Hernández-Lemus and Siqueiros-García *Complex Adaptive Systems Modeling* 2013, **1**:8
http://www.casmodeling.com/content/1/1/8

Complex Adaptive Systems Modeling
a SpringerOpen Journal

RESEARCH Open Access

Information theoretical methods for complex network structure reconstruction

Enrique Hernández-Lemus[1,2*] and Jesús M Siqueiros-García[3]

*Correspondence:
ehernandez@inmegen.gob.mx
[1]Computational Genomics Department, National Institute of Genomic Medicine, Mexico City, México
[2]Complexity in Systems Biology, Center for Complexity Sciences, National Autonomous University of México, Mexico City, Mexico
Full list of author information is available at the end of the article

Abstract

Purpose: Complex networks seem to be ubiquitous objects in contemporary research, both in the natural and social sciences. An important area of research regarding the applicability and modeling of graph- theoretical-oriented approaches to complex systems, is the probabilistic inference of such networks. There exist different methods and algorithms designed for this purpose, most of them are inspired in statistical mechanics and rely on information theoretical grounds. An important shortcoming for most of these methods, when it comes to disentangle the actual structure of complex networks, is that they fail to distinguish between direct and indirect interactions. Here, we suggest a method to discover and assess for such indirect interactions within the framework of information theory.

Methods: Information-theoretical measures (in particular, Mutual Information) are applied for the probabilistic inference of complex networks. Data Processing Inequality is used to find and assess for direct and indirect interactions impact in complex networks.

Results: We outline the mathematical basis of information-theoretical assessment of complex network structure and discuss some examples of application in the fields of biological systems and social networks.

Conclusions: Information theory provides to the field of complex networks analysis with effective means for structural assessment with a computational burden low enough to be useful in both, Biological and Social network analysis.

Keywords: Complex networks structure, Probabilistic network inference, Feature selection, Information theory

MSC: 94A15, 62B10, 91D30, 05C82

Background

Complex networks, no doubt constitute one of the cornerstones of contemporary research in many branches of science (Barabási 2012; Newman 2003). From systems biology to the study of the role of friendship in the spread of diseases, connections between individuals at different organizational levels are outlined using complex graphs.

Alongside with the statistical analysis of large complex networks, the need for robust methodologies to infer such networks from empirical data has risen. Most of these methods rest on the domain of probabilistic inference and computational learning (Bickel and Doksum 2007; Hernández-Lemus and Rangel-Escareño 2011) and as such, they are subject to expectation errors and asymptotic constraints. Apart from the problem of inferring

large networks from noisy data sources (Bansal et al. 2007; de Jong 2002; Hernández-Lemus et al. 2009), complex network researchers in general, are confronted with some subtler structural challenges in network reconstruction. One of such challenges lies in the capacity to assess direct from indirect interactions (Chua et al. 2008; Tresch et al. 2007).

The assessment of direct and indirect interactions may play an important role in understanding network navigability, community structure and information flow, specially when considering the range of influence of given nodes within a network, the strength of the interactions and how these interactions shape the whole correlation structure, as well as the topology and dynamics of the system. For instance, in gene regulatory networks, when it comes to the functional role of subnetworks forming either motifs or pathways, there is an important distinction between genes locally involved in the regulation of a small set of highly specific targets and some other genes that are involved in the transcriptional control of a large number of targets, often by means of a chain of indirect interactions. The first set of genes is responsible for the fine tuning processes involved in environment-specific genomic control, while the second set (known as master regulators) is related with long range, large scale control of genome expression used by the cell mechanisms of growth and proliferation (Baca-López et al. 2012).

In the case of social networks, there is also a growing interest in the role that indirect interactions may play in information and influence flow among nodes (Fowler and Christakis 2007, 2010). In some instances (such as the social epidemiology of obesity) (Fowler and Christakis 2007) it has been shown that indirect connections (i.e. second degree links) within a social network may, under some conditions, exert a greater influence than direct interactions that however shape the global structure of the network. Such structural determination may be one of the keystones to discern between diverse features of influence in social networks such as homophily, social contagion and covariation (Shalizi and Thomas 2011). We may envisage other instances in the realm of complex networks where the distinction between direct and indirect links may be of some relevance, such as ecological or economic-financial networks, etc (Beltrán et al. 2012; Callaway and Howard 2007; Tsatskis 2012).

Methods for direct and indirect interactions assessment found in the reviewed literature were designed *ad hoc* for too dense or too small networks (Nawrath et al. 2010; Yan et al. 2007), and most of them require additional information in order to estimate or tune parameters to differentiate the two kinds of interactions (Chua et al. 2008; Yan et al. 2007). Noteworthy is to say, a great deal of importance was expressed in direct and indirect interactions assessment, regardless the particular field of research (Systems Biology, Economics or Ecology). Most efforts invested in distinguishing between these two kinds of interactions among nodes were either done manually (for instance, see (Beltrán et al. 2012; Callaway and Howard 2007)) or rely on extremely specific issues of the underlying networks (Baldazzi et al. 2010; Nawrath et al. 2010), hence the relevance of the method we introduce in this paper.

We suggest a general method to discover and assess for direct and indirect interactions within the framework of information theory. The method we submit in this paper allows to reconstruct the basic structure of complex networks. Since our method rests on the comparison of Mutual Information (MI) among nodes in a triangle, it is not affected by directionality between links, directionality is detected instead, once the basic structure of the network is already in place. In what follows we will discuss some methods based

in probabilistic inference and information theory by means of which researchers may be able to infer and assess complex network structure from the probability distributions of some empirical quantitative features.

Methods

Information theoretical approaches in network inference

Information theory (IT) offers a powerful theoretical foundation that is well-fit to contribute to the development of computational methodologies intended to deal with network inference problems as applied to real data in several branches of complex systems theory (Hernández-Lemus and Rangel-Escareño 2011). IT also provides an analogy with statistical mechanics (SM), that can be useful for inferring network interactions (links) from between-node correlation measures, thus enabling to use (although in a quite non-trivial manner) the huge arsenal of tools of this science. There are, however a number of open questions in the application of IT to the probabilistic complex network inference. The applied algorithms may be able to return *intelligible* models relying on scarce *a priori* information while dealing with a potentially large number of variables. IT methods may also detect non-linear dependencies in highly noisy non-independent probability distributions. The best benchmarking options for such kind of complex network inference, for us, seems to be the use of sequential search algorithms (instead of stochastic search, typically involving the assignment of structures for large constrained datasets, since these procedures have a high computational complexity, even NP-hard" -exponentially large search-space-) and performance measures based on IT, since this makes `feature selection` fast end efficient, and it also provides an easy way to communicate results.

Information theoretical measures have been applied intensively to infer interactions in complex networks, in particular in the field of computational biology (Bansal et al. 2007; de Jong 2002; Fleuret 2004; Hernández-Lemus et al. 2009; Margolin et al. 2006; Peng et al. 2005; van Someren et al. 2002) but also in social network studies (Crowley-Riddey 2009; Dong 2011; Mislove 2009; Mislove et al. 2010; Zhao et al. 2011). A group of correlation measures including mutual information, Markov random fields and Kullback-Liebler divergences, amongst others are considered appropriate to perform probabilistic network inference (Hernández-Lemus and Rangel-Escareño 2011). However, since conditional probabilities obey the so-called *tower property*, a number of *false positives* links may appear as a consequence of indirect correlations (Hernández-Lemus and Rangel-Escareño 2011).

For instance, if node (or agent) A has a high value of conditional correlation (say, mutual information) with node B, and B is also highly correlated with node C, most common algorithms would predict (with a marginal probability p^{ind}) the presence of a (possibly non-existent) link between processes A and C. In order to correct for the presence of indirect links we may implement some methods from IT, such as bounds in the information-theoretical probability measures and the use of the Data Processing Inequality (DPI) (Sehgal et al. 2007). DPI can provide a bound to the extent on which *signal processing* may optimize probabilistic inference. We will discuss these and other ideas in the framework of network inference and structure assessment. We will also discuss some of their implications, and potential applications in the contemporary complex systems scenario.

Some of the essential notions of IT that will be used in this work are: (information-theoretical) entropy, mutual information and other related measures. To do so, let X and Y denote two discrete random variables having the following features:

- Finite alphabet $\mathcal{X}$ and $\mathcal{Y}$ respectively
- Joint probability mass distribution $p(X, Y)$
- Marginal probability mass distributions $p(X)$ and $p(Y)$
- Conditional probability mass distributions $p(X|Y)$ and $p(Y|X)$

Following Shannon (1949), it is possible to define the *information theoretical entropy H* of such distribution as follows

$$H = -K_s \sum_{\nu} p_\nu(X) \log p_\nu(X) \tag{1}$$

here H is called Shannon-Weaver's entropy, K_s is a constant, useful the determine the units in which entropy is measured (bits, nats, and so on, depending on the base of the log used) and $p_\nu(X)$ is the mass probability density for state ν of the random variable given by $X = x$. IT entropy is a measure of the amount of uncertainty associated to the value of X, hence relating the *predictability* of an outcome to the probability distribution. Let us now consider two discrete random variables (Y, X) with a Joint Probability Distribution (JPD) $p(Y, X)$. For these random variables the *joint entropy* $H(Y, X)$ is:

$$H(Y,X) \; = \; -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(y,x) \log p(y,x) \tag{2}$$

The maximal joint entropy corresponds to independence conditions of the random variables Y and X *i.e.* when the JPD is factorized $p(Y, X) = p(Y)p(X)$. The entropy of the JPD is then just the sum of their respective entropies. An inequality theorem could be stated as an upper bound for the joint entropy:

$$H(Y,X) \; \leq \; H(Y) \; + \; H(X) \tag{3}$$

Equality holds `iff` X and Y are statistically independent.

Also, given a Conditional Probability Distribution (CPD), the corresponding *conditional entropy* of Y given X is given by:

$$H(Y|X) \; = \; -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(y,x) \log p(y|x) \tag{4}$$

Conditional entropies measure the uncertainty of a random variable once another one (the conditioner) is known. It can be proved (Cover and Thomas 1991) that:

$$H(Y,X) \; = \; H(X) \; + \; H(Y|X) \leq \; H(Y) \; + \; H(X) \tag{5}$$

Or:

$$H(Y|X) \leq \; H(Y) \tag{6}$$

Again, equality holds `iff` X and Y are statistically independent. Equation 6 is useful in the inference/prediction scenario as follows: if Y is a target variable and X is a predictor, adding variables can only decrease the uncertainty on target Y. As it will be shown later, this is essential for network inference when applying IT methods. *Entropy reduction by conditioning* can be accounted if we consider a measure called the *mutual information*, $I(Y, X)$ which is a symmetrical measure (i.e. $I(Y, X) = I(X, Y)$) that is written as:

$$I(Y,X) = H(Y) - H(Y|X) \quad \text{or} \quad I(X,Y) = H(X) - H(X|Y) \tag{7}$$

If we resort to Shannon's definition of entropy (equation 1) (Shannon and Weaver 1949) and substitute it into equation 7 we get:

$$H(Y,X) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{8}$$

A comprehensive catalogue of algorithms to calculate diverse information theoretical measures (including mutual information) has been developed for [R], the statistical scientific computing environment (INFOTHEO). We will analyze the special role that MI has in the field of complex networks inference from quantitative feature data. MI has been applied successfully as a measure to infer 2-way interactions in complex networks (quite specially in the field of Gene Regulatory Networks or GRNs) (Andrecut and Kauffman 2006a; Andrecut and Kauffman 2006b; Madni and Andrecut 2007; Margolin et al. 2006). As we have seen, MI quantifies the degree of statistical dependency between two random variables (say α and β). One can see that $MI(\alpha,\beta) = 0$ `iff` α and β are statistically independent.

Hence, if we measure some quantitative-feature of interest ϑ (say expression level of genes in GRNs), by studying its profile (and more specifically the mutual correlation profile for a set of nodes) we may find interactions conforming a network. A pair of agents characterized by feature distributions ϑ_i and ϑ_j for which $MI(\vartheta_i,\vartheta_j) \neq 0$ are said to interact with each other. Since MI is *reparametrization invariant*, one usually calculates the normalized mutual information. In this case $MI(\vartheta_i,\vartheta_j) \in [0,1]$, $\forall i,j$.

Distinguishing between direct and indirect interactions

With these definitions in mind, let us consider two random variables, X and Y, whose mutual information is $MI(X,Y)$. Now consider a third random variable, Z, that is a (probabilistic) function of Y only. It can be shown that $P_{Z|XY} = P_{Z|Y}$, which in turn implies that $P_{X|YZ} = P_{X|Y}$, as follows from Bayes' theorem.

An information-theoretical theorem called the Data Processing Inequality (DPI) states that Z cannot have more information about X than Y has about X; that is $MI(X;Z) \leq MI(X;Y)$. We can see that $MI(X;Z) = H(X) - H(X|Z) \leq H(X) - H(X|Y,Z) = H(X) - H(X|Y) = MI(X;Y)$. Inequality follows because conditioning on an extra variable (in this case Y as well as Z) can only *decrease* entropy (in a similar way to what occurs in statistical physics when adding constraints to a thermal system), A formal definition of such a theorem would be:

Definition 1. *Three random variables X, Y and Z are said to form a* ***Markov chain*** *(in that order) denoted $X \to Y \to Z$, if the conditional distribution of Z depends only on Y and is independent of X. i.e. if we know Y, knowing X doesn't add anything new to what we already know about Z than if we know only Y.*

If X, Y and Z form a Markov chain, then the Joint Probability Distribution can be written as follows:

$$P(X,Y,Z) = P(X)P(Y|X)P(Z|Y) \tag{9}$$

Theorem 1. Data Processing Inequality: If X, Y and Z form a Markov chain, then

$$MI(X;Z) \leq MI(X;Y) \tag{10}$$

Proof. By the chain rule for mutual information we can state that:

$$\begin{aligned} MI(X;Y,Z) &= MI(X;Z) + MI(X;Y|Z) \\ &\quad MI(X;Y) + MI(X;Z|Y) \end{aligned}$$

By the Markov property, since X and Z are independent, given Y, $MI(X;Z|Y) = 0$, then, since $MI(X;Y,Z) \geq 0$ we have: $MI(X;Z) \leq MI(X;Y)$ c.q.d. □

In reference (Margolin et al. 2006), the application of DPI has shown that if nodes ϑ_1 and ϑ_3 interact only through a third node, ϑ_2 within a given network, $MI(\vartheta_1,\vartheta_3) \leq \min[MI(\vartheta_1,\vartheta_2); MI(\vartheta_2,\vartheta_3)]$. Hence, the least of the three MIs values may come from indirect interactions. The proposed algorithm examines each triplet vertex for which all three MIs are measured and compared to some threshold value MI_0. If there is an edge with an MI value below the threshold, then it is removed (see Figure 1). DPI is thus useful to quantify efficiently the dependencies among a large number of nodes. The DPI algorithm is useful in the problem of complex network structure assessment as well, since it eliminates those statistical dependencies that might be of an indirect nature.

In some cases, however, it may happen that the Markov chain structure is not absolutely fulfilled. Say when nodes ϑ_1 and ϑ_3 interact not only through a third node, ϑ_2, but also by means of a direct interaction. Hence ϑ_1 and ϑ_3 may be two-fold connected, in this case pruning-out one of the links may render an inaccurate version of the actual interaction pattern. This scenario can be accounted for by means of establishing a threshold for removing a link, i.e. the link with the lesser MI measure would be removed only if its

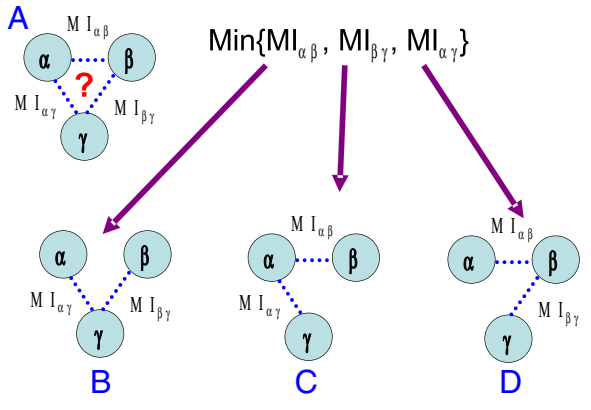

Figure 1 Absolute application of the DPI inequality to assess the direct interaction structure of a three node sub-network. Panel **A** depicts a three node subnetwork composed by nodes α, β and γ. Interactions were inferred by means of mutual information calculations for a selected quantitative feature. The interaction strength between the nodes i,j is given by the respective MI_{ij} ($i,j = \alpha,\beta,\gamma$) values. Panels **B**, **C**, and **D** result from pruning the subnetwork from the link with the lowest MI_{ij}: $MI_{\alpha,\beta}$ (B), $MI_{\beta,\gamma}$ (C), and $MI_{\alpha,\gamma}$ (D).

value is below a certain threshold (that we hereon call DPI_{tol}) -to be determined in a case-by-case basis by close examination of the network and also by considering its intended applications- and it stays in the network otherwise (Figure 2).

There are similar approaches to the one just presented, for instance the ones in reference (Liang and Wang 2008) and in reference (Zhang et al. 2012). Both approaches are based in *conditional mutual information* (i.e. the degree of information a variable X and a variables Y share, given a third variable (or group of variables) Z). These algorithms try to account for indirect links by means of conditioning the associated mutual information distributions. In reference (Zhang et al. 2012), the authors further improve the performance of the algorithm by using an adaptive computation framework. In addition to these good performance general methods (references (Liang and Wang 2008; Zhang et al. 2012) were developed for gene regulatory networks although with minimal adjustments can ba applied to any other probabilistically inferred networks), there are also more specific approaches based on somehow *Ad Hoc* considerations. We can mention, for instance the `MARINa` algorithm (Lefebvre et al. 2010; Lefebvre et al. 2007; Mani et al. 2008) developed specifically for the assessment and reconstruction of gene regulatory networks based on *statistical enrichment* of certain signatures (Subramanian et al. 1554), an approach close in philosophy of that of conditioning variables that, however requires for additional information (i.e. the signatures themselves) to be useful, hence is more restricted to its scope and applications as are approaches relying on additional phenotypic information (Wu et al. 2009; Yu et al. 2006).

Data and algorithms

In order to introduce the importance of telling direct network interactions from indirect ones, we performed a topological analysis on the Gene Network of the fruit fly, *Drosophila melanogaster* (D.m.). The fruit fly Gene Network is a paradigmatic system for genetic studies and one of the best annotated organisms in genomics databases. It also presents a high genomic similarity to that of mammals -humans included- (about 61% of disease-associated genes in humans have a D.m. counterpart) and there is open access to its high-throughput inferred biological network (Costello et al. 2009). By discussing

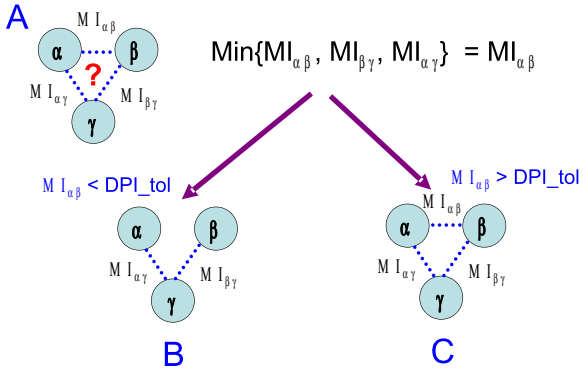

Figure 2 Relative application of the DPI inequality to assess the direct interaction structure of a three node sub-network. Panel **A** depicts a three node subnetwork composed by nodes α, β and γ. Panels **B**, and **C** result from pruning (or not) the subnetwork from the link with the lowest MI_{ij}, in this case $MI_{\alpha,\beta}$. The α-β link is removed only when its respective *MI*-value is lower than a predetermined threshold *DPI_tol* and left in the network otherwise.

some features of the network structure of this highly studied species we introduce the problem of finding direct and indirect interactions in complex networks inferred from experimental data. Once this problem has been outlined, we proceed to illustrate how the methods of information theory may be appropriate to distinguish between direct and indirect interactions in order to sketch (at least partially), the network structure on a gene regulatory network inferred from experimental data obtained from 1191 whole genome gene expression experiments in breast tissue from breast cancer patients/controls and on a social network inferred from researchers at Mexico's National Institute of Genomic Medicine coauthorship collaborations data, retrieved from the *PubMed* database.

As explained before, the methods of information theory used here correspond to the implementation of MI calculations and DPI to infer and prune respectively such networks. There is a number of different methods for computing this quantities in the literature (Hernández-Lemus and Rangel-Escareño 2011) and most of them are quite functional and almost equivalent in performance. Here we used the C++ implementation of the `aracne` algorithm (in particular we resort to `aracne 1.0` even if there is a new version `2.0` with an improved algorithmic complexity performance, because version `2.0` uses a bootstrapping method that we have found to be still a little bit unstable) (Margolin et al. 2006) for Biological Networks and `Python` scripts (some customized and others from the `NetworkX` library) for the Social Networks. The `aracne 1.0` algorithm is useful for our purposes since it is based on crystal clear MI calculations (Hernández-Lemus and Rangel-Escareño 2011), it is possible to implement DPI thresholds and its algorithmic complexity and performance are quite good (we have benchmarked `aracne 1.0` against other information-theoretical methodologies such as Information Based-Similarity (`ibs`) and linear correlation predictors in the past (Hernández-Lemus et al. 2009) with very acceptable results). `Cytoscape` and `Python`'s library `NetworkX` were used to depict and analyze the networks (Assenov et al. 2008).

Microarray pre-processing of the data was performed by using the `affy` library in `BioConductor` running under [R] on a 128 Gb RAM 8-Power5+ dual core-processor, symmetric multiprocessing (SMP) unit by IBM. All statistical tests were performed on a Dell Precision Series 16 Gb RAM QuadCore Workstation by using `limma` package in [R]/BioConductor. Information theoretical measure calculations for biological systems were performed by the `aracne` v 1.0 program in the IBM SMP machine. `Python` scripts were used instead for Social network calculations. Graphical depiction and network analyses were performed on a MacBook Pro 8 Gb i7.

The Drosophila melanogaster GRN used to highlight the presence of hierarchical structure was not further used here to demonstrate IT methods of network assessment. The reason for this is that it was inferred (Costello et al. 2009) by using Pearson correlation metric, which is a *linear* measure, thus unable to capture the whole statistical dependency spectrum. Let us recall that for two statistically independent random variables Pearson correlation coefficient is 0. However, the converse is not always true, because Pearson correlation coefficient detects only linear dependencies between two variables. Null Pearson correlation coefficients only implies statistical independence for the special case of *jointly normal* distributions. Since this is not the general case in gene expression distributions, values of linear correlations are not enough to determine statistical dependency (Hernández-Lemus et al. 2009).

Results and discussion

Network assessment in biological networks

In order to introduce the need for complex network assessment of indirect interactions, let us consider the case of the gene regulatory network of the fruit fly (*Drosophila melanogaster*, D.m.), which is one of the best curated biological networks that has been constructed ever. By means of computational integration in a probabilistic model of high throughput data sets from DNA, RNA and protein assays, combined with data mining for individual genes to phenotypes treats, such network has been assembled with the goal of providing a *...meaningful, functional gene network and to draw new and unforeseen connections...*(Costello et al. 2009). D.m. is of course a model organism for too many instances in genetics and genomics. One particular feature that is long known for the genetic structure of D.m., is the fact that gene regulation is largely determined by the action of a relatively small set of molecules that are able to exert transcriptional control in the highly active stages of development and proliferation (Baker 2001; Brennecke et al. 2007; Harrison et al. 2011; Wells 2009). For instance, aspects of regeneration are often regulated by complex mechanisms of activation for several growth regulatory pathways in damaged tissue. It has been proved that every pathway involved is being regulated by the *p53* molecule in D.m., (dP53). dP53 is thus a critical master regulator in the GRN of D.m., in particular with regards to cell growth, proliferation and differentiation. Interestingly enough, the human homolog of this very molecule is known to play a quite important role in most human cancers. dP53 is by no means the only master regulator in drosophila; an extremely important gene that acts as a universal master regulator in D.m. is the molecule called eyeless (*ey*). This gene was first study in relation to eye development (hence the name, larvae with knocked-up eye does not develop eyes). However, the role of *ey* is not restricted to eye development. Abnormal expression of the gene is able to convert other tissues into eye-cells, including legs, wings and antennae. *ey* is also involved in controlling the processes that coordinate differentiation of several cell types in a very precise way in order to develop the fly eye. The *ey*-homolog *Pax6* and homologs of other eye determination genes from D.m., are also required for tissue development in vertebrates. It has been shown that *ey* becomes a master regulator of development deep until later stages by means of *regulation by signaling* through the Notch and EGFR signaling pathways (two outstanding important generalistic signaling pathways) (Baker 2001).

Other master regulator genes are known to play special roles in D.m. Such is the case of the zinc-finger protein *Zelda* (*ZLD*) that plays a key role in transcriptional activation and as a master regulator of genome activation in the earliest stages of D.m. development (Harrison et al. 2011). The ZLD transcription factor bounds to thousands of sites across the genome at all developmental stages of D.m., with relatively small changes in binding between stages. The number and range of ZLD targets - around 2,000 genes have ZLD bound to their promoters and/or enhancers - demonstrate that it plays a major role in maternal-to-zygotic transition activation. Hence Zelda *should* be an important hub in the transcriptional regulation network for D.m. However, probably the most important group of master regulators in the fruit fly GRN is formed by the so-called *piwi* family of transcription factors. The piwi class of genes was identified long ago as encoding regulatory proteins that are responsible for the maintenance of incomplete differentiation in stem cells and also in preserving the stability of cell division rates in germ line cells, a quite important set of processes related again to all stages of development, growth and

differentiation (Brennecke et al. 2007). There are plenty of other examples (*Drosophila melanogaster* is known to have some 50 or more master regulators in its GRN (Costello et al. 2009) but we think that these may be sufficient to establish the point that within the GRN for D.m. the presence of such highly connected nodes (hubs) should be evident.

If we refer to Figure 3, in panel A we can see a rendering of the whole D.m., genome GRN (Costello et al. 2009). A clustering algorithm (Bader and Hogue 2003 was used to color-code genes (white-to-red) according to their degree of clustering (red genes are strongly clustered whereas white ones are largely isolated). The relative importance of master regulator genes -that may be measured, for instance through their associated centrality degree- is not evident from the network structure. These genes are also important for biological reasons, since they determine to a large extent the regulation patterns of the entire network. In panel B we present the same network but we color-coded individual nodes according to their degree distribution. Some important *clusters* seem to appear,

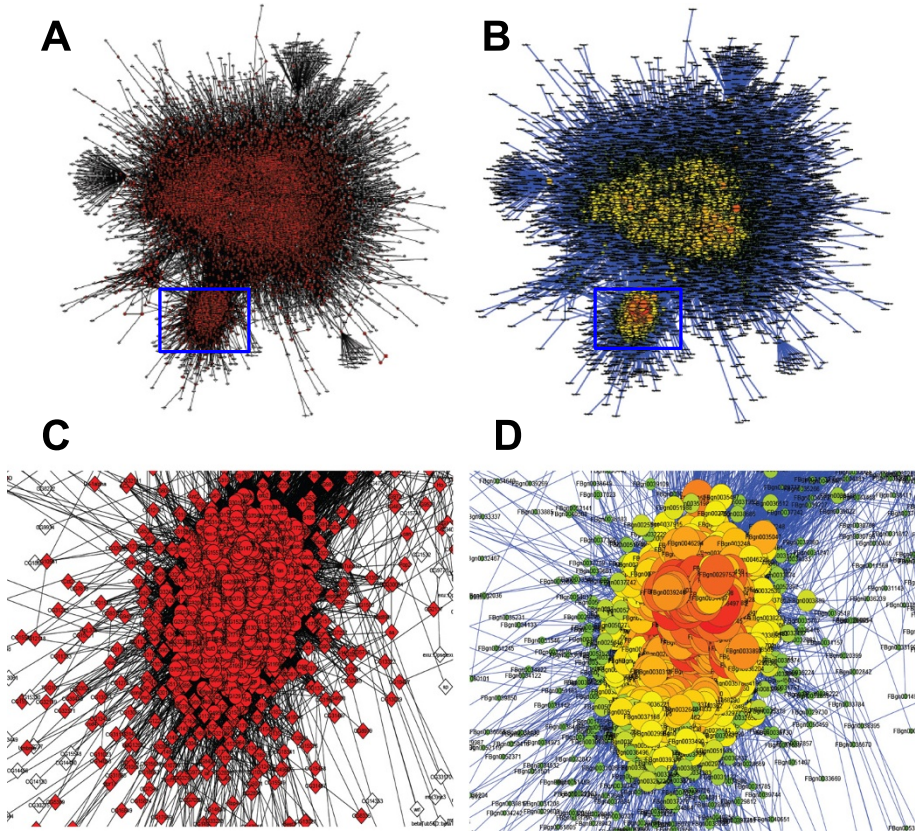

Figure 3 Largest island in the gene regulatory network of drosophila melanogaster. Gene Network constructed from top 200,000 gene pairs (Costello et al. 2009). Panel **A**: Nodes are color coded (white-to-red) according with their MCode scores (Bader and Hogue 2003) for clustering. Red nodes are genes tightly clustered. It can be seen that there is strong clustering between nearly all genes. However, no definite structural features are evident. Panel **B**: Same gene network as in panel A but with node-coloring and sizing determined dy connectivity degree: Big red nodes are highly connected genes whereas small green ones are lowly connected genes. Panels **C** and **D** present zoomed-in renderings of the boxes in panels **A** and **B** respectively. As in panel A, panel C does not permit to observe network structure. Panel **D**, in the other hand, shows the presence of some hyper-connected genes (or hubs, here in red) that are densely linked to highly connected nodes (dark orange) passing through less connected ones (light orange) until reaching barely connected genes (yellow and finally green) suggesting some kind of hierarchical structure in the network *wiring*.

however it is difficult to asses the relevance of individual nodes. In panels C and D we can see zoomed-in renderings of the rectangular regions in panels A and B respectively. As in panel A, panel C is inadequate to highlight the relative importance in the very same aspect as in panel A of individual nodes. In panel D, highly connected genes are highlighted (big red nodes), however the complex entanglement of direct and indirect interactions makes impossible to detect the actual relative importance of these, since they are surrounded by many medium to medium-high connected nodes. This situation is precisely the one calling for a method to assess for direct and indirect interactions. In what follows we will show a proposal for such method, based in the tenets of information theory as applied to both a biological network (a GRN for primary breast cancer) and (in the next subsection) a social network (a scientific collaboration network based in co-authorship probabilities).

Gene regulatory network for primary breast cancer

To show the application of IT methods to infer and assess complex biological networks, let us consider the GRN for primary breast cancer as inferred by means of MI-calculations (Margolin et al. 2006) (see Methods) in the whole genome gene expression levels (for differentially expressed genes between biopsy-captured primary breast cancer and healthy breast tissues as controls) (Baca-López et al. 2012). In Figure 4 we present the aforementioned network, where a complex entangled structure reflecting the intricate regulatory relationships driving the cancer phenotype is displayed. As in the D.m. case, in panel B we show a color- and size-coded (see Figure 4 caption) rendering based in individual degree values for every node. The relative importance of a number of genes becomes more evident, but still is not clear. It is now established that there are some (few) genes in such network acting as master regulators (Baca-López et al. 2012). This is still not evident from the topology/visualization in panel B. If we analyze the network topological structure we can see some reasons behind this. In panel C we plot the degree distribution that show an almost homogeneous behavior for two-plus orders of magnitude in the degree: from genes with a few interactions to nodes with more than a hundred connections. This scenario may be consistent with a highly structured hierarchic structure instead of the dominion of a few master regulator genes. The average clustering coefficient distribution in panel D follows a power-law-like behavior, indicating the relative importance of second neighbor and higher order interactions in the structure of the network as opposed to a neighbor dominated by a few hubs. Panel E displays the Shortest-Path length distribution that turned-out to be a short-tailed bimodal (with maxima at distances 2 and 4), indicating high network navigability.

If we now consider the MI-inference of a GRN network with the same gene expression profiling data as the one in Figure 4, but constrained by the absolute (i.e. $DPI_{tol} = 0$) application of the Data Processing Inequality (as in Figure 1), the network depicted in Figure 5 is generated. Panels A-E are the same as in Figure 4. Panel A highlights the relative importance genes as master regulators. This is even more evident in panel B where the role of *MEF2C* (big red node) and *MNDA* (big dark-orange node) (known master regulators (Baca-López et al. 2012) becomes evident, with important but bounded interrelationships with other genes. Panel C displays the degree distribution, a monotonic-decaying power-law like function highlighting the importance of a few highly connected nodes, whereas the vast majority of genes present low connectivity. Panel D shows the average clustering coefficient distribution that has obviously dropped to zero since all

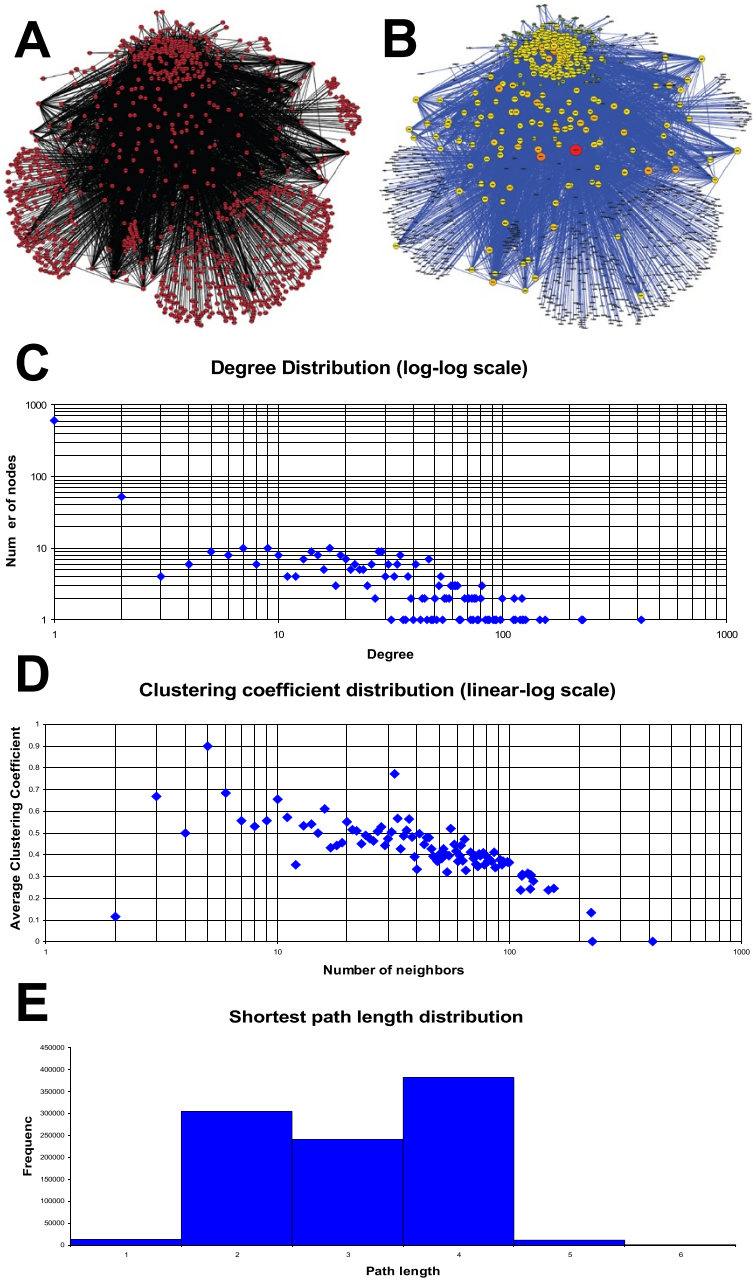

Figure 4 Gene regulatory network associated with proliferation in primary breast cancer. Gene Regulatory Network (GRN) inferred from differential gene expression profiling in 1191 whole genome expression experiments for biopsy samples from breast cancer patients/controls (Baca-López et al. 2012). Panel **A** depicts the associated GRN. The relative importance of highly connected genes is not evident. Panel **B** depicts the same network, nodes are size-coded and color-coded according with their connectivity degree (big red nodes correspond with highly connected genes, whereas small green nodes are lowly connected genes). Some genes apparently stand-out as *relevant*, however the intricate network structure does not permit to tell indirect connections from direct ones, also a number of medium-level connected nodes add complexity to the analysis. In panel **C** we can see the connectivity degree distribution: no definite trend is evident -e.g. a power law, a stretched exponential, etc.- but the distribution remains somehow homogeneous for the range between a few connections and more than a hundred connections. In panel **D** we can see how the average clustering coefficient almost follow a power-law ($R^2 = 0.85$ for the power-law fit) indicating that there may be a hierarchy related with how nodes associate. This could be an indication that a number of not-so-strong interactions are present. Panel **E** presents the short path length distribution that in this case is a bimodal (with maxima at distances of 2 and 4 links) with a short tail.

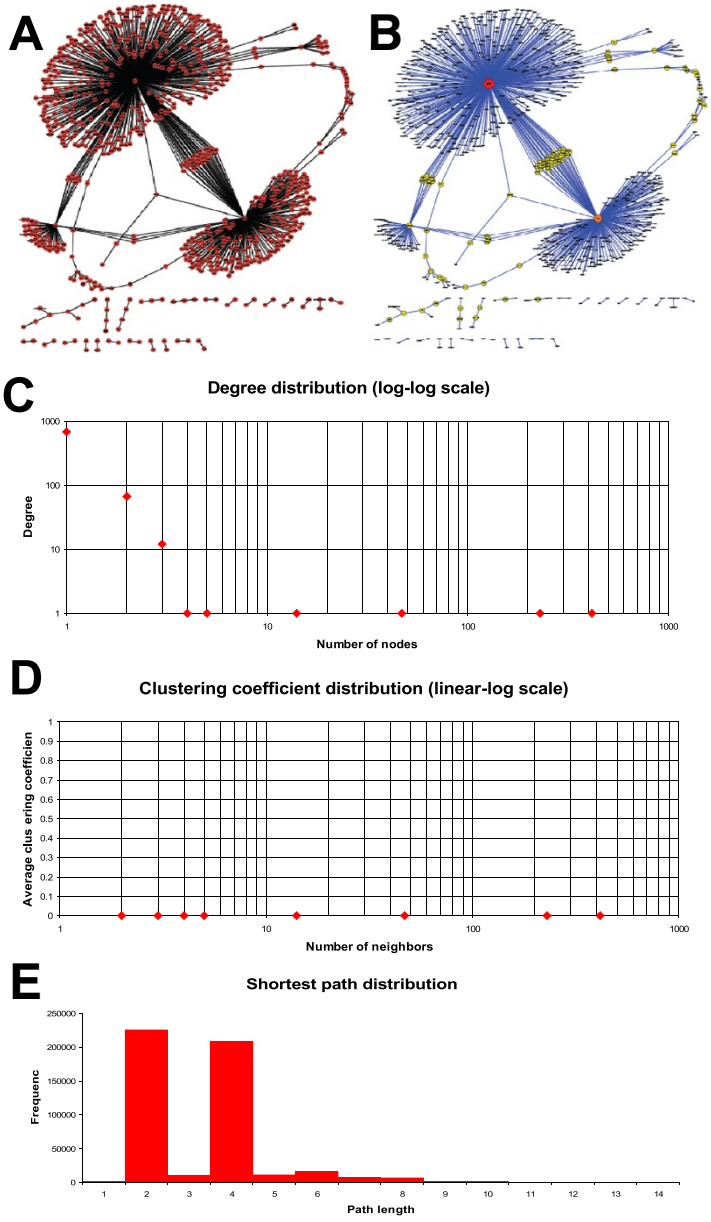

Figure 5 Gene regulatory network associated with proliferation in primary breast cancer after DPI pruning of indirect interactions. Panel **A** depicts the Breast Cancer associated GRN (Baca-López et al. 2012) (same as in Figure 4)after eliminating all indirect interactions by means of the application of the Data Processing Inequality (DPI). The relative importance of highly connected genes is now more evident. The network is basically founded on the role of two major hubs (that are also connected by means of intermediate nodes) and a couple of medium-high connected nodes. Panel **B** depicts the same network, nodes are size-coded and color-coded according with their connectivity degree. Two major hubs appear corresponding to *MEF2C* (red node) and *MNDA* (dark orange node). These two genes have been recognized as *transcriptional master regulators* in breast cancer (Baca-López et al. 2012). In panel **C** we can see the connectivity degree distribution that now resembles a power law distribution (too few nodes, however to have a reliable statistic for the fit) indicating a relatively high importance of few nodes and a low importance for most nodes. In panel **D** we can see how the average clustering coefficient has drop to zero. This is a clear effect of pruning for indirect interactions, since the relative importance of a node is now given in terms of its direct connections and not because of neighbor-by-neighbor influence. Panel **E** presents the short path length distribution that again is a bimodal (with maxima at the same distances of 2 and 4 links) that however shows a somehow larger tail than in Figure 4. This may be due to the fact that navigability in the network was easier in the presence of indirect interactions.

indirect interactions were pruned-out of the network. Interestingly enough, the Shortest-Path length distribution grossly remains a bimodal (with the same maxima at distances 2 and 4) that however exhibits a long tail, a clear indicator of diminished network navigability that results from eliminating shortcuts given by indirect interactions.

Network assessment in social networks

Scientific collaboration networks

In order to test the value of IT methods for community inference in the context of social networks, we describe a Scientific Collaboration Network (SCN from now on) based on the coauthorship history of the researchers of the National Institute of Genomic Medicine of Mexico (referred as INMEGEN). INMEGEN is one of the National Institutes of Health in Mexico and it was created in 2005, being so the second youngest institute of all 13.

The study of SCN dates back to the Erdös Number Project (Cardillo et al. 2006) and due to the accessability of data through the Internet, it has become lately a common place for those interested in social networks (Newman 2004). In this context, a very intuitive way of understanding scientific collaboration is by means of coauthorship, that is, when two scientists have worked together in one or more publications. In order to build the network, we used data of the publications reported by INMEGEN (as retrieved from *PubMed*) from 2005 to the beginning of 2012. The network includes collaborations among researchers of INMEGEN with themselves and with scientists from other institutions, as well as those collaborations between the latter as long as they have also coauthored publications with researchers from INMEGEN within the same network. The edges of the network are weighted according with their corresponding co-authorship probability (the value of every link depends on the strength of collaboration between two scientists -a MI-like function inferred from the number of collaborations with a maximum value of 11 and a minimum of 1-). This network is called *Network 1* (Figure 6). From *Network 1* we obtained a subnetwork (*Network 2* in Figure 7). *Network 2* is also a collaboration network of the researchers of INMEGEN but in which collaborators external to INMEGEN that otherwise are not connected with each other have been left out. Finally, by applying IT methods for network assessment to *Network 2*, and eliminating the weakest link between the nodes of a triangle, we obtained one last network, (*Network 3* shown in Figure 8). From *Network 3* we were able to identify four research groups and we built the collaboration network for each one of them. Groups were defined by the fact that all their members have coauthored at least two papers.

We paid special attention to general values such as number of nodes N, clustering coefficient C, centrality degree, network centralization, and characteristic pathways $\langle l \rangle$. *Network 1* has $N = 847$, a very high clustering coefficient $\langle C \rangle = 0.925$ (see Figure 6C) and a network centralization of 0.221. The characteristic path length $\langle l \rangle$ is 3.09 (Figure 6E), and the average number of neighbors is 24.8. *Network 2* has $N = 535$, a $\langle C \rangle = 0.411$ (Figure 7C), and a network centralization of 0.199. The characteristic path length $\langle l \rangle$ is 3.36 (Figure 7E), and the average number of neighbors is 4.3. *Network 3* displayed the following values, $N = 535$, $\langle C \rangle = 0.341$ (Figure 8C), and a network centralization of 0.197. The characteristic path length $\langle l \rangle$ is 3.43 (Figure 8E), and the average number of neighbors is 4.1.

Due to the shortness of the characteristic path length and its clustering coefficient, the structure of *Network 1* is close to that of a *small-world* network, which is not at all strange

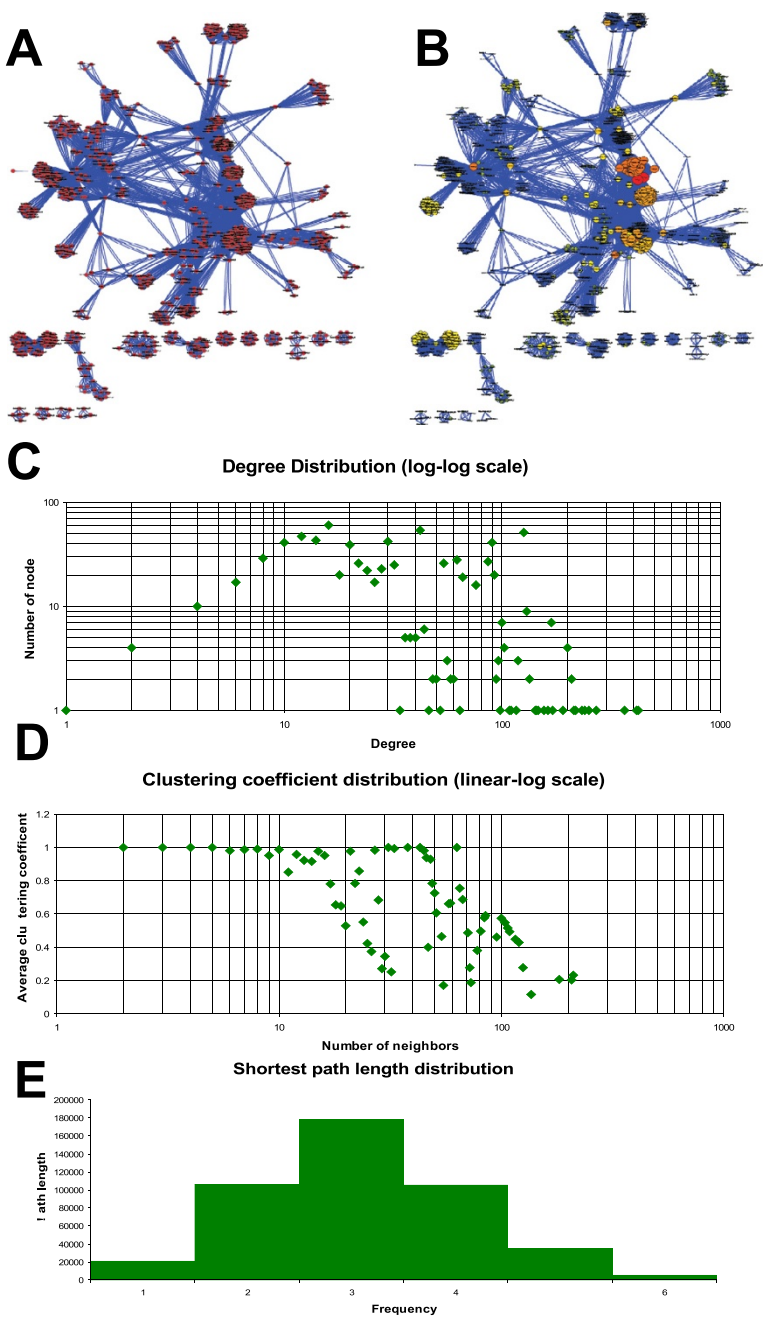

Figure 6 Scientific collaboration network of researchers at the National Institute of Genomic Medicine (INMEGEN) and their extended partners - network 1. Panel **A** presents a SCN that includes collaboration links among researchers of INMEGEN with each other and also with scientists from other institutions, also collaborations between the latter as long as they have also coauthored publications with researchers from INMEGEN within the same network. Panel **B** depicts the same network, nodes are size-coded and color-coded according with their connectivity degree. The presence of a couple of well-connected individuals (bigger red nodes) as well as a number of medium-high connected ones (orange mid-sized nodes) points out to the existence of some kind of hierarchic structure. In panel **C** we can see the connectivity degree distribution that shows a somehow anomalous behavior in the very low degree region, and then displays a typical power-law behavior. This anomaly (a very low number of barely connected nodes) may be due to incidental collaboration. In panel **D** we can see the average clustering coefficient that also presents a *left-hand* tail, most likely also due to incidental collaboration and after this a power-law like behavior. Panel **E** presents the short path length distribution, which is a quasi-symmetric unimodal with an average length of three steps.

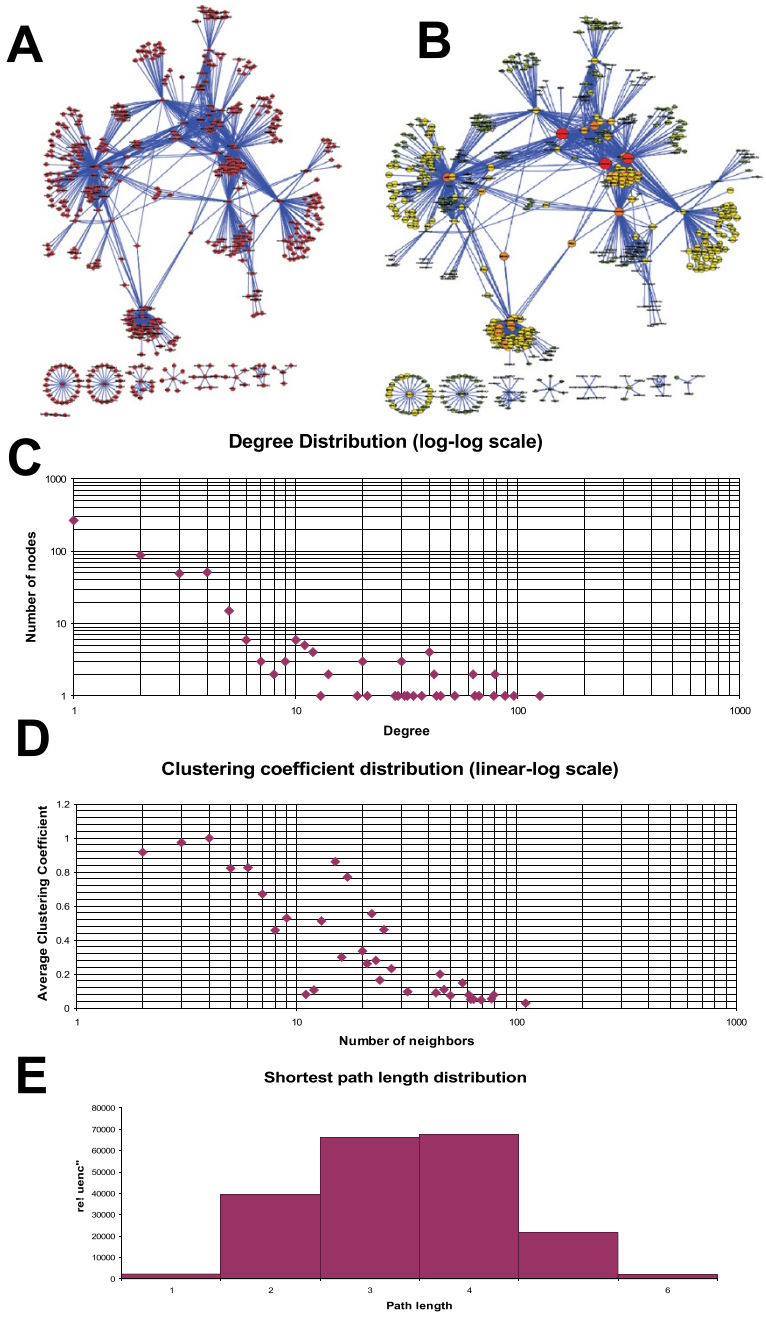

Figure 7 Scientific collaboration network of researchers at the National Institute of Genomic Medicine (INMEGEN) and their extended partners - network 1. Panel **A** presents a SCN that includes collaboration of INMEGEN researchers but in which collaborators external to INMEGEN -otherwise not connected with each other- have been left out. We can see that this network presents a topology resembling that of Network 1 but decimated in the number of links. Panel **B** depicts the same network, nodes are size-coded and color-coded according with their connectivity degree. The presence of a couple of well-connected individuals (bigger red nodes) as well as a number of medium-high connected ones (orange mid-sized nodes) points out to the existence of some kind of hierarchic structure. In this Network, the presence of localized hubs (that we may later identify as group leaders) is more evident. In panel **C** we can see the connectivity degree distribution that shows a power-law behavior with no further appearance of the incidental collaboration anomaly. In panel **D** we can see the average clustering coefficient displaying again a power-law like behavior. Panel **E** presents the short path length distribution, which is also a quasi-symmetric unimodal but with a less-defined expected value (between three and four steps) for the separation length.

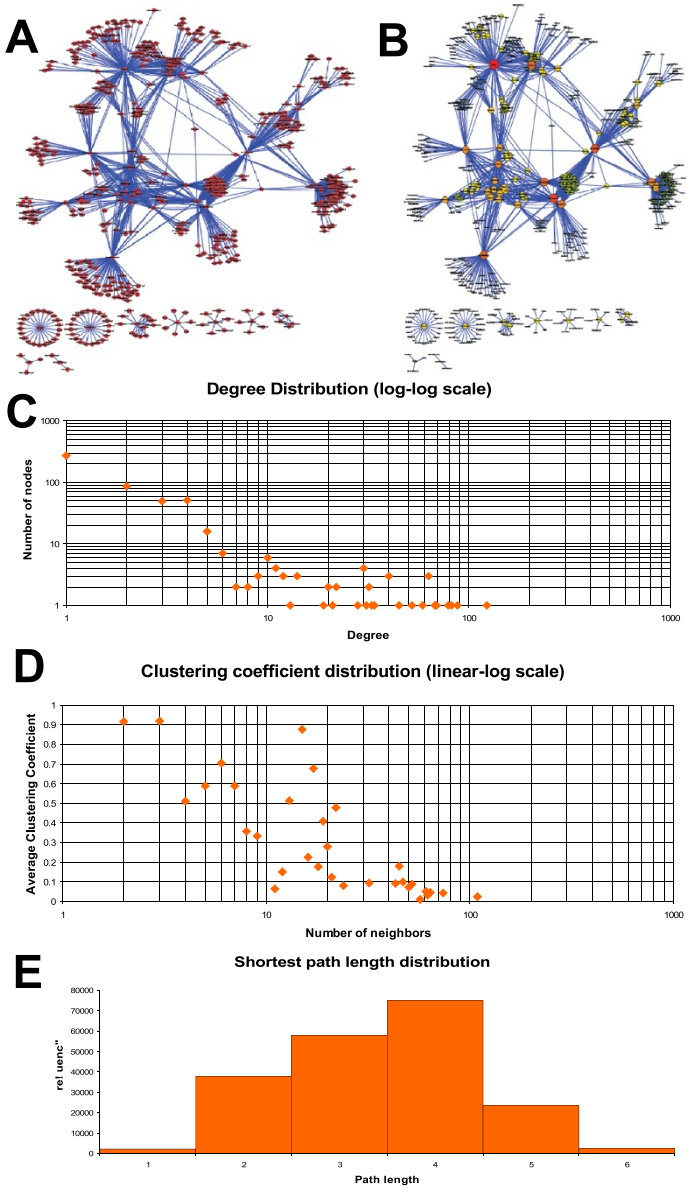

Figure 8 Scientific collaboration network of researchers at INMEGEN after DPI pruning of indirect interactions - network 3. Panel **A** presents a SCN similar to Network 2 but indirect interactions have been pruned-out, (same as in Figure 7) after eliminating all indirect interactions by means of the application of the DPI. Network 3 is similar to Network 2, i.e. the effect of eliminating *indirect* interactions has a small impact. In many triadic interactions (triangles in the network) two-out of three interactions presented the same value so that no edge was eliminated. Panel **B** depicts the same network, nodes are size and color-coded according to their connectivity degree. The presence of a couple of well-connected individuals (big red nodes) as well as a number of medium-high connected ones (orange mid-sized nodes) points out to some kind of hierarchic structure. Research groups are easily identified, yet a certain degree of collaboration between most large groups is found. In panel **C** the connectivity degree distribution shows a power-law like behavior but with a smaller high-degree tail than in Network 2. In panel **D** the average clustering coefficient displays again a power-law like behavior. The reason for finite (non-zero) clustering coefficients is that a number of triangles remained in the network for two of its edges presented equal strength of interaction. Panel **E** presents the short path length distribution, which is unimodal but with a sharp expected separation length of 4. It is noticeable that eventhough DPI-pruning did not eliminated all the triadic connections in the network, the effect of network assessment to knock-out clear indirect interactions affected network navigability (average path length went from a clear value of three in Network 1, to between three and four in Network 2 and then to a clear value of three in Network 3).

to collaboration networks and to SCN, as has been noted before (Yousefi-Nooraie et al. 2008). When compared to *Network 2*, its clustering coefficient decreases abruptly compared to its value in *Network 1*, and the characteristic path length increased slightly. Given the difference between the clustering coefficients of these two networks, it lead us to suggest that external collaboration plays a great deal in defining the small-world structure of INMEGEN's collaboration network, and most of all, it gives it a closure that makes it easily navigable. By closure we mean that INMEGEN, being such a young institution, it may still depend, to a degree, on external collaborators. Such dependency might be fueled by the fact that INMEGEN doesn't provide clinical services which means that it doesn't have the possibility to systematically recruit research subjects from where to get biological samples, depending so, on other institutions for this purpose. This idea still remains to be tested and is part of our future work.

The differences between *Network 2* and *Network 3*, are minimal. *Network 3* was inferred by IT Methods for network assessment, even though, its clustering coefficient remained almost unaffected -decreased from C= 0.411 in *Network 2*, to C= 0.341 in *Network 3*- and the characteristic path length was also basically the same. This is an important result because it means that most of the clustering was due to external collaboration as it was noted already, and for *Networks 2* and *3*, triangle formation depends on having at least two nodes from INMEGEN.

IT Methods applied to *Network 2*, eliminated what was left from the 'unbalanced' triangles and the weakest link between two nodes in a triangle was deleted, generating *Network 3*. This method made possible to identify more clearly the existence of four communities that we could recognize as INMEGEN's main research groups, with a leader (laboratory or researcher leader) easily detectable. It is important to mention that there were other, rather small groups, that have not been articulated to the main network and with a small weight value in their collaborations links. These are hypothesized to be groups in process of consolidation.

In order to analyze the structure of the four groups, we created a subnetwork for each one, based on researchers that have collaborated in at least two publications. The subnetworks made also possible to recognize different collaborative strategies among them. Group 1 and 2, with an $N = 79$ and $N = 139$ respectively, were very similar, both displayed a high value for network centralization, over 0.750 and low clustering coefficient, having Group 2 the highest $\langle C \rangle = 0.356$. For these groups, there was a leader with a very high k, and one or two very close collaborators. The difference between first and second order collaborators was large, with the highest weight (corresponding approximately to k = 9) between the leader and its first order collaborator and a lesser weight (implying k = 2) with the lowest, with an average connectivity difference of 7. The structure of these two networks could be interpreted as having a tendency towards a star-shaped structure.

In group 3, with N= 106, network centralization was not as high as in groups 1 and 2, although it remained important (Network centralization= 0.681). Noteworthy was the clustering coefficient over $\langle C \rangle = 0.500$. According to these numbers, group 3 behaves more like a group and the weight of collaborations is similar among its members.

Group 4 had the highest network centralization 0.938 as well as the highest clustering coefficient $\langle C \rangle = 0.746$. The network of group 4 depends on two members of INMEGEN on which all collaborations are fixed. The situation for group 4 is that from all, this is the extreme case of group behavior since there is a minimal connectivity degree difference

of 1 between first and second order collaborators, that is, for most nodes, the weight of collaboration among each other is the same (weight 2) and only few have a weight of 1, and none of weight 3 or higher. It is important to mention that this group is different from the other three because is the newest and the smallest ($N = 64$).

We created these networks -i.e. *Networks 1, 2, and 3-* in such a way because we wanted to have a complete image of the collaborative network of INMEGEN, and to assess INMEGEN's internal community structure. Overall, there are some emerging groups and senior researchers are still quite scarce. This may be the consequence of genomics being a new field (as compared, for instance to other biomedical disciplines), and INMEGEN is still in its infancy. If this is true, we will be able to see it in our future work when we compared INMEGEN's SCN with the networks of other National Institutes of Health, some of them with more than 50 years of history. We would also like to model the collaboration strategies followed by the different groups using agent based modeling.

Conclusions

When reconstructing the basic structure of a network, to being able to assess direct and indirect interactions among nodes can be very useful and informative. One important aspect when analyzing complex networks is to being able to distinguish and assess direct from indirect interactions (even more, with regards to local interaction levels that may shape the large scale structure and the functional features of such networks). In this paper we have shown how an application of simple theorems of information theory (Hernández-Lemus and Rangel-Escareño 2011) makes possible for researchers to grasp the nature of the links in a clear-cut way. We discuss this in the context of both biological and social networks. However, we believe that this same general method arguments may apply to any network inferred by means of mutual information measures, and to some extent to other networks inferred by other quantitative interaction measures. What is more, such a general method may serve to the general purpose of unveiling similarities and differences between networks that are as different as it is a transcriptional network from a collaborative one. It is important to recall that DPI-pruned networks may be considered along with non-DPI pruned (and even with relatives degrees of pruning as given by different values of DPI_{tol}) in order to assess for structural features of the network. This is specially relevant if one is to consider the *de-Novo* functional discovery of the role of specific individuals or the reassurance of already envisioned hypothesis along the same lines.

The material presented here is intended to show some useful features of the application of IT methods for network inference and assessment. The biological and social networks chosen here were of different nature: different in size and, more important, in structure. This may be seen by the fact that DPI-pruning of the biological network (that was much larger than the social one) resulted in a tree-like structure (with no triangles and hence with a null clustering coefficient), whereas in the social network, DPI-pruning uncover different structural properties due to the presence of undecidable cases where three edges were equally weighted. However, in both cases one important outcome of direct-vs-indirect interaction analysis resulted, i.e. the presence (or better the *visibility* of the presence) of *key players*: in one case transcription factor genes acting as master regulators and in the other researchers identified as *group leaders*. Both studies revealed some surprising elements -like the presence of novel master regulators, or some *in-consolidation*

researchers that are really emerging as *emerging group leaders*- given the structure of the direct interaction networks.

The study of the relationship between topological structure and functional organization in complex networks is, of course, still in a very early stage of development. The problems and challenges that arise include not only the determination of direct and indirect interactions and the role that some privileged nodes may play in network structure and navigability. A thorough study may also considered the issues of community structure and local connectedness. In the case of networks probabilistically inferred from experimental data one may also take into account the role that inference errors, noise and asymptotics may play. Further on, since complex networks are often dynamically adaptive systems, driven both by their inner structure and their environmental constraints, the role that dynamic evolution, fluctuations and adaptability may play will most surely also be fundamental to understand their behavior. We can only envisage what lies in the future: too many challenges should still be tackled before reaching a complete understanding of the behavior complex networks. However, we believe that information theoretical concepts and tools may play a fundamental role when facing such complexities.

Competing interests
The author declare that he have no competing interests.

Authors' contributions
EHL conceived the study, EHL and JMSG performed simulations, analyses and calculations. EHL and JMSG wrote the paper. All authors read and approved the final manuscript.

Authors' information
EHL is a theoretical physicist trained in chemical physics, probability theory and non-equilibrium statistical mechanics; his main area of research is the interface of systems biology, statistical physics and complex system's theory. JSG is an anthropolgist/ethnologist trained in theoretical biology and philosophy; his current research is in social networks; science, technology and society- studies; and genome-bioethics.

Acknowledgements
The authors gratefully acknowledge support by grants: 179431/2012 (CONACYT) and PIUTE10-92 (ICyT-DF) [Contract 281-2010], as well as federal funding from the National Institute of Genomic Medicine (México).

Author details
[1]Computational Genomics Department, National Institute of Genomic Medicine, Mexico City, México. [2]Complexity in Systems Biology, Center for Complexity Sciences, National Autonomous University of México, Mexico City, Mexico. [3]Ethics, Legal and Social Studies Department, National Institute Genomic Medicine, Mexico City, México.

Received: 12 October 2012 Accepted: 19 February 2013
Published: 8 April 2013

References

Andrecut, M, Kauffman SA: **A simple method for reverse engineering causal networks.** *J Phys Math Gen* 2006, **39:**L647–L655.

Andrecut, M, Kauffman SA: **Mean-field model of genetic regulatory networks.** *New J Phys* 2006, **8:**148.

Assenov, Y, Ramírez F, Schelhorn S, Lengauer T, Albrecht M: **Computing topological parameters of biological networks.** *Bioinformatics* 2008, **24:**282–284.

Barabási, AL: **The network takeover.** *Nat Phys* 2012, **8:**14–16.

Baca-López, K, Hidalgo-Miranda A, Mayorga M, Gutiérrez-Nájera N, Hernández-Lemus E: **The role of master regulators in the metabolic/transcriptional coupling in breast carcinomas.** *PLoS ONE* 2012, **7**(8):e42678.

Bader, GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4**(1):2.

Baker, NE: **Master regulatory genes; telling them what to do.** *Bioessays* 2001, **23**(9):763–766.

Baldazzi, V, Ropers D, Markowicz Y, Kahn D, Geiselmann J, de Jong H: **The carbon assimilation network in Escherichia coli is densely connected and largely sign-determined by directions of metabolic fluxes.** *PLoS Comput Biol* 2010, **6**(6):e1000812. doi:10.1371/journal.pcbi.1000812.

Bansal, M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles.** *Mol Syst Biol* 2007, **3:**78.

Beltrán, E, Valiente-Banuet A, Verdú M: **Trait divergence and indirect interactions allow facilitation of congeneric species.** *Ann Bot* 2012, **110:**1369–1376.

Bickel, PJ, Doksum KA: *Mathematical Statistics: Basic, Ideas and Selected Topics* Vol. 1. 2nd. New Jersey: Pearson-Prentice Hall; 2007.

Brennecke, J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ: **Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila.** *Cell* 2007, **128**(6):1089–1103.

Cardillo, A, Scellato S, Latora V: **A topological analysis of scientific coauthorship networks.** *Physica A* 2006, **372**(2):333–339.

Chua, HN, Ning K, Sung W-K, Leong HW, Wong L: **Using indirect protein-protein interactions for protein complex prediction.** *J Bioinform Comput Biol* 2008, **6**(3):435–466.

Callaway, RM, Howard TG: **Competitive networks, indirect interactions, and allelopathy: a microbial viewpoint on plant communities.** *Prog Bot* 2007, **68:**317–335.

Costello, JC, Dalkilic MM, Beason SM, Gehlhausen JR, Patwardhan R, Middha S, Eads BD, Andrews JR: **Gene networks in Drosophila melanogaster: integrating experimental data to predict gene function.** *Genome Biol* 2009, **10**(9):R97.

Cover, TM, Thomas JA: *Elements of, Information Theory*. New York: John Wiley & Sons; 1991.

Crowley-Riddey, L: **An information-theoretic approach to finding community structure in networks.** *Ph.D. Dissertation* 2009. Trinity College Dublin, School of Mathematics.

de Jong, H: **Modelling and simulation of genetic regulatory systems: a literature review.** *J Comp Biol* 2002, **9**(1):67–103.

Dong, W: **Mutual Information: inferring tie strength and proximity in bipartite social network data with non-metric associations.** 2011. M.Sc. Dissertation, University of Illinois at Urbana-Champaign, USA.

Fowler, JH, Christakis NA: **The spread of obesity in a large social network over 32 Years.** *N Engl J Med* 2007, **357**(4):370–379.

Fowler, JH, Christakis NA: **Cooperative behavior cascades in human social networks.** *Proc Natl Acad Sci USA* 2010, **107**(12):5334–5338.

Fleuret, F: **Fast binary feature selection with conditional mutual information.** *J Mach Learn Res* 2004, **5:**1531–1555.

Harrison, MM, Li X-Y, Kaplan T, Botchan MR, Eisen MB: **Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition.** *PLoS Genet* 2011, **7**(10):e1002266.

Hernández-Lemus, E, Rangel-Escareño C: **The role of information theory in gene regulatory network inference.** In *Information Theory: New Research*. Edited by Deloumeaux, P, Gorzalka JD: Mathematics Research, Developments Series, Nova Publishing; 2011:109–144.

Hernández-Lemus, E, Velázquez-Fernández D, Estrada-Gil JK, Silva-Zolezzi I, Herrera-Hernández MF, Jiménez-Sánchez G: **Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms.** *Physica A* 2009, **388:**5057–5069.

INFOTHEO: A collection of information theoretical tools based on several entropy estimators. [http://cran.r-project.org/web/packages/infotheo/index.html] Published 2012-12-10, Last accesed 2013-03-17.

Lefebvre, C, Lim WK, Basso K, Dalla-Favera R, Califano A: **A context-specific network of protein-DNA and protein-protein interactions reveals new regulatory motifs in human B cells.** *Lect Notes Bioinform* 2007, **4532:**42–56.

Lefebvre, C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A: **A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers.** *Mol Syst Biol* 2010, **6:**377. doi:10.1038/msb.2010.31.

Liang, KC, Wang X: **Gene regulatory network reconstruction using conditional mutual information.** *EURASIP J Bioinform Syst Biol* 2008:253894. doi:10.1155/2008/253894.

Madni, AM, Andrecut M: **Design and implementation of a gene network reverse engineering method based on mutual information.** *J Integr Des Process Sci* 2007, **11**(3):55–68.

Mani, KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, Califano A: **A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas.** *Mol Syst Biol* 2008, **4:**169.

Margolin, AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNe: An Algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7**(Suppl I):S7.

Mislove, AE: **Online social networks: measurement, analysis, and applications to distributed information systems.** *Ph.D. Dissertation* 2009. Rice University, Department of Computer Science.

Mislove, AE, Viswanath B, Gummadi KP, Druschel P: **You are who you know: Inferring user profiles in online social networks.** In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining WSDM'10*; 2010:251–260.

Nawrath, J, Romano MC, Thiel M, Kiss IZ, Wickramasinghe M, Timmer J, Kurths J, Schelter B: **Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales.** *Phys Rev Lett* 2010, **104:**038701.

Newman, MEJ: **The structure and function of complex networks.** *SIAM Rev* 2003, **45:**167–256.

Newman, MEJ: **Coauthorship networks and patterns of scientific collaboration.** *Proc Nat Acad Sci* 2004, **101**(Supl1):5200–5205.

Peng, H, Long F, Ding C: **Feature selection based on mutual information: criteria for max-dependency, max-relevance and min-redundancy.** *IEEE Trans Pattern Anal Mach Intell* 2005, **27**(8):1226–1238.

Sehgal, MSB, Gondal I, Dooley L, Coppel R, Mok GK: **Transcriptional gene regulatory network reconstruction through cross platform gene network fusion.** In *Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science* 2007, **4774:**274–285.

Shalizi, CR, Thomas AC: **Homophily and contagion are generically confounded in observational social network studies.** *Sociol Meth Res* 2011, **40**(2):211–239.

Shannon, CE, Weaver W: *The Mathematical, Theory of Communication*. Urbana: The University of, Illinois Press; 1949.

Subramanian, A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approachfor interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 1554, **102:**5–15550.

Tsatskis, I: **Systemic losses in banking networks: indirect interaction of nodes via asset prices.** *SSRN 2062174* 2012. [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2062174]

Tresch, A, Beissbarth T, Saltmann H, Kuner R, Poustka A, Buness A: **Discrimination of direct and indirect interactions in a network of regulatory effects.** *J Comput Biol: J Comput Mol Cell Biol* 2007, **14**(9):1217–1228.

van Someren, EP, Wessels LFA, Backer E, Reinders MTJ: **Genetic network modelling.** *Pharmacogenomics* 2002, **3**(4):507–525.

Wells, BS: **Drosophila p53: A master regulator of DNA and tissue damage repair.** 2009. Ph.D. Dissertation, Columbia University, Genetics, Department.

Wu, X, Liu Q, Jiang R: **Align human interactome with phenome to identify causative genes and networks underlying disease families.** *Bioinformatics* 2009, **25:**98–104.

Yan, KK, Maslov S, Mazo I, Yuryev A: **Prediction and verification of indirect interactions in densely interconnected regulatory networks.** *arXiv preprint arXiv:0710.0892* 2007. [http://arxiv.org/abs/0710.0892]

Yousefi-Nooraie, R, Akbari-Kamrani M, Hanneman RA, Etemadi A: **Association between co-authorship network and scientific productivity and impact indicators in academic medical research centers: a case study in Iran.** *Health Res Policy Syst* 2008, **6:**9.

Yu, H, Xia Y, Trifonov V, Gerstein M: **Design principles of molecular networks revealed by global comparisons and composite motifs.** *Genome Biol* 2006, **7:**R55.

Zhang, X, Zhao XM, He K, Lu L, Cao Y, Liu J, Hao JK, Liu ZP, Chen L: **Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information.** *Bioinformatics* 2012, **28**(1):98–104. doi:10.1093/bioinformatics/btr626.

Zhao, K, Karsai M, Bianconi G: **Entropy of dynamical social networks.** *PLoS ONE* 2011, **6**(12):e28116.

doi:10.1186/2194-3206-1-8
Cite this article as: Hernández-Lemus and Siqueiros-García: Information theoretical methods for complex network structure reconstruction. *Complex Adaptive Systems Modeling* 2013 **1**:8.