Unexpected links reflect the noise in networks

Authors: Anatoly Yambartsev[1], Michael Perlin[2], Yevgeniy Kovchegov[3], Natalia Shulzhenko[4], Karina L. Mine[5], Andrey Morgun[2].

AY & AM equally contributed to this work.

Correspondence to: andriy.morgun@oregonstate.edu

Affiliations:

(1) Institute of Mathematics and Statistics, Department of Statistics, University of Sao Paulo, SP, Brazil.

(2) College of Pharmacy, Oregon State University, Corvallis, OR, United States

(3) Department of Mathematics, College of Science, Oregon State University, Corvallis, OR, United States

(4) College of Veterinary Medicine, Oregon State University, Corvallis, OR, United States

(5) Instituto de Imunogenética - Associação Fundo de Incentivo à Pesquisa (IGEN-AFIP), São Paulo SP, Brazil

Abstract

Gene regulatory networks are commonly used for modeling biological processes and revealing underlying molecular mechanisms. The reconstruction of gene regulatory networks from observational data is a challenging task, especially, considering the large number of involved players (e.g. genes) and much fewer biological replicates available for analysis. Herein, we proposed a new statistical method of estimating the number of erroneous edges that strongly enhances the commonly used inference approaches. This method is based on special relationship between correlation and causality, and allows to identify and to remove approximately half of erroneous edges. Using the mathematical model of Bayesian networks and positive correlation inequalities we established a mathematical foundation for our method. Analyzing real biological datasets, we found a strong correlation between the results of our method and the commonly used false discovery rate (FDR) technique. Furthermore, the simulation analysis demonstrates that in large networks, our new method provides a more precise estimation of the proportion of erroneous links than FDR.

1. Unexpected correlations.

1.1. Introducing the concept of unexpected correlations

It is quite common, especially in biology, that in order to understand how system transitions from one state to another (e.g. from health to disease) scientists compare how parameters such as gene expressions, protein levels, or metabolite abundances differ between these states. One result of such a comparison is a list of parameters up- or down-regulated (that is, some numerical value attributed to the parameter has either increased or decreased) from the first state to the second. The parameters are not regulated independently from each other; rather, they make up regulatory networks each with a limited number of key drivers that govern the transition. A common approach to the reconstruction of regulatory network structure is the inference of a correlation network build from these parameters. In particular, correlation (or, for the purposes of this paper, co-variation) networks are widely used in gene expression analysis (see, for example, Butte et al., 2000, Opgen-Rhein and Strimmer, 2007, and references within). Any co-variation network inference implies that any edge in the network (corresponding to correlations between parameter/nodes) is an empirical result of either direct or indirect causal relationships unless they edge is erroneously drawn. The primary question that drove this study was thus whether the causal nature of gene expression networks has any specific implication for their structure and organization. Furthermore, in the case that this relation (causality-network structure) exists, we ask whether it can be used to improve gene network analysis.

In order to address this question we look to basic principles connecting correlation and causality. Causal effects have to follow Reichenbach's principles (Reichenbach, 1956; Pearl, 2009) which, in the example at hand, imply that if there is a correlation between two genes expressions g_1 and g_2, *provided that it is not a statistical artifact*, at least one of three must hold: 1) g_1 regulates g_2; 2) g_2 regulates g_1; or 3) there is common cause (perhaps another gene, g_3) that regulates (directly or indirectly) both g_1 and g_2 (Figure 1). Thus, in the particular situation under discussion, namely a system with two equilibrium states with two types of regulation (stimulation and inhibition) we propose a scheme in which a sign (positive or negative) of correlation coefficient is associated with direction of regulation of correlated genes. Sign association follows a simple set of rules:

- If there is a correlation between two mutually "up" or "down" regulated genes, the corresponding sign associated with the link is positive.
- If there is a correlation between an "up" regulated gene and a "down" regulated gene, the corresponding sign associated with the link is negative.

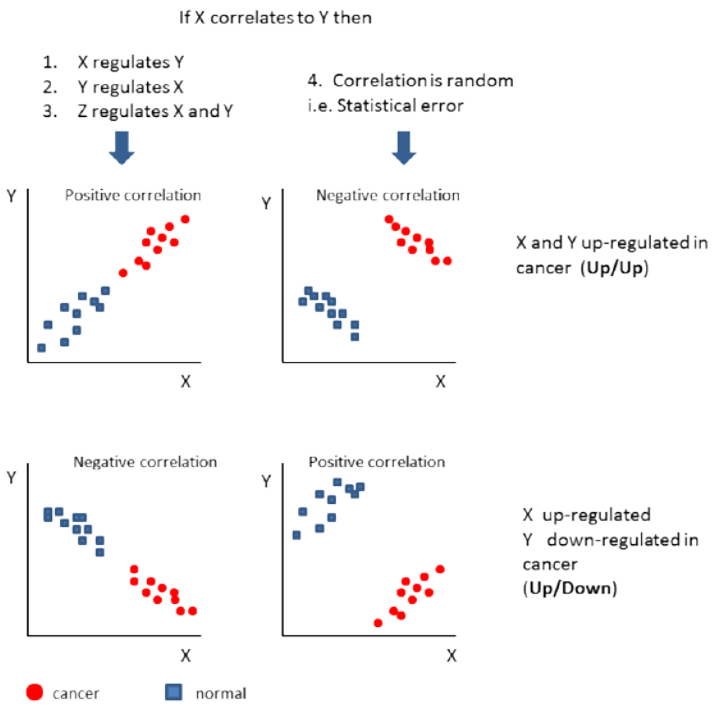


Figure 1: We hypothesize that monotonic regulation mechanisms cannot account for relationships such as the two shown on the right. Correlations exhibiting such regulation between states are thus marked unexpected and hypothesized to result from statistical error.

We hypothesize that correlations whose sign disagrees with that associated with the corresponding link are erroneous (i.e. the result of noise or statistical error rather than causal relationships). We will hereafter call such correlations unexpected, and their rough proportion we abbreviate as PUC (the Proportion of Unexpected Correlations).

The fundamental reasoning motivating this hypothesis is that regulation mechanisms in biological systems (as well as many other systems) are not generally a function of biological state. Though gene expression levels in a cancerous cell may vary from that in a healthy cell, gene function and regulation schemes in most cases remain constant. The differences in gene expression levels between two biological states should reflect the nature of their regulatory pathways. We expect positively correlated genes to mutually increase or decrease in expression, and negatively correlated ones to be regulated in opposite directions (i.e. up/down or down/up). Note that correlations are evaluated within each state independently, while differences in gene expression is evaluated between the two states. A deviation from this behavior suggests that a particular correlation between the expression levels of two genes is NOT due to a causal link.

A straightforward way to empirically test whether, as we hypothesize, unexpected correlations are erroneous is to analyze some real-world data and compare PUC, which we

believe to be a measure of error in a correlation network, to a standard measure of network error, the false discovery rate (FDR) [Benjamini&Hochberg, 1995]. For a proof-of-concept comparison, we used gene expression data from our recently published paper on network analysis in cervical cancer (Mine and Shulzhenko et al., 2013).

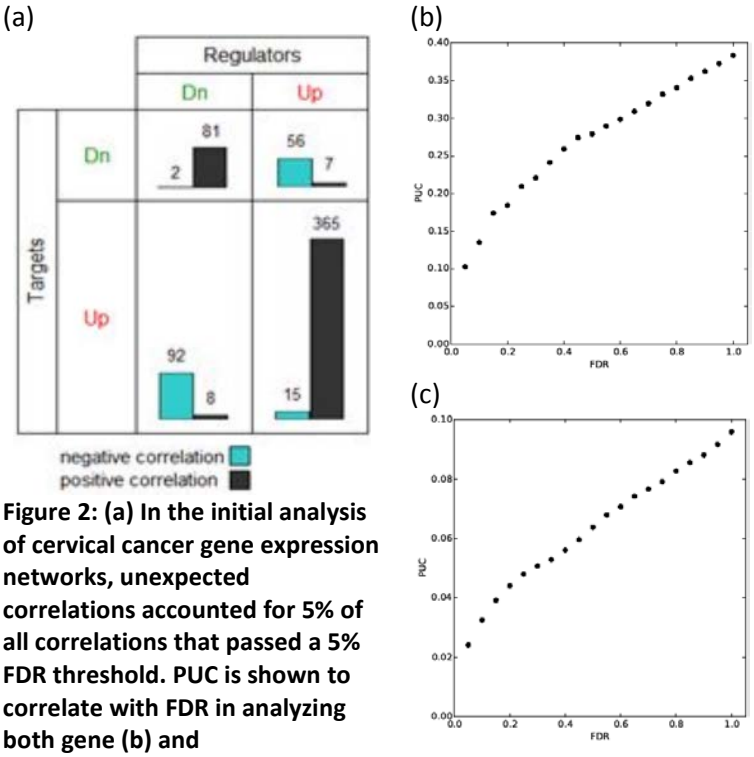


Figure 2: (a) In the initial analysis of cervical cancer gene expression networks, unexpected correlations accounted for 5% of all correlations that passed a 5% FDR threshold. PUC is shown to correlate with FDR in analyzing both gene (b) and macroeconomic (c) data.

We felt that this network should provide excellent real data to analyze our prediction, as it was constructed from a robust meta-analysis of five cancer gene expression datasets and thus validated by large, independent datasets. To our great satisfaction and some surprise, under an FDR threshold of 5% we observed an identical PUC of 5% in this gene expression network (Figure 2, see section I.1. of the supporting material and Figure S1).

The fact that we observed similar levels of unexpected correlations and of erroneous edges in the network reconstructed form cervical cancer data suggests that it can be extrapolated to the whole field of gene-gene regulation and that PUC can potentially be used as a measure of error.

Encouraged by this result, to better understand the properties of this new metric (PUC) we went further to establish a mathematical framework for its application. Indeed, although concept of PUC can be formulated and tested empirically without mathematical theory, a rigorous mathematical formalization of PUC is necessary for its establishment as a widely applicable and powerful method of analysis.

1.2. Mathematical formalism relating causation and the sign of correlation

Our hypothesis that unexpected correlations are erroneous can be rigorously proven for systems that transit between two stable states with two types of relations between parameters: stimulation and inhibition. Herein, we provide a proof of our hypothesis in

the domain of Bayesian networks (Pearl, 2009) with two equilibrium states and linear dependences between nodes (see proof for more general case in Supplementary Material, section II.2). In order to formulate our results we need to introduce some mathematical notation.

Consider some regulatory network, directed without loops (i.e. a directed acyclic graph, DAG), represented by a graph $G = (V, E)$. Any edge $e \in E$ is an oriented pair of vertices (nodes) $e = (v, w) \in V^2$. The orientation of an edge represents the direction of causality in a regulatory network (that is, an orientation (v, w) implies that v regulates w). For any node v we associate the set of its parents as $pa(v) \coloneqq \{u \in V : (u, v) \in E\}$. We define the set of grandfathers $gf(G)$ for the graph G as the set of all nodes without parents: $gf(G) \coloneqq \{v \in V : pa(v) = \emptyset\}$.

The graph G will be weighted graph. It means that every edge $e = (v, w) \in E$ has a label (weight) $c_{vw} \in \mathbb{R}$. With any node $v \in V$ we associate a random variable M_v. The distribution of random variables is given by their respective structural linear equations $M_v = \sum_{w \in pa(v)} c_{wv} M_w + \varepsilon_v$, where ε_v are mutually independent and identically distributed with mean 0 and variance σ^2.

In the previously discussed biological framework, a graph G represents the entire gene expression network. A node v represents some gene, which has an expression level M_v. An edge $e = (v, w)$ represents a causal link between two genes v and w in which the expression of w is regulated by v. The sign of w reflects the direction of regulation: negative sign and positive sign correspond to inhibition and stimulation, respectively. The parents of v are simply all genes which regulate v and the grandfathers of G are the primary regulators of the entire network, the genes at the top of the regulatory chain.

For simplicity, we consider a regulatory network with only one grandfather ($|gf(G)| = 1$), denoted by the vertex o. Let $M_o^{(P)}$ and $M_o^{(Q)}$ denote the expressions of node o in two distinct equilibrium states P and Q. For any v we denote the changes in expression between states as $\Delta_v = \mathbb{E}M_v^{(P)} - \mathbb{E}M_v^{(Q)}$, where $\mathbb{E}$ denotes the expectation value (mean) of corresponding variable.

The mathematical definition of expected and unexpected links, given heuristically in the introduction, is formally expressed in the following way:

<u>Definition</u>. An edge $e \in E$ is called an expected link between nodes $v, w \in V$ if and only if $\Delta_v \Delta_w cov\left(M_v^{(P)}, M_w^{(P)}\right) > 0$ and $\Delta_v \Delta_w cov\left(M_v^{(Q)}, M_w^{(Q)}\right) > 0$. Any edge which is not an expected link constitutes an unexpected link.

This definition states that the directions of regulation of two genes between two states should agree with the sign of the correlation between them within each state.

It is straightforward to prove the following lemma (proven in section II.1 of the supporting material):

Lemma 1. For any finite DAG with linear structural equations there exists some σ_0^2 *such that for any variance* $\sigma^2 < \sigma_0^2$ *there are no unexpected links in the graph.*

Lemma 1 implies that in regulatory networks unexpected correlations must have appeared as a result of noise within the network. Thus, the proportion of unexpected correlation thus reflects the noise level in a network.

As a side note, the linear relations between variables can be generalized by the expression $M_v = f_v\left(\{M_w\}_{w\in pa(v)};\ \varepsilon_v\right)$, where f_v is some monotonic function over its variables and ε_v is the internal network noise. If the functions f_v are not linear but monotonic, then the lemma still holds.

1.3. Unexpected correlations reflect the noise in real and simulated networks.

Mathematical models are restricted by the domain of their assumptions, which may sometimes correspond to only a fraction of real world situations, making them exceedingly limited in applicability. Thus, although we have empirically observed an appropriately small PUC at a low FDR threshold in cervical cancer data, we wanted to verify whether this correspondence would still hold in the gene regulation of an entirely different biological process.

For this we chose a more mundane physiological process than cancer: we analyzed the gene expression network perturbed as a result of colonization of intestinal tissue with normal microbiota (i.e. the mix of microorganisms that live in the gut). In these data, we again found that a low FDR threshold corresponds to a low PUC. Furthermore, PUC is highly correlated with FDR (Figure 2), which provides additional support for our prediction that PUC, similarly to FDR, quantitatively reflects network error.

An important question, however, is whether PUC brings any advantage over the standard approach to measuring the proportion of erroneous edges in a reconstructed regulation network (i.e. FDR). Real data makes such a comparison difficult because though both methods of analysis will return values for network error, there is not necessarily any obvious way to determine which is more accurate; i.e. in real data, the “correct” level of network error is not known.

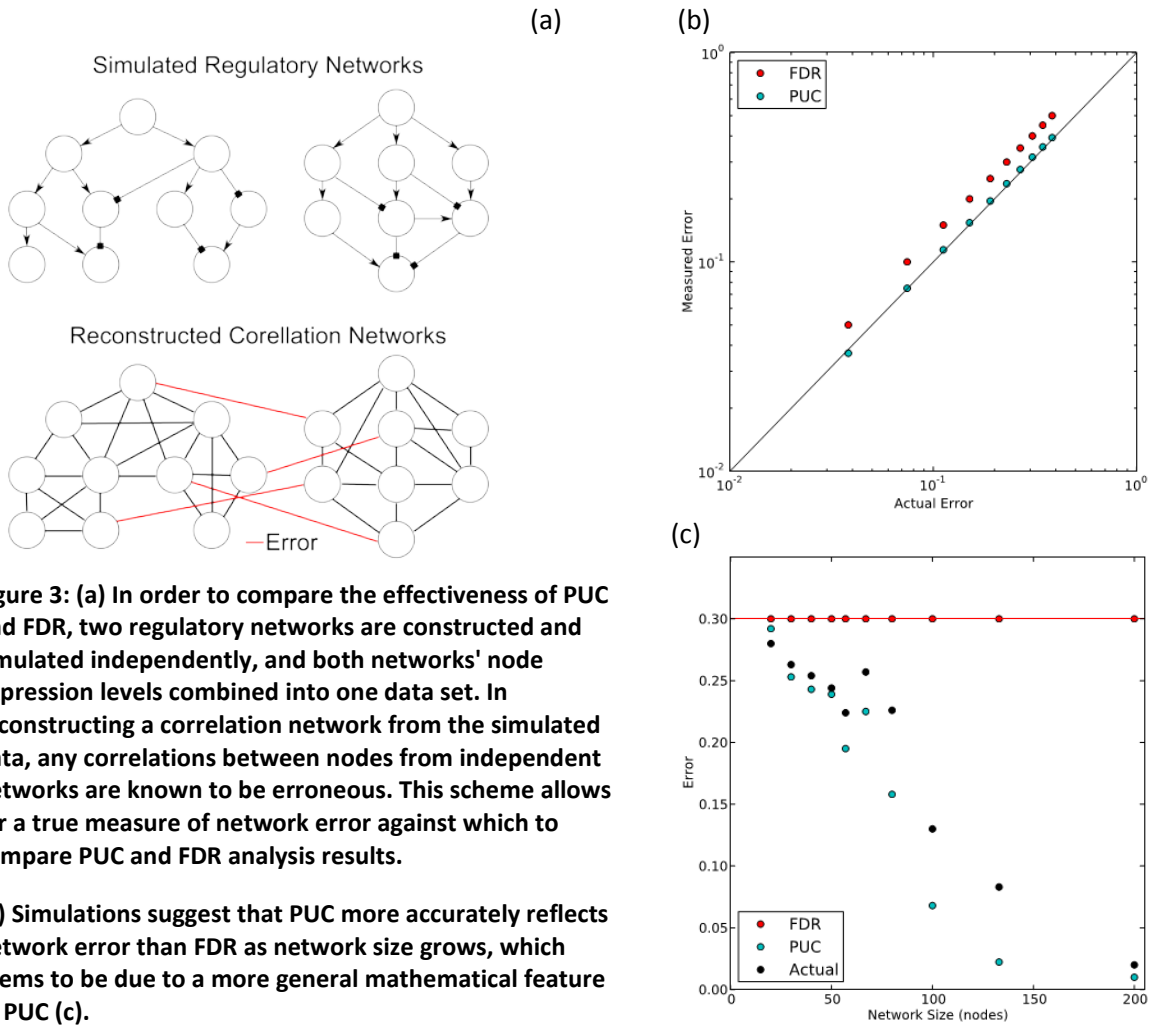


Figure 3: (a) In order to compare the effectiveness of PUC and FDR, two regulatory networks are constructed and simulated independently, and both networks' node expression levels combined into one data set. In reconstructing a correlation network from the simulated data, any correlations between nodes from independent networks are known to be erroneous. This scheme allows for a true measure of network error against which to compare PUC and FDR analysis results.

(b) Simulations suggest that PUC more accurately reflects network error than FDR as network size grows, which seems to be due to a more general mathematical feature of PUC (c).

To investigate the behavior of PUC in a "controlled environment" we simulated Bayesian networks as a model of gene regulation. We define as "true error" any correlation found between the nodes of disjoint, independent networks (Figure 3a).

In order to determine which method (FDR or PUC) better quantifies error, we look at all three measures of error (FDR, PUC, and the true error) and compare the accuracies of FDR and PUC relative to true error (Figure 3b). Simulation results demonstrate that PUC is more accurate than FDR in estimating true error.

It is known that FDR is an overly conservative approach (i.e. it overestimates the number of false positives) in cases when the hypotheses of an analysis are inter-dependent. In the case of regulatory networks, each edge constitutes a hypothesis; interdependency of regulatory network hypotheses manifests in indirect regulation between genes. Indeed, this is exactly the case with co-variation networks, in which it is possible to find numerous indirect pathways with only a few direct links. Using PUC as a measure of error, however,

does not require any assumption of hypothesis independence. PUC may thus be more applicable than FDR for reconstruction of networks with a large number of interconnected nodes. The degree of dependency between hypotheses also depends on the size and number of sub-networks that compose a network. A network made up of ten sub-networks consisting of ten nodes each should have a lower degree of hypothesis interdependency than a single network consisting of one hundred nodes lacking any well-defined sub-networks. PUC may thus similarly be more applicable than FDR for analyzing networks with a large edge density. In agreement with these presumptions, we found in simulation analyses that FDR initially provides an accurate estimate of real false positives for small networks (approximately 20-50 nodes, Figure 3c), but diverges from true error as the sizes of networks grow.

We hypothesize that PUC is expected to reflect error independently of size of the network. In order to test this prediction, we performed the same comparisons between the accuracies of FDR and PUC for networks of varying size. The results demonstrated that PUC is more accurate than FDR for larger networks, with differences in accuracy becoming negligible at network sizes of approximately 20 nodes (Figure 3c).

1.4. Noise estimation and error correction.

Another very important property of PUC is that it represents approximately half of all erroneous correlations:

$$2\mathbb{E}(PUC) \approx \mathbb{E}(total\ proportion\ of\ false\ positive\ links).$$

A formal proof of this statement is given in section III.3 of the supporting material, as well as an explanation for why it should make intuitive sense.

The identification of unexpected correlations has two primary impacts. Firstly, it provides a new method to estimate the proportion of erroneous links in a network. Secondly, it allows for the *removal* of approximately half of the erroneous edges in the network (namely, those that are unexpected), decreasing their proportion by a factor of two, thus improving the overall accuracy of the reconstructed network. The final value of network error consists of an estimated proportion of remaining false positive correlations.

The entire procedure for a correlation network is as such: first, all correlations in a differential expression list are ranked by p-value. A network is constructed with edges consisting of correlations within an arbitrary p-value threshold (e.g. 0.01). Unexpected links are identified, counted, and removed from the network. The final error in the remaining network is given by $u/(t-u)$, where u is the number of unexpected correlations and t is the total number of correlations within the p-value threshold.

1.5. PUC in a non-biological system.

The fact that we could mathematically prove the relationship between unexpected correlations and network error suggests that this principle could be widespread beyond gene interactions in various biological systems. As a proof-of-concept of PUC's generality, we turned our attention to economics. The basis for this interest was the presumption that economy, similarly to biology, is ruled by cause-effect relationships and, by extension, can be described with regulatory networks. We analyzed 1503 parameters (retrieved from World Bank economic databases) for the year 2008 in 193 countries in such areas as business, education, health, etc. Parameters with bimodal distributions (such as expenditure on primary education as a percent of GDP per capita) defined distinct states of economic networks for any given country. As expected, these networks also demonstrated a high concordance between the network errors given by PUC and FDR (Figure 2C, Figure S2). This result supports the idea that the concept of unexpected correlations can be extrapolated to a large variety of causal networks and that measurement of the proportion of unexpected correlations (PUC) can improve network analysis in many different fields of science.

Discussion

The growth of molecular biology has advanced such that we can measure the expression of thousands of genes simultaneously. Simply measuring the expression of multiple individual genes, however, is insufficient to describe a systems issue such as complex diseases. To relate gene expression to physiological states (e.g. disease) and other variables in an organism's environment we utilize gene expression networks. These networks enable more intelligent identification of molecular subtypes of diseases and molecular targets for treatment. The reconstruction of gene expression networks, however, is not easily accomplished. Constructing reliable gene expression networks with current methods requires obtaining large data sets and/or discarding sizeable portions of data to reduce false positive deductions.

Although the False Discovery Rate (FDR - Benjamini-Hochberg, see Benjamini and Hochberg, 1995) is the most popular multiple hypothesis correction procedure, its application for network inference is a conservative procedure and makes the often unfitting assumption of the independence between correlations in gene networks. There are less popular versions of FDR (for example Benjamini-Yukateli) which take into account various dependence structures between the hypotheses under consideration, but the usage of these corrections does not demonstrate any significant advantage over PUC (data not shown). Consequently, these corrections tend to have a rate of high false negative discovery (i.e. low power) and require vast sample sizes in order attain desirable degrees

of certainty about reconstructed networks. There is thus a critical need for more powerful methods of estimation of false positive connections between genes in co-expression networks.

In this study we have revealed and mathematically proved a new feature of causal networks. This feature is based on the notion that any correlation has causal and noise components. In the case that causal components prevail over noise, the sign of a correlation between two genes should be related to their up- or down- regulation of the genes between two states (Figure 1). We proposed using this relation for identifying false connections in co-variation networks, increasing network accuracy, an estimating total network error. This approach demonstrates clear advantage over the classic method (FDR) not only by providing better estimates of error in large reconstructed networks, but also by allowing the removal of approximately half of all erroneous edges. The fact that PUC demonstrates similar behavior to standard methods of analysis (i.e. PUC has a strong correlation with FDR) in both real and simulated Bayesian networks further supports the use of this adopted modeling approach. Indeed, certain questions can only be answered using a modeled system. We had to use simulated networks where we know the exact number of false links to compare FDR and PUC.

The concept of expected and unexpected correlations that we introduced is closely related to the concept of monotone causal effects and the covariance between them. The rules we proved for linear relations should therefore hold for any monotone relationships; this idea is expanded in section II.2. of the supporting material, and the framework of PUC extended to a broader class of networks than those mentioned thus far.

We must also address how non-monotonicity affects the notion and application of unexpected correlations. The concept of non-monotonicity can be exemplified for our problem as different types of relationships in two network states, such as a negative correlation between parameters in one biological state and a positive correlation in another. In such cases, despite violation of monotonicity, we expect unexpected correlations to arise primarily due to noise, rather than the change in relationships. Nonetheless, we demonstrated (see section II.4. of the supporting material) that there is no evidence for non-monotonicity to suggest that these exceptionally rare non-erroneous correlations are in fact responsible for the observed changes in gene expression between states of a biological system. Therefore, because the ultimate goal of network inference is actually to model and understand the transition of biological system from one state to another, we can safely remove these unexpected correlations from the reconstructed network for independent reasons (i.e. that they do not have causal contribution to system state transition).

We believe that this work introduces an entirely new way of dealing with error in regulatory network reconstruction. Indeed, statistical methods employed for such problems normally estimate an error, but cannot detect erroneous edges. We propose a method that besides (according to simulations, potentially superior) error estimation allows for identification and removal of approximately half of total network error. Thus, the identification and removal of unexpected correlations decreases the proportion of irrelevant and erroneous connections and strongly increases the power of network inferences.

Finally, our study provides a good example of the success of a systems approach. The collaboration between biologists and mathematicians resulted in the integration of fundamental principles of causality with real world findings (e.g. Figure 2a cervical cancer) to provide the scientific community with a powerful technique that improves the traditional task of network inference from observational data.

Acknowledgements: We thank Eric Zubriski, Chris Sullivan, and Xiaoxi Dong from Oregon State University for help in setting-up computation infrastructure at CGRB (Center for Genomic Research and Biocomputing).

Supporting Material:

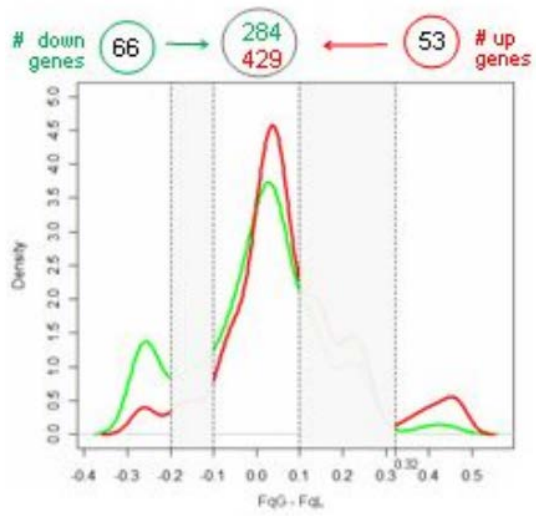
I. Experimental procedures

I.1. Statistically significant correlations between differentially expressed genes (DEGs) show expected signs

In our recent study (Nature Commun. 2013;4:1806) we have shown that key drivers of cervical carcinogenesis are located in regions of frequent chromosomal aberrations and that these genes cause most of the alteration in gene expression in cervical cancer. Therefore, in order to evaluate whether statistically significant correlations between DEGs which result from known causal relations follow our prediction we performed the following analysis:

First, we selected two groups of genes from DEGs discovered in our previous study: 1) genes in which it has been determined that chromosomal aberrations are responsible for the change in regulation; and 2) genes located in regions in which aberrations are rare, defined by FqG – FqL between -0.1 and 0.1 (Figure S1). Next, we analyzed gene co-expression in tumors samples in order to find correlations between those two groups of DEGs. We found 626 correlated gene-gene pairs with FDR 5%. The results provided support to our hypothesis that significant correlations should to have "expected" signs. Indeed, 95% (594 of 626 total pairs) of significant correlations had expected signs.

(a)



(b)

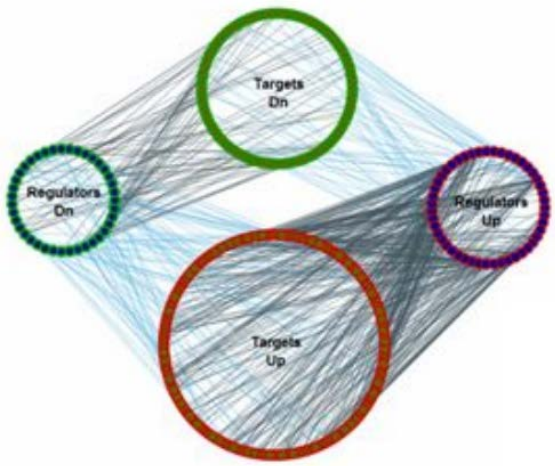


Figure S1: Genes directly regulated by chromosomal aberrations can also in turn regulate genes located outside of the aberrations. (a) Genes regulated by chromosomal aberrations in the expected direction (located in the regions $FqG - FqL < -0.2$ or $FqG - FqL > 0.3$) were considered as potential regulators, and genes located within the regions of very rare aberrations ($|FqG - FqL| \leq 0.1$) were considered to be potential targets. The green (red) line represents up-regulated (down-regulated) genes. (b) The reconstructed regulatory network with correlations in agreement with gene expression. The two green (red/purple) circles are made of up down-regulated (up-regulated) nodes, the middle (side) circles are made up of targets (regulators), and the black (cyan) lines represent positive (negative) correlations.

II. Theoretical basis.

Here we provide some formal definitions of concepts used in the paper and all necessary proofs. This section consists of four parts: 1) we introduce the mathematical machinery for PUC using Bayesian networks; 2) we generalize the previous formalism to handle a broader set of cases; 3) we demonstrate that PUC reflects half of total network error; and 4) we address concerns with network non-monotonicity.

II.1. PUC on Bayesian networks.

In order to apply the new concept of noise estimator we use Bayesian Networks as a convenient model for gene expression. Let $G = (V, E)$ be some network, which is directed acyclic graph (DAG). Any edge $e \in E$ is an oriented pair of vertices $e = (v, w)$: and direction of edge is from the first vertex v to the second vertex w. We assume that the graph is weighted graph – any edge $e = (v, w)$ has its labels (weight), c_{vw} , which is some real number $c_{vw} \in \mathbb{R}$. For any node v we associate the set of parents of the node v:

$$pa(v) \coloneqq \{w \in V : (w, v) \in E\} \quad (1)$$

We define the set of grandfathers for the graph G:

$$gf(G) \coloneqq \{v \in V : pa(v) = \emptyset\} \quad (2)$$

With any node (gene) $v \in V$ we associate the random variable (gene expression) M_v. The random variables satisfy the following linear relations (structure equations): for any $v \notin gf(G)$

$$M_v = \sum_{w \in pa(v)} c_{wv} M_w + \varepsilon_v, \quad (3)$$

where ε_v are i.i.d. random variable (intrinsic noise) with mean 0 and variance σ^2. Moreover, for simplicity we suppose that there exist only one grandfather $|gf(G)| = 1$ and let us denote it as a vertex o.

A path $\pi(v, w)$ of length n from a vertex v to a vertex w is a sequence of edges $e_i = (v_i, v_{i+1}), i = 1, \dots, n-1$, with $v_0 = v$ and $v_n = w$. The weight of the path $W(\pi(v, w))$ is the product of weights of edges from this path:

$$W\big(\pi(v, w)\big) \coloneqq \prod_i c_{v_i, v_{i+1}} \quad (4)$$

Let $\Pi(v, w)$ be the set of all paths connecting nodes v and w. And let

$$W(v, w) \coloneqq \sum_{\pi \in \Pi(v,w)} W\big(\pi(v, w)\big) \quad (5)$$

The graph coupled with expressions we consider as a model of regulatory signaling paths system. The distribution of expressions within the system is determined by the topology of graph, weights and the distribution of expressions of grandfathers. Fixed graph and weights the state (joint distribution of variables $M_v, v \in V$) will be defined by the distribution of grandfather.

For example, let o be the grandfather vertex and $M_o^{(P)}$ and $M_o^{(Q)}$ its expressions in these two different states. Same indexes we add for expression for any node in two different states. Denote d^2, d the variance and standard deviation for grandfather expression in two states, suppose that they do not depend on the state: $d^2 := \mathbb{V}\mathrm{ar}\left(M_o^{(P)}\right) = \mathbb{V}\mathrm{ar}\left(M_o^{(Q)}\right)$. Denote the mean changes in expression of grandfather´s gene as $\Delta_o = \mathbb{E}M_o^{(P)} - \mathbb{E}M_o^{(Q)}$. Expression for any non-grandfather vertex v can be expressed by formula: for any state $S \in \{P, Q\}$

$$M_v^{(S)} = X_o^{(S)} W(o,v) + \sum_{w \in V \setminus o} \varepsilon_w^{(S)} W(w,v) \qquad (6)$$

The mean change in expression of gene $v \in V \setminus o$ is

$$\Delta_v := \mathbb{E}M_v^{(P)} - \mathbb{E}M_v^{(Q)} = \Delta_o W(o,v). \qquad (7)$$

Moreover, for any $S \in \{P, Q\}$

$$cov\left(M_v^{(S)}, M_w^{(S)}\right) = d^2 W(o,v) W(o,w) + \sigma^2 \sum_{v' \in V \setminus o} W(v',v) W(v',w) \qquad (8)$$

*Definition. We say that a pair of genes $v, w \in V$ satisfy **expected correlation inequality** if and only if*

$$\Delta_v \Delta_w \, cov\left(M_v^{(P)}, M_w^{(P)}\right) \geq 0, \;\; \Delta_v \Delta_w \, cov\left(M_v^{(Q)}, M_w^{(Q)}\right) \geq 0 \qquad (9)$$

*If (9) holds then we say that the two gene expressions $M_v^{(P)}, M_w^{(P)}$ or $M_v^{(Q)}, M_w^{(Q)}$ have **expected correlations**. If one or both expected correlations inequalities are not satisfied, we say that $M_v^{(P)}, M_w^{(P)}$ or $M_v^{(Q)}, M_w^{(Q)}$ have **unexpected correlations**.*

Note that the considered model, by (8) the covariations in (9) do not depend on a state: $cov\left(M_v^{(P)}, M_w^{(P)}\right) = cov\left(M_v^{(Q)}, M_w^{(Q)}\right)$. It means that in definition we can use only covariation in one state. In this case the following statement takes place.

Lemma 1. For any finite DAG network with linear relations between variables there exists some σ_0^2 such that for any $\sigma^2 < \sigma_0^2$ there are no unexpected correlations into the network.

Proof. Direct from formulas (7), (8). By definition (9) and by representations (7), (8) we have

$$\Delta_v \Delta_w \, cov\left(M_v^{(P)}, M_w^{(P)}\right)$$
$$= \Delta_o^2 W(o,v)W(o,w)\left(d^2 W(o,v)W(o,w) + \sigma^2 \sum_{v' \neq o} W(v',v)W(v',w)\right)$$
$$= \Delta_o^2 d^2 W^2(o,v)W^2(o,w) + \Delta_o^2 \sigma^2 W(o,v)W(o,w) \sum_{v' \neq o} W(v',v)W(v',w)$$

The second sum can be made as less as possible because of σ^2. It proves the Lemma.

The formula (8) shows that any link/correlation between two nodes in a network can be represented as a sum of two parts: *causal propagation* from causal node and *noise propagation* part:

$$Cov\left(M_v^{(S)}, M_w^{(S)}\right) = \underbrace{d^2 W(o,v)W(o,w)}_{causal\ propagation} + \underbrace{\sigma^2 \sum_{v' \neq o} W(v',v)W(v',w)}_{noise\ propagation} \qquad (10)$$

Here, it is easy to see that if the grandfather variance d^2 increase, then the causal propagation will determine the sign of the covariance after some threshold. It means that it determines a link to be expected or unexpected.

Moreover, Lemma says that if we observe in such regulation networks (DAGs with linear relationships between variables) unexpected correlations, it means that they appeared as a result of noise propagation within the network. Thus the proportion of unexpected correlation reflects the noise level on a network.

Note 1. The concept of expected correlations was also observed in VanderWeele and Robins, 2010, as a rule governing the relationship between monotonic links and the sign of covariance between variables.

Note 2. The linear relations between variables can be generalized: the expression $X_v = f_v\left(\{X_{v'}\}_{v' \in pa(v)};\ \varepsilon_v\right)$*, where* f_v *is a monotone function, and* ε_v *is internal network noise. If structural functions are monotonic function, then the lemma holds also.*

Estimation of noise. The error estimation based on the following. If two genes belong to two unrelated subnetworks (see Figure 3a), then the correlation between their respective expression levels has to be equal to 0. However, observable correlation can be significantly different from 0 due to noise, in which case, the observable correlation is positive (or negative) in close to 50% of the cases (see formula (20)). Then, on average,

half of all random correlations between any pair of genes from unrelated subnetworks can be classified as unexpected, as in (9). Thus $2 \cdot PUC$ can be utilized as an error estimator.

Moreover, it is possible to prove for tree like graphs that within one network the noise propagation (see the formula (11)) has the same property as stated in formula (20). Indeed, the representation (6) means that any variable $M_v^{(S)}$ can be decomposed into the causal component $X_o^{(S)} W(o,v)$ and the noise component $\xi_v^{(S)} := \sum_{w \in V \setminus o} \varepsilon_w^{(S)} W(w,v)$. Then the covariance between $\xi_v^{(S)}$ and $\xi_w^{(S)}$ can be calculated exactly (compare with formula (10))

$$cov\left(\xi_v^{(S)}, \xi_w^{(S)}\right) = \sigma^2 \sum_{u \in V} W(u,v) W(u,w). \qquad (11)$$

If c_{vw} are mutually independent, identically distributed, with positive probabilities for being positive or negative, then the covariance (11) for any $S \in \{P, Q\}$ will be negative approximately in half of cases.

II.2. Definitions and generalization.

Here we study the concept of unexpected links in a more general framework. The positive and negative correlation inequalities are an active research direction in the field of probability and statistical mechanics. We believe these inequalities will allow us to generalize the concept of unexpected correlations in the PUC method. The following framework connects FKG (Fortuin–Kasteleyn–Ginibre) inequality in Statistical Mechanics to the concept of expected and unexpected links.

Let Ω be the underlying sample space of a biological system, as an example of a biological system we consider a gene regulatory network, and Ω can be considered as a set of all possible gene expression configurations. We can suppose that the state space Ω has an ordering (or partial ordering) "$\prec$" assigned to pairs of its elements. Here, if $\omega, \omega', \omega'' \in \Omega$, and if $\omega \prec \omega'$ and $\omega' \prec \omega''$, then $\omega \prec \omega''$.

In statistics and in statistical mechanical models the notion of an increasing random variable is remarkable.

Definition. *A random variable $X = X(\omega)$ is said to be increasing if $\omega \prec \omega'$ implies $X(\omega) \leq X(\omega')$. Similarly, a random variable is decreasing if $\omega \prec \omega'$ implies $X(\omega) \geq X(\omega')$. Both types of random variables, increasing and decreasing, are said to be monotone random variables.*

In the field of statistical mechanics and probabilistic combinatory, the FKG inequality (Fortuin–Kasteleyn–Ginibre inequality) explains most of the results involving monotone random variables and monotone (increasing or decreasing) events. It states that for two increasing random variables X and Y,

$$\mathbb{E}(XY) \geq \mathbb{E}(X)\mathbb{E}(Y) \qquad (12)$$

In some applications, such as percolation models, partial ordering of Ω is sufficient for the FKG to hold. See reference [1]. Many important results in applied mathematics and physics, such as the exact value of critical probability in two-dimensional percolation models, would have been impossible without the FKG inequality.

Let $G = (V, E)$ be a graph (network) with vertices (nodes) V and edges E. Nodes $v \in V$ represent the genes. Let $X_v(\omega)$ be monotone functions (random variables) assigned to each node $v \in V$. Here X_v represents the noiseless gene expressions. In this framework it is convenient represent the state system as a probability measure. Consider two probability measures P and Q over Ω such that

$$P(\sigma \in \Omega: \sigma \prec \omega) \geq Q(\sigma \in \Omega: \sigma \prec \omega) \qquad (13)$$

$f$or all $\omega \in \Omega$. Here P and Q correspond to the two states of a biological system. Let us denote, as before, $\Delta_v := \mathbb{E}_P[X_v] - \mathbb{E}_Q[X_v]$. Here the variables not have anymore the indices for variables but for expectations with respect to the corresponding measures. We repeat the definition of expected and unexpected links.

Definition. We say that random variables X_v and X_u modeling gene expressions in a pair of genes satisfy ***expected correlation inequality*** *if and only if*

$$\Delta_v \, \Delta_u \, cov_P(X_v, X_u) \geq 0, \;\; \Delta_v \, \Delta_u \, cov_Q(X_v, X_u) \geq 0, \qquad (14)$$

in which case we say that the two gene expressions X_v and X_u have ***expected correlations****. If one or both expected correlations inequalities are not satisfied, we say that X_v and X_u have* ***unexpected correlations****.*

Lemma 2. If X_v and X_u are monotone functions, and probability measures P and Q satisfy the condition (13), then X_v and X_u satisfy expected correlation inequality (or X_v and X_u have expected correlations).

Proof. Indeed, if X_v is increasing (decreasing) variable, then $\Delta_v \leq 0$ ($\Delta_v \geq 0$). Now, if both X_u and X_v are either increasing or decreasing the FKG inequality (12) implies non-negative correlations, so that for any state $S \in \{P, Q\}$

$$cov_S(X_u, X_v) := \mathbb{E}_S[X_u X_v] - \mathbb{E}_S[X_u]\mathbb{E}_S[X_v] \geq 0, \ \forall\, u, v \in V, \quad (15)$$

which implies expected correlation inequalities (14).

Similarly, if one of the two variables (i.e. X_u or X_v) is increasing while the other is decreasing, the FKG inequality (12) implies non-positive correlations, such that for any state $S \in \{P, Q\}$,

$$cov_S(X_u, X_v) := \mathbb{E}_S[X_u X_v] - \mathbb{E}_S[X_u]\mathbb{E}_S[X_v] \leq 0, \ \forall\, u, v \in V \quad (16)$$

implying (14) hold once again. It proves the Lemma 2. □

Next, let ξ_v denote the errors for each node $v \in V$. We assume that the random variables $\xi_v, v \in V$ are functions over a probability space Ξ, independent from any probability measure over Ω, such as P and Q. Let μ be the joint distribution of $\xi_v, v \in V$ and $\mathbb{E}_\mu[\xi_v] = 0$ for any $v \in V$. The measured gene expression we quantify as a random variable

$$M_v = X_v + \xi_v, \quad v \in V, \quad (17)$$

over the product space $\Omega \times \Xi$, and the two different states of a biological system correspond to two different probability product measures, $P \times \mu$ and $Q \times \mu$. Note that for any gene v:

$$\mathbb{E}_{P\times\mu}[M_v] - \mathbb{E}_{Q\times\mu}[M_v] = \mathbb{E}_P[X_v] - \mathbb{E}_Q[X_v] =: \Delta_v \quad (18)$$

The following Lemma is an analogous of the Lemma 1 for the general framework.

Lemma 3. If the variances of errors $\sigma_v^2 = Var(\xi_v)$ are small enough for all $v \in V$, then the pairs of measured gene expression M_v will also satisfy the inequalities (14). Thus in the noiseless networks we foresee no unexpected correlations.

Proof. The proof is direct consequence of the covariance calculation.

$$cov_{S\times\mu}(M_u, M_v) = cov_{S\times\mu}(X_u + \xi_u, X_v + \xi_v) = cov_S(X_u, X_v) + cov_\mu(\xi_u, \xi_v)\,. \quad (19)$$

By Cauchy-Schwarz inequality

$$|cov_S(\xi_u, \xi_v)| \leq \sigma_u \sigma_v$$

the second covariance in (19) can be made so small that the sign of $cov_S(M_u, M_v)$ and the sign of $cov_S(X_u, X_v)$ will coincide. This proves Lemma. □

However in the noisy networks, the expected correlations rule (14) can be violated. Here the fraction of edges (u, v) violating (14) that we call the Proportion of the Unexpected Correlations (PUC) becomes an estimator of the frequency of false edges.

II.3. PUC represents 50% of erroneous.

For any $u, v \in V$; $S \in \{P, Q\}$; and $\mu \in \Xi$, let us assume that the variables ξ_v are random such that, asymptotically, $cov_\mu(\xi_u, \xi_v)$ is positive for half of the $\binom{|V|}{2}$ edges (u, v), and negative for the rest of the pairs:

$$\lim_{|V|\to\infty} \frac{\#\{(u,v):\, cov_\mu(\xi_u,\xi_v)>0\}}{\binom{|V|}{2}} = \lim_{|V|\to\infty} \frac{\#\{(u,v):\, cov_\mu(\xi_u,\xi_v)<0\}}{\binom{|V|}{2}} = \frac{1}{2}. \quad (20)$$

If the covariance $cov_{S\times\mu}(M_u, M_v)$ is of a different sign than $cov_S(X_u, X_v)$ (i.e. if a particular correlation (u, v) is unexpected), it must hold that (see (20)):

$$\frac{cov_{S\times\mu}(M_u,M_v)\; cov_S(X_u,X_v)}{\left(cov_\mu(\xi_u,\xi_v)\right)^2} = \left(\frac{cov_S(X_u,X_v)}{cov_\mu(\xi_u,\xi_v)}\right)^2 + \frac{cov_S(X_u,X_v)}{cov_\mu(\xi_u,\xi_v)} < 0. \quad (21)$$

This condition is of the form $R^2 + R < 0$, where $R = \frac{cov_S(X_u,X_v)}{cov_\mu(\xi_u,\xi_v)}$, which trivially has the solution:

$$\frac{1}{R} = \frac{cov_\mu(\xi_u,\xi_v)}{cov_S(X_u,X_v)} < -1. \quad (22)$$

The resulting inequality is satisfied under two conditions, which are thus requisite for a correlation to be unexpected, namely:

$$|cov_\mu(\xi_u, \xi_v)| > |cov_S(X_u, X_v)| \quad (23)$$

$$cov_\mu(\xi_u, \xi_v) cov_S(X_u, X_v) < 0 \quad (24)$$

The first condition (23) is interpreted as a drowning out of the causal link between two nodes by error; that is, the magnitude of error in the correlation between two nodes' expressions is greater than the magnitude of real correlation between them. The second condition (24) is interpreted as a counteracting of error to causal connections: the contribution to the empirical correlation between two nodes due to error must counteract the contribution due to causal mechanisms.

Condition (24) implies that, given a condition (20) for error distribution, PUC will statistically detect 50% of total false correlations for which the causal contribution is

negligibly small, as the signs of the error and causal contribution are equally likely to be the same as they are to be opposite.

II.4. Unexpected correlations under non-monotonicity.

Here we prove the proposition in the conclusion about non-monotonic links. The statement says that a non-monotonic link between two nodes with an unexpected correlation cannot cause a transition between two distinct states of a network. We provide an extreme example of non-monotonicity, in which the dependence between two nodes changes in sign in the two states of a network (e.g. stimulation in one state of a biological system and inhibition in the other).

Assume we are given $n+2$ gene expressions in two biological state P and Q: $X_P, Y_P, X_{1,P}, \dots, X_{n,P}$ and $X_Q, Y_Q, X_{1,Q}, \dots, X_{n,Q}$. We assume linear (or almost linear) dependence of Y on X within any one given biological state, stated as follows: $Y_P = \alpha_P X_P + \xi_P$ and $Y_Q = \alpha_Q X_Q + \xi_Q$, where ξ_P is a function of $X_{1,P}, \dots, X_{n,P}$, and ξ_Q is a function of $X_{1,Q}, \dots, X_{n,Q}$, and $\alpha_P \alpha_Q \neq 0$. We suppose that X_P (X_Q) and ξ_P (ξ_Q) are independent. Recall that all gene expression values are positive and remember that $\Delta X \coloneqq \mathbb{E}_P[X] - \mathbb{E}_Q[X] = \mathbb{E}[X_P] - \mathbb{E}[X_Q]$.

Lemma 4. Suppose $\alpha_P \alpha_Q < 0$ (implying that the relation between X and Y is non-monotonic), then:

- *(a) X and Y have unexpected correlations.*
- *(b) The sign of ΔY may not depend on the sign of ΔX, but instead mostly depends on the sign of $\Delta \xi$.*

Proof. Observe that, due to independence of X_P (X_Q) and ξ_P (ξ_Q):

$$cov_P(X,Y) = cov(X_P, Y_P) = \alpha_P Var[X_P],$$

$$cov_Q(X,Y) = cov\big(X_Q, Y_Q\big) = \alpha_Q Var[X_Q].$$

Therefore, $cov_P(X,Y) cov_Q(X,Y) < 0$ (so that the expected correlation inequalities do not hold simultaneously) if and only if $\alpha_P \alpha_Q < 0$. This proves the item (a) of the lemma.

Let us prove (b). Without loss of generality, $cov_P(X,Y) < 0$, implying $\alpha_P < 0$ and $\alpha_Q > 0$. Hence:

$$\Delta Y = \mathbb{E}\big(Y_P - Y_Q\big) = \mathbb{E}\big(\alpha_P X_P - \alpha_Q X_Q\big) + \mathbb{E}(\xi_P - \xi_Q)$$

Note that $\mathbb{E}(\alpha_P X_P - \alpha_Q X_Q) < 0$ regardless of the values of X_P and X_Q (both of which are strictly positive). Thus in the case $\Delta\xi > 0$ the change ΔY will still be negative. The sign of ΔY will be positive only if $\Delta\xi \gg 0$. □

References:

Benjamini, Yoav and Hochberg, Yosef. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society.Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289-300.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks*. Proc Natl Acad Sci USA 2000, 97:12182-12186.

Mine K.L.; Shulzhenko N.; Yambartsev A.; Sanson G. F.O.; Varma S.; Volfovsky N.; Brenna S. MF; Carvalho C. R.N.; Ribalta J. C.L.; Skinner J.; Lyng H.; Silva I.D.C.G.; Gerbase-DeLima M.; Morgun A. *Reconstruction of an integrative gene regulatory meta-network reveals cell cycle and antiviral response as major drivers of cervical cancer*. Accepted Nature Communications in March, 2013.

Oldham M, Horvath S, Geschwind D: *Conservation and evolution of gene coexpression networks in human and chimpanzee brains*. Proc Natl Acad Sci USA 2006:17973-17978.

Opgen-Rhein R. and Strimmer K. *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data*. BMC Systems Biology, 1:37, 2007.

Pearl, Judea. Causality: Models, Reasoning and Inference 2nd ed. Cambridge University Press Sep. 2009.

Reichenbach, Hans. Direction of Time. University of California Press, Berkley 1956.

Steuer R: *On the analysis and interpretation of correlations in metabolomic data*. Brief Bioinform 2006, 151:151-158.

T.J.VanderWeele and J.M.Robins. *Signed directed acyclic graphs for causal inference*. J.R.Statist. Soc. B (2010), 72, Part 1, pp. 111-127.

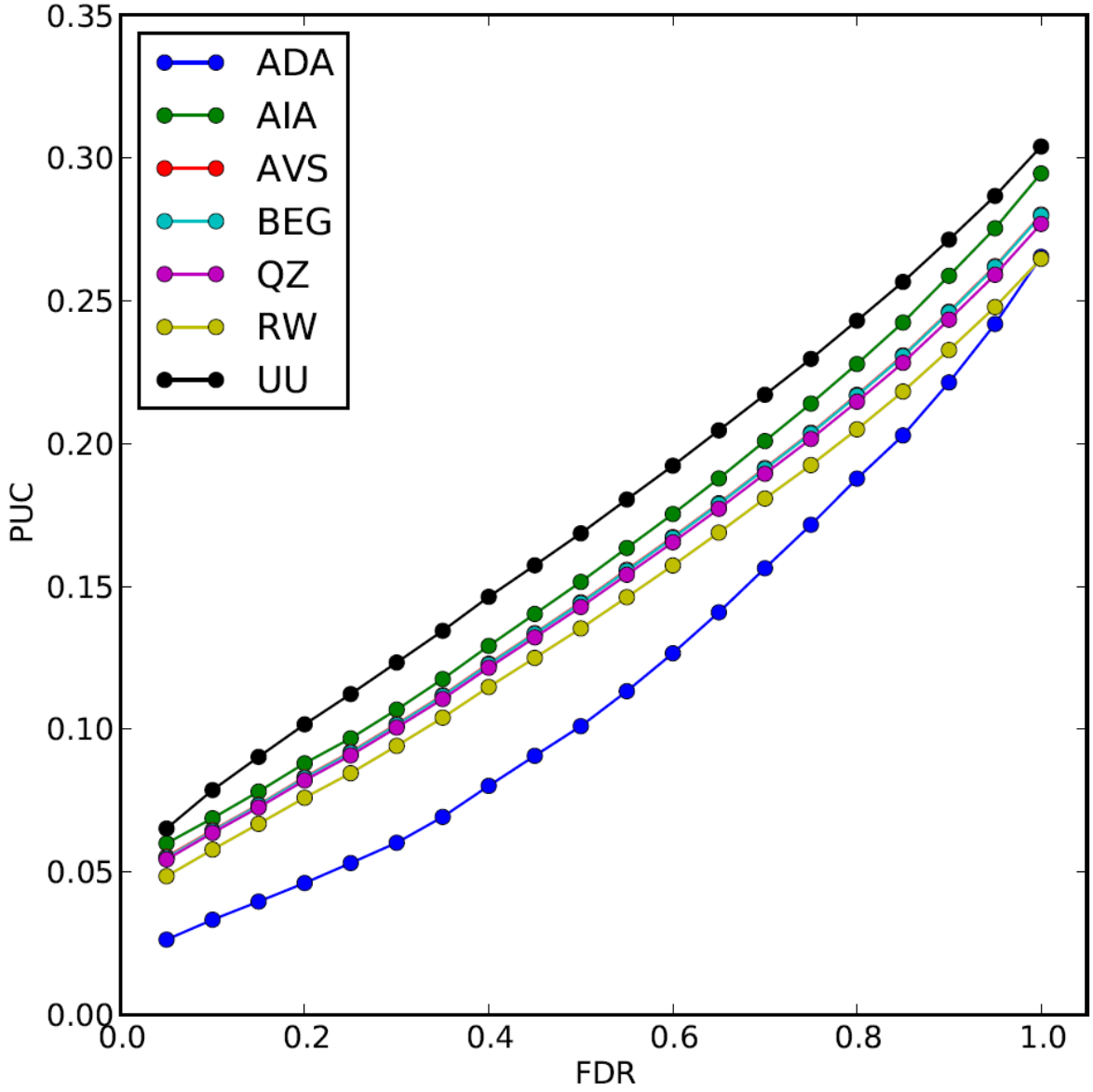


Figure S2: PUC and FDR correlate strongly when reconstructing macroeconomic networks using various bimodal parameters to define system states. Parameters shown are: ADA - Duration of compulsory education; AIA - Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total); AVS - Manufactures exports (% of merchandise exports); BEG - Educational expenditure in pre-primary as % of total educational expenditure; QZ - Private credit bureau coverage (% of adults); RW - Strength of legal rights index; UU Passenger cars (per 1,000 people)