Functional Modularity in a Large-Scale Mammalian Molecular Interaction Network

Andreas Krämer, Daniel R. Richards, James O. Bowlby, and Ramon M. Felciano
Ingenuity Systems
1565 Charleston Road
Mountain View, CA 94043
akramer@ingenuity.com

Abstract

The Ingenuity™ Pathways Knowledge Base (IPKB) contains over one million findings manually curated from the scientific literature. Highly-structured content from the IPKB forms the basis for a large-scale molecular network of direct interactions observed between mammalian orthologs, which is used in Ingenuity's Pathway Analysis (IPA) system. In this study we explore the relationship between this global network and known functional annotations of genes. In particular we show that (a) subnetworks formed by genes annotated with the same functional category have significantly more edges than equivalent random subnetworks, and (b) highly-interconnected subnetworks are significantly enriched in genes with specific functional annotations.

1. Introduction

The idea that biological function correlates with locally dense interactions in complex molecular networks has been explored in a number of recent publications. Subnetworks of highly-interconnected nodes that are less connected with the rest of the network have been identified as functional modules in protein-protein interaction [5], metabolic [2], and transciption regulation networks [1].

Ingenuity's Pathway Analysis[1] (IPA) algorithm, used to construct biologically relevant subnetworks from a list of user-provided genes of interest, is also based in part on the assumption that a high density of network interactions is an indicator for coherent biological function. In order to show that biological function is in fact related to dense subnetworks in IPA's underlying large-scale mammalian molecular interaction network we perform a quantitative statistical analysis using known functional annotations of genes. This statistical analysis is based on a null model of random graphs which explicitly preserves the expectation values of node degrees. Highly-interconnected subnetworks are identified by maximizing the network's modularity in replicated simulated annealing runs, and subsequently applying a hierarchical clustering method.

2. Method

Let G be an undirected graph representing the global network with V nodes and E edges. The null model is given by an ensemble of random graphs G' defined over the same set of nodes as G where edges are chosen independently at random such that for each node (with index i) the expectation value of the node degree in G' corresponds to the node degree d_i in G. Both G and G' shall contain no self-edges. It can be shown that the edge probability p_{ij} ($i \neq j$) of G' is then approximately given by

$$p_{ij} = \frac{V}{V-1}\frac{d_i d_j}{2E}. \qquad (1)$$

Let $G[S]$ and $G'[S]$ be the subgaphs induced by a given set S of nodes in G and G'. The number of edges X in $G'[S]$ is a random variable with expectation value $\mathbb{E}(X) = \sum_{i>j} p_{ij}$ where the sum runs over all possible edges between nodes in S. Let S_A be the set of all nodes that are annotated with a specific functional category A. Since the number of possible edges between nodes in S_A is large in most cases ($= |S_A|(|S_A|-1)/2$) and all edges are chosen independently, the probability distribution $P(x)$ of the number of edges X in $G'[S_A]$ is well approximated by a Poisson distribution

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad (2)$$

where $\lambda = \mathbb{E}(X)$. Let E_A be the number of edges in $G[S_A]$. We can then calculate the right-tailed p-value

$$p = \sum_{x \geq E_A} P(x) \qquad (3)$$

[1] http://www.ingenuity.com

as a measure of significance for the enrichment in edges in the subnetwork formed by genes with annotation A.

Highly-interconnected subnetworks are identified by the following method: For any given partitioning of the network G into K subsets of nodes S_k (called "modules" or "communities") the modularity $\mathcal{M}$ is defined as [3]

$$\mathcal{M} = \frac{1}{E} \sum_{k=1}^{K} \left[E_k - \mathbb{E}(X_k) \right], \tag{4}$$

where E_k is number of edges in $G[S_k]$ and the random variable X_k is the number of edges in $G'[S_k]$. The goal is to maximize $\mathcal{M}$ in the space of all possible partitionings. An equivalent formulation of (4) uses the Hamiltonian of a q-state Potts model

$$H = -\sum_{i>j} \left(a_{ij} - p_{ij} \right) \delta_{\sigma_i \sigma_j}, \tag{5}$$

where σ_i are Potts spins, and a_{ij} denotes the adjacency matrix of G. Here, modules correspond to sets of nodes carrying the same spin, and the partitioning with maximal modularity corresponds to the ground state of (5). The number of Potts spin states q is irrelevant as long as q is larger than the number of modules. As pointed out in [4] the Hamiltonian (5) contains a ferromagnetic and an anti-ferromagnetic contribution and the ground-state will in general exhibit spin glass-like behavior, i.e. there are many low-lying energy states. This means that modules are in general "fuzzy" [4], and boundaries between modules are not well-defined.

In order to circumvent this problem and to remove ambiguity in the choice of optimal modules we have developed a method to locate clusters of highly-interconnected nodes that consistently appear together in the *same* module in *many* partitionings represented by local minima of (5). We perform a number of independent simulated annealing runs, each starting from a different initial condition, to generate a set P of locally optimal partitionings of the network G. We then calculate the spin-spin correlation matrix $g_{ij} = \left\langle \delta_{\sigma_i \sigma_j} \right\rangle$, where the average is taken over all samples in P. For pairs of nodes (i,j) that most of the time appear in the same module, i.e. have the same Potts spin, we expect that g_{ij} will be close to 1 while for all other pairs g_{ij} is expected to be small. In fact, it is found that g_{ij} can be transformed into an approximate block-diagonal form using a hierarchical clustering method with an appropriately defined metric in the space of row (or column) vectors of g_{ij}. Clusters of highly-interconnected nodes determined this way turn out to be insensitive to details of the hierarchical clustering method. Based on 12 independently replicated simulated annealing runs we found 16 highly-connected subnetworks of G with sizes larger than 50 nodes.

	p [Eq. (3)]	E	$\mathbb{E}(X)$
TRPS/TK signaling pathway	$3.631{\cdot}10^{-69}$	71	3.1
processing of RNA	$5.800{\cdot}10^{-100}$	112	5.8
cytolysis	$4.589{\cdot}10^{-12}$	53	17.3
response to biotic stimulus	$4.405{\cdot}10^{-97}$	2626	1696.0
adhesion of cells	$2.231{\cdot}10^{-88}$	1005	498.5
GPCRP signaling pathway	$2.616{\cdot}10^{-203}$	321	30.7
protein kinase cascade	$2.007{\cdot}10^{-65}$	229	57.3
secretory pathway	$3.278{\cdot}10^{-65}$	92	7.6
metabolism of DNA	$1.359{\cdot}10^{-86}$	1256	680.8
cell cycle progression	$6.815{\cdot}10^{-77}$	2833	1957.7
synaptic transmission	$2.930{\cdot}10^{-52}$	152	32.4
ion transport	$9.309{\cdot}10^{-38}$	52	4.3
biosynthesis of protein	$9.464{\cdot}10^{-28}$	133	43.4
regulation of apoptosis	$3.054{\cdot}10^{-31}$	370	189.5
binding of cells	$5.301{\cdot}10^{-120}$	592	189.9
TRPTK signaling pathway	$2.780{\cdot}10^{-68}$	130	16.9
metabolism of DNA	$1.359{\cdot}10^{-86}$	1256	680.8
transcription	$3.488{\cdot}10^{-115}$	3253	2120.3

Table 1. P-values [Eq. (3)], number of edges, and expected number of edges for various functional annotations (TRPS/TK = transmembrane receptor protein serine/threonine kinase, GPCRP = G-protein coupled receptor protein, TRPTK = transmembrane receptor protein tyrosine kinase).

3. Results

The results of this analysis are shown in Table 1 and Figure 1. We calculate network p-values according to Eq. (3) for 1694 biological process annotations that involve at least 20 genes in the global network. These annotations were derived from the IPKB findings and annotations from the Gene Ontology [6]. We find that 39% of these annotations have a network p-value that is smaller than 10^{-10}. Table 1 lists network p-values for 18 functional categories along with the actual and randomly expected number of edges in the corresponding subnetworks. In all cases we find a significantly increased number of edges in subnetworks defined by genes with the same functional annotation when compared to a random network.

We determined a second set of p-values (shown in Figure 1) that measure the significance of annotations in the context of each of the 16 highly-interconnected clusters identified with the method described above. These annotation p-values were calculated using Fisher's exact test based on the size of the network, the size of the clusters, and the number of annotated genes in the network and in the cluster. We found that all of the clusters, except for cluster 5 and 7, can be unambiguously assigned to one or two specific, high-level functional categories with the lowest annotation p-values. These are the 18 functional categories listed in Figure 1 and Table 1. The numbers of annotated genes (between 50 and 887) and cluster sizes (between 50 and

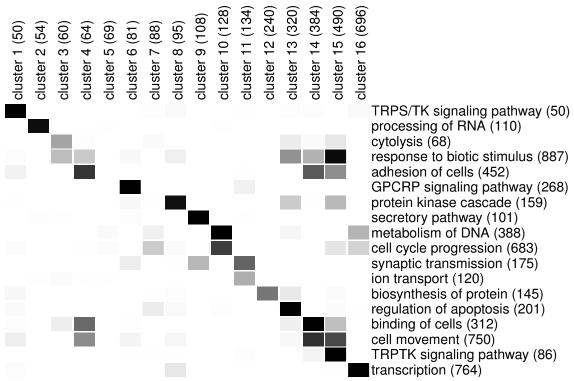

Figure 1. P-values of functional annotations in highly-interconnected subnetworks (black: $< 10^{-30}$, white: 1, logarithmic grayscale in-between).

696) are shown in parantheses adjacent to the corresponding functional annotation or cluster index. Three of the clusters (cluster 5, 7, and 12) consist of few interconnected hubs (E2F1/E2F4, TP53/TP73, and MYC/MYCN/JRK) and their leafs. As an example, the subnetwork corresponding to cluster 2 is shown in Figure 2 with genes carrying its dominant functional annotation *processing of RNA* highlighted.

4. Conclusion

In this analysis we have examined characteristics of the global mammalian direct molecular interaction network computed from Ingenuity's Pathways Knowledge Base (IPKB). We have shown for a number of functional categories that subnetworks formed by genes annotated with the same biological function have significantly more edges than equivalent random subnetworks based on a node-degree preserving null model. We have also shown that highly-interconnected subnetworks (clusters), which were detected by maximizing modularity, are significantly enriched in genes with specific functional annotations. In particular we found 14 clusters that can be unambiguously assigned to one or two dominant functional categories. These findings suggest that network clustering algorithms that optimize for densely-connected subnetworks are likely to identify genes that participate in coordinated biological function.

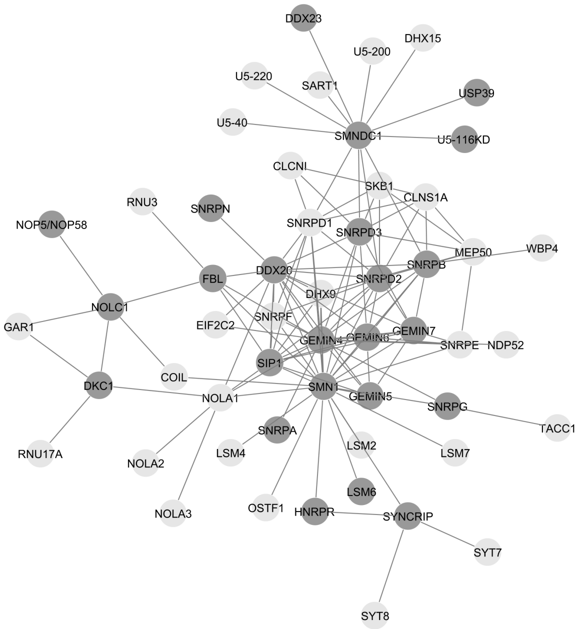

Figure 2. Subnetwork corresponding to cluster 2 in Figure 1 with genes annotated *processing of RNA* **highlighted.**

References

[1] R. Boscolo, B. A. Rezaei, V. P. Roychowdhury, and P. O. Boykin. Functionality encoded in topology? Discovering macroscopic regulatory modules from large-scale protein-DNA interaction networks. *arXiv:q-bio.MN/0501039*, January 2005.

[2] R. Guimerà and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, February 2005.

[3] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[4] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.*, 93(21):218701, 2004.

[5] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, October 2003.

[6] The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.