International Journal of Biophysics 2012, 2(2): 18-25
DOI: 10.5923/j.biophysics.20120202.02

Transcriptional Network Structure Assessment Via the Data Processing Inequality

Enrique Hernández-Lemus[1,2]

[1]Computational Genomics Department, National Institute of Genomic Medicine, México City, 14610, México
[2]Complexity in Systems Biology, Center for Complexity Sciences, National Autonomous University of México, Mexico City, 04510, México

Abstract Whole genome transcriptional regulation involves an enormous number of physicochemical processes responsible for phenotypic variability and organismal function. The actual mechanisms of regulation are only partially understood. In this sense, an extremely important conundrum is related with the probabilistic inference of gene regulatory networks. A plethora of different methods and algorithms exists. Many of these algorithms are inspired in statistical mechanics and rely on information theoretical grounds. However, an important shortcoming of most of these methods, when it comes to deconvolute the actual, functional structure of gene regulatory networks lies in the presence of indirect interactions. We present a proposal to discover and assess for such indirect interactions within the framework of information theory by means of the data processing inequality. We also present some actual examples of the applicability of the method in several instances in the field of functional genomics.

Keywords Gene Regulatory Networks, Inference And Assessment, Data Processing Inequality, Information Theory

1. Introduction

One important problem in contemporary computational biology and biophysics, is that of reconstructing the best possible set of (physicochemical) regulatory interactions between genes (a so called gene regulatory network -GRN) from partial knowledge, as given for example by means of gene expression analysis experiments. This has turned to be so, since most common pathologies are not caused by the mutation of a single gene, rather they are complex diseases that arise due to the dynamic interaction of many genes and environmental factors. In order to perform GRN inference, we need to understand on a quantitative (or, at least, semi-quantitative) level the functional interplay between thousands of genes and their related proteins.

Way too many issues arise in the analysis of whole genome gene expression. Current challenges include the nature of the experimental set-ups, since microarray technology generates highly noisy signals. The so-called high dimensionality problem also arise given the fact that there are far more variables involved (number of genes and interactions among them) than experimental samples. Finally, we must deal with a great complexity due the highly nonlinear character of the biochemical dynamics underlying whole genome translation, RNA processing and regulation.

* Corresponding author:
ehernandez@inmegen.gob.mx (Enrique Hernández-Lemus)
Published online at http://journal.sapub.org/ biophysics

Information theory (IT) offers a powerful theoretical foundation, which is useful to develop algorithms and computational techniques able to deal with network inference problems as applied to real data. In the case of inferring physical interactions from correlation measures, IT even provides a useful (but far from trivial!) analogy with thermal systems and statistical mechanics. There are, however open questions and shortcomings related with the application of IT to transcriptional network inference. The applied algorithms should return intelligible models (i.e. they must result understandable); they should also rely on little *a priori* knowledge. The methods may be able to deal with thousands of variables and detect non-linear dependencies. Currently, all of these features should be accomplished even when we start with tens (or at most few hundreds) of highly noisy samples.

There are several alternate ways proposed in the literature in order to accomplish such a task. In our opinion, after having been consider a number of methods[1-3], the best benchmarking options for the GRN inference scenario, are the use of sequential search algorithms[4] (as opposed to stochastic search; that typically involve finding assignment structures for large constrained datasets, hence have a high computational complexity, even NP-hard -exponentially large search-space-) and performance measures based on IT[1, 5], since this made feature selection fast end efficient, and also provide an easy means to communicate the results to non-specialists (e.g. molecular biologists, geneticists and physicians)[6,7].

2. The Gene Network Inference Problem

Information theoretical measures have been applied to infer gene-gene interactions in transcriptional networks[8,9]. In particular, the family of probability measures that includes mutual information, Markov random fields and Kulback-Liebler divergences, has established itself as a sound and robust alternative for this task[1].

However, due to the fact that conditional probabilities obey a tower property (i.e. if X and Y are random variables with compact support on the same probability space (triple), then the expected value of the conditional expected value of X given Y is the same as the expected value of X), a number of *false positive* links appear (due to the fact that conditional correlations for chains of events obey the tower property), in some instances as a consequence of indirect correlations[10,11].

For instance, if process A has a high value of conditional information (say, mutual information) with process B, and process B is also highly correlated with process C, most common algorithms would predict also a (possibly non-existent) link between processes A and C.

One way to assess and correct for these indirect links is -as we will show later- by use of the Data Processing Inequality (DPI) which is a simple but useful theorem that states that no matter what processing you do on some data, you cannot get more information (in the sense of Shannon[10]) out of a set of data than was there to begin with. DPI then provides a bound on how much can be accomplished with signal processing.

We will outline an algorithmic implementation of the DPI within the framework of GRN inference and structure assessment and discuss some of its applications in the contemporary molecular biophysics of gene regulation.

2.1. The Joint Probability Distribution Approach (Guilt by Association)

A growing number of deconvolution methods (also called reverse engineering methods) for the probabilistic inference of gene regulatory networks, have been proposed[2,3]. In general, the goal of such methods is to provide a defined representation of the cellular network topology of the transcriptional interactions as it is revealed by, for instance, gene expression measurements, either by means of whole genome microarray expression data or, more recently by means of RNA-sequencing experiments (RNA-seq) aimed also at determining cellular gene expression patterns. Expression levels are then treated as samples taken from a joint probability distribution.

Deconvolution methods look to discover GRNs based on statistical dependence structure within this joint distribution[4]. The central aim is to develop a methodology to decompose the Statistical Dependency Matrix into a series of well defined contributions coming from interactions of several orders of complexity.

There are two major challenges related to the feature selection and network inference procedures: i) non-linearity and ii) large number of variables. Information theoretical methods are often efficient techniques to deal with such drawbacks[5-9]. Most of these methods rely on some form of mutual information (MI) metric. MI is a model independent information-theoretic measure of dependency which has been used to define (and also to quantify) relevance, redundancy and interaction in large noisy datasets.

If we resort to the standard practice of defining mutual information in terms of information-theoretical entropies (or uncertainties), then for two random variables X and Y, MI can be understood just as the amount of uncertainty in X which is removed by knowing Y, that corresponds with the accepted meaning of mutual information as the amount of information (that is, reduction in uncertainty) that knowing either variable provides about the other[1].

In fact, it is easy to see, that the mutual information is just the Kullback-Leibler distance between the joint distribution, $P_{XY}(x,y)$, and the product of the independent –marginal-distributions, $P_X(x)\ P_Y(y)$, thus, MI is an extended measure of statistical dependency[1]. MI is also able to capture non-linear dependencies[8,9] and it is also rather fast to compute. For such reasons, it can be calculated a high number of times in a still reasonable amount of time, an explicit requirement in whole-genome transcription analysis[9].

Deconvolution of a GRN based on maximum entropy optimization of the JPD of gene-gene interactions as given by gene expression experimental data is implemented as follows[9]. The Joint Probability Distribution (JPD) for the stationary expression of all genes $P(\{g_i\})$, $i=1,\ldots,N$ could be written as follows[8]:

$$P(\{g_i\}) = \frac{1}{Z}\, exp^{H_{gen}} \tag{1}$$

$$H_{gen} = [-\sum_{i}^{N}\Phi_i(g_i) - \sum_{i,j}^{N}\Phi_{i,j}(g_i,g_j) - \sum_{i,j,k}^{N}\Phi_{i,j,k}(g_i,g_j,g_k) - \ldots] \tag{2}$$

Here N is the number of genes, Z is a normalization factor (i.e. the statistical mechanics partition function), the Φ's are interaction potentials. A set of variables (genes) Ω, interacts with each other if and only if the potential Φ_Ω between such set of variables is non-zero. The relative contribution of Φ_Ω is taken as proportional to the strength of the interaction between this set.

Equation 2 does not define the potentials uniquely, thus, additional constraints should be provided in order to avoid ambiguity. A usual approach to do so is *specify* Φ's using maximum entropy (MaxEnt) approximations consistent with the available information on the system in the form of marginals. In the case of the gene network inference problem, the use of marginals is closely related with a class of methods, commonly termed hidden Markov models (HMMs)[1]. As in the case of HMMs the rationale behind marginals is in recognizing that, eventhough some priors are given, there remains a (probably quite large) set of unknown parameters that may affect the inference process and should be taken into account even if by an indirect treatment.

Hidden Markov models and MaxEnt approaches differ in the marginalizing procedure, since in HMMs the hidden states take the place of the unknown variables, whereas in MaxEnt approximations these are *marginalized* instead. A common way to do so, is by considering that *interaction potentials* (already *marginalized*, or to use the language of statistical physics, *coarse-grained*) are in some sense equivalent to correlation measures.

To be more precise; two highly correlated genes (say in their mRNA expression levels) are believed to be physically interacting (by means of some still undisclosed -but probably physically complex-mechanisms) in the transcriptional regulation network[9]. Hence, the interaction potentials $\Phi_{i,j}(g_i,g_j)$ are approximated by correlation measures, say mutual information, i.e. $\Phi_{i,j}(g_i,g_j) \approx MI(g_i,g_j)$.

2.2. Direct and Indirect Interactions: How to tell? the Data Processing Inequality

As stated before, DPI provides a bound on how much can be accomplished with signal processing[11]. More quantitatively speaking, 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.

The DPI simply states that Z cannot have more information about X than Y has about X; that is $MI(X;Z) \leq MI(X;Y)$. This inequality, which is a property of Shannon's information, can be proved. The 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 thermodynamic entropy can only decrease, conversely when removing constraints, say by allowing an irreversible process to take place, thermodynamic entropy can only increase), and the second to last equality follows because $P_{X|YZ} = P_{X|Y}$ [8,12]. More formally,

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. That is, if we know Y, knowing X does not tell us any more 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:

$$P(X,Y,Z) = P(X)P(Y \mid X)P(Z \mid Y) \tag{3}$$

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

$$MI(X;Z) \leq MI(X;Y) \tag{4}$$

Proof: By the chain rule for mutual information we can write:

$$MI(X;Y,Z) = MI(X;Z) + MI(X;Y \mid Z)$$
$$MI(X;Y) + MI(X;Z \mid Y)$$

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

In reference[8] the application of DPI has shown that if genes g_1 and g_3 interact only through a third gene, g_2 within a given GRN; we have that $MI(g_1,g_3) \leq \min[MI(g_1,g_2);MI(g_2,g_3)]$.

Hence, the least of the three MIs can come from indirect interactions only so that the proposed algorithm examines each gene triplet for which all three MIs are greater than some threshold value MI_0 and removes the edge with the smallest value. DPI is thus useful to quantify efficiently the dependencies among a large number of genes. The DPI algorithm eliminates those statistical dependencies that might be of an indirect nature, such as between two genes that are separated by intermediate steps in a transcriptional cascade. Such genes will very likely have non-linear correlated expression profiles which may result in in high MI, and otherwise would be selected as candidate interacting genes.

In fields such as developmental biology and cancer genetics, there is a growing need to place the vast number of newly identified gene variants into well-ordered genetic and molecular pathways. This will require efficient methods to determine which genes interact directly and indirectly. In this sense a methodology such as DPI-characterization will result extremely useful indeed.

For instance, the role of transcriptional cascades in development is becoming evident. Well-known examples may include, the hierarchical interactions underlying hematopoiesis and adipogenesis in vertebrates and the ecdysone and segmentation gene pathways in Drosophila[25].

In such cases, "...*gene expression in such cascades is predominantly controlled at the level of transcript initiation, and is based on interactions between sequence-specific transcription factors and their cis-acting response elements.*

Two types of regulatory relationships, direct and indirect, can be defined. Direct interactions occur independently of intermediary gene regulation but need not involve direct molecular contact between the regulator and its target gene promoter. Indirect interactions require the activation or repression of intermediary genes, the products of which act on the target gene in question...."[25].

This is precisely the scenario in which a methodology such as DPI-prunning becomes relevant to distinguish between these two different (but often indistinguishable) conditions with aims to discern the actual functional mechanisms behind them.

For instance, intron-regulation of transcription has been elucidated. Introns are able to affect gene expression significantly, both in plants and also in many other eukaryotes in a variety of ways. Some introns may contain enhancer elements or other types of promoters, whereas others function by elevating mRNA accumulation by a process called intron-mediated enhancement (IME). The intron-regions causing IME must be inside transcribed sequences near the start of a gene and in their natural

orientation in order to increase expression. Detection of IME activity by sequencing is not easy, however by observing DPI-curated networks, we may be able to infer some candidate genes, and perform deeper studies just in this reduced set.

2.3. Sufficient Statistics and Minimal Networks

The data processing inequality lies also behind some minimal representations. In particular DPI is the foundation behind the idea of *sufficient statistics*.

Definition 2 *Suppose that you have observations* $x_1, x_2, \ldots, x_n$ *for a random variable* X *distributed according to some empirical distribution* $f_\theta(x)$. *A statistic* $T(X)$ *extracts some of the information in your observed sample* $X \to T(X)$, *by the DPI,* $MI(\theta, T(X)) \leq MI(\theta, X)$. *In the cases in which equality holds, we call* T *a sufficient statistic for* θ. *That is to say, a sufficient statistic for some distribution* $f_\theta(x)$ *extracts all of the information within your data (samples)* $x_1, x_2, \ldots, x_n$ *about the value of* θ.

Let $f(x,\theta)$ be a parametric family of probability distribution functions for X. A statistic $T(X)$ is a sufficient statistic for the parameter θ iff for all sample points x and for all the parameters θ holds that:

$$f(x,\theta) = g[T(x) \mid \theta]\, h(x) \qquad (5)$$

with g and h non-negative functions, $h \neq h(\theta)$. We call equation 5 a *factorization theorem[*13] and it is a necessary and sufficient condition for sufficient statistics. If no such factorization exists for T (in the support under consideration), then T is not a sufficient statistic (in that support). Factorization theorems are important in minimal network estimation since they provide a somehow independent way of *sufficient statistics* assessment to DPI inference.

With this in mind, we can see that DPI (via the sufficient statistics argument) may be useful to infer *Minimal networks*, i.e. the smaller GRNs that are able to capture μ-almost all information content of the correlation structure of the actual (larger) biological network.

3. Applications

3.1. Minimal networks

Definition *3* *A minimal network in the context of transcriptional regulation, is the GRN that spans the statistically significant pathways -defined by a threshold in, say, a hypergeometric test of known pathways- with the minimum number of nodes and links. It is thus a concept informally related with network navigability, but instead of being defined by its topology, it refers to biological functional features.*

Minimal networks are important due to economic, logistical and analytical constraints. Nowadays, it is possible to infer extremely large and comprehensive gene regulatory networks with a certain degree of reliability for a number of cellular conditions. Such networks have been, of course, studied in their global topological features[14] and have been also the object of statistical and data mining analyses to search for biological functions and pathways[15-18]. However, detailed functional studies about the biological behavior of such large GRNs is not plausible neither experimentally nor by means of simulations. For that reason, research in functional genomics in terms of GRNs should be bounded to the minimum sized networks that one can find[9].

In order to exemplify the biological relevance of the use of the DPI to find out minimal networks, let us consider the gene regulatory network related with papillary thyroid cancer (PTC-GRN)[9]. In Figure 1 we can see two different instances of PTC-GRN. Panel A displays a GRN with 134 genes and 384 regulatory interactions. DPI was not applied in the inference of this network. Panel B displays the same GRN as panel A, however DPI was applied in this case to prune-off for indirect regulatory interactions. The network in panel B consists in 75 genes and 170 regulatory interactions among them[9].

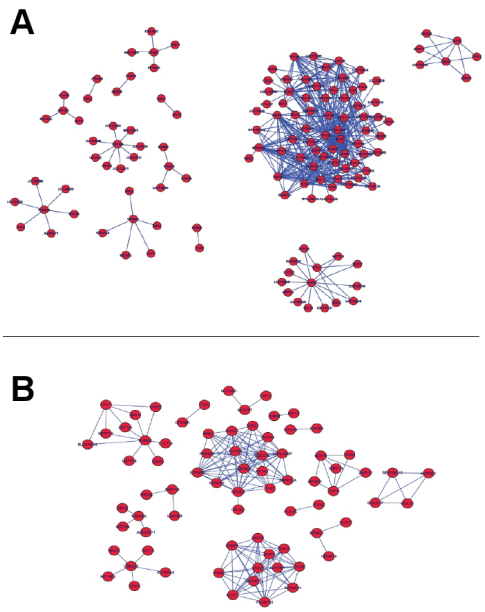

Figure 1. Gene regulatory networks associated with papillary thyroid cancer. Panel A corresponds to a network consisting in 134 genes and 384 interactions as inferred in[9] with no use of the DPI. Panel B corresponds to a network consisting in 75 genes and 170 interactions, panel B is the same network as in panel A except that DPI has been applied to prune-off indirect interactions

Network B is smaller than network A (about half-size indeed) for this very reason, network B is easier to validate and analyze, either experimentally or computationally. Of course, this issue in itself is not an advantage, unless a clearer biological picture could be extracted from the smaller network. We will show that this is the case. First of all, a large number of the nodes that were pruned correspond to

either hypothetical proteins or poorly annotated genes. For the biological researcher these molecules contribute almost nothing to their understanding of the underlying biochemical phenomena.

Moreover, given the fact that they formed mostly indirect interactions, we could hypothesize that their role (in case of not being false positives of the probabilistic inference) would be likely as *background* environmental tuning of the regulatory processes and not central to them. One obvious change in the network topology is that the biggest cluster in panel A is broken down in two smaller clusters. This fact reduced the centralization of the network (from 0.178 to 0.145) and increased slightly the clustering coefficient (from 0.409 to 0.429), thus making network B more *modular*. A detailed biological function analysis reveal that these forming modules possess indeed a clear interpretation.

Gene ontology (GO) statistical enrichment analysis [Hypergeometric tests of the networks against GO Biological Processes, corrected for multiple testing by the FDR Benjamini-Hochberg algorithm with corrected p-value < 0.05] showed that after performing DPI statistically significant GO terms arise.

Among these are the following categories: *Inflammatory response*, *Platelet- derived growth factor receptor signaling pathway*, *Stem cell maintenance* and *Hydrogen transport*. Other GO terms were conserved (that is, are significant in network A as well as network B) but its statistical significance increased, being this the case of *NAD metabolic processes* and *Regulation of RabGTPase activity*. Other (rather generalistic) processes lost their statistical significance after DPI, being this the case of *cell projection biogenesis* and *peptidyl-aminoacid modifications*.

Biochemical pathways determined by these networks also were *fine tuned* by application of the DPI. We performed Statistical enrichment analysis of biochemical pathways in both networks[Hypergeometric tests of the networks against pathways in the Reactome database[15], also corrected for multiple testing by the FDR Benjamini-Hochberg algorithm with corrected p-value < 0.05]. We found the following results of the contrast: the lower FDR-corrected p-value for network A is 1.1×10^{-3}, whereas the lower FDR-corrected p-value for network B is 2.7×10^{-5}. DPI thus improved p-value performance by almost two orders of magnitude. DPI assessment also prompted new significant biochemical pathways, some of the more important are: *urokinase plasminogen activation* and the related *plasmin synthesis and activation*; *innate immune system*, *cell junction organization* and *HNP1-4/CD4/Defensin signaling*.

As we can see, global topological features pointing out to greater modularity –hence robustness-; clearer functional mechanisms related to inflammation and growth receptor signaling (two hallmark processes in Cancer); as well as stronger statistics were attained after careful DPI-prunning of the network. This means that, at least in this case DPI methodology presents itself as an efficient tool for the analysis (both functional and modular) of biological networks.

3.2. Master Regulators Discovery: DPI+ non-DPI

DPI is also a useful method when looking to discover genes coding for transcription factors that are acting as Master Regulators[19]. Master regulator (MR) genes control a multitude of specific cellular processes and transcriptional regulation of proteins in large complexes in so-called *context-dependent* manner. Once we located the most highly connected genes -hubs- in this non-DPI pruned network, we proceeded to look up for these genes in a DPI-version of the same network.

Transcription factors acting as Master Regulators (i.e. TFs that are at the top of the transcriptional cascade) are known to display many indirect correlations with other genes. For instance, if a Master Regulator (say gene A) is a TF for another gene (B) that in turn is a TF for a third gene (C); non-DPI network will display a link between A and C whereas DPI-network will not. By analyzing both versions of a network alongside with topological parameters such as connectivity degree distributions, it is possible to look up for genes that may be Master Regulators.

If we refer to Figure 2 (which displays a non-DPI version of the GRN for the root of Arabidopsis thaliana) in panel A we can see a non-DPI pruned version of the complete transcriptional regulatory network, whereas in panel B we can observe a *zoom-in* rendering of a small region of the network in panel A in which bigger red genes are highly connected while smaller green ones have lesser number of connections.

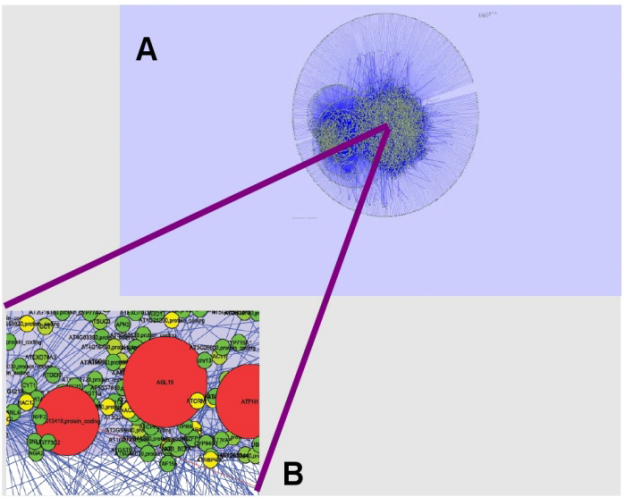

Figure 2. Gene regulatory network in Arabidopsis thaliana. Panel A displays the complete transcriptional network for the root of Arabidopsis thaliana with no-DPI implemented. Panel B shows a zoom-in of a small region of that network displaying genes color-coded and size coded according to their connectivity degree. Bigger red genes are highly connected while smaller green ones have lesser number of connections

We can see in the insert that genes that, in principle, are Master Regulators in this network (such as AGL19 and ATPH1) appear as hubs (bigger red nodes in Figure 2B). By comparison between the connectivity degree distribution in this network as well as in a DPI-pruned version, in particular with regards to these candidate Master regulators, it is possible to figure out if they are acting as such. Those of

these genes that show a drastic decrease in their connectivity can be further analyzed to determine, whether they are indeed master regulators.

Further investigation revealed the role of such genes in the morphogenesis and development in Arabidopsis thaliana. AGL 19 is now known to be an important upstream regulator. Ectopic expression of AGL19 has been demonstrated to strongly accelerate flowering. In fact, AGL19 mutants have shown a decreased response to promote flowering by prolonged cold. Epistasis analyses unveiled that AGL19 does not require SOC1 to function. Elevated AGL19 levels activate LFY and AP1 -and also by these means, activate their corresponding pathways- and eventually cause flowering[20].

With regards to ATPH1, it is structurally a planthomolog ue of human pleckstrin. As such, facilitates protein-protein interaction, in addition to protein- phosphoinositide interaction, to regulate cellular signalling. As in the case of the human pleckstrin, its role in gene regulation may be related with the fact that it can bind phosphatidylinositol lipids within biological membranes. It is then, not only a transcription factor but also a second messenger molecule involved in cell signaling[21].

Detailed analysis of gene regulatory networks[8,9, 14] has shown convincing evidence on their scale-free nature. The behavior of the whole network is dominated by a relatively small number of nodes with a large degree of connectivity. The genes corresponding to those nodes are known as *master regulators* and collectively drive the regulatory program of the underlying cellular phenotypes. Although comprehensiv e computational genomics techniques have been developed to analyze the behavior of master regulators[8], most of these rely on vast *a priori* knowledge in the form of *gene signatures*. One alternative, that may be used in a first stage of analysis, is the one above: i.e. set-theoretical difference between DPI and non-DPI networks.

In the example just considered, a number of master regulators have been discovered. For some of these, it has long been known their role as key transcription factors, while for others, only indirect evidence have been available. After this analysis, it is possible to prioritize the list of candidate master regulators in order to design RNA interference experiments to validate their functional role.

3.3. Transcription Factor Interactions in Large Networks

Given a transcription factor, application of the DPI will generate predictions about other genes that may be its direct transcriptional targets or its upstream transcriptional regulators[22,23]. The use of the DPI may result not only in a greater assessment of the results but also in a significant reduction of the computational burden associated with network inference. Zola, et al.[24] presented a parallel method integrating mutual information, data processing inequality, and statistical testing to detect significant dependencies between genes, and efficiently exploit parallelism inherent in such computations. They developed a method to carry out permutation testing for assessing statistical significance of interactions, while reducing its computational complexity by a factor of $O(n^2)$, where n is the number of genes.

The problem of inference (usually consuming thousand of computation hours) at the whole genome network level by constructing a 15,222 gene network of the plant Arabidopsis thaliana from 3,137 microarray experiments in 30 minutes on a 2,048-CPU IBM Blue Gene/L, and in 2 hours and 25 minutes on a 8-node Cell blade cluster[24].

4. Conclusions and Perspectives

In this work we have shown the relevance of the use of a theorem form information theory, the data processing inequality (Theorem 1) in the context of primary assessment of gene regulatory networks. Due to the many challenges – both, experimental and computational- involved in whole genome gene regulatory networks. Assessment methods and validation procedures are required steps in any GRN analysis. Machine learning and statistical bootstrapping techniques are commonly used, as is Montecarlo modeling and Expectation-Maximization algorithms. However, all these methods validate all genes treating them in an equivalent way, without taking into account *a priori* information about their function or their role in network topology.

In contrast, we are presenting an alternative (and additional) validation tool developed in information theoretical grounds and based in the tenets of signal analysis. DPI is one of several probability bounds on the limits in signal processing. As such it applies equally to artificial signal processing as well as to, for instance biosignals. The mathematical applicability requirements are extremely general (integrability, compact support, etc.) and are thus, almost always met in biological data such as, for instance, gene expression experiments.

If we consider whole genome gene expression patterns as signals (as we do when we reverse-engineer GRNs from expression data), then DPI states a bound on the mutual information measures between two genes. In this way, DPI enables us to distinguish -with a certain inference error- between direct and indirect transcriptional interactions (or more properly between highly correlated and correlated sets of gene expression measurements). As we have just stated, this distinction happens to be quite an important one in functional genomics studies.

As we have shown, DPI is useful in many instances in the field of GRN inference and analysis as it can be applied to large as well as to small and medium sized networks. We also show its applicability in three different problems. Modularity studies in small sized networks via sufficient statistics, as well as master regulator search and transcription factor interactions in large networks.

These are just a handful of examples amidst the broad variety of situations in which the application of so simple an idea as DPI uncovers interesting network properties or helps

to assess the validity of the inferred regulatory interactions. If applied with a completely stringent threshold, DPI may render an originally cycle-containing graph into a DAG thus making possible to compare the results of probabilistic models based on Bayesian networks with others inferred by means of information theory or non-Bayesian statistics. Bayesian networks are extremely important tools to evaluate regressive models, often useful in clinical settings. For this very reason, having a computational tool that allows comparison with such Bayesian models results important. Even in those cases in which DPI-asymptotics render incomplete information about the original networks, we can still reach approximate conclusions about the underlying systems.

DPI may also become important when analyzing physical interactions such as in protein-protein networks. In such case DPI may provide some bounds on the strength of the interactions. Due to the fact that DPI calculations possess a low algorithmic complexity (i.e. is computationally cheap) it is possible to apply it, in the computational chemistry inference of protein interactions (whose computational burden is high) as some kind of sieve in preliminary results before proceeding to more detailed calculations.

In brief, DPI is a useful, easy to implement, computational method for the assessment of the probabilistic inference of complex networks that may become important for the computational analysis of complex biophysical systems.

ACKNOWLEDGEMENTS

We gratefully acknowledge support by grant: PIUTE10-92 (ICyT-DF)[Contract 281-2010], as well as federal funding from the National Institute of Genomic Medicine (México).

We also acknowledge Professor Elena Álvarez-Buylla Roces for access to the transcriptional database for the root of Arabidopsis thaliana.

REFERENCES

[1] Hernández-Lemus, E., Rangel-Escareño, C., "The role of information theory in gene regulatory network inference" en Information Theory: New Research, Pierre Deloumeaux, Jose D. Gorzalka (Ed.), Mathematics Research Developments Series, Nova Publishing (2011) ISBN: 978-1-62100-395-3.

[2] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.; "How to infer gene networks from expression profiles", Molecular Systems Biology 3:78 (2007)

[3] van Someren, E.P., Wessels, L.F.A., Backer, E., Reinders, M.T.J., "Genetic Network Modeling", Pharmacogenomics, 3, 4, 507-525, (2002)

[4] de Jong, H., "Modeling and simulation of genetic regulatory systems": a literature review, J. Comp. Biol., 9, 1, 67-103 (2002)

[5] Peng, H., Long F., Ding, C., "Feature selection based on mutual information: criteria for max-dependency, max-relevance and min-redundancy", IEEE Trans. Pattern Analysis and Machine Intelligence 27, 8, 1226-1238, (2005)

[6] Fleuret, F., "Fast binary feature selection with conditional mutual information", Journal of Machine Learning Research 5, 1531-1555, (2004)

[7] Faith, J., Hayete, B., Thaden, J., Mogno, I, Wierzboski, J., Cottarel, G., Kasif, S., Collins, J., Garner, T., "Large scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles", PLoS Biology 5, xii, (2007)

[8] Margolin, A.A., 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, 7 (Suppl I):S7, (2006) doi:10.1186/1471-2105-7-S1-S7

[9] Hernández-Lemus, E., Velázquez-Fernández, D., Estrada-Gil, J.K., Silva-Zolezzi, I., Herrera-Hernández, M.F., Jiménez-Sánchez, G., "Information Theoretical Methods to Deconvolute Genetic Regulatory Networks applied to Thyroid Neoplasms", Physica A 388, 5057-5069, (2009)

[10] Shannon, C.E., Weaver, W., "The Mathematical Theory of Communication", The University of Illinois Press, Urbana, Illinois, (1949)

[11] Cover T. M., Thomas J.A., "Elements of Information Theory", New York: John Wiley & Sons; (1991)

[12] Sehgal, M.S.B., Gondal, I., Dooley, L., Coppel, R., Mok, G.K., "Transcriptional Gene Regulatory Network Reconstruction Through Cross Platform Gene Network Fusion", in Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science,4774/2007, 274-285, (2007) doi: 10.1007/978-3-540-75286-8_27

[13] Bickel, P.J., Doksum, K.A. , "Mathematical Statistics: Basic Ideas and Selected Topics", Vol. 1., 2nd ed., Updated Printing, Pearson, Prentice Hall, NJ, (2007)

[14] Assenov, Y., Ramírez, F., Schelhorn, S., Lengauer, T. & Albrecht, M., "Computing topological parameters of biological networks", Bioinformatics 24: 282-4, (2008).

[15] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. (2011) "Reactome: a database of reactions, pathways and biological processes", Nucleic Acids Res. (Database issue):D691-7; as well as, Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P. (2009) "Reactome knowledgebase of human biological pathways and processes", Nucleic Acids Res. 37:D619-22.
for the database see http://www.reactome.org/ReactomeGWT/entrypoint.html

[16] http://www.geneontology.org/

[17] Huang DW, Sherman BT, Lempicki RA. "Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources", Nature Protoc. 2009; 4 (1):44-57 and also, Huang DW, Sherman BT, Lempicki RA. "Bioinfor-

matics enrichment tools: paths toward the comprehensive functional analysis of large gene lists", Nucleic Acids Res. 2009;37 (1):1-13

[18] http://www.ingenuity.com/products/pathways_analysis.html

[19] Lefebvre, C., Rajbhandari, P., Alvarez, M.J., Bandaru, P., Lim, W.K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B.C., 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. 6:377. PMID: 20531406, (2010)

[20] Schönrock, N., Bouveret, R., Leroy, O., Borghi, L., Köhler, C., Gruissem, W., Hennig, L., "Polycomb-group proteins repress the floral activator AGL19 in the FLC-independent vernalization pathway", Genes Dev.20, 12, 1667-78, (2006).

[21] Mikami, K., Takahashi, S., Katagiri, T., Yamaguchi-Shinozaki, K., Shinozaki, K. "Isolation of an Arabidopsis thaliana cDNA encoding a pleckstrin homology domain protein, a putative homologue of human pleckstrin", J. Exp. Bot.50, 334, 729-730, (1999)

[22] Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., Califano, A., "Reverse engineering cellular networks", Nat Protoc.,1, 2, 662-71, (2006)

[23] He, F., Balling, R., Zeng, A-P; "Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present methods and future perspectives", Journal of Biotechnology 144, 3, 190-203, (2009)

[24] Zola, J.; Aluru, M.; Sarje, A.; Aluru, S., "Parallel Information-Theory-Based Construction of Genome-Wide Gene Regulatory Networks", IEEE Transactions on Parallel and Distributed Systems 21, 12, 1721-1733, (2010)

[25] Nasiadka, A.; Krause, H.M.; "Kinetic analysis of segmentation gene interactions in Drosophila embryos", Development 126, 1515-1526 (1999)