FEBS Letters 509 (2001) 230–234 FEBS 25552

Mining DNA microarray data using a novel approach based on graph theory

Gabriel del Rio[a], Theodore F. Bartley[b], Heberto del-Rio[c], Rammohan Rao[a], KunLin Jin[a], David A. Greenberg[a], Mark Eshoo[a], Dale E. Bredesen[a,d,*]

[a] *The Buck Institute for Age Research, 8001 Redwood Blvd., Novato, CA 94945, USA*
[b] *Undergraduate student from the University of California, Santa Barbara, CA, USA*
[c] *Centro de Investigaciones en Matemáticas, CIMAT, Guanajuato, Guanajuato, Mexico*
[d] *University of California, San Francisco, CA, USA*

Received 17 October 2001; revised 26 October 2001; accepted 26 October 2001

First published online 20 November 2001

Edited by Julio Celis

Abstract The recent demonstration that biochemical pathways from diverse organisms are arranged in scale-free, rather than random, systems [Jeong et al., Nature 407 (2000) 651–654], emphasizes the importance of developing methods for the identification of biochemical nexuses – the nodes within biochemical pathways that serve as the major input/output hubs, and therefore represent potentially important targets for modulation. Here we describe a bioinformatics approach that identifies candidate nexuses for biochemical pathways without requiring functional gene annotation; we also provide proof-of-principle experiments to support this technique. This approach, called Nexxus, may lead to the identification of new signal transduction pathways and targets for drug design. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Functional genomics; DNA microarray; Graph theory; Yeast cell cycle; Yeast metabolism; Cerebral ischemia

1. Introduction

During evolution, living organisms have developed intricate molecular mechanisms to respond to diverse environmental conditions. Virtually every cellular function in multicellular organisms is thought to be dependent on signaling molecules [2,3], and therefore on signal transduction pathways. In order to understand fully the molecular mechanisms involved in biological processes, it is necessary to identify each molecule participating in the mechanisms. Addressing the complexity of biological systems, high-throughput approaches have emerged in the past decade, intended to determine the molecular details of the cellular responses to diverse stimuli. Some of these approaches are intended to analyze the most well-studied biomolecules (i.e. nucleic acids and proteins), and accordingly are classified as genomics [4] and proteomics [5]. By using these approaches, we now have access for the first time to the full list of genes present in diverse organisms, as well as protein–protein interaction maps for some of them [6]. Determination of how those genes/proteins work coordinately, and which are the crucial genes/proteins in the phenomena studied, are important goals of functional genomics/proteomics, and should help lead to a better understanding of cellular mechanisms, allowing (among other things) the design of novel therapeutics.

Previous studies in proteomics have been carried out to trace protein interaction maps, but the limitation of those approaches is the interpretation of the maps generated [6]. That is, given that different protein interactors are detected for a protein, this does not necessarily mean that they all act in a protein complex. Alternatively, in the genomics field, microarray technology provides a tool to detect genes/proteins involved in the transcriptional response of a cell to a particular stimulus (or set of stimuli). From cluster analysis [7] it is possible to detect those genes participating in a common mechanism. The limitation of this approach is that, although clusters of genes may be identified, the mechanisms by which these genes function together, and the determination of which of the associated gene products are optimal for potential drug discovery programs, are not addressed by cluster analysis. Additionally, the functional analysis of genomics and proteomics data usually depends on the annotated function of the genes and proteins. Although this is a powerful way of analyzing data, it has certain limitations: for example, many genes have no currently assigned function (and to establish gene function is often laborious).

Recognizing these limitations, we sought to develop a bioinformatics approach and a conceptual framework for predicting genes/proteins controlling biomolecular mechanisms. We refer to this approach as Nexxus, and it is designed to detect key proteins controlling biochemical mechanisms, where regulation of protein–protein interactions is crucial for the system studied, by combining data from genomics (microarray technology) and proteomics (protein–protein interactions). This approach is based on the prediction of protein paths using a mathematical algorithm derived from the theory of graphs [8] and intended to search for the shortest path connecting two vertices in a graph. In this way, we reconstruct the minimum protein–protein interaction map that includes the genes detected in a microarray experiment (see Fig. 1C). Since these paths are not necessarily the ones present in the cellular event studied (see below), any prediction based on the proteins constituting the paths that were not detected in the microarray experiment may lead to false predictions.

*Corresponding author. Fax: (1)-415-209 2230.
E-mail address: dbredesen@buckinstitute.org (D.E. Bredesen).

0014-5793/01/$20.00 © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
PII: S0014-5793(01)03165-9

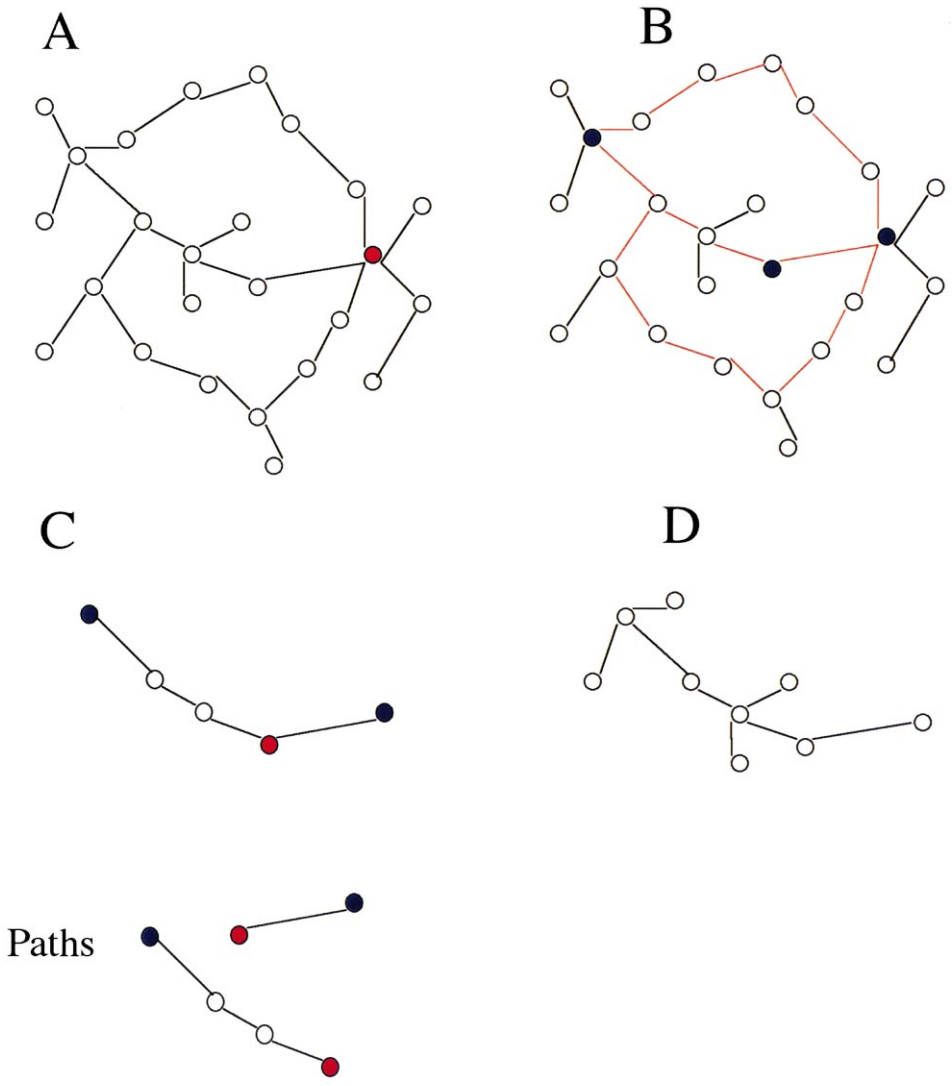

Fig. 1. Protein–protein interaction maps as graphs. A protein–protein interaction map is depicted using lines and circles to indicate protein–protein interactions and proteins, respectively. This representation can be classified as a graph (A) or a tree (D). A red circle represents a protein nexus (A, C). Proteins detected from a microarray are represented in blue (B). The red lines indicate the different protein paths that can be traced connecting the circles in blue. The smallest protein–protein interaction map (C) produced from B with the corresponding protein paths is also indicated.

Alternatively, we considered the paths as an indication that two proteins can be linked through protein–protein interactions.

Based on earlier work [1] it is reasonable to hypothesize that in protein graphs, nexuses play a key role in maintaining the organization and function of the molecular network described by each graph. Studying diverse biological systems, it has been shown that biochemical pathways are not arranged with random numbers of connectivities; instead, they display a scale-free character [1]. A crucial characteristic of this type of system is that it is largely controlled by a few hubs or nexuses, which dominate the overall connectivity and stability of the graph. By implication, identifying these nexuses is likely to represent an important step in functional genomic studies. When all of the cellular proteins are considered to generate a protein graph, nexuses can be detected that are likely to be important evolutionarily and functionally [9]. However, cell specialization may create subsets of these rather complete protein graphs. In these subsets, different nexuses may be detected (see Fig. 1B,C). We therefore introduce the notion of dynamic nexuses (DN) in order to: (a) define the nexuses used in a particular cellular state; (b) distinguish DN from static nexuses (SN) detected from the complete protein–protein interaction map of an organism. Our approach, Nexxus, is intended to detect both DN and SN without relying on the annotated function of the genes/proteins, instead utilizing the data provided by microarray (gene expression) and protein–protein interaction maps. As protein interaction maps and microarray studies become more extensive, the predictive power of Nexxus, as well as the accuracy of the predictions generated, should increase concomitantly.

Considering the nature of the data used in this work (protein–protein interactions and gene expression patterns), Nexxus is expected to be useful in the analysis of systems in which the regulation of protein–protein interactions is critical (i.e. signal transduction), but is less likely to be useful in detecting proteins involved in critical enzymatic reactions (e.g. proteolysis, metabolism).

2. Materials and methods

2.1. Protein–protein interaction data

Two data sets of protein–protein interactions were used in this study. The ProNet database (http://pronet.doubletwist.com) was used to trace protein complexes/paths in the ischemia model. Alternatively, the protein–protein interactions reported in the PathCalling Yeast Interaction database (http://portal.curagen.com) [10] were used for detecting nexuses in yeast. In this latter case, every protein–protein interaction reported at this site was downloaded, consisting of 2547 protein–protein interactions and 1376 proteins.

The ProNet database contains information on protein–protein interactions that have been described in the published literature for human proteins identified using the yeast two-hybrid system. We obtained a list of all of the protein–protein pairs that could be linked to the set of gene products regulated during the ischemia experiment. These included 820 protein–protein interactions and 611 proteins.

2.2. Detecting static nexuses from the connectivity distribution $P(k)$ of protein interaction maps

For each protein, the number of interactors was determined (k) and then the numbers of proteins having $k = 0,1,2...$ number of interactors were determined. Dividing each value obtained in this way by the total number of proteins in the graph provided the value $P(k)$, or the probability that the protein will have a number k of interactors. SN are proteins with the larger k values. These are referred to as static because these can be detected without tracing paths.

2.3. Detecting dynamic nexuses

In order to predict protein paths based on protein–protein interactions, we adapted a previously described algorithm [11] to detect the shortest path connecting two vertices. In this way we traced the minimum protein–protein interaction map that includes the gene products detected by the microarray experiments and protein–protein interaction data (see Fig. 1C).

Given a minimum protein–protein interaction map, DN can be detected. DN are proteins connecting the largest number of protein paths. This definition differs from the one used to define SN in the sense that the connectivity in DN refers to paths rather than to individual proteins in the graph. Two types of DN can be identified. One is determined considering only the genes detected in the microarray experiment, and the other type is obtained when considering all of the proteins predicted in the paths. We refer here to DN as the first type. It was initially conceivable that this approach would detect only those proteins with the most protein interactors reported in the protein–protein interaction maps. In order to exclude this possibility, we reported the number of interactors for each protein included in detecting DN (see Tables 1A,B and 2).

3. Results and discussion

In proof-of-principle experiments, we first determined the topology of the protein graphs used in these studies, derived from studies of apoptosis [12] and yeast [10]. Fig. 2A,B shows that in both cases, the protein–protein interaction maps display a characteristic distribution of the heterogeneous class of graphs referred to as scale-free, for which the connectivity distribution $P(k)$ follows a power law, $P(k) \approx k^{-\gamma}$ [1]. The estimated γ values for apoptosis and yeast graphs were 2.41 and 2.28, respectively.

The scale-free character of these graphs implies that nodes of highest connectivity should be identifiable. Beyond simple connectivity, we analyzed all possible protein interaction paths to identify the proteins most frequently involved in these paths. These proteins were dubbed nexuses. In order to identify DN, i.e. those derived by tracing protein interaction paths using the algorithm described herein, and SN, which are the proteins with highest reported simple connectivity, we used the yeast as a model. In the group of genes regulated in the S phase of the yeast cell cycle [7] (65 proteins, 15 of which have known interactors), we identified CDC5 and MOB1 as DN. In the groups of genes regulated during phosphate metabolism [13] (37 proteins, eight of which have known interactors), we identified PHO85 and MSN5 as DN (see Table 1A,B). DN are proposed to play key roles in the mechanism that utilizes the network in which these are located. Proteins posited to be critical in the S phase or phosphate metabolism are thus hypothesized to be associated with an inviable null phenotype. In agreement with this hypothesis, the null phenotypes for the DN detected were indeed inviable [14,15]: all DN predicted in these proof-of-principle experiments displayed inviable null phenotypes. Conversely, all proteins not predicted to be DN displayed viable null phenotypes, implying that these are not required for survival, with a very few exceptions – two of 15 for the cell cycle study, and zero of eight for the phosphate metabolism study – exceptions that proved to be informative (Table 1): CDC20 was the only protein predicted to be neither a DN nor a SN yet still displayed a non-viable null phenotype, but importantly the function of CDC20 is not restricted to the cell cycle. TEM1 was the other 'outlier' in that it was not predicted to be a DN; however, it was predicted to be a SN; furthermore, it displayed the highest dynamic connectivity after CDC5 and MOB1, suggesting that when complete protein interaction maps are available, it may indeed prove to be a DN.

Table 1A
Detecting protein nexuses for cell cycle-regulated genes in S phase

Protein name	DC	SC	Null phenotype
MET28	11	2	viable
NUM1	11	3	viable
KIP2	11	1	viable
ACE2	0	1	viable
MET17	11	1	viable
MET14	0	1	viable
MYO1	11	2	viable
CDC5	14	3	inviable
SWI5	11	4	viable
CDC20	11	2	inviable
MOB1	14	4	inviable
TEM1	12	24	inviable
CLB2	11	8	viable
CLB1	11	2	viable
HST3	0	1	viable

Table 1B
Detecting protein nexuses for regulated genes in phosphate metabolism

Protein name	DC	SC	Null phenotype
PHO4	1	2	not reported
PHO85	6	11	inviable
PHO80	1	1	viable
PSE1	5	11	viable
MSN5	6	5	viable/inviable
PHO13	5	1	viable
PHO12	5	1	not reported
CTF19	5	1	viable

Protein name: the name used to identify the gene product in the PathCalling database. DC: dynamic connectivity, i.e. the number of protein interactors found in the protein paths predicted by Nexxus. Those proteins with the greatest dynamic connectivity are predicted to be DN. SC: static connectivity, i.e. the number of protein interactors reported in the PathCalling database for each protein. Those proteins with the greatest static connectivity are predicted to be SN. Null phenotype: as described in the PathCalling database.

These initial studies therefore suggested that it is likely to be important to identify both DN and SN. It is important to note that CDC5 and MOB1 did not demonstrate high simple connectivity (number of interactors) and were therefore not predicted to be SN, yet were predicted to be DN based on the path tracing algorithm, and indeed proved to be required for yeast survival. Our results may be used to enhance the understanding of the relationship between the physical connectivity and lethal functionality of proteins [9]. Overall, then, combining the cell cycle study with the phosphate metabolism study, five of six non-viable null phenotypes were predicted by considering dynamic and static connectivity, with the one not predicted (CDC20) functioning outside the cellular systems studied; and 14 of 15 viable phenotypes were predicted on the basis of lower connectivities (PSE1 displayed high static connectivity (Table 1) yet was reported to be viable; note that PHO4 and PHO12 were reported to be neither viable nor non-viable). To estimate the statistical significance of these results, we determined the 95% confidence interval (CI) for Nexxus in detecting critical genes based on these initial studies. This calculation was made with the assumption that the data were independent of each other. The CI was 80–100% for specificity; that is, with 95% confidence, Nexxus will detect in $>80\%$ of the cases non-critical genes as non-critical. Due to the lower number of non-viable phenotypes in these initial studies, the 95% CI for sensitivity was larger – 37–100% – and additional studies will be required to narrow this interval.

Table 2
Predicting functional nexuses from microarray data generated by studies of cerebral ischemia

ProNet name	DC	SC	Function
E2F1	10	3	transcription factor
Bcl2-α	8	15	Bcl-2, alt. transcript α (239)
CASP3	6	1	caspase 3
JUN	5	7	proto-oncogene c-Jun
RB1	6	7	retinoblastoma 1
VDR	4	7	vitamin D3 receptor
CASH-α	6	5	caspase-like, alt. transcript 1
NCOA1a(v)	8	4	nuclear receptor coactivator 1
CDK2	4	4	cyclin-dependent kinase
UBE3A	1	2	E6-AP ubiquitin-protein ligase
F1A-α	6	1	F1A-α
NFKB1	6	2	transcription factor
TRIP11	6	1	thyroid receptor interactor 11
PGR	1	1	progesterone receptor
SMN1	6	4	survival of motor neuron 1
TP53BP2	5	4	tumor protein p53-binding protein 2
AF	6	2	antisecretory factor 1
TP53	5	9	tumor protein p53
UBE2I	10	11	ubiquitin-conjugating enzyme
LYL1	6	1	lymphoblastic leukemia sequence 1
RBL2	4	6	retinoblastoma-like protein 2
XPB	6	1	xeroderma pigmentosum gene
14-3-3e	2	3	14-3-3 protein, ε
TNFR1	6	7	tumor necrosis factor receptor 1
14-3-3b	6	5	14-3-3 protein, β
GADD34	0	1	growth arrest and DNA damage-inducible
UBE2L1	1	3	ubiquitin-conjugating enzyme
E2A(E12)	6	3	transcription factor
PTP1E	2	2	tyrosine phosphatase
BCR	6	1	breakpoint cluster region
RAD23B	6	2	xeroderma pigmentosum, repair
c-Cbl	1	9	proto-oncogene c-Cbl
FADD	6	4	FADD
RXRa	6	22	retinoid X receptor, α
DP1	6	5	transcription factor
c-Raf1	5	15	viral oncogene homolog C1
E2F5	1	2	transcription factor
APO-1	6	3	apoptosis antigen 1

ProNet name: the name used to identify the gene product in the ProNet database. DC: dynamic connectivity, as described above. SC: static connectivity, as described above. Function: the function assigned to the protein in the ProNet database.

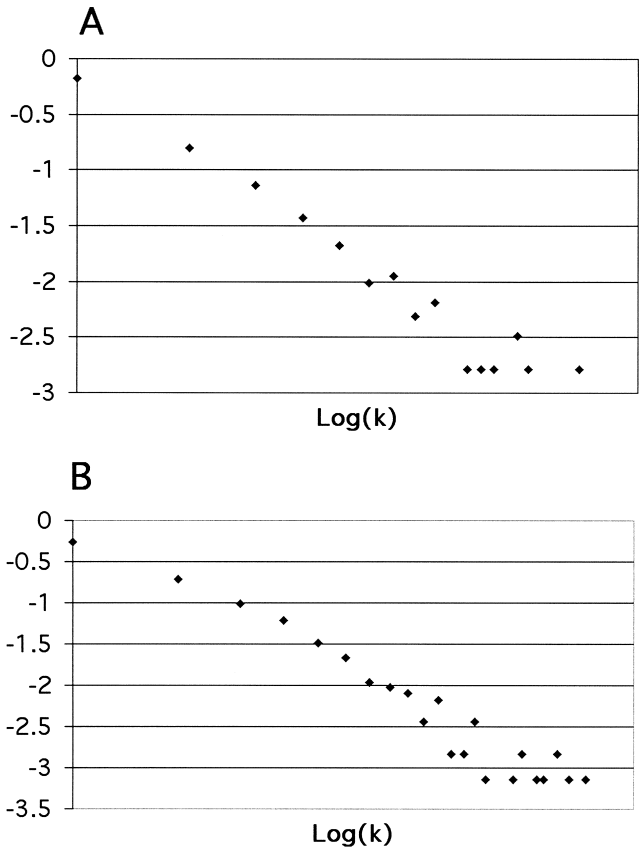

Fig. 2. Topology of protein–protein interaction maps. Connectivity distributions (Log($P(k)$) vs. Log(k)) for the apoptotic (A) and yeast (B) protein–protein interaction maps.

We next carried out a microarray experiment to detect SN and DN from upregulated apoptosis-associated genes during an in vivo global cerebral ischemia experiment [7]. We were able to trace protein paths connecting eight genes out of 36 upregulated genes. The protein E2F1 was predicted to be a DN (see Table 2). None of the other upregulated genes detected in the microarray experiment were predicted to be DN or SN (data not shown). Interestingly, previous studies have shown that an increased level of protein E2F1 is observed during neuronal cell death after an ischemic insult [16]. In that study it was reported that inhibiting an upstream effector of E2F1 reduced protein levels of E2F1 and reduced cell death by 80%, supporting a key role for E2F1 in ischemic cell death.

In summary, Nexxus is a bioinformatics approach that utilizes microarray and protein–protein interaction data to identify proteins with key roles in biomolecular mechanisms (SN and DN) in which regulation of protein–protein interactions is crucial. One advantage of this approach is the ability to detect genes crucial for biomolecular mechanisms from microarray data without any knowledge about their functions. In proof-of-principle studies, we have demonstrated the use of this

approach by determining static and dynamic connectivities, thus detecting SN and DN in three different cellular roles. As information on gene expression and protein–protein interactions accumulates, the accuracy and completeness of analysis by Nexxus should increase concomitantly, making Nexxus an increasingly valuable tool for the identification of nodal control points in biochemical pathways.

Acknowledgements: This work was supported by a Fogarty Fellowship to G.d.R. and by NIH Grants NS33376 and AG12282 to D.E.B. We appreciate Evan Hermel's suggestion of Nexxus as the name for the approach described, and the statistical assistance of Dr. Peter Bacchetti from the department of epidemiology and biostatistics at UCSF.

References

[1] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) Nature 407, 651–654.
[2] Pawson, T. and Nash, P. (2000) Genes Dev. 14, 1027–1047.
[3] Bredesen, D.E., Ye, X., Tasinato, A., Sperandio, S., Wang, J.J., Assa-Munt, N. and Rabizadeh, S. (1998) Cell Death Differ. 5, 365–371.
[4] Lockhart, D.J. and Winzeler, E.A. (2000) Nature 405, 827–836.
[5] Pandey, A. and Mann, M. (2000) Nature 405, 837–846.
[6] Walhout, A.J. and Vidal, M. (2001) Nat. Rev. Mol. Cell. Biol. 2, 55–62.
[7] Spellman, P.T. et al. (1998) Mol. Biol. Cell 9, 3273–3297.
[8] Chartrand, G. (1977) in: Introductory Graph Theory, pp 294, General Publishing Company, Boston, MA.
[9] Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Nature 411, 41–42.
[10] Uetz, P. et al. (2000) Nature 403, 623–627.
[11] Weiss, M. (1999) in: Data Structures and Algorithm Analysis in Java, pp. 291–349, Addison-Wesley, Redong, MA.
[12] Jin, K., Mao, X.O., Eshoo, M., Nagayama, T., Minami, M., Simon, R.P. and Greenberg, D.A. (2001) Ann. Neurol. 50, 93-103.
[13] Ogawa, H., Sata, Y., Takeshita, I., Tateishi, T. and Kitamura, K. (1985) Development 18, 133–141.
[14] Johnston, L.H., Eberly, S.L., Chapman, J.W., Araki, H. and Sugino, A. (1990) Mol. Cell. Biol. 10, 1358–1366.
[15] Kitada, K., Johnson, A.L., Johnston, L.H. and Sugino, A. (1993) Mol. Cell. Biol. 13, 4445–4457.
[16] Osuga, H. et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10254–10259.