A systems-biology approach to modular genetic complexity

Gregory W. Carter,[1] Cynthia G. Rush,[1,2] Filiz Uygun,[1,3] Nikita A. Sakhanenko,[1] David J. Galas,[1] and Timothy Galitski[1]
[1]*Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA*
[2]*Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, USA*
[3]*Computer Science and Communications Research Unit, University of Luxembourg, Luxembourg L-1359, Luxembourg*

(Received 4 March 2010; accepted 26 May 2010; published online 30 June 2010)

Multiple high-throughput genetic interaction studies have provided substantial evidence of modularity in genetic interaction networks. However, the correspondence between these network modules and specific pathways of information flow is often ambiguous. Genetic interaction and molecular interaction analyses have not generated large-scale maps comprising multiple clearly delineated linear pathways. We seek to clarify the situation by discerning the difference between genetic modules and classical pathways. We review a method to optimize the discovery of biologically meaningful genetic modules based on a previously described context-dependent information measure to obtain maximally informative networks. We compare the results of this method with the established measures of network clustering and find that it balances global and local clustering information in networks. We further discuss the consequences for genetic interaction networks and propose a framework for the analysis of genetic modularity. © *2010 American Institute of Physics*. [doi:10.1063/1.3455183]

Systematic genetic perturbation is a powerful tool for inferring gene function in model organisms. Functional relationships between genes can be inferred by observing the effects of combined genetic perturbations. The study of these relationships, generally referred to as genetic interactions, is a classic technique for ordering genes in pathways, thereby revealing genetic organization and information flow paths among genes and their products. Large-scale genetic interaction studies based on this technique have provided substantial evidence of modular organization in genetic interaction networks. However, the correspondence between these network modules and specific pathways of information flow is often ambiguous in that the scaling up of genetic interaction analysis has not generated large-scale maps comprising distinct linear pathways. We seek to clarify the situation by defining genetic modules independent of classical pathways and vice versa. We propose that a genetic module is a more general construct than the molecular pathway concept and define a module as a set of coinformative genes that may or may not be involved in the same linear molecular sequence. We review a recently proposed method to optimize information extraction that consequently led to the discovery of these modules in genetic interaction data. We contrast this method to other measures of network clustering and discuss its relationship to alternate methods of genetic interaction analyses.

I. INTRODUCTION

Genetic interaction analysis is rapidly becoming a prominent tool for inferring the function and structure of genetic networks. To date, genome-scale studies have involved primarily the baker's yeast *Saccharomyces cerevisiae* due to its genetic manipulability, short life cycle, and potential for high-throughput phenotyping. Large-scale studies performed with both engineered strains[1–7] and yeast intercross strains[8] have revealed the power of genetic interactions to map genetic networks and to understand gene function.

The use of genetic interactions to understand the structure and flow of biological information is derived from the classical analysis of comparing the effects of two individual genetic mutations with the effects of the combination of those mutations. Historically, targeted genetic interaction analysis has been an effective tool for mapping biological pathways.[9] As data collection grows in scale, the mapping of individual pathways has become increasingly intractable due to the functional and structural complexity inherent in biological systems. Networks that represent the interactions of multiple genetic variants typically form a dense web of numerous potential pathways and molecular mechanisms. The concept of *genetic modularity* provides a powerful paradigm for the analysis of such large and dense networks.[10] A modular representation allows a substantial reduction of genetic complexity,[11] making detailed genetic modeling of key system elements tractable. Since modular analysis is not constrained by the concept of sparsely connected linear pathways, it is more suitable to data-driven mapping of dense, large-scale genetic networks.

However, it is not clear how to define modularity in genetic interaction networks. While metabolic reaction networks and protein-protein interaction networks often exhibit modularity as regions of high connectivity,[11] genetic interaction networks encode more abstract information and can generate modules of genes that function together in diverse ways to inform phenotype. These modules can be defined as groups of genes with interaction coherence across a large

network;[1,2,7,12] however, the resulting modularity can depend on how genetic interactions are defined.[13] Here, we expand on previous works[2,14,15] to show how an unsupervised method of finding the most informative mapping of genetic interactions tends to yield networks with modular architecture. These modules, furthermore, were shown to make significant biological sense. Given that modularity was a result rather than an assumption of this analysis, we propose that this method reveals inherent modularity in genetic data.

II. MODULES VERSUS PATHWAYS

We draw a key distinction between a genetic pathway and a genetic module. A pathway is a specific information-flow conduit, usually a sequence of molecular interactions. In contrast, a module is an information-processing unit with a self-contained emergent function. Modules therefore can contain multiple pathways, and pathways can operate between modules to form intermodule connections.

Intermodular pathways serve as lines of communication and coordination between distinct biological processes that combine to regulate cellular function. For example, cell differentiation from yeast-form to filamentous growth in budding yeast requires a pathway linking a mitogen-activated protein (MAP)-kinase signal transduction module to the cell-cycle control module in order to regulate cell elongation.[16] The intermodular biomolecular pathway responsible for this linkage is mediated by the Ste12-Tec1 transcription complex, which is activated by the MAP-kinase Kss1 to transcriptionally activate the cyclin-encoding gene *CLN1*. Indeed, the definition of a module as a functional cellular subunit requires such coordinating connections, and these connections often correspond to the classical definition of a pathway.

By contrast, intramodular pathways are often the central features of modules. In some cases, a module can be operationally defined as a collection of connected molecular-interaction pathways. In addition to information-flow lines, intramodular pathways involve feedback and feedforward loops, scaffolds and tethers, regulators, and other interfaces that combine to produce a distinct functional unit. Thus, modules can be viewed as a level of organization above biomolecular pathways but below phenotypes.

The distinction between modules and pathways is particularly relevant when one seeks to analyze biological processes with large-scale data sets. Using the early tools of the biochemist (e.g., radioactive tracers) or the developmental geneticist (e.g., gene/protein ordering through epistasis testing), one can decipher biochemical sequences. These methods, by their nature, tend to reveal distinct biomolecular pathways, and from such early studies, the concept of biomolecular pathways arose. Observational biases and low experimental throughput necessitated a focus on a modest number of major information-flow trunk lines. From this perspective, it is not surprising that early molecular network maps feature sparsely connected pathways. However, analyzing a high-throughput collection of phenotype observations across multiple genetic backgrounds reveals functional organization involving many genes that are often not directly involved in shared biomolecular pathways. Modern high-throughput technologies for molecular network cartography generate densely connected networks with numerous possible pathways, but a relatively modest number of interaction clusters. Had such high-throughput experiments been the first look at these networks, the module would probably be the most prominent organizational concept rather than the pathway.

This module-versus-pathway framework provides a promising strategy for understanding large-scale genetic data. The immediate challenge, however, is to develop technologies that infer and characterize genetic modules systematically, and that complement the proven techniques for pathway mapping. Recent studies in genetic cartography (mapping interactions between genes on a large scale) have developed analytical methods to infer genetic modules. These modules comprise of cofunctional sets of genes and are derived primarily from phenotypic observation[1,2,6,7,15,17] or computational analysis.[12] A modular representation (by definition) substantially reduces the complexity of the genetic data. Key pathways, operating within or between modules, can be identified and mapped in terms of specific information flows. In cases where large-scale molecular data are available, these information conduits can then be translated into specific molecular hypothesis.[18]

The inference of genetic modularity is ideally pursued without preconceptions of the extent or even existence of such modularity. In developing a technique to maximize the extraction of biological knowledge from genetic data, we recently found that the most informative network analysis also yielded highly connected clusters of coinformative genes. We identified these clusters as gene modules.[15] Thus, the study of genetic modularity might fruitfully be viewed from the perspective of information theory. In this light, modular architecture inferred in a genetic network maps how information is distributed throughout a biological system or, more specifically, a particular genetic data set derived from that system. This proposition requires a method to measure the information content of a system, and we proposed using set complexity as a measure.[14,15] By maximizing this complexity in genetic network analysis by finding the most informative rules of interaction, we were able to identify genetic modules and thereby optimize the biological information obtained from data derived from a set of genetic perturbations. Each module contained genes with shared functional annotations unique to that module, providing strong evidence that these gene sets are precisely the gene modules we have defined above. The modules overlapped with known pathways but also allow for an interpretation of cofunctionality that is complementary to specific molecular sequences of information flow. Furthermore, the genetic interaction rules that maximized set complexity often did not correspond to rules commonly used in pathway analysis. These complexity-based rules were interpreted as those that govern how genes are organized into functional groups, taking into account the full content (and limitations) of the analyzed data set. This was contrasted with the pathway analysis of genetic interactions, in which the rules are interpreted in terms of information flow through individual gene pairs. Thus, we conclude that the most fruitful application of the complexity-based algorithm is the identification of gene

modules rather than linear gene pathways. As a corollary, we conclude that methods designed to order genes into molecular-interaction sequences (pathways) are not ideal for the discovery of modules.

In this work, we further demonstrate that these modular structures are optimally defined using the set complexity method described previously[15] in a way that best balances general and specific information within a network. We show that naïve clustering measures are often not functionally informative, particularly as networks become very dense and involve multiple modes of interaction between nodes. Since genetic interaction networks can become very dense, especially when one considers many genes involved in a given function, a clustering measure that reflects functional modularity is necessary. We provide evidence that set complexity maximizes nontrivial, functional modularity.

III. MODULARITY IN GENETIC INTERACTION DATA

Genetic interaction is a general term to describe phenotypic nonindependence of two or more genetic perturbations. However, it is generally unclear how to define this independence.[2,13,19] Therefore, it is useful to consider a general approach to the analysis of genetic interaction. We have developed a method to systematically encode genetic interactions in terms of phenotype inequalities.[2] This allows the modes of genetic interaction to be systematically analyzed and formally classified. Consider a genotype X and its cognate observed phenotype P_X. The phenotype could be a quantitative measurement or any other observation that can be clearly compared across mutant genotypes (e.g., slow versus standard versus fast growth, or color or shape of colony, or invasiveness of growth on agar, etc.). The genotype is usually labeled by the mutation of one or more genes, which could be gene deletions, high-copy amplifications, single-nucleotide polymorphisms, or other allele forms. With genotypes labeled by mutant alleles, a set of four phenotype observations can be assembled which defines a genetic interaction: P_A and P_B for gene A and gene B mutant alleles, P_{AB} for the AB double mutant, and P_{WT} for the wild type or reference genotype. The relationship among these four measurements defines a genetic interaction. For example, if we follow the classic genetic definitions described above, $P_{AB}=P_A<P_{\mathrm{WT}}<P_B$ describes one type of epistatic interaction, while $P_{\mathrm{WT}}<P_{AB}=P_A=P_B$ is an example of asynthesis. There is a total of 45 distinct inequalities that can be constructed from four phenotypes.

Although this procedure reduces the data to a limited set of experimental outcomes, there is still the potential for substantial complexity.[20] One strategy to reduce this complexity is to group these inequalities into rules of genetic interaction, with each inequality within a rule representing different instances of the same biological relationship. For example, inequalities $P_{AB}<P_A=P_B=P_{\mathrm{WT}}$ and $P_A=P_B=P_{\mathrm{WT}}<P_{AB}$ might both be considered instances of synthetic interaction, defined as the occurrence of two genetic perturbations without individual effects on the phenotype combining to cause an effect. Different groupings have been proposed and examined in literature.[2,4] The goal of any such analysis is to obtain the most biologically informative set of rules for genetic interaction. Placed in this context, seeking the most informative analysis is a problem of finding the groupings of interaction inequalities that best resolve the underlying biology. A set complexity measure, based in information theory and discussed in detail below, provides an agnostic solution to this problem. Namely, this set complexity measure can be maximized to find the most informative inequality grouping. This procedure depends only on the genotype and phenotype data, requiring no additional prior information. We then assessed these networks for biological meaning using two published methods (Fig. 1).[2]

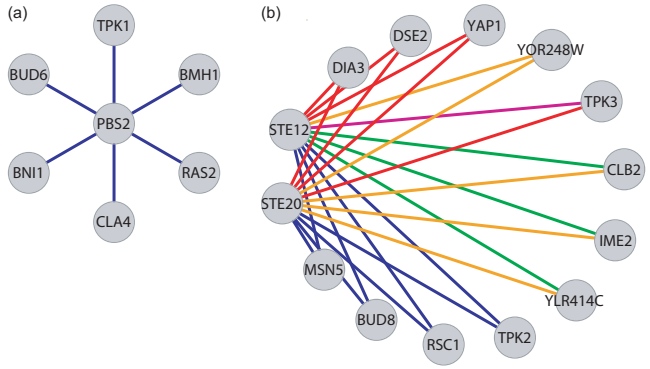

FIG. 1. (Color) Examples of biological information in genetic interaction networks. (a) A biological statement showing the interactions of a gene deletion (*PBS2*) with perturbations of genes with a common function (signal transduction) via a common interaction rule (blue edges). (b) Mutually informative gene perturbations of *STE12* and *STE20* show large-scale patterns of genetic interaction. Both panels adapted from Drees *et al.* (Ref. 2).

The first method we have used to assess biological information is finding statistically significant associations between genes and functions [Fig. 1(a)]. The genomes of model organisms have been well annotated for gene function and these annotations have been organized into the Gene Ontology database.[21] We generated and assessed a genetic interaction network for biological statements, defined as a particular gene nonrandomly interacting via a single rule with multiple genes annotated with a shared biological function.[2] The significance of statements can be computed with Fisher's exact test and we defined valid statements as those that meet a significance criterion (e.g., $p<0.01$ in Ref. 15). The result was a computer-generated list of biological statements relating genes, interaction modes, and target annotations, with entries such as: "A deletion of gene *PBS2* interacts additively with deletion mutations of signal transduction genes ($p=0.001$)." The number of such existing statements is highly sensitive to the interaction rules in the network and thus served as a measure of how informative each classification scheme was in a biological sense.

The second method we have used to extract biological information from genetic interaction networks is the computation of mutually informative allele pairs within the network [Fig. 1(b)]. These calculations revealed global patterns of gene association and distilled a complex genetic interaction network down to modules of coinformative genes. These mutually informative pairs of alleles exhibited an improbably high degree of mutual information with common interaction partners such that knowing the interactions of one allele may

allow one to know the interactions of another. In genetic interaction networks this pairwise property can be quantified by the Shannon mutual information scores used to compute the context-dependent complexity metric. We identified pairs of alleles with statistically significant mutual information and these pairs were mapped in mutual information networks. We found that clusters or cliques of genes in a mutual information network identify genes with similar effects on biological processes. These groups of genes clustered by mutual information correspond to specific modules. Therefore, a larger number of mutually informative pairs correspond to a more comprehensive module mapping.

After an initial analysis based primarily on pathway mapping,[2] we later found that analyzing genetic interaction networks by maximizing set complexity[14] yielded a greater amount of biological information.[15] In particular, networks with maximal set complexity contained many more gene pairs with significant mutual information in their interaction patterns across common neighbor nodes. Representing these pairs as a network of coinformative alleles yielded large interconnected subnetworks, which segregated the Ras-cyclic adenosine monophosphate (Ras-cAMP) and filamentation MAP-kinase signaling networks involved in yeast invasion. From this, we concluded that these gene subnetworks represent gene modules or sets of genes that somehow cofunction to produce a phenotype.[1,12,17] We further speculated that maximizing our set complexity measure served to find the most modular representation of the data set, which the modularity hypothesis would associate with the best representation of the cell's functional organization that could be obtained from the limited set of genetic perturbations.

IV. MODULARITY AND SET COMPLEXITY

The set complexity measure used to optimize the analysis of genetic interaction data led to substantial modularity in the genetic interaction network. However, it is unclear how this modularity relates to other definitions of modularity and network structure. Here, we review the definition of set complexity, investigate its relationship with global and local clustering measures, and highlight some aspects of set complexity that are especially suited to genetic interaction analysis.

The set complexity metric applied in Ref. 15 was defined and developed in Ref. 14. It is based on the normalized information distance function between two strings as derived by Li *et al.*,[22] which is a metric satisfying the three criteria of identity, symmetry, and the triangle inequality. This metric is universal in that it discovers all computable similarities between strings.[22] As shown by Galas *et al.*,[14] a simple relationship between the universal information distance and the pairwise mutual information allows the set complexity Ψ to be computed with mutual information.

For network analysis, for which the sample space is well-defined in terms of nodes and possible edges, we compute the set complexity using single and mutual Shannon entropies. The set complexity for a network is thus defined as follows. Consider a network of N nodes with M types of edges that connect the nodes. For simple binary networks, $M=2$, commonly corresponding to the presence or absence of an edge. For the ith node in a network, we first compute the Shannon information K_i based on its interactions with all other nodes. This is done by computing the fraction of nearest neighbors within each class of interaction, denoted as $p_i(a)$ for the ath interaction class, with the frequency of these connections defining effective probabilities. Summing over all interaction types yields the single-node complexity,

$$K_i = -\frac{1}{\ln(M)}\sum_{a=1}^{M} p_i(a)\ln p_i(a), \tag{1}$$

where M is the number of interaction classes and the sum is over all interaction classes. The normalization ensures that this quantity is always between 0 and 1. Edge directionality can be considered where relevant, with outgoing edges considered a different interaction type than incoming edges, although here we consider only nondirectional edges. We next compute the mutual information for every pair of nodes in the network using the Shannon approach. This can be written as

$$m_{ij} = \frac{1}{\ln(M)}\sum_{a=1}^{M}\sum_{b=1}^{M} p_{ij}(a,b)\ln\left(\frac{p_{ij}(a,b)}{p_i(a)p_j(b)}\right), \tag{2}$$

where $p_{ij}(a,b)$ is the joint probability of node i interacting with a third node with rule a and node j interacting with the same third node with rule b. This expression is also normalized to the interval $[0,1]$.

With these normalized quantities we compute the context-dependent complexity of a network with N nodes by summing over all node pairs as

$$\Psi = \frac{4}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Max}(K_i,K_j)m_{ij}(1-m_{ij}). \tag{3}$$

This complexity measure is normalized to yield values between 0 and 1. Any network can be scored in terms of set complexity Ψ. As edge mapping varies for different analysis schemes, the single-node entropies (K_i) and pairwise mutual information values (m_{ij}) differ and lead to variations in Ψ.

Substantial insight can be gained by considering the simple case of $M=2$, corresponding to Erdős–Rényi graphs of nodes connected by one undirected and unweighted edge type without any self-interactions. We previously found that for such graphs maximal complexity arises from nearly bimodular or near-bipartite graphs.[14] These graphs appear to balance the requirement of maximal complexity for each single node with the requirement of uniform mutual information between all node pairs. Figure 2 shows an example of such a graph, representing the maximally complex graph found for $N=20$. The set complexity of this graph is $\Psi=0.92$. While the modular structure of this network is apparent, the intermodular connections are critical for a high complexity score. For example, the union of two complete graphs with ten nodes has a Ψ of only 0.017.

The two most striking aspects of the maximally complex graphs are the apparent modularity coupled with the presence of a limited number of linkages between the two graph modules. To further explore this architecture, we systematically compared set complexity to standard measures of graph properties across an ensemble of networks. We first consid-

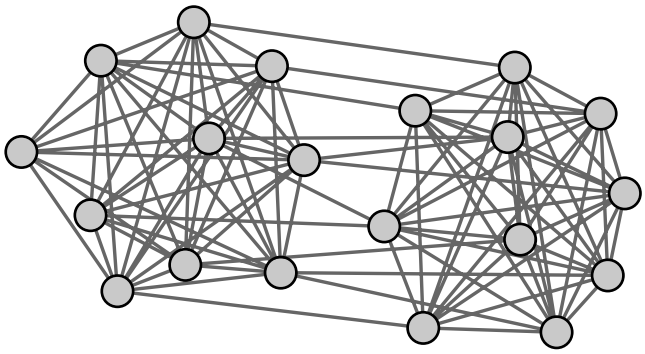

FIG. 2. The maximally informative undirected, unweighted graph with $N=20$.

ered the global clustering coefficient,[23] a simple measure of graph modularity defined as the number of three-node cliques (fully connected subgraphs) divided by the number of three-node subgraphs with at least two edges. The ratio is denoted by C and varies from 0 (nonclustered network) to 1 (fully clustered network). We also consider the more sophisticated measure of modularity proposed by Clauset,[24] which defines a measure of local modularity denoted by R. This metric arises from an algorithm that infers a hierarchy of communities by considering the neighborhood of each vertex in a graph. Greater values of R correspond to more community structure, with $0<R<1$. Finally, we consider the importance of intermodular links by computing the betweenness centrality of each node in a graph. For a given node A, this is defined as the fraction of the shortest paths linking two other node pairs that pass through A, summed over all node pairs. A node with high betweenness centrality is therefore a node that lies on many shortest paths connecting node pairs across the graph. Of particular interest to us here is the maximum betweenness centrality in the network, denoted as $B_{\max}$, which represents the presence or absence of a small number of central linking nodes.

We first compared the maximally complex graph (Fig. 2) to increasingly random graphs with a fixed density (101 edges, equal to 0.53 of all possible edges). Beginning with the maximally complex graph, we randomly reassigned edges one at a time until graphs became fully random. This procedure was repeated 200 times, and the mean graph statistics are shown in Fig. 3. The maximally complex graph is the most modular graph in terms of both global clustering [Fig. 3(a)] and local modularity [Fig. 3(b)]. Although the maximally complex graph features a limited set of linkages between the two major modules (Fig. 2), this does not lead to particular nodes having more betweenness centrality than a random network [Fig. 3(c)]. So while the most complex networks are substantially more modular than random networks, they do not contain specific nodes that bridge the modules. This result is further supported by the fact that power-law or scale-free networks[25] are not substantially more complex, on the average, than random networks (data not shown). These results follow from the observation that Ψ is greatest when information is shared throughout the network.

Although these results reinforce the association between complexity and modularity, comparing the maximally complex network to random networks of fixed density omits an important feature of genetic interaction networks. Namely, the definition of genetic interaction is often ambiguous because of the nature of a given data set.[13] A single genetic data set can yield sparse, dense, or intermediately dense networks depending on the criteria used to define interactions, the size of the data set, and the inherent noise.

It is therefore of interest to consider how Ψ is related to C, R, and $B_{\max}$ across a range of network densities. To this end, we calculated these quantities for a sequence of 20-node networks ranging from an empty network (no edges) to a complete network (all nodes linked by an edge), averaging over an ensemble of 200 independent sequences that each traverse the maximally complex network. This is equivalent to the edgewise construction of the maximally complex network (Fig. 2) from an empty network, followed by the filling of the remaining edges to a complete network. The mean graph statistics plotted in Fig. 4 reveals some substantial differences between Ψ and three modularity measures. Since the global clustering coefficient is the ratio of three-node cliques to potential cliques, it varies from 0 in an empty network to 1 in a complete network. Thus, the most-clustered configuration according to this measure is a complete network, which is a fairly trivial statement of clustering. Furthermore, in the context of biology such networks are unlikely to be informative of how individual pairs of nodes are related since all pairs are similarly related. The complexity measure Ψ avoids this simplification by quickly decreasing as the edge density approaches 1 [Fig. 4(a)]. While the local modularity measure R also vanishes for a complete network, it maximizes for very sparse configurations [Fig. 4(b)] that correspond to the early steps in building the maximally complex network. On the average, these networks feature small, localized edge groups that are reflected in the large local

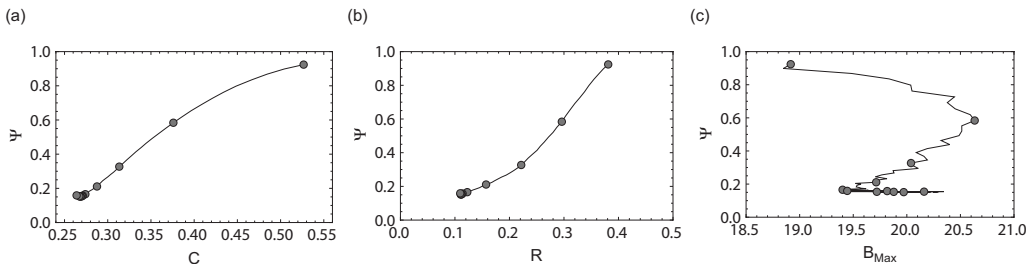

FIG. 3. Set complexity vs (a) global clustering coefficient, (b) local modularity, and (c) maximum betweenness centrality for a sequence of 20-node networks ranging a random network to the maximum-Psi network with the number of edges fixed. Results have been averaged over 200 paths, and dots represent every tenth network configuration.

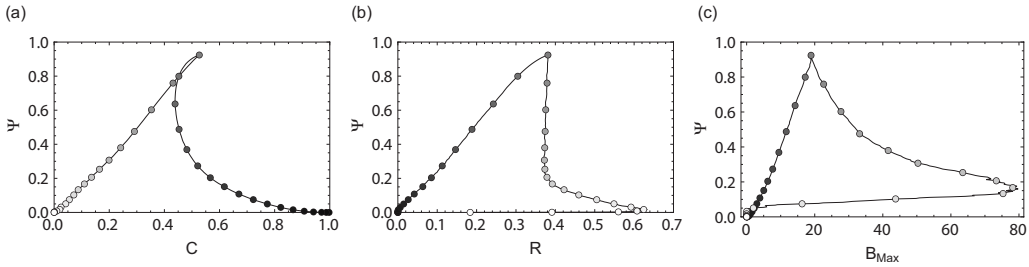

FIG. 4. Set complexity vs (a) global clustering coefficient, (b) local modularity, and (c) maximum betweenness centrality for a sequence of 20-node networks ranging from an empty network to a complete network, averaged over 200 paths that traverse the maximum-Psi network. Dots represent every tenth network configuration and are shaded according to network density ranging from an empty network (white) to a complete network (black).

modularity measure. However, in a biological context such a sparse network will often not be the most informative as it may miss many biologically important features in the data. Similar results were observed for other measures of modularity that are maximal for localized network clusters.[26] An analogous behavior is seen for the maximum betweenness centrality B_{max} [Fig. 4(c)] as sparse networks are more likely to feature a single node with very large B. In contrast, networks with higher Ψ feature a few nodes of moderate B and distribute the betweenness centrality over multiple nodes that bridge modules. Thus, the network complexity metric Ψ is a good candidate for balancing the global and local aspects of modularity, allowing nodes to be characterized on a global scale in a way that retains potentially meaningful local information. These findings agree well with our previous interpretation.[15]

These properties of set complexity Ψ extend to networks with multiple edge types, although the lack of well-established clustering measures for multimodal networks makes exact, comparative analysis impossible. Such networks with multiple edge types, which are essential to represent gene interactions, are readily computable with Ψ. The primary difference we find is that a network of M edge types with maximum complexity exhibits M modules, each comprised of nodes that exhibit a large degree of mutual information. An example of a maximum-Ψ network is shown in Fig. 5. This network has 3 edge types, 12 nodes, and a complexity $\Psi=0.81$. It exhibits the similar features to the binary network of Fig. 2, with near-perfect modularity disrupted by a small number of alternate edge types. The key feature of this network is the separation of otherwise identical nodes by the edges, and permutations of the specific edge colors yield equally complex networks.

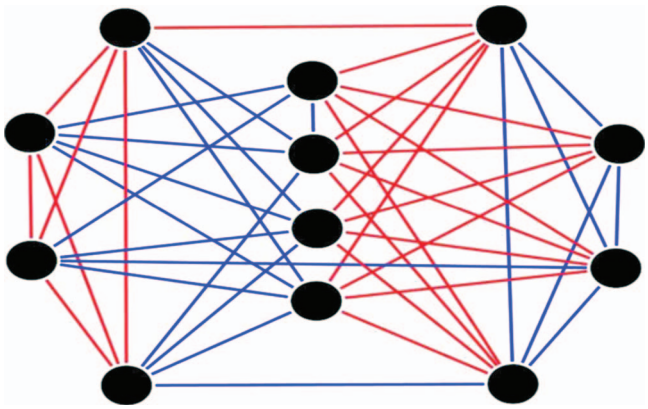

FIG. 5. (Color) The maximally informative graph with 12 nodes and 3 edge types (red, blue, and no edge). The graph layout is chosen to illustrate edge monochromaticity between node sets.

V. DISCUSSION AND CONCLUSIONS

Genetic interactions have a successful history of mapping pathways of information flow in biological systems, and contemporary high-throughput technologies allow such interactions to be assayed on large scales. The resulting data sets provide a resource for mapping not only isolated pathways but also large-scale genetic architecture. There is a growing body of evidence that this architecture is modular, and these genetic modules are traversed and connected by molecular pathways. Furthermore, there is substantial evidence that genetic modules comprise of sets of cofunctional genes. This allows for the generation of functional hypotheses for incompletely annotated genes that fall within a module containing many other genes of a common function. It additionally enables the identification of novel biological process associations with broader phenotypes and candidate genes for the control of that process.

Here, we have shown that this modularity can arise as a consequence of maximizing set complexity, which provides a flexible basis for effectively determining the most biologically informative analysis of a given genetic data set. The modularity results from an unsupervised assessment of biological complexity, which itself is agnostic to the presence of modular network architecture. Thus, the degree of modularity observed can be viewed as the inherent modularity of a data set that has been analyzed in a way that optimally resolves general and specific information. We further propose that these networks maximized for complexity exhibit a nontrivial modularity that balances global and local clustering to yield the most information from a given data set. We emphasize that although the calculations presented here address purely theoretical network architectures and real biological data exert strong constraints on possible networks derived from those data, the general results will apply. Given the possible networks derivable from a specific data set, maximizing for set complexity will select the network with the greatest nontrivial modularity. Although the full space of possible networks is computationally intractable for most data sets, Ψ can serve as an optimization metric for determining the most informative analysis without the requirement of any prior biological knowledge.

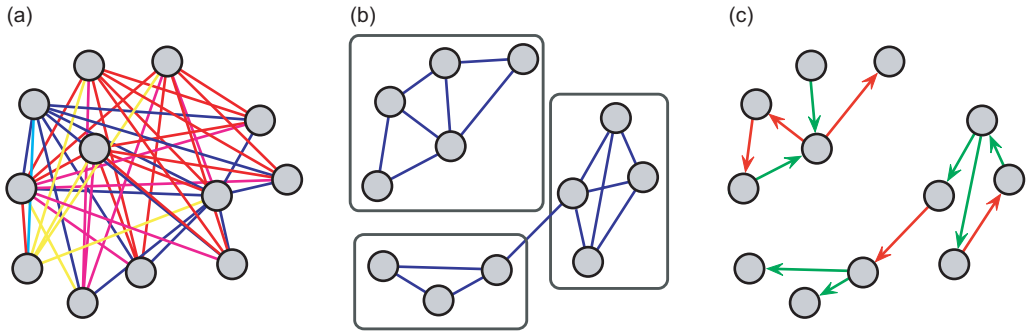

FIG. 6. (Color) Modular analysis of a hypothetical genetic interaction network. (a) Multimodal network representing pairwise genetic interactions. (b) Reduced network of gene pairs with significant mutual information and the resulting modular structure. (c) Network of gene-gene information flow paths derived from further analysis based on the modular network (b).

While many early high-throughput genetic interaction studies were confined to two edge types,[1,3] the analysis of genetic interaction networks often involves multiple interaction types.[4,7,12,27] The appropriate choice of edge type, or rule of genetic interaction, is often ambiguous and is likely to depend on the system under study, the phenotype that is measured, and the specific genetic perturbations underlying the phenotypic diversity. The spectrum of genetic interaction types depends crucially on how genetic interactions are defined. Recent work by Mani *et al.*[13] defines genetic interactions as being deviations from genetic independence, measured on an additive, multiplicative, or binary scale. This analysis has been extended by Gao *et al.*[19] with a maximum-likelihood approach to determine which of these interaction models best captures epistasis. These studies both assess interactions in growth rates of yeast strains. While a summary statistic characterizing genetic interactions (denoted epsilon in many studies) might well be sufficient for assessing growth rate variation, in many cases additional information may be needed. For example, in the genetic study of molecular signaling it is often useful to know which of two mutant phenotypes masks the other when combined in a double mutant.[9] Maximizing the complexity of genetic interaction networks based on phenotype inequalities allows such information to be retained and, furthermore, can judge its biological meaning relative to the analyzed phenotype. Additionally, the inequality-based strategy does not rely so strictly on quantitative data, as phenotype inequalities can often be determined from semiquantitative or qualitative data that can be arranged on a comparative scale. However, when detailed quantitative data are available, the complexity-maximization technique might be applied to the statistical assessment of interaction parameters, as performed in the maximum-likelihood approach of Gao *et al.*[19] Finally, the complexity-based strategy does not restrict genetic analysis to a set of model classes, although it could if such constraints are known to be appropriate.

Our results align well with the concept of monochromaticity in genetic interaction, first hypothesized by Segré *et al.*[12] The maximization of complexity naturally yields networks with monochromatic interactions separating modules (Fig. 5). Experimental data are rarely expected to have such a simple structure, as real outcomes often contain redundancy, random noise, and biological complexity that are insufficiently probed in a single data set. However, maximally complex networks derived from real data show evidence of systematic blocks of uniform interaction type between gene modules. Assessing the complexity of the computational metabolic network originally studied by Segré *et al.*[12] might further elucidate the relationship between monochromaticity and complexity.

In addition to providing functional hypotheses, modular network abstraction can substantially reduce the complexity in genetic interaction networks. This concept is illustrated in Fig. 6. Beginning with a network of genetic interactions [Fig. 6(a)], gene pairs with high mutual information can be extracted to map a simplified network of coinformative genes [Fig. 6(b)]. Genes, and perturbations thereof, that function together in an emergent process are naturally grouped into cofunctional modules, which can then be assessed and modeled in relation to other multigene modules. This greatly reduces the number of system elements and the combinatorial complexity and allows the identification of key network nodes. This, in turn, enables the prioritization of important genes for further study. For example, additional experimentation and analysis can be used for fine mapping of information flow paths within this limited set of genes and between genes that bridge modules [Fig. 6(c)]. The formulation of such models is a critical task in systems biology and one that, so far, has been less vigorously pursued than genetic cartography. Such efforts are often hindered by the overwhelming number of possible paths, the lack of data specific to a given condition or phenotype, or insufficient congruence between functional (e.g., genetic) and physical (e.g., molecular) data. Reducing the genetic complexity to a set of key system elements coupled with methods that map information flow between a limited number of genetic actors[18,28] might resolve these difficulties, thereby enabling the inference of models of system function with substantial predictive power.

The identification of key nodes that connect modules may be of particular interest in understanding how multiple biological processes are coordinated. For example, the complexity-maximization analysis of the yeast invasion network[2,15] yielded two major gene modules that represented the cAMP and filamentation MAP-kinase signaling networks. By identifying the best candidate gene pairs that connect these modules by identifying the pair with the most mutual information relative to expectation (lowest likeli-

hood), we found a possible link between deletions of the nuclear kinase genes *IPK1* and *SNF4*. This gene pair thus serves as a hypothetical mechanism for signal integration in the nucleus. The identification of such key bridge nodes can greatly constrain and/or prioritize the space of possible models.

Generating gene modules by maximizing set complexity might be particularly useful when addressing natural genetic variance across populations. The number of relevant genetic mutations within a given population is limited, making pathway identification and mapping particularly challenging since many links within a molecular path will not vary across the population. The result is a series of fragmented pathways and an incomplete association of cofunctional genes. Modular analyses provide a more general basis for associating groups of genes than linear pathway analysis. Modular analysis flexibly groups genes based on clues at the phenotype level instead of imposing the constraint of linear connections. Complexity-based methods of inferring genetic modules, however, are particularly suited to extracting the most biological information from a given data set. In this way, the analysis is tuned to the resolution of the genetic variation that resides in a given sample population.

Combining the inference of genetic modules with predictive network modeling might be of particular use in the analysis of natural genetic variations with sparse prior annotation. For example, genetic modularity may be used to classify rare disease-related gene variants into sets of mutually informative genetic perturbations. Modules of rare variants that coinform phenotypes such as cancer susceptibility[29] might represent multiple biological processes involved in disease etiology and progression. Candidate modules would provide a basis for identifying biological processes relevant to the disease outcome, and key nodes connecting distinct modules would represent candidate paths of intermodular communication and regulation. These nodes could then be analyzed at greater resolution to infer a model of system function at the genetic level. The result would be a two-level model of system elements and relevant interactions rather than ambiguous lists of gene candidates. Such models have the potential to predict the outcomes of genetic and/or therapeutic interventions at the molecular level, aiding in the development of personalized and predictive medicine.

ACKNOWLEDGMENTS

We thank the editors of Chaos for their invitation to submit an article. This work was supported by Grant No. P50 GM076547 from NIH and Grant No. FIBR EF-0527023 from NSF and by the ISB-University of Luxembourg Program. G.W.C. was supported by Grant No. K25 GM079404 from NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIMGS or the NIH.

[1]A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Ménard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, Science **303**, 808 (2004).
[2]B. L. Drees, V. Thorsson, G. W. Carter, A. W. Rives, M. Z. Raymond, I. Avila-Campillo, P. Shannon, and T. Galitski, Genome Biol. **6**, R38 (2005).
[3]X. Pan, P. Ye, D. S. Yuan, X. Wang, J. S. Bader, and J. D. Boeke, Cell **124**, 1069 (2006).
[4]R. P. St. Onge, R. Mani, J. Oh, M. Proctor, E. Fung, R. W. Davis, C. Nislow, F. P. Roth, and G. Giaever, Nat. Genet. **39**, 199 (2007).
[5]L. Decourty, C. Saveanu, K. Zemam, F. Hantraye, E. Frachon, J.-C. Rousselle, M. Fromont-Racine, and A. Jacquier, Proc. Natl. Acad. Sci. U.S.A. **105**, 5821 (2008).
[6]D. Fiedler, H. Braberg, M. Mehta, G. Chechik, G. Cagney, P. Mukherjee, A. C. Silva, M. Shales, S. R. Collins, S. van Wageningen, P. Kemmeren, F. C. P. Holstege, J. S. Weissman, M.-C. Keogh, D. Koller, K. M. Shokat, and N. J. Krogan, Cell **136**, 952 (2009).
[7]M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St. Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. San Luis, E. Shuteriqi, A. H. Y. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pál, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A.-C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, and C. Boone, Science **327**, 425 (2010).
[8]J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, Nat. Genet. **40**, 854 (2008); J. Gerke, K. Lorenz, and B. Cohen, Science **323**, 498 (2009).
[9]L. Avery and S. Wasserman, Trends Genet. **8**, 312 (1992).
[10]T. Galitski, Annu. Rev. Genomics Hum. Genet. **5**, 177 (2004).
[11]A. W. Rives and T. Galitski, Proc. Natl. Acad. Sci. U.S.A. **100**, 1128 (2003); S. Prinz, I. Avila-Campillo, C. Aldridge, A. Srinivasan, K. Dimitrov, A. F. Siegel, and T. Galitski, Genome Res. **14**, 380 (2004).
[12]D. Segré, A. Deluna, G. M. Church, and R. Kishony, Nat. Genet. **37**, 77 (2005).
[13]R. Mani, R. P. St. Onge, J. L. Hartman IV, G. Giaever, and F. P. Roth, Proc. Natl. Acad. Sci. U.S.A. **105**, 3461 (2008).
[14]D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, and I. Shmulevich, IEEE Trans. Inf. Theory **56**, 667 (2010).
[15]G. W. Carter, D. J. Galas, and T. Galitski, PLOS Comput. Biol. **5**, e1000347 (2009).
[16]J. M. Gancedo, FEMS Microbiol. Rev. **25**, 107 (2001).
[17]P. Ye, B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader, Mol. Syst. Biol. **1**, 26 (2005).
[18]G. W. Carter, S. Prinz, C. Neou, J. P. Shelby, B. Marzolf, V. Thorsson, and T. Galitski, Mol. Syst. Biol. **3**, 96 (2007).
[19]H. Gao, J. M. Granka, and M. W. Feldman, Genetics **184(3)**, 827 (2009).
[20]R. J. Taylor, D. Falconnet, A. Niemisto, S. A. Ramsey, S. Prinz, I. Shmulevich, T. Galitski, and C. L. Hansen, Proc. Natl. Acad. Sci. U.S.A. **106**, 3758 (2009).
[21]M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, Nat. Genet. **25**, 25 (2000).
[22]M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, IEEE Trans. Inf. Theory **20**, 1 (2004).
[23]R. D. Luce and A. D. Perry, Psychometrika **14**, 95 (1949).
[24]A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[25]A. L. Barabási and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[26]M. E. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[27]M. Schuldiner, S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, J. S. Weissman, and N. J. Krogan, Cell **123**, 507 (2005).
[28]C. T. Workman, H. C. Mak, S. McCuine, J.-B. Tagne, M. Agarwal, O. Ozier, T. J. Begley, L. D. Samson, and T. Ideker, Science **312**, 1054 (2006); E. Chaibub Neto, C. T. Ferrara, A. D. Attie, and B. S. Yandell, Genetics **179**, 1089 (2008); C. J. Vaske, C. House, T. Luu, B. Frank, C.-H. Yeang, N. H. Lee, and J. M. Stuart, PLOS Comput. Biol. **5**, e1000274 (2009).
[29]A. Galvan, J. P. Ioannidis, and T. A. Dragani, Trends Genet. **26**, 132 (2010).