Trey Ideker
Vineet Bafna (Eds.)

LNBI 4532

Systems Biology and Computational Proteomics

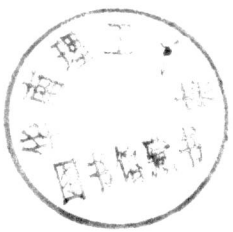
Joint RECOMB 2006 Satellite Workshops on Systems Biology and on Computational Proteomics San Diego, CA, USA, December 2006 Revised Selected Papers

Springer

Trey Ideker Vineet Bafna (Eds.)

Systems Biology and Computational Proteomics

Joint RECOMB 2006 Satellite Workshops
on Systems Biology and on Computational Proteomics
San Diego, CA, USA, December 1-3, 2006
Revised Selected Papers



Springer

Volume Editors

Trey Ideker
University of California
Department of Bioengineering
San Diego, CA 92093, USA
E-mail: tideker@ucsd.edu

Vineet Bafna
University of California
Computer Science and Engineering Dept.
San Diego, CA 92093, USA
E-mail: vbafna@cs.ucsd.edu

Library of Congress Control Number: 2007931338

CR Subject Classification (1998): F.2, G.3, E.1, H.2.8, J.3

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN	0302-9743
ISBN-10	3-540-73059-1 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-73059-0 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12076215 06/3180 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4532

Edited by S. Istrail, P. Pevzner, and M. Waterman

Editorial Board: A. Apostolico S. Brunak M. Gelfand
T. Lengauer S. Miyano G. Myers M.-F. Sagot D. Sankoff
R. Shamir T. Speed M. Vingron W. Wong

Subseries of Lecture Notes in Computer Science

Preface

The RECOMB Satellite Conferences on Systems Biology and Computational Proteomics were held December 1–3, 2006, at La Jolla, California. The Systems Biology meeting brought researchers together on various aspects of systems biology, including integration of genome-wide microarray, proteomic, and metabolomic data, inference and comparison of biological networks, and model testing through design of experiments. Specific topics included:

- Pathway mapping and evolution in protein interaction networks
- Inference of protein signaling networks for understanding cellular responses and developmental programs
- Model prediction of drug mechanism of action and toxicity
- Multi-scale methods which bridge abstract and detailed models
- Systematic design of genome-scale experiments
- Modeling and recognition of regulatory elements
- Identification and modeling of *cis*-regulatory regions
- Modeling the structure and function of regulatory regions
- Comparative genomics of regulation

With the sequencing of the genome, and subsequent identification of the parts list (the gene and their protein products), there is a renewed emphasis on studying the proteome. This year, the computational proteomics meeting focused on on computational mass spectrometry. Mass spectrometry is emerging as a key technology for proteomics. The last few years have seen tremendous improvement in the quality and quantity of available peptide mass spectrometry data, as well as the realization that advanced computational approaches are critical to the success of this technology. The conference explored the use of this technology in various proteomic applications, including, but not limited to: protein identification and quantification in specific cellular environments; structural genomics; networks of protein interaction; post-translational modifications; and others.

We received approximately 50 full paper submissions to the joint workshops. After review, a total of 20 were invited for oral presentations, adding to 14 plenary talks. These papers appear either as extended abstracts in this volume or are published in the journal *Molecular Systems Biology*.

Finally, we gratefully acknowledge support from our sponsors: the International Society for Computational Biology, RECOMB Steering Committee, the California Institute for Telecommunications and Information Technology (Calit2), the UC Discovery Program, and Pfizer La Jolla.

December 2007

Vineet Bafna
Trey Ideker

Organization

Program Committee

Annette Adler	Agilent
John Aitchison	Institute of Systems Biology
Gary Bader	Memorial Sloan-Kettering Cancer Center
Vineet Bafna (Co-chair)	University of California, San Diego
Ron Beavis	University of British Columbia
Marshall Bern	Palo Alto Research Center
Tim Chen	University of Southern California
Eric Davidson	California Institute of Technology
Nathan Edwards	University of Maryland
Keith Elliston	Genstruct
Eleazar Eskin	University of California, San Diego
Tim Galitski	Institute of Systems Biology
Mark Gerstein	Yale University
Jeff Hasty	University of California, San Diego
Ralf Herwig	Max Planck Institute for Molecular Genetics
Leroy Hood	Institute for Systems Biology
Trey Ideker (Co-chair)	University of California, San Diego
Janette Jones	Unilever SEAC
Peter Karp	Bioinformatics Research Group, SRI Intl
Stuart Kim	Stanford University
Edda Klipp	Max Planck Institute for Molecular Genetics
Oliver Kohlbacher	Universität Tübingen
Douglas Lauffenburger	Massachusetts Institute of Technology
Mike Levine	University of California, Berkeley
Bin Ma	University of Western Ontario
Edward Marcotte	University of Texas
Andrew McCulloch	University of California, San Diego
Satoru Miyano	University of Tokyo
Alexey Nesvizhskii	University of Michigan
William Noble	University of Washington
Bernhard Palsson	University of California, San Diego
Dana Pe'er	Harvard Medical School
Pavel Pevzner	University of California, San Diego
Tzachi Pilpel	Weizmann Institute of Science
Teresa M. Przytycka	NIH/NLM/NCBI
Ben Raphael	Brown University
Knut Reinert	Freie Universität Berlin
Cenk Sahinalp	Simon Fraser University
Ron Shamir	Tel Aviv University

Roded Sharan	Tel Aviv University
Alfonso Valencia	Centro Nacional de Biotecnologia
Guy Warner	Unilever SEAC
Christopher Workman	Technical University of Denmark
John Yates	The Scripps Research Institute
Ralf Zimmer	Institut für Informatik

RECOMB Systems Biology Steering Committee

Trey Ideker (Chair)	University of California, San Diego
Ron Shamir	Tel Aviv University
Satoru Miyano	University of Tokyo
Douglas Lauffenburger	Massachusetts Institute of Technology
Leroy Hood	Institute for Systems Biology

RECOMB Computational Proteomics Steering Committee

Vineet Bafna (Chair)	University of California, San Diego
John Yates	The Scripps Research Institute
Tim Chen	University of Southern California
Pavel Pevzner	University of California, San Diego

Organizing Committee

Vineet Bafna	University of California, San Diego
Trey Ideker	University of California, San Diego
Nuno Bandeira	University of California, San Diego
Thomas Lemberger	MSB
BJ Morrison McKay	ISMB
Samantha Smeraglia	University of California, San Diego
Shaojie Zhang	University of California, San Diego

Sponsoring Institutions

The International Society for Computational Biology
Molecular Systems Biology
The UC Discovery Grant
Pfizer Inc.
The UCSD Jacobs School of Engineering
Calit2: California Institute for Telecommunications and Information Technology

Lecture Notes in Bioinformatics

Vol. 4544: S. Cohen-Boulakia, V. Tannen (Eds.), Data Integration in the Life Sciences. XI, 282 pages. 2007.

Vol. 4532: T. Ideker, V. Bafna (Eds.), Systems Biology and Computational Proteomics. IX, 131 pages. 2007.

Vol. 4463: I. Măndoiu, A. Zelikovsky (Eds.), Bioinformatics Research and Applications. XV, 653 pages. 2007.

Vol. 4453: T. Speed, H. Huang (Eds.), Research in Computational Molecular Biology. XVI, 550 pages. 2007.

Vol. 4414: S. Hochreiter, R. Wagner (Eds.), Bioinformatics Research and Development. XVI, 482 pages. 2007.

Vol. 4366: K. Tuyls, R.L. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007.

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4316: M.M. Dalkilic, S. Kim, J. Yang (Eds.), Data Mining and Bioinformatics. VIII, 197 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

Vol. 4216: M.R. Berthold, R.C. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006.

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 4023: E. Eskin, T. Ideker, B. Raphael, C. Workman (Eds.), Systems Biology and Regulatory Genomics. X, 259 pages. 2007.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R.C. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Table of Contents

Not All Scale Free Networks Are Born Equal: The Role of the Seed Graph in PPI Network Emulation 1
Fereydoun Hormozdiari, Petra Berenbrink, Nataša Pržulj, and Cenk Sahinalp

Probabilistic Paths for Protein Complex Inference 14
Hailiang Huang, Lan V. Zhang, Frederick P. Roth, and Joel S. Bader

Markov Additive Chains and Applications to Fragment Statistics for Peptide Mass Fingerprinting 29
Hans-Michael Kaltenbach, Sebastian Böcker, and Sven Rahmann

A Context-Specific Network of Protein-DNA and Protein-Protein Interactions Reveals New Regulatory Motifs in Human B Cells 42
Celine Lefebvre, Wei Keat Lim, Katia Basso, Riccardo Dalla Favera, and Andrea Califano

Identification and Evaluation of Functional Modules in Gene Co-expression Networks 57
Jianhua Ruan and Weixiong Zhang

A Linear Discrete Dynamic System Model for Temporal Gene Interaction and Regulatory Network Influence in Response to Bioethanol Conversion Inhibitor HMF for Ethanologenic Yeast 77
Mingzhou (Joe) Song and Z. Lewis Liu

A Computational Approach for the Identification of Site-Specific Protein Glycosylations Through Ion-Trap Mass Spectrometry 96
Yin Wu, Yehia Mechref, Iveta Klouckova, Milos V. Novotny, and Haixu Tang

De Novo Signaling Pathway Predictions Based on Protein-Protein Interaction, Targeted Therapy and Protein Microarray Analysis 108
Derek Ruths, Jen-Te Tseng, Luay Nakhleh, and Prahlad T. Ram

Alignment of Mass Spectrometry Data by Clique Finding and Optimization 119
Daniel Fasulo, Anne-Katrin Emde, Lu-Yong Wang, Karin Noy, and Nathan Edwards

Author Index 131

Not All Scale Free Networks Are Born Equal: The Role of the Seed Graph in PPI Network Emulation

Fereydoun Hormozdiari[1], Petra Berenbrink[1], Nataša Pržulj[2], and Cenk Sahinalp[1]

[1] School of Computing Science, Simon Fraser University, Canada
[2] Department of Computer Science, University of California, Irvine, USA

Abstract. The (asymptotic) degree distributions of the best known "scale free" network models are all similar and are independent of the seed graph used. Hence it has been tempting to assume that networks generated by these models are similar in general. In this paper we observe that several key topological features of such networks depend heavily on the specific model and the seed graph used. Furthermore, we show that starting with the "right" seed graph, the *duplication model* captures many topological features of publicly available PPI networks very well.

1 Introduction

In the past few years protein-protein interaction (PPI) networks of several organisms have been derived and made publicly available. Some of these networks have interesting topological properties; e.g. the degree distribution of the Yeast PPI network is heavy tailed (i.e. there are a few nodes with many connections). It has been argued that the degree distribution of these networks are in the form of a *power-law*[14], [24].[1] Since well known random graph models also have power-law degree distributions [3], [8], [25] it has been tempting to investigate whether these models agree with other topological features of the PPI networks.

There are two well known models that provide power law degree distributions (see [10], [9], [4]). The *preferential attachment* model [2], [8], was introduced to emulate the growth of naturally occurring networks such as the web graph; unfortunately, it is not biologically well motivated for modeling PPI networks. The *duplication model* on the other hand [7], [22], [18] is inspired by Ohno's hypothesis on genome growth by duplication. Both models are iterative in the sense that they start with a *seed graph* and grow the network in a sequence of steps.

The degree distribution is commonly used to test whether two given networks are similar or not. However, networks with identical degree distributions can have very different topologies.[2] Furthermore, it was observed in [23] that given two networks with substantially different initial degree distributions, a partial (random) sample from

[1] Some recent work challenge this by attributing the power law like behavior to sampling issues, experimental errors or statistical mistakes [23], [16], [21], [19], [12].

[2] Consider, for example, an infinite two dimensional grid vs a collection of cliques of size 5; in both cases all nodes have degree 4.

T. Ideker and V. Bafna (Eds.): Syst. Biol. and Comput. Proteomics Ws, LNBI 4532, pp. 1–13, 2007.

those networks may give subnetworks with very similar degree distributions. Thus the degree distribution can not be used as a sole measure of topological similarity.

In the recent literature two additional measures have been used to compare PPI networks with random network models. The first such measure is based on the *k-hop reachability*. The 1-hop reachability of a node is simply its degree (i.e. the number of its neighbors). The k-hop reachability of a node is the number of distinct nodes it can reach via a path of $\leq k$ edges. The k-hop reachability of all nodes whose degree is ℓ is the average k-hop reachability of these nodes. Thus the k-hop reachability (for $k = 2, 3, \ldots$) of nodes as a function of their degree can be a used to compare network topologies. An earlier comparison of the k-hop reachability of the Yeast network with networks generated by certain duplication models concluded that the two network topologies are quite different [5]. The second similarity measure is based on the *graphlet distribution*. Graphlets are small subgraphs such as triangles, stars or cliques. In [16] it was noted that certain "scale free" networks are quite different from the Yeast PPI network with respect to the *graphlet distribution*. This observation, in combination with that on the k-hop degree distribution seem to suggest that the known PPI networks may not be scale free and existing scale free network models may not capture the topological properties of the PPI networks.

There are other topological measures that have been commonly employed in comparing social networks etc. but not PPI networks. Two well known examples are the *betweenness* distribution and the *closeness* distribution [26]. Betweenness of a vertex v is the number of shortest paths between any pair of vertices u and w that pass through v, normalized by the total number of such paths. Closeness of v is the inverse of the total distance of v to all other vertices u. Thus one can use betweenness and the closeness distributions, which respectively depict the number of vertices within a certain range of betweenness and closeness values can be used to compare network topologies.

2 Network Generation Models

The two network network models we study here both start with a small seed graph and add one node to it in each iteration. Let $G(t) = (V(t), E(t))$ be the graph at the end of time step t, where $V(t)$ is the set of nodes and $E(t)$ is the set of edges/connections. Let v_t be the node generated in time step t. Given a node v_τ, we denote its degree at the end of time step t by $d_t(v_\tau)$.

Preferential attachment model. The preferential attachment model was analyzed in [2], [6], [8] ,[10]. In step t it generates v_t and connects it to every other node v_τ independently with probability $c \cdot d_{t-1}(v_\tau)/2|E(t-1)|$, where c is the average degree of a node in G; i.e. v_t prefers to connect itself to high degree nodes.

Duplication model. This model is based on Ohno's hypothesis of genome evolution [7], [18], [22]. In iteration t, a node v_τ of $G(t-1)$ is picked uniformly at random and "duplicated", i.e. an exact copy of v_τ as v_t is generated. The model then updates v_t's edges, first by deleting each of its edges with probability $(1-p)$, then by connecting each node $v_{t'}$ (except the neighbors of v_τ) to v_t independently with probability $r/|V(t)|$. Here, p and r are user defined parameters. Much of the earlier work on the duplication model

aim to maintain a constant average degree throughout the generation of the network; this is achieved by setting $r = (1/2 - p).a$.

As mentioned earlier, the degree distribution of the preferential attachment model as well as the duplication model asymptotically approaches a power law [2], [8], [10], [9]. More specifically, in the log-log scale, it forms a straight line (this is valid for only "high degree" nodes) whose slope is independent of the seed graph and a function of the values of p and r for the duplication model or c for the preferential attachment model. Thus, the two iterative models are equivalent with respect to the degree distribution.

Both the preferential attachment and the duplication model produce many *singletons* [3] [4]. Singletons are nodes which are not connected to any other node. Unfortunately there are no known bounds on the number of generated singletons in the duplication model. In the duplication model, for the special case $r = 0$, $p = 1/2$, the proportion of singletons asymptotically approaches 1. However, the number of singletons in known PPI networks is very small.

Modified duplication model. It is well known that the number of singletons in PPI networks are quite limited. This does not come as a surprise as genes with no functionality are not conserved during evolution. Thus a slightly modified duplication model which deletes each singleton node as soon as it is generated may better emulate the growth of PPI networks. This model has also been shown to achieve a power law degree distribution [4].

Unfortunately, similar to the number of singletons in duplication model, in modified model the total number of generated nodes is not known. Moreover, it is not known which values of p and r ensure that the expected average degree is constant through all iterations. In Section 2.1 we derive conditions on p and r that are necessary for having a constant expected degree. We later use the derived relationship between p and r so that the modified duplication model can well approximate the desired average degree as well as the degree distribution of the PPI networks under investigation.

2.1 The Parameters of the Modified Duplication Model

Here we show how to determine conditions on deletion probability $1 - p$ and insertion probability r so that the expected average degree of the network can be set to any given value. For this, we make the the assumption that the degree frequency distribution and the average degree of nodes are fixed asymptotically once the values of p and r are determined. Let $G(t) = (V(t), E(t))$ be the network generated by the modified duplication model and let $n(t) = |V(t)|$ and $e(t) = |E(t)|$. Also, let $n_k(t)$ be the number of nodes in time step t with degree k and $a(t)$ be the average degree of nodes in $G(t)$. Finally let $P_k(t) = n_k(t)/n(t)$, the frequency of nodes with degree k at time step t. We assume that $P_t(k)$ is asymptotically stable, i.e. $P_k(t) = P_k(t+1)$ for all $1 \leq k \leq t$ for sufficiently large values of t. In other words we assume that $P_k(t) = d_k$

[3] We also note that the known PPI networks have several self loops. Both the preferential attachment and the duplication models can be modified slightly to produce such self loops(homodimers).

for some fixed d_k. By definition

$$a(t) = \sum_{k=1}^{t} k \cdot \frac{n_k(t)}{n(t)} = \sum_{k=1}^{t} k \cdot P_k(t) = \sum_{k=1}^{t} k \cdot d_k.$$

Now we can calculate the average degree $a(t+1)$ under the condition that degree frequency distribution is stable and $a(t) = a$, a constant.

$$Exp[e(t+1)] = e(t) + \sum_{k=1}^{t} k \cdot P_k(t) \cdot p + r = \frac{n(t) \cdot a(t)}{2} + p \cdot a(t) + r.$$

Let $Pr_s(t)$ be the probability that v_{t+1} ends up as a singleton.

$$Pr_s(t) = \sum_{k=1}^{t} P_k(t) \cdot (1-p)^k \cdot \left(1 - \frac{r}{n(t)}\right)^{n(t)-k} \approx \sum_{k=1}^{t} d_k \cdot (1-p)^k \cdot \frac{1}{e^r}.$$

Since this probability does not depend on t asymptotically, we can set $Pr_s(t) = Pr_s$. Now we can calculate the expected number of nodes and the expected number of edges in step $t+1$.

$$Exp[n(t+1)] = Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1).$$

$$Exp[e(t+1)] = Exp\left[\frac{n(t+1) \cdot a(t+1)}{2}\right] = \frac{a}{2} \cdot Exp[n(t+1)]$$

$$Exp[e(t+1)] = \frac{a}{2} \cdot (Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1)).$$

Comparing the above equation with the first equation for $Exp[e(t+1)]$ we get

$$\frac{a}{2} \cdot (Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1)) = \frac{n(t) \cdot a(t)}{2} + p \cdot a(t) + r = \frac{n(t) \cdot a}{2} + p \cdot a + r.$$

Solving the above equation results in $a = 2r/(1 - Pr_s - 2p)$ where Pr_s is a function of p, r and d_k only.

The discussion above demonstrates that the two key parameters p and r of the (modified) duplication model are determined by the degree distribution (more specifically the slope of the degree distribution in the log-log scale) and the average degree of the PPI network we would like to emulate. Perhaps due to the strong evidence that the seed network does not have any effect on the asymptotic degree distribution [5], the role of the seed network (the only free parameter remaining) in determining other topological features of the duplication model has not been investigated.

3 Measures for Comparing Networks

There are several topological features that can be used to test whether two networks are similar or not, starting at very rigorous measures like isomorphism, to very relaxed characteristics like the degree distribution. In this paper we focus on five such properties, namely the *degree distribution*, the *k-hop reachability*, the *graphlet frequency*, the *betweenness distribution* and the *closeness distribution*.

Isomorphy. Two networks G and G' are called *isomorphic* if there exists a bijective mapping F from each node of G to a distinct node in G', such that two nodes v and w are connected in G if and only if $F(v)$ and $F(w)$ are connected. G and G' are called *approximately isomorphic* if by removing a "small" number of nodes and edges from G and G' they could be made isomorphic. Ideally, a random graph model that aims to emulate the growth of a PPI network should produce a network that is approximately isomorphic to the PPI network under investigation. Unfortunately there is no known polynomial algorithm for testing whether two networks are (exactly or approximately) isomorphic or not.

k-hop reachability. Let $V(i)$ denote the set of nodes in V whose degree is i. Given a node v, denote by $d(v,k)$ its k-hop degree, i.e., the number of distinct nodes it can reach in $\leq k$ hops. Now we define $f(i,k)$, the k-hop reachability of $V(i)$ as

$$f(i,k) = \frac{1}{|V(i)|} \sum_{w \in V, d(w)=i} d(w,k).$$

Thus $f(i,k)$ is the "average" number of distinct nodes a node with degree i can reach in k hops; e.g. $f(i,1) = i$ by definition.

Graphlet frequency. The graphlet frequency was introduced in [16] to compare the topological structure of networks. A graphlet is a small connected and induced subgraph of a large graph, for example a small triangle or a small clique. The *graphlet count* of a given graphlet g with r nodes in a given graph $G = (V, E)$ is defined as the number of distinct subsets of V (with r nodes) whose induced subgraphs in G are isomorphic to g. In this paper we consider all 141 possible graphlets/subgraph topologies with $3, 4, 5, 6$ nodes. Additionally, we consider cliques of sizes $7, 8, 9, 10$. We enumerate these graphlets as shown in Figure 6.

Betweenness distribution. The betweenness of a fixed node of a network measures the extend to which a particular point lies 'between' point pairs in the network $G = (V, E)$. The formal definition of betweenness is as follows. Let $\sigma_{x,y}$ be the number of shortest path from $x \in V$ to $y \in V$ for all pairs $x, y \in V$. (Note that $\sigma_{x,y} = \sigma_{y,x}$ in undirected graphs). Let $\sigma_{x,y}(v)$ be the number of shortest path from $x \in V$ to $y \in V$ which go through node v. The betweenness $\mathrm{Bet}(v)$ of node v is now defined as follows.

$$\mathrm{Bet}(v) = \sum_{(i,j) \in V, i,j \neq v} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}}.$$

Closeness. For all $x, y \in V$, we define $d_{x,y}$ as the length of the shortest path between x and y. The closeness of a node $v \in V$ is defined as

$$\mathrm{Cls}(v) = \frac{|V| - 1}{\sum_{i \in V} d_{v,i}}.$$

4 Results and Discussion

As mentioned above, scale free network generators such as the preferential attachment model and the duplication model can have very similar degree distributions under appropriate choice of parameters. Moreover, the degree distribution of these models converge to a power law degree distribution whose shape is determined solely by the edge deletion and edge insertion probabilities and not by the initial "seed" graph [10]. Hence, it has been tempting to assume that networks generated by these models are similar in general and the effect of the seed graph in shaping the topologies of these networks has largely been ignored in recent literature.

Unfortunately two networks with very similar degree distributions may have very different topologies. For example, a network generated by the preferential attachment and another generated by the duplication model may have very different k-hop reachability, graphlet, betweenness and closeness distributions while having almost identical degree distributions (see Section A.1). Furthermore two networks generated by the same duplication model (and hence have very similar degree distributions) can differ substantially in terms of the above topological measures, if their seed networks are different (see Section A.2).

If the seed selection makes such a difference in shaping the topology of the generated network, is it possible to select the "right" seed network so that all interesting topological features of the PPI networks in question can be captured? We answer this question positively by demonstrating that carefully chosen seeds can result in a network that is very similar to PPI networks we considered in terms of all of the above distributions.

The PPI networks we tested include (the largest connected component of) the complete Yeast PPI network [20] with 4902 proteins and 17200 edges (as of Jul 2006). We also tested the more accurate but much smaller CORE Yeast network [11] and the lesser developed Worm network [20] (see Section A.3).

The seed graph we used for capturing the Yeast PPI network basically has two highly connected cliques of respectively 10 and 7 nodes. There are a few additional nodes sparsely connected to the cliques in a random fashion (the total number of nodes was 50). This ensured that the (normalized) degree distribution of the Yeast PPI network as well as its clique frequency distribution (which turns out to be an important determinant of the overall graphlet distribution) were similar to that of the seed graph.

There are two additional parameters associated with the duplication model: p, the edge maintenance probability and r, the edge insertion probability. These two parameters alone determine the (asymptotic) degree distribution and the average degree of the generated network. We chose $p = 0.365$ and $r = 0.12$ so that the degree distribution of the duplication model matches that of the Yeast PPI network (see Section 2.1 for the exact mathematical expressions for p and r).

We used the duplication model to generate 4 independent networks each with 4902 vertices. The resulting networks are compared to the Yeast PPI network in terms of the k-hop reachability, the graphlet , betweenness, and closeness distributions in Figure 1.

Under all these measures, the Yeast network is very similar to those produced by the duplication model. In fact the duplication model we consider here provides much

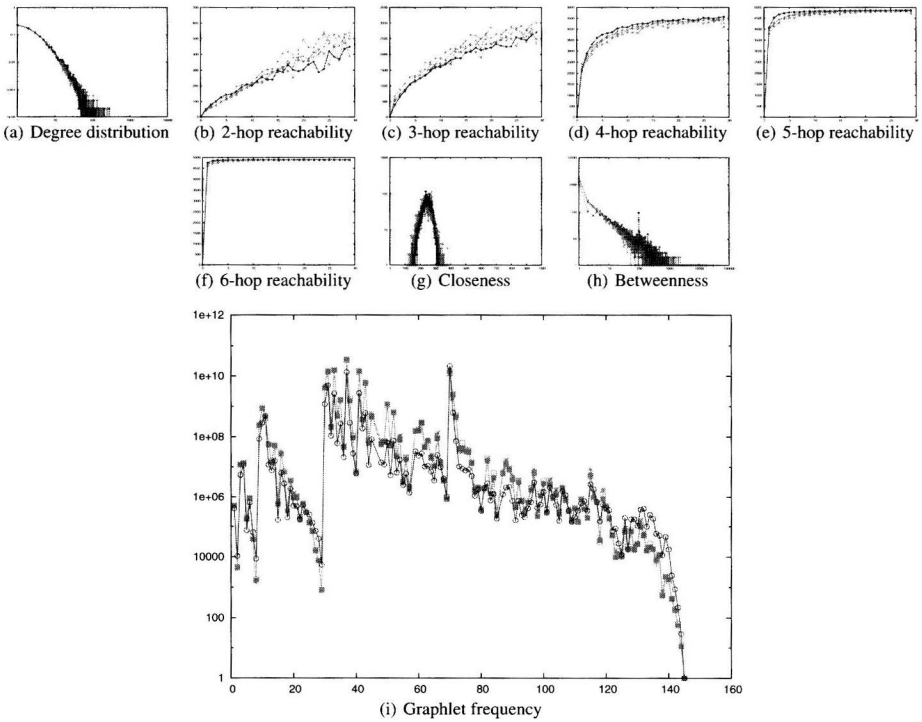


Fig. 1. The degree distribution, the k-hop reachability, the graphlet, closeness and betweenness distributions of the Yeast PPI (Red) network against four independent runs of the duplication model (Green)

better fits to both the k-hop degree distribution and the graphlet distribution of the Yeast network than the random graph models described in of [5] and [16] - which were specifically devised to capture the respective features of PPI networks.

References

1. Alfarano, C., et al.: The biomolecular interaction network database and related tools. Nucl Acids Res. 33(Database Issue), 418–424 (2005)
2. Aiello, W., Chung, F., Lu, L.: Random graph model for power law graphs. In: Proc ACM STOC, pp. 171–180 (2000)
3. Barabási, A.-L., Albert, R.A.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
4. Bebek, G., Berenbrink, P., Cooper, C., Friedetzky, T., Nadeau, J., Sahinalp, S.C.: The degree distribution of the general duplication models, Theor Comp Sci. (to appear)
5. Bebek, G., Berenbrink, P., Cooper,C., Friedetzky, T., Nadeau, J., Sahinalp, S.C.: Topological Properties of proteome networks. In: Proc RECOMB Sat. Mtg. on Sys. Bio. (LNBI) (2005)
6. Berger, N., Bollobás, B., Borgs, C., Chayes, J., Riordan, O.: Degree distribution of the FKP network model. In: Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginger, G.J. (eds.) ICALP 2003. LNCS, vol. 2719, pp. 725–738. Springer, Heidelberg (2003)