Reverse engineering of regulatory networks in human B cells

Katia Basso¹, Adam A Margolin², Gustavo Stolovitzky³, Ulf Klein¹, Riccardo Dalla-Favera¹,⁴ & Andrea Califano²

Cell phenotypes are determined by the differential activity of networks linking coregulated genes. Available methods for the reverse engineering of such networks from genome-wide expression profiles have been successful only in the analysis of lower eukaryotes with simple genomes. Using a new method called ARACNe (algorithm for the reconstruction of accurate cellular networks), we report the reconstruction of regulatory networks from expression profiles of human B cells. The results are suggestive of a hierarchical, scale-free network, where a few highly interconnected genes (hubs) account for most of the interactions. Validation of the network against available data led to the identification of MYC as a major hub, which controls a network comprising known target genes as well as new ones, which were biochemically validated. The newly identified MYC targets include some major hubs. This approach can be generally useful for the analysis of normal and pathologic networks in mammalian cells.

¹Institute for Cancer Genetics and ²Joint Centers for Systems Biology, 1300 St. Nicholas Avenue, Room 912, New York, New York 10032, USA. ³IBM T.J. Watson Research Center, Yorktown Heights, New York, New York 10598, USA. ⁴Departments of Pathology and Genetics & Development, Columbia University, New York, New York 10032, USA. Correspondence should be addressed to A.C. (califano@c2b2.columbia.edu).

Published online 20 March 2005; doi:10.1038/ng1532
transmission theory, the ‘data processing inequality’ (DPI), which had not been previously applied in this context. Hence, relationships included in the final reconstructed network have a high probability of representing either direct regulatory interactions or interactions mediated by post-transcriptional modifiers that are undetectable from gene-expression profiles.

An essential requirement of any reverse engineering method is the availability of a large set of gene-expression profile data representative of perturbations of the cellular systems, leading to the analysis of a broad range of cellular states and associated gene-expression levels. This is necessary because genetic interactions are best inferred when the genes explore a substantial dynamical range. Traditionally, this has been achieved by systematic perturbations in simple organisms (e.g., by large-scale gene knockouts or exogenous constraints), which are not easily obtained in more complex cellular systems. We show here that an equivalent dynamic richness can be efficiently achieved by assembling a considerable number of naturally occurring and experimentally generated phenotypic variations of a given cell type. To this end, we applied ARACNe to genome-wide expression profiles from a panel of 336 B cell phenotypes representative of a wide selection of normal, transformed and experimentally manipulated human B cells related to the germinal center, a structure in which selection of normal, transformed and experimentally manipulated B cells is achieved. This has been achieved by systematic perturbations in simple organisms (e.g., by large-scale gene knockouts or exogenous constraints), which are not easily obtained in more complex cellular systems. We show here that an equivalent dynamic richness can be efficiently achieved by assembling a considerable number of naturally occurring and experimentally generated phenotypic variations of a given cell type. To this end, we applied ARACNe to genome-wide expression profiles from a panel of 336 B cell phenotypes representative of a wide selection of normal, transformed and experimentally manipulated human B cells related to the germinal center, a structure in which B cells are selected on the basis of their ability to produce antibodies with high affinity for the antigen and from which many types of B cell lymphomas are derived.

ARACNe reconstructed a network suggestive of a hierarchical, scale-free organization, with a power-law relationship between the number of genes and their connectivity. In the inferred network, a relatively small number of highly connected genes interact with most other genes in the cell, either directly or hierarchically, through other highly connected subhubs. The proto-oncogene MYC emerged as one of the largest hubs in the network. In-depth analysis of this gene uncovered a hierarchically organized subnetwork, which we used to validate the predictions against an extensive number of previously identified targets of the MYC transcription factor. ARACNe recapitulated known MYC target genes and identified new candidate targets, which we then validated by biochemical analysis. The success of ARACNe in the genome-wide identification of gene networks indicates that the method can be useful in the dissection of both normal and pathologic mammalian phenotypes.

**RESULTS**

**Validation on a synthetic network**

As a first test of ARACNe’s performance, we compared it to Bayesian networks, which are among the best reverse engineering methods available, by analyzing a synthetic genetic network model that is part of a project aimed at integrating the songbird brain. ARACNe correctly identified more interactions than Bayesian networks (13 versus 11 true positives) and incorrectly identified substantially fewer relationships (2 versus 11 false positives; Fig. 1a). We then systematically compared the sensitivity (percentage of correctly inferred true interactions) and precision (percentage of correct interactions among all inferred ones) of the two methods as a function of the number of available synthetic samples (ranging between 100 and 2,000; Fig. 1b). Overall, ARACNe was comparable to Bayesian networks in sensitivity and largely superior in precision.

**The B cell network has hierarchical scale-free behavior**

We then used ARACNe to deconvolute cellular networks from a set of 336 expression profiles representative of perturbations of B cell phenotypes, including normal B cell subpopulations, various subtypes of B cell tumors and experimentally manipulated B cells (Fig. 2 and Supplementary Table 1 online). Normal cells included resting pre–germinal center naive B cells, proliferating germinal center B cells (centroblasts and centrocytes) and post–germinal center memory B cells. Transformed cells were represented by panels of cell lines and biopsies representative of the more than ten subtypes of B cell malignancies. Experimentally manipulated B cells included cell lines that were treated in vitro to induce specific signal transduction pathways or engineered for the conditional expression of several transcription factors.

ARACNe inferred a network with ~129,000 interactions (Supplementary Table 2 online), which may include those specific to normal and malignant B cells. Given the relatively small number of known phenotypes derived from normal mature B cells, an
equivalent network for normal B cells could not be produced with acceptable accuracy. Because the inferred network is too complex to be shown in its entirety, we summarized its global connectivity properties (Fig. 3). The results show a power-law tail in the relationship between the number of genes, $n$, in the network and their number of interactions, $k$. This tail extends over slightly more than one order of magnitude. This is suggestive of a scale-free underlying network structure. The deviation of the curve at low values of connectivity ($k < 12$) from the power-law extrapolated by linear fit from larger connectivity values ($k \geq 12$) is probably a consequence of the limited number of available genes. In fact, only $\sim 6,000$ genes on the microarray have enough dynamic range in our expression profiles to allow the inference of at least one interaction. According to the theoretical curve, however, $\sim 100,000$ genes would be needed to assess the power-law behavior down to $k = 1$.

**Construction of the MYC subnetwork**

As expected given the scale-free nature of the network, a small percentage of genes accounts for most of the connections. We defined as major hubs the largest $5\%$ of hubs in the network; they collectively participate in $\sim 50,000$ interactions, almost as many as the remaining $95\%$ of genes combined. Using the Gene Ontology classification and the GOMiner tool, we analyzed the biological processes affected by the top $5\%$ of hubs and identified substantial enrichment for genes involved in essential cellular processes such as cell cycle regulation, protein synthesis and catabolism, RNA processing and metabolism, and transcription (Supplementary Table 3 online).

The ARACNe-generated network could be used to identify the subnetworks associated with any gene of interest. We chose to study the proto-oncogene MYC, which emerged as one of the top $5\%$ of largest cellular hubs. Because MYC is extensively characterized as a transcription factor, we could compare the ARACNe-inferred interactions with those previously identified by biochemical methods. Because ARACNe should be able to identify genes both upstream and downstream of MYC, we expected the subset of genes directly connected to MYC by an edge in the network (first neighbors) to be significantly enriched in transcriptionally regulated targets. To identify the MYC subnetwork, we selected from the complete network only those genes that had statistically significant mutual information with any of the available probe sets for MYC ($P < 10^{-7}$). The inferred subnetwork structure included 2,063 genes, 56 of which were directly connected to MYC (Fig. 4 and Supplementary Table 4 online). Previously reported MYC target genes were identified in the subnetwork using as reference the MYC database.

The network has several limitations: (i) edges lack directionality (i.e., they do not indicate which gene is ‘upstream’ or ‘downstream’); (ii) some direct connections may involve unknown intermediates, as not all biochemical species participating in cellular interactions are represented on the microarray (e.g., missing probes or post-transcriptionally modified intermediates); (iii) some direct interactions may have been incorrectly removed by the DPI (e.g., any regulatory loop formed by three interacting genes would result in the weakest of the three interactions being removed, unless its strength is within $15\%$ of that of the intermediate interaction).

**The in silico network is enriched in known MYC targets**

Twenty-nine of the 56 ($51.8\%$) predicted first neighbors were previously reported (or biochemically validated here) as MYC targets. This represents a highly significant enrichment ($P = 1.8 \times 10^{-22}$ by
Figure 4 The MYC subnetwork. (a) A MYC-specific subnetwork was obtained by including all the genes that have \( P < 10^{-7} \) based on their pairwise mutual information with MYC. The faster bin-counting estimator was used with an error tolerance \( \epsilon = 0.15 \). The MYC subnetwork includes 56 genes directly connected to MYC (first neighbors; represented by larger circles) and 2,007 genes connected through an intermediate (second neighbors). For representation purposes, only the first 500 genes are shown, including all 56 first neighbors and the 444 most statistically significant second neighbors. Red or pink nodes represent first neighbor target genes for which ChIP data is available or not available, respectively; yellow and light yellow nodes represent second neighbor target genes for which ChIP data is available or not available, respectively; MYC is shown in green; white nodes represent genes for which no MYC-related information is available. The complete list of genes, including gene symbol, Affymetrix ID and LocusLink ID, is given in Supplementary Table 4 online.

(b) The first neighbors of the MYC subnetwork. The size of each circle is proportional to the number of the gene interactions. For hubs with more than 100 interactions, the exact number of first neighbors is shown beside the gene symbol.
$\chi^2$ test) with respect to the expected 11% of MYC targets among randomly selected genes\(^{24}\). In addition, known MYC target genes were significantly more enriched among first neighbors than second neighbors (51.8\% versus 19.4\%; $P = 1.4 \times 10^{-9}$), indicating that ARACNe is effective at separating direct regulatory interactions from indirect ones (Fig. 5). Moreover, 37.5\% of first neighbors (as opposed to 12\% of second neighbors) were also validated by chromatin immunoprecipitation (ChIP) assay, either previously or as part of this study. In total, 419 of the 2,063 (20\%) genes in the predicted MYC subnetwork were known MYC targets. Some bona fide direct interactions may be identified by ARACNe as second neighbors for two reasons: (i) if the gene is involved in a three-gene loop and the edge representing the interaction with MYC is the weakest, it is removed by the DPI; and (ii) if the interaction between the two genes requires a cofactor expressed only in a subset of the data, mutual information may be too low to be detected. Overall, we identified $\sim$40\% of the genes reported in the MYC database and validated in human B cells.

**Candidate new MYC targets are biochemically validated**

We validated new candidate MYC targets biochemically by ChIP assay to show direct binding of MYC to their promoter regions in vivo. We identified the 34 first neighbors present among the top 100 genes ranked based on mutual information and examined 2 kb of genomic DNA upstream and downstream from the transcription initiation site for the presence of canonical MYC binding sites (E boxes, CACGTG; Supplementary Table 5 online). After eliminating 9 genes as previously reported MYC direct targets, we selected a representative set of 12 genes for validation by ChIP in the Ramos B cell lymphoma line. A large fraction (11 of 12) of the tested genomic sequences could be immunoprecipitated by specific MYC antibodies but not by control antibodies (Fig. 6), indicating that MYC was bound to these regions in vivo. Among them there are genes associated with programs known to be affected by MYC\(^{23,25,26}\), such as purine biosynthesis (ATIC), folding of newly translated protein (TCP1), mitochondrial ribosomal protein (MRPL12), ribosomal function and regulation of protein synthesis (HRMT1L3), and genes whose function has not yet been fully investigated (ZRF1, BYSL, RSL1D1). Taken together, these results indicate that ARACNe can identify MYC targets with $\sim$90\% accuracy.

**MYC has a hierarchical control structure**

The identification of the MYC subnetwork allowed further investigation of its extended connections. Seventeen of 56 MYC first neighbors (30\%) are themselves very large genetic hubs, with more than 100 first neighbors in the complete network (range 106–362; Fig. 4b). By comparison, only a small minority ($<$1\%) of the genes represented in the HG-U95Av2 array have more than 100 first neighbors. This is suggestive of a hierarchical structure of the network, such that MYC can modulate a substantial percentage of all genes in the cell through a relatively small number of highly connected subhubs.

Analysis of the genes representing MYC larger subhubs using the Gene Ontology categories confirmed the involvement of MYC in multiple cellular processes (Supplementary Table 6 online). Notably, MYC directly controls BYSL, a gene whose function is poorly characterized but which is the largest hub in the whole B cell network.

**DISCUSSION**

A key result of this study is the ability to infer genetic interactions on a genome-wide scale from gene-expression profiles of mammalian cells. This objective has so far been elusive except for few instances involving limited number of genes, as shown by the study on the early development of spinal cord in mice\(^{27}\), which elucidated interactions between 112 genes.

In general, mammalian networks have presented a formidable algorithmic challenge for most optimization- and regression-based algorithms, and integrative approaches are not yet fully applicable.
given the very scattered nature of the available mammalian cell information. For instance, although the analysis of conservation of coexpression patterns across human, fly, worm and yeast expression profiles28 identified key genetic modules common to all of these organisms with reasonable accuracy, it produced ≤10% accuracy in the dissection of human cellular networks in isolation. Similarly, a genome-wide study of cis-acting elements across coexpressed genes in mammalian cell lines produced only nine interactions among six genes23. The success of ARACNe, on the other hand, is probably a direct result of its inherent simplicity, low algorithmic complexity and lack of empirical rules (‘ad hoc heuristics’) such as adding penalties to more complex networks. Consistent with these observations, the comparative analysis (Fig. 1) of ARACNe versus Bayesian networks shows that ARACNe offers substantially higher precision at an equal or better sensitivity, even in the analysis of relatively simple connectivity (more than 400 interactions) could be used to further reproduce a large fraction of the data on MYC targets that were collected over a decade of research using traditional molecular biology methods and to identify a substantial number of new targets.

The results from the synthetic data analysis (Fig. 1) were further confirmed by the high success rate for the validation of MYC targets in human B lymphocytes using real gene-expression profile data. In particular, the reported precision on the synthetic networks (80–90% at ~340 data points) is comparable to the success rate of the biochemical validation step on the MYC putative targets (>90%). Moreover, the percentage of *bona fide* interactions among first neighbors will probably increase as additional upstream regulators and downstream targets of MYC are identified among the untested genes. Such high precision is a particularly desirable feature of a reverse engineering algorithm, suggesting that the method could be applied to identify candidate interactions without the need for an extensive set of biological validations, which are time- and resource-consuming.

A second result of this study is that gene-expression profile data for systematic experimental perturbations, such as gene inactivations, can be successfully substituted by expression profiles for a wide variety of naturally occurring cellular phenotypes, such as the ones represented by normal and transformed B cells. Using such data implies that the expression of many genes changes over relatively broad dynamic ranges, allowing many regulatory constraints to be identified by algorithmic means. This approach is particularly relevant for studies involving mammalian cells, in which experimental gene perturbations are more technically challenging and time-consuming than in lower organisms. Its systematic application could make reverse engineering techniques broadly applicable to a variety of existing expression profile data for human and mouse tissue.

Substantial evidence indicates that the structure of both protein–protein interactions and metabolic networks in lower organisms is of a scale-free nature1–5, meaning that the network topology is dominated by a few highly connected hubs and an increasing number of less-connected nodes. Although hierarchical, scale-free networks may represent a common blueprint for all cellular constituents, evidence for this has not been obtained for higher-order eukaryotic cells. ARACNe’s ability to reconstruct networks with a broad range of connectivity (more than 400 interactions) could be used to further investigate this hypothesis. Figure 3 shows a power-law tail in the log-log connectivity plot, spanning slightly more than one order of magnitude, with a saturation effect for connectivity ranges below 12. Similar saturation effects are reported for other scale-free networks when the maximum connectivity range is below 1,000 (ref. 29). The power-law should ideally extend over two or more orders of magnitudes, but these results nevertheless support the hypothesis that mammalian networks are scale-free. Furthermore, the marked enrichment of large genetic hubs among first neighbors of other large hubs and the partial overlap of their first neighbors, as observed for MYC, support the notion of a hierarchical control mechanism, providing both redundancy and fine-grain combinatorial control of the cellular genetic programs by a handful of genes. The main functional features of hierarchical scale-free networks are their robustness (i.e., the high redundancy of their pathways), their modularity and their high degree of error tolerance. These properties have important implications for the ability of cells to react to physiologic stimuli and to resist mutations. In addition, the understanding that cellular phenotypes are controlled modularly, by a relatively small number of key genetic hubs, should lead to efforts to identify them as key biological determinants and, possibly, therapeutic targets in disease.

We chose to investigate the MYC subnetwork because the large amount of available data might allow a validation of ARACNe’s power and limitations. The generated network includes ~40% of previously identified MYC target genes in human B cells. There are several reasons for the absence in the network of previously reported MYC target genes. First, ARACNe cannot produce reliable predictions for genes that are expressed at very low levels, given the substantial impact of measurement noise. Second, the stringency used in our analysis may hinder the detection of genes that have lower mutual information, including those genes for which MYC is not the principal regulator. Third, a substantial number of genes are not represented in the microarray used. Nonetheless, ARACNe was able both to reproduce a large fraction of the data on MYC targets that were collected over a decade of research using traditional molecular biology methods and to identify a substantial number of new targets.

The results generated by ARACNe provide a wealth of information on MYC target genes that require further analysis exceeding the scope of this study. Although the idea that MYC is a major hub was perhaps expected, on the basis of evidence that MYC can bind to the promoter region of a large number of genes24, the hierarchical structure of the MYC-dependent network has new and important conceptual implications for the function of this gene. This structure suggests that the large number of targets may increase further and be ‘molded’ by the use of subhubs. Most of these major subhubs do not seem to encode DNA binding transcription factors, suggesting that MYC may function primarily by inducing or suppressing the expression of molecules, such as transcriptional cofactors or kinases, that can indirectly control transcription. Overall, the structure of the network indicates that the subhubs are crucial mediators of MYC function. Therefore, future studies of MYC should consider these subhubs as priority targets.

One of the most notable aspects of the MYC network is that one of the MYC subhubs and a new biochemically validated MYC target is *BYSL*, which is the most connected gene in the entire cellular network generated by ARACNe. *BYSL* is a highly evolutionarily conserved gene whose function is poorly understood30. Studies in human cells have suggested that it has a cytoplasmic localization and a role in cell adhesion31; the yeast *BYSL* homolog, *ENP1*, is nuclear and has a putative role in ribosomal RNA splicing and ribosome biogenesis32. *Drosophila* *BYSL* (*bys*) has nuclear localization, with a pattern of
expression mirroring that of a number of MYC target genes, and a suggested role in cell growth. A preliminary analysis of the BYSL subnetwork suggests that many genes involved in nucleic acid metabolism, cell proliferation and ribosome biogenesis are represented (Supplementary Table 6). Although additional studies are needed to elucidate its function, the high number of connections of BYSL indicates that the encoded protein may be an important cellular molecule and a critical effector of MYC function.

The approach and results shown here have broad general applicability. First, the data set generated in B cells can be used to identify the network of any gene whose range of expression is sufficiently broad in these cells, leading to accumulation of reliable data that usually requires extensive experimental studies. Second, the network can be used as a basis to identify changes in its structure in other tissues and in specific subtypes of B cell malignancy. Third, the efficiency of ARACNe in identifying cellular connections suggests that it can be readily applied to available gene-expression profile data sets, provided that they have adequate size and complexity.

METHODS

Degree of connectivity. We define the degree of connectivity of a node (gene), k, as the number of nearest neighbors of that gene in the directed graph that represent the genetic interaction network. This is equal to the sum of the in-degrees (inbound edges) and out-degrees (outbound edges) of a node. The connectivity structure of a network can be summarized by plotting the number of genes having degree of connectivity k against k in log-log scale. A decreasing linear dependency in this plot indicates that the network has a scale-free topology, in which a fraction of nodes have a very large number of edges, while the majority of nodes have only a few edges. In our networks, the degree distribution follows a power-law with an exponent of approximately 3. We used a value of \( 0.15 (15\%) \) for the synthetic network and \( e = 0.15 (15\%) \) for the B cell network. The higher value used for B cells is associated with the presence of regulatory interactions in this tissue.

Mutual information. Mutual information for a pair of discrete random variables, x and y, is defined as \( I(x,y) = \sum_i \sum_j p(x_i,y_j) \log \left( \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \right) \), where \( p(t_i) = \frac{t_i}{N} \) is the probability associated with each discrete state or value of the variable. If the variable is continuous, the entropy is replaced by the differential entropy, which has the same definition as the entropy for a discrete variable, with the exception that the sum is replaced by an integral. The mutual information between any subpair of the three variables is zero. This does not prevent the latter from occurring in the predicted topology. This is the case for the MYC network, in which several three-node loops are observed. We used \( e = 0.1 (10\%) \) for the synthetic network and \( e = 0.15 (15\%) \) for the B cell network. The higher value used for B cells is associated with the presence of regulatory interactions in this tissue.

Synthetic model. We tested ARACNe on a simulated genetic network to study dynamic Bayesian network inference. We used a 20-gene network containing 14 gene regulatory interactions with one negative feedback loop. Regulatory interactions are defined to affect the transcriptional rate of the target gene linearly, as modeled in a discrete time step simulation by the formula \( Y_{t+1} = f(Y_t) = A(Y_t - T) + e \), where \( e \) is a noise term drawn uniformly from the interval \([-10,10]\). We studied performance by choosing precision,

Statistical threshold for mutual information. For each value of \( N_m \), the synthetic data analysis, we obtained the \( P \) value associated with a given value of mutual information in the null hypothesis by Monte Carlo simulation using 10,000 iterations. The null hypothesis corresponds to pairs of nodes that are disconnected from the network and from each other, such as nodes 12 to 19 in the model. These follow a random-walk dynamic, in the range \([1,100]\), with a noise term drawn from a uniform probability density over the interval \([-10,10]\), as previously reported. For the MYC analysis, we computed the \( P \) value by Monte Carlo simulations using one million iterations. Because a null-hypothesis dynamical model is not available, it is defined as a pair of existing genes whose values are randomly shuffled at each iteration with respect to the microarray profile in which they were observed.
over specificity, as a metric for two reasons: (i) precision translates directly into the expected success rate of a subsequent biological validation step; and (ii) the large number of potential interactions in a network, proportional to the square of the number of genes, makes the specificity a less relevant metric.

**Bayesian networks.** A Bayesian network is a representation of a joint probability distribution as a directed acyclic graph whose vertices correspond to random variables \( \{X_1, \ldots, X_n\} \) and whose edges correspond to dependencies between variables. The most likely graph \( G \) for a given data set \( D \) can be inferred by searching for the optimal graph based on a statistically motivated scoring metric. In this study, we used the Bayesian Scoring Metric, defined as
\[
S(G : D) = \log P(G|D) = \log P(D|G) + \log P(G) - \log P(D),
\]
where \( \log P(D) \) is independent of \( G \) and can be treated as a constant. \( P(G) \) is the prior over graphs, for which we use a uniform prior, following a previously described method. The results of the analysis (Fig. 1) were produced with the Bayesian Networks software LibB2.1 (http://www.cs.huji.ac.il/labs/compbio/LibB/). The MYC network software LibB 2.1 (http://www.cs.huji.ac.il/labs/compbio/LibB/). The ARACNe platform is available at http://amdec-bioinfo.cu-genome.org/html/caworkbench.htm.

**GEO accession number.** Gene-expression profiles, GSE2350.

Note: Supplementary information is available on the Nature Genetics website.

**ACKNOWLEDGMENTS**

We thank I. Nemenman for his expertise in Information Theory, M. Mattioli for contributing to the generation of the B-cell gene expression database and V. Millonig for help with the microarray hybridizations. K.B. is supported by a fellowship from the American-Italian Cancer Foundation, A.M. by the National Library of Medicine Medical Informatics Research Training Program at Columbia and U.K. by a fellowship from the Human Frontiers Science Program. This study was supported by a US National Institutes of Health grant to R.D.-F. and by the computational resources of the AMDeC Bioinformatics Core at Columbia University.

**COMPETING INTERESTS STATEMENT**

The authors declare that they have no competing financial interests.

Received 14 December 2004; accepted 8 February 2005

Published online at http://www.nature.com/naturegenetics/


