BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 19 2011, pages 2707–2713
doi:10.1093/bioinformatics/btr469

Systems biology

Advance Access publication August 10, 2011

Contextual analysis of RNAi-based functional screens using interaction networks

Orland Gonzalez* and Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, 80333 München, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Considerable attention has been directed in recent years toward the development of methods for the contextual analysis of expression data using interaction networks. Of particular interest has been the identification of active subnetworks by detecting regions enriched with differential expression. In contrast, however, very little effort has been made toward the application of comparable methods to other types of high-throughput data.

Results: Here, we propose a new method based on co-clustering that is specifically designed for the exploratory analysis of large-scale, RNAi-based functional screens. We demonstrate our approach by applying it to a genome-scale dataset aimed at identifying host factors of the human pathogen, hepatitis C virus (HCV). In addition to recovering known cellular modules relevant to HCV infection, the results enabled us to identify new candidates and formulate biological hypotheses regarding possible roles and mechanisms for a number of them. For example, our analysis indicated that HCV, similar to other enveloped viruses, exploits elements within the endosomal pathway in order to acquire a membrane and facilitate assembly and release. This echoed a number of recent studies which showed that the ESCRT-III complex is essential to productive infection.

Contact: gonzalez@bio.ifi.lmu.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 24, 2011; revised on July 18, 2011; accepted on August 7, 2011

1 INTRODUCTION

An important challenge in the analysis of high-throughput datasets, such as transcriptomic and proteomic data, is the integration of prior knowledge in their interpretation. One class of methods that has arisen to address this challenge is gene set analysis (GSA) (Dinu *et al.*, 2009; Goeman and Buhlmann, 2007; Liu *et al.*, 2007). Motivated by the increasing body of evidence that cellular responses are usually organized into pathways or active modules (Hartwell *et al.*, 1999; Vidal, 2001), GSA methods try to assess the overall association between the phenotype of interest and known groupings of genes, by comparing the latter with the data. Although hugely successful, as evidenced by a large body of literature, GSA suffers from a number of drawbacks. Probably foremost of these, and somewhat ironic, is its reliance on predetermined gene sets. As a consequence of this dependency, GSA cannot be expected to discover novel pathways. In addition, with few exceptions (see Massa *et al.*, 2010; Tarca *et al.*, 2009), GSA algorithms treat pathways simply as sets; i.e. they completely ignore information on the individual interactions between entities (e.g. topology). Such information could potentially also be of use.

To complement the shortcomings of GSA, methods that try to more generally integrate interaction with expression data, outside the scope of canonical pathways, have emerged. We collectively refer to these as significant area search (SigArSearch) methods. One of the earliest of these was the co-clustering approach by Hanisch *et al.* (2002). In their paper, they proposed to cluster genes based on a distance function that integrated measures of both expression correlation and interaction network proximity. Almost at the same time, Ideker *et al.* (2002) presented a method that framed the contextual interpretation task as an optimization problem. Specifically, they defined a scoring function on subnetworks—based on the data values of the genes included in a subgraph—and employed a search algorithm to find high scoring regions. Because the underlying combinatorial problem, i.e. that of finding the maximal-scoring connected subgraph with additive scores defined on the vertices, was proven to be NP-hard, the authors proposed a heuristic strategy based on simulated annealing.

Similar to GSA, the interpretative power of SigArSearch is also limited by the scope of the background knowledge/information that is used (i.e. the interaction network). However, a key difference between the two paradigms exists, owing to the different roles that gene groups play in them. Whereas GSA requires that these be specified beforehand, SigArSearch actually tries to infer such modular relationships between genes directly (at least the ones relevant to the current expression data), subject to the binary relations encoded in the network. Thus, SigArSearch is potentially able to discover new groupings, which could, for example, correspond to uncharacterized protein complexes or pathways. This capability is particularly relevant if the background network used consists of, or includes, interactions taken from unbiased, large-scale interaction screens (e.g. using Y2H or TAP/MS; see Shoemaker and Panchenko, 2007 for a review of such methods).

The optimization framework introduced by Ideker *et al.* (2002) for SigArSearch later inspired the development of several improvements. For example, a number of groups proposed the use of alternative heuristic search strategies, such as the greedy approaches of Sohler *et al.* (2004) and Nacu *et al.* (2007). Others proposed extensions to the scoring system, such as including scores defined on the edges (interactions) (Cabusora *et al.*, 2005) instead of just the vertices (genes). Subsequently, Dittrich *et al.* (2008) pointed out that the existing methods (based on the optimization framework)

*To whom correspondence should be addressed.

were heuristic in nature, and as such could not guarantee to identify the maximally scoring subgraph. In light of this, they presented an exact approach that was described as being able to deliver provably optimal solutions in reasonable time, which worked by reducing instances of the maximal subgraph problem to an integer-linear program. Other recent developments include extensions to the scoring system that incorporate measures of gene coexpression (whereas previous methods only used the individual data values of genes) (Guo *et al.*, 2007; Ulitsky and Shamir, 2007), and an approach based on random walks that forgoes the optimization paradigm altogether (Komurov *et al.*, 2010). Although these methods are already quite popular for use with expression data, applications to other types of high-throughput datasets are still limited. In this study, we present a new method, inspired by the co-clustering approach of Hanisch *et al.* (2002), that is specifically intended for use with RNAi-based functional screens.

Functional studies in cultured cells were hampered in the past by the lack of a powerful method for perturbing gene activities (Echeverri and Perrimon, 2006). This changed with the discovery of RNA interference (RNAi) and the subsequent development of siRNA libraries aimed at targeting complete genomes for a number of organisms (Birmingham *et al.*, 2009). Indeed, RNAi screens have proven to be effective at identifying genes associated with various biological processes, including cellular differentiation (Chia *et al.*, 2010; Hu *et al.*, 2009; Zhao and Ding, 2007), cancer (Bauer *et al.*, 2010; Wurdak *et al.*, 2010; Zender *et al.*, 2008), signaling (Berns *et al.*, 2004; DasGupta *et al.*, 2005) and host–pathogen interactions (Brass *et al.*, 2008, 2009; Li *et al.*, 2009; Tai *et al.*, 2009; Zhou *et al.*, 2008). Although there are now well-defined computational approaches for the various stages of the primary analysis of RNAi screens, including quality control (Zhang *et al.*, 1999; Zhang, 2007b), normalization (Malo *et al.*, 2006; Wiles *et al.*, 2008) and hit selection (Chung *et al.*, 2008; König *et al.*, 2007; Zhang, 2007a), the potential for methods that integrate interaction information into the interpretation of the data is only starting to be realized (see Wang *et al.* 2009 and Bankhead *et al.* 2009 for some examples). In this study, we propose a new method for the contextual interpretation of RNAi screens that works by co-clustering together with interaction data.

2 METHODS

2.1 Data sources

We demonstrate our approach by applying it to a genome-wide RNAi screen for host factors of the human pathogen, hepatitis C virus (HCV) (Tai *et al.*, 2009). The screen provided ‘fold change’ data, which represents the ratio by which viral growth was inhibited or enhanced as a consequence of the knockdown of a particular gene, as well as *P*-values. We were able to map ∼18 000 of the knockdowns to ENTREZ records.

To serve as context for the analysis of the RNAi data, we assembled a network by collecting interactions defined in the STRING database (Jensen *et al.*, 2009). Only those rated with at least a medium level of confidence (combined score≥0.4) were included. As with before, all identifiers were mapped to ENTREZ records. Mapping information (Ensembl protein id to Entrez gene id) was retrieved from ENSEMBL BioMart. In situations where more than one STRING record could be mapped to a gene pair, the genes were considered to interact if at least one of the records fulfilled the minimum required combined score. This resulted in a network composed of 13 104 vertices (genes) and 330 523 edges (interactions).

2.2 Co-clustering

As mentioned earlier, the task of analyzing expression data in the context of an interaction network has often been framed as an optimization problem (Dittrich *et al.*, 2008; Ideker *et al.*, 2002; Nacu *et al.*, 2007; Sohler *et al.*, 2004). Roughly speaking, this means defining a scoring function on subnetworks, and then using a search algorithm to look for high scoring, connected regions. Although a similar approach is possible for RNAi data, such a formulation of the problem is probably not optimal. At the very least, it should not be the only one investigated. For one thing, simply searching for high scoring, connected subnetworks, where the score is based only on the individual data values of the genes included in a subgraph, would all too often return regions that are, although connected, not functionally coherent. This is especially a problem when using interaction networks that have a small world property, which is the case for many biological networks (at least in an approximate sense). Although it is possible to ameliorate this problem in the case of expression data by including measures of coexpression in the scoring function (Guo *et al.*, 2007; Ulitsky and Shamir, 2007), a similar concept unfortunately does not exist in RNAi datasets. As such, functional coherency needs to be enforced through some other means.

As an alternative (or complement) to viewing the joint interpretation task as an optimization problem, we propose to frame it as one of clustering. The general idea of our approach is to ‘cocluster’ by simultaneously considering the two types of data, such that genes which are both near each other in the interaction network and at the same time showing strong links to the phenotype of interest (i.e. the RNAi data) tend to be clustered together.

We represented an interaction network as an undirected graph $G=(V,E)$, where the set of vertices V correspond to genes, and the set of edges E to interactions. Given an edge $e_{u,v}$ which connects two vertices u and v, a weight, which we view as a distance, was assigned to the edge in such a manner that it is smaller the stronger the u and v are linked to the phenotype (i.e. the stronger the data values of u and v in the RNAi screen, the shorter the edge connecting them, if it exists). Specifically, if p_u and p_v are the *P*-values in the RNAi screen of the genes u and v, respectively, then we can view

$$h_{u,v}=(1-p_u)(1-p_v) \qquad (1)$$

as a measure of the ‘probability’ that both genes u and v are relevant to the phenotype of interest. Following Equation (1), we calculate the edge weight $w(e_{u,v})$ as the probability of observing $h_{u,v}$ or higher. Formally,

$$w(e_{u,v})=P\left(h_{u,v}\leq(1-X)(1-Y)\right) \qquad (2)$$

where X and Y are independent random variables with a standard uniform probability distribution. The right-hand side of Equation (2) can be expressed as

$$\int_0^{1-h_{u,v}}\left(1-\frac{h_{u,v}}{1-X}\right)dX$$

which can then be simplified to

$$1+h_{u,v}\cdot\left(\ln\left(h_{u,v}\right)-1\right) \qquad (3)$$

Note that $w(e_{u,v})$, which is visualized in Figure 1, can be viewed as a *P*-value for how high $h_{u,v}$ is. Roughly speaking, the more significant p_u and p_v are, the smaller $w(e_{u,v})$ becomes.

Equation (3) was derived with the assumption that the *P*-values associated with the two vertices linked by an edge are independent. This, however, is not always true in actual biological data. Indeed, it is very reasonable to expect related genes, for example those that belong to the same protein complex, to induce similar phenotypes on the virus (if indeed they do induce one). Nevertheless, this does not hurt the intended purpose of the edgeweight, which is simply to bring two genes closer to each other in the network if they both have strong data values (i.e. they both induce strong phenotypes).

With weights assigned to all the edges of the interaction network, we then construct the $|V|\times|V|$ distance matrix M, where $m_{i,j}$ is the length of the shortest path from vertex i to j. We use the well-known Floyd–Warshall algorithm to solve the all-pairs-shortest path problem. Using M as input, we then cluster the genes using a version of *average linkage clustering*.

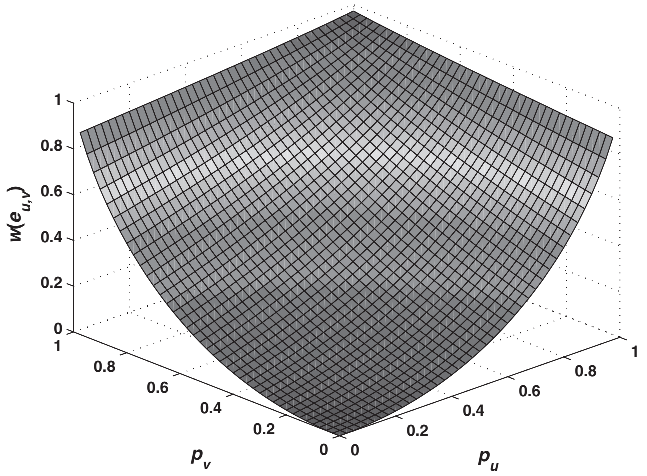

Fig. 1. Edge weights. The surface corresponds to the weight (distance) assigned to an edge as a function of the *P*-values of the two vertices that it connects.

Briefly, the method starts by treating the set of vertices (genes) as singleton clusters, and then successively joins clusters with the smallest average pairwise distance. The result of the clustering procedure is a binary tree (or dendogram), wherein each leaf corresponds to a gene (vertex), and each internal node to a cluster consisting of all the genes (leaves) that descend from it.

The dendogram by itself could already be used for biological hypothesis generation. Starting from seed nodes, which could be genes of prior interest or simply genes with the strongest data values in the RNAi screen, one could gain insight into the mechanism of action of the seed gene by looking at the clusters that it belongs to. This can be done by exploring the path in the dendogran which originates from the corresponding leaf, and then moving progressively upwards. Because of the manner in which the clustering procedure was conducted, the clusters along this path should give an impression as to which genes are both associated with a strong screening value and at the same time near the seed in the interaction network. To aid in this exploration, we use an additional scoring system to estimate the amount of signal in the different clusters. One possibility for this purpose is the subnetwork score used by Ideker *et al.* (2002) for expression data:

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \tag{4}$$

where A is the set of vertices in the subnetwork, k is the size of A and z_i is the z-score of vertex (gene) i. However, the problem with Equation (4) is that it is far too 'tolerant'; such that search strategies using this scoring function often return huge subnetworks, even if there is actually no signal in the data (Dittrich *et al.*, 2008; Rajagopalan and Agarwal, 2005). The reason for this is that even under the null hypothesis, where z_i are normally distributed, approximately half of them could still contribute to a positive score.

Because of the shortcomings of Equation (4), we defined an alternative scoring function, similar to that employed by Dittrich *et al.* (2008), which allows the control or estimation of the amount of signal in the clusters. Specifically, we calculated the score of a cluster C as

$$S_{q_0}(C) = \sum_{v \in C} \log \left(\frac{\frac{(1-q_v)}{q_v}}{\frac{(1-q_0)}{q_0}} \right) \tag{5}$$

where q_v is the q-value of vertex v, and $q_0 \in (0, 1.0]$ is a reference q-value parameter. Since the q-value of a test provides a measure of the false discovery rate (FDR) when that particular test is called significant, then the main numerator $r_v = (1-q_v)/q_v$ of Equation (5) can be roughly interpreted as a signal to noise ratio for vertex v. Inspired by a log-odds score, the final contribution of v to S_{q_0} was calculated by dividing r_v by the analogous ratio for the reference q-value, q_0, and then taking the logarithm. This results in a scoring system wherein only vertices associated with an FDR better than q_0 would be able to contribute positively. Consequently, even though clusters can contain negative scoring vertices, only those that include genes that actually have a chance of being significant can get a non-negative S_{q_0}. We calculated q-values using the method of Storey (2002).

In addition to exploring the dendogram by tracing paths from seed nodes (leaves), an alternative strategy for biological hypothesis generation is to directly extract 'active regions' from the network. We do this by scanning the whole tree, and then creating a super cluster consisting of all clusters that have a positive score given q_0 and larger than a minimum size. In this way, the parameter q_0 controls both the size of, and the amount of signal in, the resulting subnetwork. Note that if the RNAi data actually has no signal, then all genes would have a q-value of at least 1.0. This means that no cluster can have a positive score regardless of q_0, and accordingly an empty subnetwork would be returned (which is the desired behavior). Control of this parameter is further discussed in the Supplementary Material.

3 RESULTS AND DISCUSSION

HCV is a positive sense single-stranded RNA virus that in humans causes its namesake disease, hepatitis C (Senecal and Morelli, 2007). About 3% of the world's population (270–300 million) is chronically infected with HCV. Of this number, ∼30% will develop cirrhosis (liver scarring) within 20 years of initial infection, a condition that could then progress to life-threatening complications, including liver failure and hepatocellular carcinoma. Like all viruses, various stages in the life cycle of HCV, including entry, uncoating, intracellular transport, replication, assembly and egress, are dependent on cellular proteins. We applied our method to a genome-wide screen aimed at identifying these host factors (Tai *et al.*, 2009). Out of the approximately 21 000 knockdowns included in the study, we were able to map 17 821 to ENTREZ records. Of these, 13 104 participate in at least one interaction in the STRING-derived network. All subsequent analyses were limited to this subset.

One of the genes with the strongest RNAi data value is COPB1. It is a subunit of the COPI complex, which coats vesicles transporting proteins from the Golgi body back to the rough endoplasmic reticulum. The early parts of the clustering path originating from the COPB1 leaf going to the dendogram root is visualized in Figure 2. In addition to COPB1, several other COPI subunits showed very strong RNAi data values. These were incrementally clustered together with the seed, culminating in cluster A (Fig. 2), which includes five of the seven COPI subunits and the genes ARF1 and ARCN1. Given that the latter two genes are also involved in vesicular trafficking, it is thus very likely that the strong phenotype induced by knockdown of COPB1 has indeed something to do with its role in the COPI complex. Note that although the subsequent additions to cluster A decreased the score (see clusters B and C), this does not mean that the added genes are likely to be noise, since a very stringent reference q-value was used ($q_0 = 0.02$). Indeed, all the genes in cluster C are associated with a $q < 0.5$. A comparison between using Equations (4) and (5) to score the clusters is made in Supplementary Figure S3.

In addition to manual exploration, the dendogram can also be used for biological hypothesis generation by directly extracting from the network regions with high activity and coherence. In the case of HCV, we do this by merging all positve-scoring clusters (for $q_0 = 0.02$) with at least four vertices. The resulting subnetwork

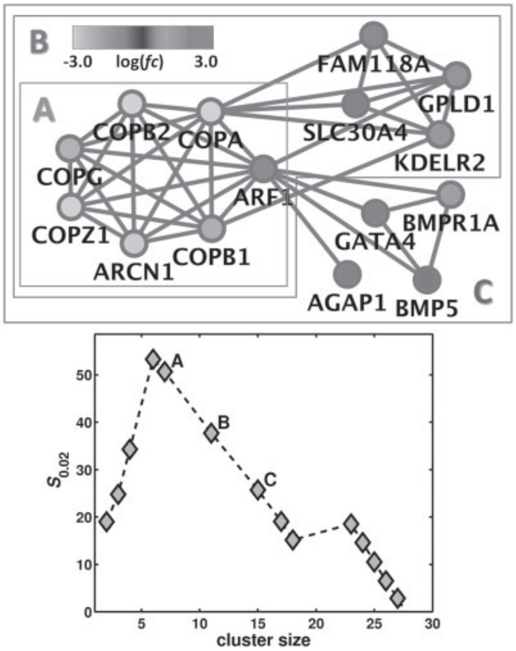

Fig. 2. Clustering for seed COPB1. The plot at the bottom shows the scores and sizes of the clusters originating from the COPB1 leaf. Three of the clusters, labeled A, B and C, are visualized on the diagram at the top (*fc* = fold change).

is visualized in Figure 3. Note that it includes the earlier described coatomer complex. Other prominent functions and/or complexes include eukaryotic translation initiation, ribosome, RNA export, inositol metabolism, rRNA and ESCRT-III. A detailed list of the genes is provided as Supplementary Material.

Chromatin-modifying protein 2A (CHMP2A) was one of the strongest hits in the original screen (Tai *et al.*, 2009). Although the gene was further validated by deconvolution of the corresponding siRNA pool, its relevance to HCV replication was not explored beyond noting that the gene has been linked to membrane trafficking. Here, the clustering (Fig. 3) gives some insights on a probable mechanism as to how CHMP2A could facilitate HCV infection. Note that the gene is clustered together with VPS4A, VPS24 and CHMP4B, all of which are associated with multivesicular bodies (MVBs). These play central roles in the endosomal pathway, which is responsible for internalizing molecules from the plasma membrane and sorting them for degradation or recycling back to the cell surface. In particular, CHMP2A, CHMP4B and VPS24 (also called CHMP3) are components of the ESCRT-III complex, which, together with other multiprotein complexes (0, I and II), act to incorporate ubiquitinated target proteins into MVBs. On the other hand, VPS4A catalyzes the dissociation of membrane-bound ESCRT-III assemblies, and redistributes them to the cytoplasm for further rounds of MVB sorting. Given that many enveloped viruses exploit this pathway in order to acquire a membrane and facilitate assembly and release, including human immunodeficiency virus type-1 (Garrus *et al.*, 2001), Ebola virus (Martin-Serrano *et al.*, 2001), hepatitis B virus (Lambert *et al.*, 2007) and herpes simplex virus 1 (Crump *et al.*, 2007), it is not difficult to imagine that the same is also true for HCV. Indeed, a very recent study concluded this to be the case (Corless *et al.*, 2010). In it, they showed that viral particle production is greatly reduced upon transfection of plasmids expressing various dominant negative forms of VPS4 and ESCRT-III components.

Edge weights, as calculated according to Equation (3), rely on the individual *P*-values of the adjacent vertices. However, it is not always possible to obtain such statistical measures of significance (i.e. *P*-values), particularly if the data are unusually distributed. Accordingly, we also developed a non-parametric version of our method that is able to deal with such cases. This is described and compared to the basic approach in the Supplementary Material.

To illustrate the difference between our method, which aims at greater functional coherency in the identified active subnetworks, and one that uses an optimization formulation, we also applied the algorithm of Dittrich *et al.* (2008) (as implemented in the BioNet R package) to the HCV data. The comparison is summarized in Figure 4. Note that for similarly sized subnetworks, the ones returned by our method (labelled as 'COCLUST - ave-linkage' in the figure) tend to have more connections between the vertices (i.e. the genes tend to have more interactions between them; see graph of Fig. 4A and C). However, this increase in functional coherency was achieved at the cost of an increase in the overall *q*-value level associated with the genes (Fig. 4B). Finer control of this trade-off is one of the directions that we intend to pursue in subsequent studies.

In addition to average-linkage, we also tried using a single-linkage criterion for the clustering aspect of our method when applied to the HCV data. The results are again summarized in Figure 4. Using this variant, the properties of the subnetworks that we identified, specifically with respect to connectivity and the mean associated *q*-value of the genes, were more similar to those of comparably sized subnetworks found by the optimization approach BioNet. This is because in contrast to average-linkage, where merge steps depend on the average distance between all pairs of genes from two clusters, single-linkage only considers the two genes that are closest to each other. That is, single-linkage, similar to BioNet (which only requires subnetworks to be connected), does not take group effects into account. In fact, for small network sizes ($n < 100$), single-linkage co-clustering even returned subnetworks with a better overall *q*-value level than the optimization approach (Fig. 4A). This latter behavior is due to the fact that, in contrast to BioNet, we do not require active regions to be fully connected; rather, we allow them to be composed of separate clusters or modules.

4 CONCLUSIONS AND FUTURE WORK

We presented a method for the contextual interpretation of RNAi-based functional screening data using interaction networks. Rather than posing the task as a maximal connected subgraph problem, as is often done with expression data, we framed it as one of clustering. In particular, we employed a co-clustering strategy where genes that are both near each other in the interaction network and at the same time show strong links to the phenotype of interest tend to be clustered together. The result of our method is a dendogram that can be used for biological hypothesis generation in two ways: (i) by manually exploring the clustering from seed genes of *a priori* interest; or (ii) by directly extracting 'active subnetworks'. We demonstrated our approach by applying it to a genome-wide screen for host factors of the human pathogen, hepatitis C virus. In addition to recovering known cellular modules relevant to HCV infection, our analysis

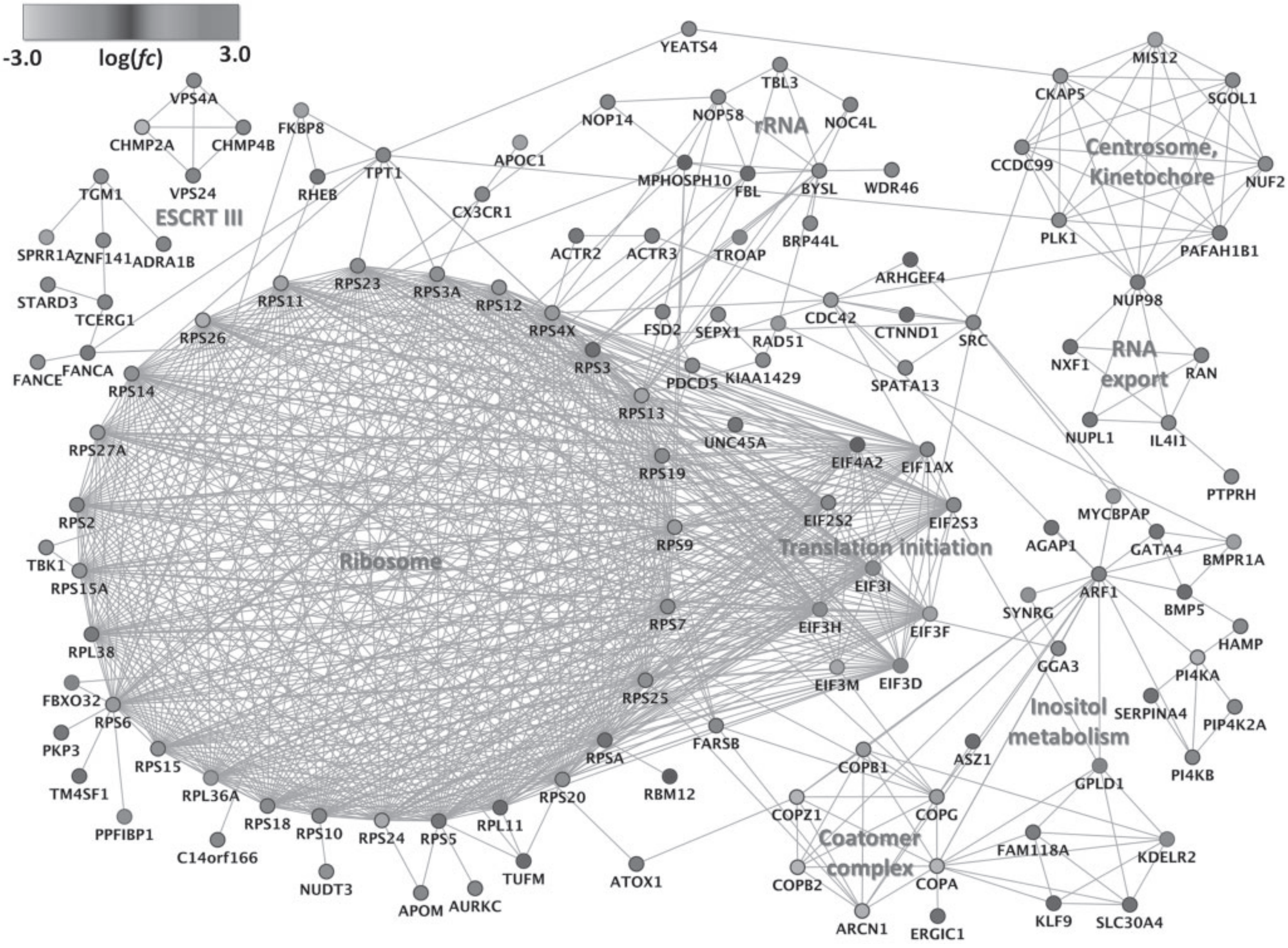

Fig. 3. Host factors of hepatitis C virus. A global view of the human proteins affecting HCV replication was formed by merging all positve-scoring clusters (for $q_0 = 0.02$) with at least four vertices (fc = fold change). In addition to the coatomer complex that was described earlier, other prominent functions and/or complexes in the subnetwork include eukaryotic translation initiation, ribosome, RNA export, inositol metabolism, rRNA and ESCRT-III.

allowed us to identify new potential candidates, and formulate biological hypotheses regarding possible roles and mechanisms for a number of them.

Our clustering approach enforces a level of functional coherency in the resulting subnetworks by considering ‘group effects’ at each merge step. As opposed to the more traditional optimization strategy usually employed for expression data, where subnetworks only need to be connected, the use of an average linkage score means that clusters would tend to comprise more of genes that are more closely connected to each other. In addition to promoting functional coherency, this also affords our method the possibility of highlighting groups of related genes that, although individually may only be moderately associated with the phenotype of interest, could collectively be more important than genes with more obvious data values. This feature is especially important for RNAi datasets since false negatives systematically arise from several properties of the experimental system. For example, even if a gene were genuinely linked to the phenotype of interest, knocking it down would not necessarily result in an observable effect if there is another gene that could also perform the respective function. Even if the alternative can only partially compensate for the loss, this could still be enough to pull down the signal to noise levels. Another potential source of false negatives is the efficacy of the RNAi reagents themselves. In most large-scale screens, there is typically no information on the extent to which knockdown of a targeted protein has been achieved. In fact, it is often not even clear whether the target has been perturbed at all, or, potentially even more problematic, whether the RNAi reagent cross-reacted with unintended targets, thus causing the so-called ‘off-target effects’. Although the latter possibility actually causes false positives instead of false negatives, it still underscores the need for functional coherency in the results in order to improve confidence.

One of the future improvements that we intend to pursue is the formulation of alternative edge weight functions [see Equation (3)]. This will have an influence on the method’s trade-off between the individual significance of the genes in the clusters and their functional coherency. In particular, we plan to investigate the use of a weighted sum of logistic curves, with shape parameters that could be used to control the aforementioned trade-off in a continuous manner. Other aspects that we also plan to investigate include the

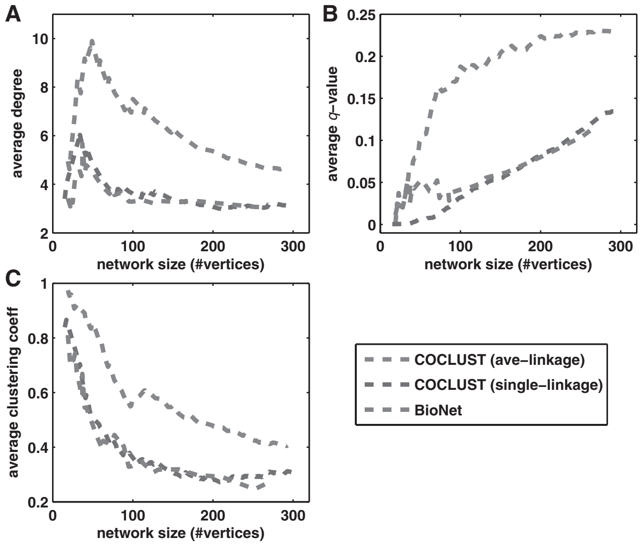

Fig. 4. Comparison with an optimization approach. We compared active regions found by our method (using both average- and single-linkage clustering) to those found by BioNet with respect to: (**A**) average degree; (**B**) average *q*-value associated with the nodes; and (**C**) average clustering coefficient.

use of more sophisticated methods for extracting active regions from the dendogram, the use of alternative clustering algorithms, and extending our method to be applicable to high-dimensional RNAi screens (i.e. screens that simultaneously monitor more than one phenotype). We will be applying the method to three new RNAi-based screens for viral host factors that we are currently analyzing.

Funding: German Research Foundation (Deutsche Forschungsgemeinschaft) through the International Research Training Group RECESS (IRTG 1563).

Conflict of Interest: none declared.

REFERENCES

Bankhead,A. *et al.* (2009) Knowledge based identification of essential signaling from genome-scale siRNA experiments. *BMC Syst. Biol.*, **3**, 80.

Bauer,J.A. *et al.* (2010) RNA interference (RNAi) screening approach identifies agents that enhance paclitaxel activity in breast cancer cells. *Breast Cancer Res.*, **12**, R41.

Berns,K. *et al.* (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, **428**, 431–437.

Birmingham,A. *et al.* (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat. Methods*, **6**, 569–575.

Brass,A. *et al.* (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, **15**, 921–926.

Brass,A.L. *et al.* (2009) The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, **139**, 1243–1254.

Cabusora,L. *et al.* (2005) Differential network expression during drug and stress response. *Bioinformatics*, **21**, 2898–2905.

Chia,N.Y. *et al.* (2010) A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, **468**, 316–320.

Chung,N. *et al.* (2008) Median absolute deviation to improve hit selection for genome-scale RNAi screens. *J. Biomol. Screen.*, **13**, 149–158.

Corless,L. *et al.* (2010) Vps4 and the ESCRT-III complex are required for the release of infectious hepatitis C virus particles. *J. Gen. Virol.*, **91**, 362–372.

Crump,C.M. *et al.* (2007) Herpes simplex virus type 1 cytoplasmic envelopment requires functional Vps4. *J. Virol.*, **81**, 7380–7387.

DasGupta,R. *et al.* (2005) Functional genomic analysis of the Wnt-wingless signaling pathway. *Science*, **308**, 826–833.

Dinu,I. *et al.* (2009) Gene-set analysis and reduction. *Brief. Bioinformatics*, **10**, 24–34.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, i223–i231.

Echeverri,C.J. and Perrimon,N. (2006) High-throughput RNAi screening in cultured cells: a user's guide. *Nat. Rev. Genet.*, **7**, 373–384.

Garrus,J.E. *et al.* (2001) Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding. *Cell*, **107**, 55–65.

Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Guo,Z. *et al.* (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, **63**, 3912–3918.

Hanisch,D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18** (Suppl. 1), S145–S154.

Hartwell,L.H. *et al.* (1999) From molceular to modular cell biology. *Nature*, **402**, C47–C52.

Hu,G. *et al.* (2009) A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes Dev.*, **23**, 837–848.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Jensen,L.J. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Komurov,K. *et al.* (2010) Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput. Biol.*, **6**, e1000889.

König,R. *et al.* (2007) A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Methods*, **4**, 847–849.

Lambert,C. *et al.* (2007) Hepatitis B virus maturation is sensitive to functional inhibition of ESCRT-III, Vps4, and gamma 2-adaptin. *J. Virol.*, **81**, 9050–9060.

Li,Q. *et al.* (2009) A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc. Natl Acad. Sci. USA*, **106**, 16410–16415.

Liu,Q. *et al.* (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

Malo,N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.

Martin-Serrano,J. *et al.* (2001) HIV-1 and Ebola virus encode small peptide motifs that recruit Tsg101 to sites of particle assembly to facilitate egress. *Nat. Med.*, **7**, 1313–1319.

Massa,M.S. *et al.* (2010) Gene set analysis exploiting the topology of a pathway. *BMC Syst. Biol.*, **4**, 121.

Nacu,S. *et al.* (2007) Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**, 850–858.

Rajagopalan,D. and Agarwal,P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Senecal,D.L. and Morelli,J. (2007) Hepatitis C virus infection: a current review. *JAAPA*, **20**, 21–25.

Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Sohler,F. *et al.* (2004) New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**, 1517–1521.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc.*, **64**, 479—498.

Tai,A. *et al.* (2009) A functional genomic screen identifies cellular cofactors of hepatitis C virus replication. *Cell Host Microbe*, **5**, 298–307.

Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.

Vidal,M. (2001) A biological atlas of functional maps. *Cell*, **104**, 333–339.

Wang,L. *et al.* (2009) A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics*, **10**, 220.

Wiles,A.M. *et al.* (2008) An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme. *J. Biomol. Screen.*, **13**, 777–784.

Wurdak,H. *et al.* (2010) An RNAi screen identifies TRRAP as a regulator of brain tumor-initiating cell differentiation. *Cell Stem Cell*, **6**, 37–47.

Zender,L. *et al.* (2008) An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell*, **135**, 852–864.

Zhang,J.H. *et al.* (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.*, **4**, 67–73.

Zhang,J.H. (2007a) A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. *J. Biomol. Screen.*, **12**, 645–655.

Zhang,J.H. (2007b) A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*, **89**, 552–561.

Zhao,Y. and Ding,S. (2007) A high-throughput siRNA library screen identifies osteogenic suppressors in human mesenchymal stem cells. *Proc. Natl Acad. Sci. USA*, **104**, 9673–9678.

Zhou,H. *et al.* (2008) Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, **4**, 495–504.