interessiert sich heutzutage zum Beispiel dafür, wie sich die Erbsubstanz in Krebszellen von der Erbsubstanz in daneben liegenden gesunden Zellen unterscheidet. Ein anderes wichtiges Forschungsgebiet ist die Resistenzforschung, wie es Viren und Bakterien durch Veränderung ihres Erbgutes schaffen, die Wirkung von an sich hoch aktiven antiviralen oder antibakteriellen Wirkstoffen ins Leere laufen zu lassen.

Das Zentrum für Bioinformatik organisiert an der Universität des Saarlandes einen grundständigen Bachelorstudiengang Bioinformatik und einen darauf aufbauenden Masterstudiengang. Basierend auf den anerkannten Forschungserfolgen, aber auch befördert durch die Anbindung an die exzellenten Forschungsinstitute der Saarbrücker Informatik und durch die hohe gesellschaftliche Relevanz der bearbeiteten Forschungsthemen, genießt die Saarbrücker Bioinformatik weltweit eine hohe Reputation. Die Absolventen der beiden genannten Studiengänge sind in nationalen und internationalen Unternehmen sowie in Forschungsinstituten sehr stark nachgefragt.

[1] www.zbi.uni-saarland.de

Preface

Gene regulatory networks are fixed determinants of cellular control and the abundance of differentially expressed regulatory proteins, called transcription factors, their driving signal. In concert with specific epigenetic marks, transcription factors define the active subset of the network to govern distinct cellular states in time and space.
Eukaryotic gene expression is generally controlled through molecular logic circuits combining regulatory signals of several transcription factors. Recently, it has been shown that complexation of regulatory proteins is a prevailing and highly conserved mechanism of signal integration within critical regulatory pathways, like body part formation or differentiation. A knowledge of potential assembly candidates could provide the basic information that is needed to infer possible target genes as well as the exerted mechanism of influence. There already exists a plethora of approaches to predict protein complexes from protein-protein interaction data. However, those are generally designed to detect large self-contained functional complexes and lack the ability to reveal dynamic and highly modular combinatorial complex assemblies, a property of crucial importance for the signal integration exerted by transcription factor complexes.

The method proposed in this thesis combines protein-protein interaction networks and domain-domain interaction networks with the well-known cluster-quality metric cohesiveness. A novel growth algorithm is described that locally maximizes the metric on the holistic level of protein interactions while sophisticated connectivity constraints are preserved. Assuming that each domain can only support one interaction, the domain topology can be utilized to account for the exclusive and thus combinatorial nature of physical interactions between proteins. During the growth process, the complex candidate is thought to be backed by a spanning tree of simultaneously possible domain interactions which restrict further expansion possibilities. Consequently, every addition of a protein requires the choice of an applicable domain interaction which again influences later steps. Often many options have to be taken into account by branching of the algorithm, which naturally allows for the justified prediction of a manifold of transcription factor complexes from a common start.

The proposed approach outperformed popular complex prediction methods by far for the prediction of transcription factor complexes in yeast. The evaluation was based on established benchmarks assessing accordance with several reference complex datasets as well as measures of biological relevance. Additionally, many of the predictions of the proposed method could be associated with target genes and a potential regulatory effect. Furthermore, predicted candidates could be mapped to distinct functions during a defined cellular state and condition by analyzing the expression coherence among their regulated genes for cell cycle expression data. Many findings were backed up by literature evidence.

The results encourage an application to higher eukaryotes where the combinatorial interplay between transcription factors is more pronounced. The knowledge of putative transcription factor complexes - DNA-binding members and recruited potentially regulatory active proteins - offers novel capabilities in the automatized modeling of gene regulatory networks which may assist to surpass nowadays models.

A condensed summary of the novel concept, the main method and the results for yeast was previously published in [1] prior to the production of this book.

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

bp	base pair(s)
CRM	cis-regulatory module
DDI(N)	domain-domain interaction (network)
EC(S)	expression coherence (score)
GO	Gene Ontology (annotation)
GTF	general transcription factor
HMM	hidden markov model
MMR	maximum matching ratio
ORF	open reading frame
PDB	Protein Data Bank
Pol II	RNA-Polymerase II
PPI(N)	protein-protein interaction (network)
SPIN	simultaneous protein(-protein) interaction network
TBP	TATA-binding protein
TF	transcription factor
TSS	transcriptional start site
YPA	Yeast Promotor Atlas