the FEBS Journal

MINIREVIEW

Utilizing logical relationships in genomic data to decipher cellular processes

Peter M. Bowers[1,2,*], Brian D. O'Connor[3,*], Shawn J. Cokus[4], Einat Sprinzak[2], Todd O. Yeates[2,3] and David Eisenberg[1,2]

1 Howard Hughes Medical Institute, University of California, Los Angeles, CA, USA
2 Institute for Genomics and Proteomics, University of California, Los Angeles, CA, USA
3 Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA
4 Department of Mathematics, University of California, Los Angeles, CA, USA

Keywords
genomic data; logic analysis; microarray expression; phylogenetic profile

Correspondence
D. Eisenberg, Howard Hughes Medical Institute, University of California,
Los Angeles, Los Angeles, CA 90095, USA
Fax: +1 310 206 3914
E-mail: david@mbi.ucla.edu

Note
*These authors contributed equally to this work

(Received 25 May 2005, revised 26 July 2005, accepted 2 August 2005)

doi:10.1111/j.1742-4658.2005.04946.x

The wealth of available genomic data has spawned a corresponding interest in computational methods that can impart biological meaning and context to these experiments. Traditional computational methods have drawn relationships between pairs of proteins or genes based on notions of equality or similarity between their patterns of occurrence or behavior. For example, two genes displaying similar variation in expression, over a number of experiments, may be predicted to be functionally related. We have introduced a natural extension of these approaches, instead identifying logical relationships involving triplets of proteins. Triplets provide for various discrete kinds of logic relationships, leading to detailed inferences about biological associations. For instance, a protein C might be encoded within an organism if, and only if, two other proteins A and B are also both encoded within the organism, thus suggesting that gene C is functionally related to genes A and B. The method has been applied fruitfully to both phylogenetic and microarray expression data, and has been used to associate logical combinations of protein activity with disease state phenotypes, revealing previously unknown ternary relationships among proteins, and illustrating the inherent complexities that arise in biological data.

Introduction

The sequencing of genomes from diverse species, small and large, has tremendous potential to impact our understanding of biology by enabling both the identification of all proteins, and subsequently the analysis of their function. Understanding the network of biological linkages utilizing genomic information is becoming a realistic goal (see, for example [1–4]). Accomplishing this, however, will require the application of computational and experimental approaches to use massive amounts of relevant data to assemble biological networks, combining inferences and observations of protein–protein interactions derived from different data sources [5–12]. The integration of these types of data helps provide a complete view of cellular pathways and regulatory networks that regulate physiological processes. It is these linkages that also provide the basis for a precise understanding of cellular pathways, and ultimately, disease mechanisms, facilitating the development of therapeutics optimized for efficacy [13–15].

Abbreviations
CDK5R2, cyclin-dependent kinase 5, regulatory subunit 2; COG, clusters of orthologous groups; GLUT10, glucose transporter 10; GMFG, gliomal maturation factor gamma; KOG, eukaryotic orthologous group; NCF2, neutrophil cytosolic factor 2; PTPRT, protein tyrosine phosphatase, receptor type; SVD, singular value decomposition; TRHDE, thyrotropin-releasing hormone degradation enzyme.

Functional linkages

Computational tools, including the phylogenetic profile method, have been developed to detect functional linkages between proteins from the set of fully sequenced genomes [16–23]. A phylogenetic profile of a protein is a vector representing the presence or absence of the protein's orthologs encoded among the fully sequenced genomes. The result of a homology search across *n* genomes is an *n*-dimensional vector of ones and zeros for each protein, where the presence of a homolog in a given genome is indicated by a one, and the absence by a zero. Given a sufficient number of fully sequenced genomes, pairs of proteins exhibiting statistically similar patterns of presence or absence are hypothesized to be associated with the same biological function [5,18].

Complete genome sequences have also facilitated the development of experimental methods for collecting genome-scale data describing cellular processes [for example 6,7,12,15,24–27]. In particular, oligonucleotide expression data, which monitors transcription levels at each gene locus, has proved to be a powerful tool for characterizing biological processes and disease mechanisms. As with the phylogenetic profile method, analysis of microarray data normally attempts to associate genes displaying similar responses to experimental conditions, or to associate noteworthy genes with their presumed pathways, disease processes, or phenotypic outcomes. In particular, examination of gene expression in various tumor cell lines has permitted new concepts relating to tumorigenesis, which in turn led to novel disease concepts [15,25].

The phylogenetic profile and related methods of computational analysis use inferences derived from genomic data to help deduce the likelihood of protein linkage in a cellular network or process, without additional experimentation. The power of this approach is the ability to produce a model of network associations that acts as a reference point for scientists to generate hypotheses explaining cellular functions, where underlying molecular mechanisms have yet to be elucidated. Although the sequences of all of the proteins encoded by the genome may be known, only a fraction of the protein functions have been annotated, and our understanding of disease mechanisms is often rudimentary at best. This suggests that our understanding of both normal and pathological mechanisms within the cell is still underdeveloped relative to the proportion of supporting biological data that currently exists.

Algorithms

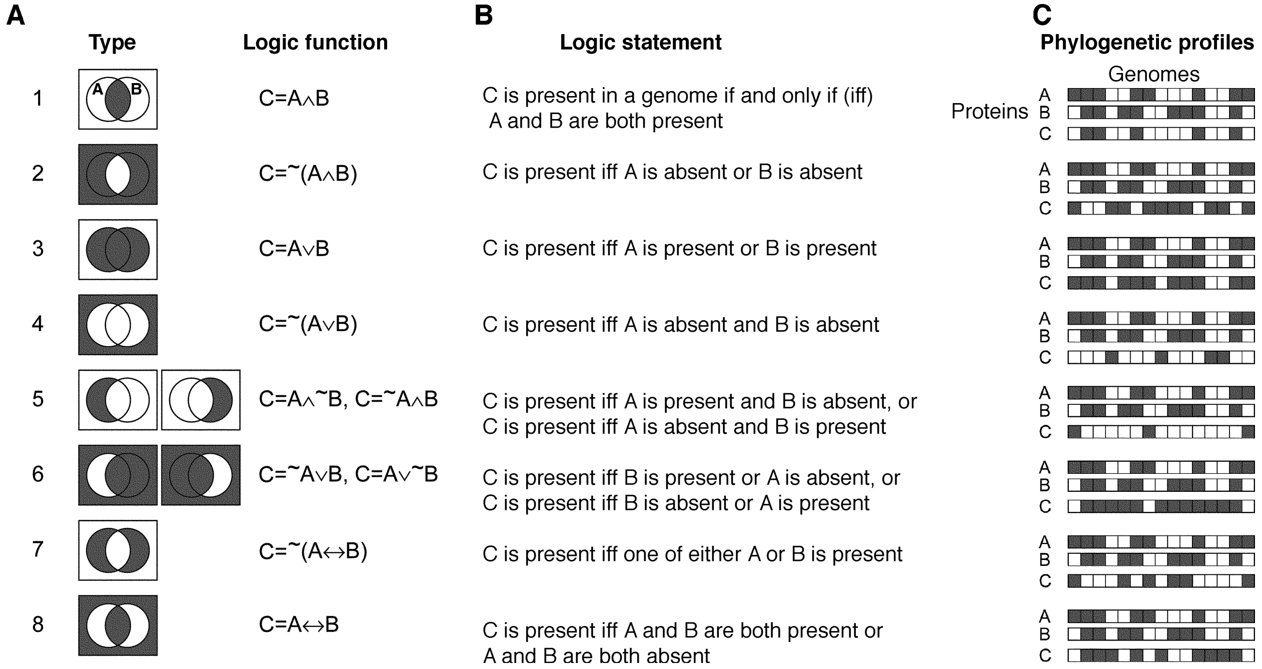
Statistical methods for associating biological entities in genome-wide data are numerous and can be described only briefly here [28]. Basic information metrics for associating data vectors include the Pearson correlation coefficient, Euclidean and Hamming distances, mutual information, the hypergeometric distribution and shortest-path anaylsis [29], to name but a few. Hierarchical clustering, employed by the software package CLUSTER developed by Eisen and colleagues [30], uses many of these metrics to organize associated proteins into a hierarchical tree, where local branches are intuitively understood to represent proteins involved in similar cellular functions or pathways [16,17,30]. Clustering of gargantuan biological data sets has also been furthered by the implementation of the K-means cluster (FUZZY K) and self-organizing maps (GENECLUSTER) methods that attempt to reduce the high dimensionality of genomic data, making its interpretation more accessible to the biologist [31,32]. Similarly, representing genomic data in terms of 'eigen-proteins' derived from singular value decomposition (SVD) can greatly aid in both noise reduction and classification of proteins into regulatory subgroups or functions [33]. An advantage of SVD analysis is that it allows a gene or experimental vectors to be described as linear combinations of 'basis' or eigenstates of the system. Expression deconvolution, developed by Marcotte and colleagues, demonstrated that cell cycle dynamics and replicative states of the cell, can be modeled as combinations of microarray expression profiles [34]. Analysis of genome data to identify associations between genes and phenotypes, cellular pathways, or clinical outcomes has also received a good deal of attention in the literature, particularly predictive analysis of cancer outcomes and phenotypes from microarray data [for example 15,25,35,36]. Analysis of genomic data, in the form of unsupervised learning, Bayesian analysis, logical regression, liquid association as well as the methods listed above, have all been applied to the identification of proteins that may predict cellular functions and disease states [35,37–40]. Logic regression analysis has been applied to single nucleotide polymorphism data to create weighted decision trees that link outcome phenotypes with sets of binary descriptors [35].

We sought to develop a method of analysis that would lead to the identification of novel biological associations and to specific hypotheses that could be experimentally tested. An ideal computational method would not only answer the question of which proteins interact, but also how these proteins might interact conditionally; for example, illuminating how they contribute to a cancer state, not simply which proteins were predictive or associated with a cancer type.

Triplets of phylogenetic profiles

We recently described methods of analysis that examine the possible logical relationships between triplets of phylogenetic profiles [41]. Rather than attempting to identify equality relationships between two protein profiles, we sought to locate instances in which the combined logical patterns embodied by two proteins determined the behavior of a third. In the context of phylogenetic analysis, a protein C might be encoded within a genome if, and only if, proteins A and B are also both encoded within the genome (denoted here as a type 1 logic relationship), from which we would infer that the function of protein C may be necessary exactly when the functions of proteins A and B are both present. Conversely, a protein C may be encoded within a genome if, and only if, either A or B (but not both) is encoded (a type 7 logic relationship), which may be seen when organisms choose between two different but functionally equivalent protein families in combination with a common third protein to accomplish some task [(A and C) or (B and C)] (Fig. 1). A software package that performs the analysis on a binary matrix can be found at http://www.doe-mbi.ucla.edu/~bowers/Triples/. Figure 1 illustrates all eight possible logic relationships combining two binary states to match a third state.

We systematically examined phylogenetic data, in the form of binary presence/absence vectors, in an attempt to identify the logic relationships described in Fig. 1 [41]. Binary-valued phylogenetic vectors were generated, describing the presence or absence of each of 4800 protein families in 67 organisms, also known as clusters of orthologous groups (COG) [42,43]. Triplet combinations of profiles were identified within the set, and rank-ordered according to the information captured in the profile triplet that was not found in each of the individual pairwise comparisons. We identified logical combinations of vectors A and B, which, when combined, were better able to describe a protein



Type	Logic function	Logic statement
1	$C=A\wedge B$	C is present in a genome if and only if (iff) A and B are both present
2	$C={}^{\sim}(A\wedge B)$	C is present iff A is absent or B is absent
3	$C=A\vee B$	C is present iff A is present or B is present
4	$C={}^{\sim}(A\vee B)$	C is present iff A is absent and B is absent
5	$C=A\wedge{}^{\sim}B$, $C={}^{\sim}A\wedge B$	C is present iff A is present and B is absent, or C is present iff A is absent and B is present
6	$C={}^{\sim}A\vee B$, $C=A\vee{}^{\sim}B$	C is present iff B is present or A is absent, or C is present iff B is absent or A is present
7	$C={}^{\sim}(A\leftrightarrow B)$	C is present iff one of either A or B is present
8	$C=A\leftrightarrow B$	C is present iff A and B are both present or A and B are both absent

Fig. 1. Detection of pathway relationships among proteins, based on a logic analysis of phylogenetic profiles (adapted from Bowers *et al.*) [41]. Triplets of proteins are considered, where the presence or absence of a third protein C across numerous genomes is a logic function of the presence or absence of two other proteins, A and B. (A) Venn diagrams and associated logic statements illustrate the eight distinct kinds of logic functions that describe the possible dependence of the presence of C on the presence of A and B, jointly. For example, logic type 1 describes the case in which protein C is present in a genome, if and only if, A and B are both present. Logic functions are grouped together if they are related by a simple exchange of proteins A and B. The symbols, '∧', '∨', '~', and '↔', indicate 'logical AND', 'logical OR', 'logical negation' and 'logical equality', respectively. (B) The meaning of each logic relationship is described in a single text sentence, and (C) hypothetical phylogenetic profiles are used to illustrate the eight possible logic functions.

vector C than either of the vectors A or B alone, such that;

$$\mathrm{U}[c, f(a,b)] \gg \mathrm{U}(c|a) \text{ and } \mathrm{U}(c|b)$$

$$\text{where } \mathrm{U}(c|a) = [\mathrm{H}(c) + \mathrm{H}(a) - \mathrm{H}(c,a)]/\mathrm{H}(c)$$

$$\text{and } \mathrm{H}(a) = \sum \mathrm{p}(a)\ln(\mathrm{p}(a))$$

$$\text{and } \mathrm{H}(c,a) = \sum\sum \mathrm{p}(c,a)\ln(\mathrm{p}(c,a))$$

where U refers to the uncertainty coefficient (referred to hereafter as an information coefficient) comparing either the logically combined vectors or individual vectors A or B with vector C, conditioned on the information available in vector C, and where f is one of eight possible logic functions. The value of U can range between 1.0 (complete information) and 0.0 (no information). We sought those triplets where the individual pairwise comparisons provided significantly less information ($\mathrm{U}(c|a) < 0.40$ and $\mathrm{U}(c|b) < 0.40$) than the logically combined vectors [$\mathrm{U}(c|f(a,b)) > 0.6$).

We found that a logic analysis of COG phylogenetic profiles revealed thousands of relationships among protein families that cannot be detected using traditional pairwise analysis. In our original manuscript [41], we provided several examples from basic sugar and amino acid metabolism. For instance, the interconversion of the 5-carbon sugar ribose to the 6-carbon sugar 6-phosphogluconate constitutes a central pathway in carbohydrate metabolism, and is accomplished by three successive enzymatic steps. The proteins are not linked using a traditional pairwise phylogenetic analysis. However, a logic analysis recognizes a type 3 logical relationship, such that when either of the terminal enzymatic steps, carried out by COG0524 (EC 2.7.1.15) and COG0362 (EC 1.1.1.44), are present in an organism, the intervening enzymatic step, carried out by ribose-5-phosphate isomerase COG0120 (EC 5.3.1.6), is also present.

Amongst the 4800 COG protein families, our logic analysis of phylogenetic profiles recovered approximately three million new links among protein families (out of a possible 62 billion), whose accuracy was validated by several benchmarking methods. The ability to recover links between proteins annotated as belonging to a major functional category has been used widely to corroborate computational inferences of protein interactions. Observed triplet relationships frequently relate three proteins all belonging to the same COG category, or involve two proteins from the same category and a third from a second category, indirectly confirming that the logical associations link proteins closely related in cellular function. Triplets with information coefficient scores $\mathrm{U} > 0.60$ were observed with a frequency $\approx 10^2$-fold greater than that observed from shuffled profiles with an equivalent information content. Finally, the eight distinct logic types occurred with widely varying frequencies, with types 1, 3, 5 and 7 being especially common. In contrast, logic types 2 and 8 are difficult to relate to simple cellular logic, and these patterns are observed much less frequently in the data.

Logic analysis of microarray expression data

Can the logic analysis technique also be applied successfully to other types of genomic data? We analyzed logical relationships within microarray expression data, with attention to identifying logical combinations of proteins that led directly to the observation of clinical outcomes. Previous work has used a binary-only representation of gene expression data to examine the mechanics of gene regulation networks [44,45]. Schmulevich *et al.* [45] have shown, for example, that glioma tumor types can be segregated using a binary representation of expression data. Because the cancer microarray dataset contains descriptors describing clinical outcomes and tumor types, we were also able to explore whether logical relationships can identify meaningful sets of genes that match clinical outcomes.

Here, we show how the triplet logic idea can be extended to treat microarray expression data. As an application of triplet logic analysis to expression data, samples were chosen from Freije *et al.*, representing 85 diffuse infiltrating gliomas quantified using oligonucleotide arrays [25]. Each tumor sample was annotated with additional information including tumor type, grade, and patient survival clustered into four prognosis groups. The dataset was converted to binary data suitable for use with the logic analysis method using the MICROARRAY SUITE 5 (MAS5) algorithm with the default presence or absence thresholds, resulting in 22 000 binary expression vectors. Once converted, the set was supplemented with 12 additional phenotype profiles that represented the annotations of disease/tumor properties, where a zero represents the absence of a phenotypic trait, and a one indicates the presence of the phenotype [25]. The resulting binary profiles were then examined using a logical analysis as previously described [41]. Logical combinations of two genes expression profiles were compared to 12 phenotype profiles using the eight possible logic types. In this way, general phenotypes and observations were related to gene expression patterns derived from the samples.

The result was 1341 logical relationships identified, for which the two separate gene profiles each have an uncertainty U < 0.4 when compared to the phenotype profile, yet when logically combined their uncertainty score is 0.6 or greater with respect to the phenotype profile.

In Fig. 2A, a set of binary expression and phenotype profiles taken from a gliomal microarray dataset illustrate the method. Under a type 1 logic relationship, phenotype C is present when gene A and gene B are also both expressed within the cancer cell line. The pairwise comparisons of profiles A and C (U = 0.33, P < 1e-9) and B and C (U = 0.39, P < 1e-8) contain less information and are statistically more likely to be observed by chance than a logical combination of proteins A and B matching the profile of phenotype C (U = 0.65, P < 1e-16). Here, the P-values associated with each information coefficient were calculated using a standard hypergeometric distribution analysis of the individual and combined vectors. Thus the information coefficient, U, is able to identify statistically significant triplet relationships from the microarray expression profiles.

The distribution of observed logic types satisfying our selection criteria, as shown in Fig. 2B, is dominated by logic type 5 (XOR) and, to a lesser extent, logic type 1 (AND). These logic types were also commonly observed in the phylogenetic profile analysis [41] and in the analysis of other microarray data sets (data not shown). Randomized trials, carried out as

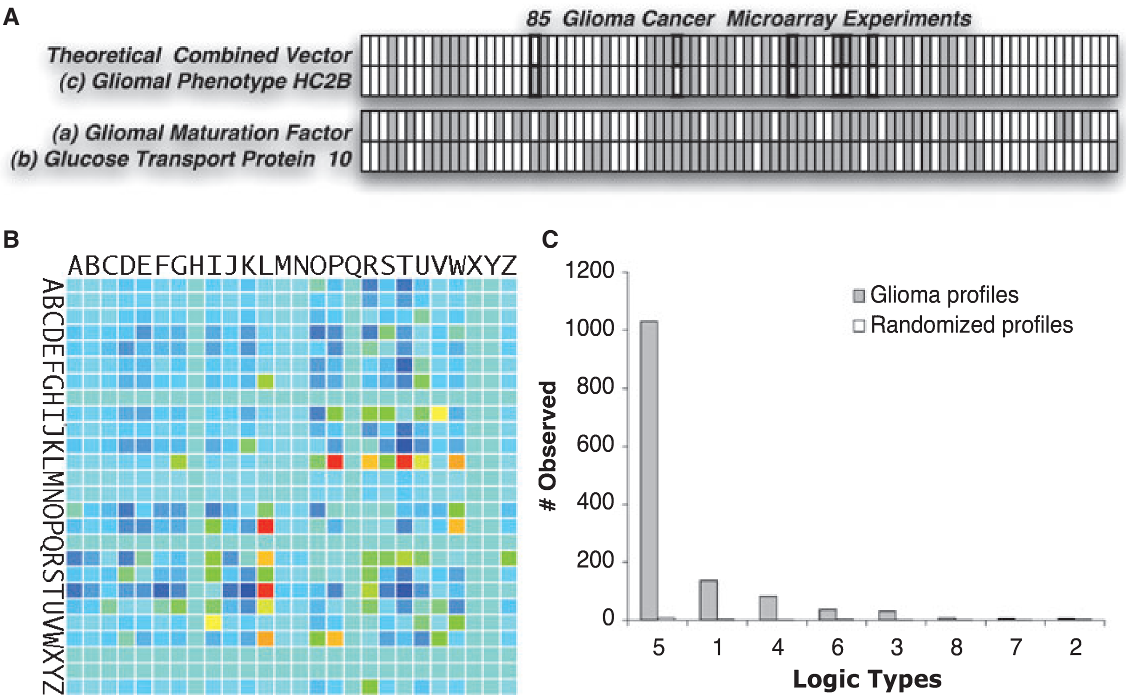


Fig. 2. Microarray experiments for 85 glioma samples were used in the logic analysis method to detect relationships in triplets of genes and phenotypes combined with one of eight logical operators. (A) Eighty-five glioma microarray experiments are shown in binary form, where ■ indicates the presence of an mRNA representing a given gene of interest, and □ indicates the absence of detected mRNA in the sample. The bottom two rows represent the binary profiles of gliomal maturation factor gamma (GMFG) (a) and glucose transporter 10 (SLC2A10) (b), respectively. When logically combined, the theoretical combined vector (top row) is produced, which closely matches the binary profile (c) of the gliomal phenotype HC_2B, a poor prognosis group, with bold boxes indicating experiments where the combined and real profiles are mismatched. (B) A heat-map showing biases in a pairwise comparison of annotations from pairs of probe-sets identified as matching a phenotype profile with a combined uncertainty U($c|f(a,b)$) > 0.6. Each gene was annotated with a KOG category and, for those pairings of two annotated genes, a tally of KOG category pairings was maintained. Observed values were normalized to a Z-score with randomized trials repeated 500 times. Red signifies a five-fold increase in the observed frequency, relative to the expected frequency, and light blue signifies no change relative to the expected frequency of category pairings. KOG categories observed with increased frequency include L (replication and repair), P (inorganic ion transport and metabolism), T (signal transduction), and W (extracelluar structures). (C) The distribution of logic relationship types in significant triplets; 1341 in total for the gliomal profiles were identified that met the selection criteria. Most were dominated by logic type 5 (XOR) and, to a lesser extend logic type 1 (AND). Trials using randomized phenotype profiles are also plotted, confirming that only a very small number of triplet profiles meeting the selection criteria would be observed by chance.

described previously, were used to ascertain whether the inferred logical relationships were statistically meaningful. Each of the 12 phenotype profiles in the dataset was randomized 100 times and analyzed. On average, fewer than four logical triplets were identified per randomized trial for each phenotype, strongly suggesting the 1341 logical triplets were not identified by chance (Fig. 2B).

To examine overall relations between the gene and phenotype profiles identified we annotated general functional categories for each gene profile and looked for biases in the distribution of annotations across profile pairs. This technique has been used previously to validate logic analysis-derived relationships between protein triplets across COGs [41]. Similar approaches have also been used to corroborate inferences of protein relationships through recovery of known protein annotations [21,22]. Each gene profile was annotated using one or more major eukaryotic orthologous group (KOG) functional categories [42]. Pairs of annotated gene profiles were then examined and the groupings of KOG category annotations were tabulated. The pairwise comparison of KOG categories for annotated probe-set pairs were then normalized to *z*-values using 500 randomized trials and plotted in Fig. 2C. Several annotations appear together in the logical relationships more often than predicted by chance. These most notably include KOG categories L (replication and repair), P (inorganic ion transport and metabolism), T (signal transduction), and W (extracelluar structures). Interestingly, the biases in these category pairings seems to be specific to a cancer dataset, as a normal tissue dataset previously examined with the logic analysis process showed less enrichment for all categories but T.

A glioma cancer phenotype corresponding to a poor prognosis outcome (HC_2B) was selected for further analysis [25]. Ideally, the proteins that logically combined to match a poor prognosis cancer phenoytype should have annotated cellular functions that might reasonably be expected to influence cancer disease mechanisms. GLUT10, a member of the facilitative glucose transporter family [46], was found to be linked in eight different logical triplets, all of which relate it, and another neuronal protein, to the HC_2B phenotype outcome from Freije *et al.* (Fig. 3). The HC_2B phenotype represents a poor prognosis group and has been linked to enrichment for genes coding for extracellular matrix components. GLUT10 is itself interesting because malignant cellular growth has been previously noted to be characterized by and dependent on increased glucose transport. A study by Matsuzu *et al.* previously identified glucose transporter 10 as being up-regulated in thyroid cancer using real-time PCR [46]. Interesting, most of the genes identified in GLUT10-containing profiles seen in Fig. 3 seem to play some potential role in cancer and are involved in informative logical combinations with GLUT10.

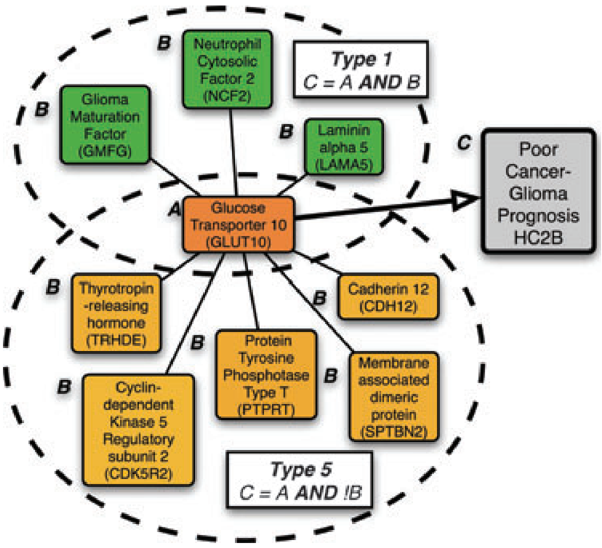

Type 1
C = A AND B
B
Neutrophil Cytosolic Factor 2 (NCF2)
B
Glioma Maturation Factor (GMFG)
B
Laminin alpha 5 (LAMA5)
C
Poor Cancer-Glioma Prognosis HC2B
A
Glucose Transporter 10 (GLUT10)
B
Thyrotropin -releasing hormone (TRHDE)
B
Cadherin 12 (CDH12)
B
Protein Tyrosine Phosphotase Type T (PTPRT)
B
Membrane associated dimeric protein (SPTBN2)
B
Cyclin-dependent Kinase 5 Regulatory subunit 2 (CDK5R2)
B
Type 5
C = A AND !B

Fig. 3. Proteins logically related to the presence or absence of the glucose transport protein GLUT10 define a poor gliomal cancer phenotype outcome. Each logical relationship related GLUT10 and one other protein to the HC_2B poor prognosis glioma cluster through either a type 1 logic (AND) or type 5 logic (XOR) relationship. Those proteins that logically related to the GLUT10 transport protein via a type 1 logic (AND) relationship (shown in green) perform growth stimulatory or growth differentiation roles within the cell. Proteins that logically combine with GLUT10 via the type 5 logic (XOR) relationship to affect a poor prognosis phenotype are believed to execute inhibitory roles (shown in orange). The model suggests that changes to multiple protein expression patterns are required to obtain an aggressive cancer phenotype, including the down-regulation of several inhibitory proteins, and the up-regulated on several known oncogenes.

Gliomal maturation factor gamma (GMFG) and neutrophil cytosolic factor 2 (NCF2) [47,48] are both related, with GLUT10, to the negative phenotype outcome with an AND logical relationship (phenotype c = a AND b), indicating that both are necessary if the sample is annotated as HC_2B. Both tumor genes have been previously linked to roles suggestive of oncogenic properties within the cell. GMFG is important for the development of glia and neurons where it seems to have a stimulatory role for growth and differentiation. Likewise, NCF2 is involved in oxidase regulation and its expression is linked to respiratory bursts during differentiation. The genes that combine with GLUT10 in an exclusive or (XOR) relationship to give the poor prognosis outcome appear to affect various inhibitory roles within the cell. For instance, thyrotropin-releasing hormone degradation enzyme (TRHDE), protein tyrosine phosphatase, receptor type (PTPRT), cadherin 12 (CDH12), and cyclin-dependent kinase 5, regulatory subunit 2 (CDK5R2) all appear to fulfil roles of inhibitory regulators of cell growth and differentiation [49–52]. TRHDE degrades thyrotropin-releasing hormone which itself is an important stimulator of hormone secretion from the pituitary. Mutations in PTPRT and other tyrosine phosphatases have been shown to be mutated in human cancers and their general inhibitory role on cell growth supports a tumor suppressor role in the cell. Finally, cadherin 12 has previously been shown to be under-expressed in ameloblastoma tumors while CDK5R2 has been implicated in mediating apoptosis in human glioblastoma multiform cells. Together these observations support a model in which a negative cancer phenotype HC_2B is logically linked to GLUT10 in combination with several proteins that either inhibit or enhance cancer progression. Most strikingly, the observations highlighted in Fig. 3 lead directly to a hypothesis regarding which proteins and protein interactions affect a change in measurable phenotypic outcome.

Conclusions

The ultimate goal of genomics research is to describe the cellular networks of molecules and interactions that govern all biological functions and disease processes. Simple pairwise associations between proteins and between proteins and disease states lack significant detail, and presumably a fully realized cellular model will contain additional temporal, spatial, directional and conditional information. Computational methods for analysis of genomic data would ideally create not only associations between data, but lead to intuitive and biologically grounded hypotheses with details as to how the proteins or entities are related. Our logical analysis begins to address these issues by identifying thousands of new, higher order associations and by providing a framework for understanding the complex logical dependencies that relate proteins to other proteins, phenotypes, single nucleotide polymorphisms, and other biological features within the cell.

In earlier work, functional relationships among cellular proteins were analyzed by combining both genomic and microarray data [21]. In that study, Marcotte *et al.* integrated these two types of data, for finding pairwise functional relations among the ≈6000 yeast *Saccharomyces cerevisiae* proteins. This analysis demonstrated that the integrative approach enabled more accurate assignment of function than using each data type separately [21]. In general, integration of different data sources helps to uncover nonobvious relationships between genes and also increases the reliability of the interpretation of experimental results. We show here that adding logical analysis can define additional types of relationships among biological data. Extension of such methods of combining genomic, microarray, and other data appears to be a fruitful area for developing more powerful bioinformatics tools.

Acknowledgements

B.O. was supported by a USPHS National Research Service Award GM07185. This work was supported by NIHGM31299 and the DOE Office of Science, Biological and Environmental Research.

References

1 Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.

2 Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543.

3 Lee I, Date SV, Adai AT & Marcotte EM (2004) A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558.

4 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736.

5 Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO & Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**, R35.

6 Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.

7 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M & Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* **98**, 4569–4574.

8 von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA & Bork P (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33**, D433–D437.

9 von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P & Snel B (2003) STRING: a database of predicted

functional associations between proteins. *Nucleic Acids Res* **31**, 258–261.

10 Yanai I & DeLisi C (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol* **3**, research0064.1–research0064.12.

11 Uetz P & Hughes RE (2000) Systematic and large-scale two-hybrid screens. *Curr Opin Microbiol* **3**, 303–308.

12 Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.

13 Crooke ST (1998) Optimizing the impact of genomics on drug discovery and development. *Nat Biotechnol* **16** (Suppl.), 29–30.

14 Weinstein JN (2002) 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. *Curr Opin Pharmacol* **2**, 361–365.

15 van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

16 Strong M, Mallick P, Pellegrini M, Thompson MJ & Eisenberg D (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol* **4**, R59.

17 Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ, Yeates TO & Eisenberg D (2003) Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res* **31**, 7099–7109.

18 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D & Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96**, 4285–4288.

19 Overbeek R, Fonstein M, D'Souza M, Pusch GD & Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* **96**, 2896–2901.

20 Overbeek R, Fonstein M, D'Souza M, Pusch GD & Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* **1**, 93–108.

21 Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO & Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86.

22 Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO & Eisenberg D (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753.

23 Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90.

24 Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.

25 Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, Mischel PS & Nelson SF (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* **64**, 6503–6510.

26 Eisen MB & Brown PO (1999) DNA arrays for analysis of gene expression. *Methods Enzymol* **303**, 179–205.

27 Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D & Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**, 41–46.

28 Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* **32** (Suppl.), 502–508.

29 Zhou X, Kao MC & Wong WH (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA* **99**, 12783–12788.

30 Eisen MB, Spellman PT, Brown PO & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**, 14863–14868.

31 Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES & Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* **96**, 2907–2912.

32 Gasch AP & Eisen MB (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol* **3**, research0059.

33 Alter O, Brown PO & Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* **97**, 10101–10106.

34 Lu P, Nakorchevskiy A & Marcotte EM (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc Natl Acad Sci USA* **100**, 10370–10375.

35 Ruczinski I, Kooperberg C & LeBlanc ML (2003) Logic Regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

36 Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA & Bork P (2005). Systematic Association of Genes to Phenotypes by Genome and Literature Mining. *PLoS Biol* **3**, e134.

37 Li KC, Liu CT, Sun W, Yuan S & Yu T (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc Natl Acad Sci USA* **101**, 15561–15566.

38 Friedman N, Linial M, Nachman I & Pe'er D (2000) Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601–620.

39 Barash Y & Friedman N (2002) Context-specific Bayesian clustering for gene expression data. *J Comput Biol* **9**, 169–191.

40 Kooperberg C, Ruczinski I, LeBlanc ML & Hsu L (2001) Sequence analysis using logic regression. *Genet Epidemiol* **21** (Suppl. 1), S626–S631.

41 Bowers PM, Cokus SJ, Eisenberg D & Yeates TO (2004) Use of logic relationships to decipher protein network organization. *Science* **306**, 2246–2249.

42 Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

43 Tatusov RL, Koonin EV & Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**, 631–637.

44 Liang S, Fuhrman S & Somogyi R (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput* 18–29.

45 Shmulevich I & Zhang W (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* **18**, 555–565.

46 Matsuzu K, Segade F, Matsuzu U, Carter A, Bowden DW & Perrier ND (2004) Differential expression of glucose transporters in normal and pathologic thyroid tissue. *Thyroid* **14**, 806–812.

47 Gauss KA, Bunger PL, Larson TC, Young CJ, Nelson-Overton LK, Siemsen DW & Quinn MT (2005) Identification of a novel tumor necrosis factor alpha-responsive region in the NCF2 promoter. *J Leukoc Biol* **77**, 267–278.

48 Inagaki M, Aoyama M, Sobue K, Yamamoto N, Morishima T, Moriyama A, Katsuya H & Asai K (2004) Sensitive immunoassays for human and rat GMFB and GMFG, tissue distribution and age-related changes. *Biochim Biophys Acta* **1670**, 208–216.

49 Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS *et al.* (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304**, 1164–1166.

50 Catania A, Urban S, Yan E, Hao C, Barron G & Allalunis-Turner J (2001) Expression and localization of cyclin-dependent kinase 5 in apoptotic human glioma cells. *Neuro-Oncol* **3**, 89–98.

51 Heikinheimo K, Jee KJ, Niini T, Aalto Y, Happonen RP, Leivo I & Knuutila S (2002) Gene expression profiling of ameloblastoma and human tooth germ by means of a cDNA microarray. *J Dent Res* **81**, 525–530.

52 Schomburg L, Turwitt S, Prescher G, Lohmann D, Horsthemke B & Bauer K (1999) Human TRH-degrading ectoenzyme cDNA cloning, functional expression, genomic structure and chromosomal assignment. *Eur J Biochem* **265**, 415–422.