From Perturbation Data to a Causal Functional Pathway Representation

Nir Yosef [‡,a], Alon Kaufman [‡,b] and Eytan Ruppin [a,c]
[‡]These authors made an equal contribution to this work.
[a]School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel,
[b]Center of Neural Computation, Hebrew University, Jerusalem, Israel,
[c]School of Medicine, Tel-Aviv University, Tel-Aviv, Israel.

1 Introduction

Which elements within a system are important for its performance? How do these elements influence the system's performance, and to what extent? Are there inter-element interactions which significantly affect the system's performance? These fundamental questions, revolving around causal relationships in the data, typically arise when attempting to analyze a system in order to understand its workings.

Building predictive models based on the states of the elements (or features) in a system is a basic stage in attempting to understand how the system works. However, models which are based simply on event correlations (and are usually sufficient to make good predictions) are insufficient to uncover causal relationships in data. To causally deduce the roles played by elements of a system in determining a function of interest, perturbation experiments are necessary [1]. In complex systems, single perturbations are not sufficient to uncover the workings of the system, due to interactions between elements, and hence multiple concomitant perturbations should be employed. The goal of the approach at the basis of this paper is to analyze such multiple perturbation data and reveal the main causal functional pathways and functional interactions.

Throughout this study we relate to biological systems such as gene networks, however the approach is suitable for any system (function) in which the values attained by the elements (features) are restricted to a small, discrete domain. Focusing on genetic networks, our goal is to obtain a causal functional description of the network, by analyzing data gathered in studies where a specific cellular function is probed using a variety of multiple perturbation experiments.

Recent work in developing multiple perturbation analysis methods include the Multi-Perturbation Shapley value Analysis (MPA) presented by Keinan et al [2], which was further applied in a framework of feature selection [3], to identify the roles/contributions of individual elements to the studied function. In this extended abstract we focus on a more advantageous goal, revealing the main causal functional pathways and interactions within the system. We briefly review our previous work addressing this goal [4, 5] (section 2) and present a novel approach which utilizes additional "modular" information about the system (section 3). We demonstrate the application of the new approach to gene knock-out data studying the DNA post replication repair pathway in the yeast *Saccharomyces cerevisiae*, and compare it to the previous results in [4].

2 The Functional Influence Network (FIN) Approach

Let the investigated system be defined by a pair (N, F). $N=\{1,\ldots,n\}$ is the set of all elements in the system, where each element can be in one of two states, either intact(1) or perturbed(0).
$F : \{0,1\}^n \rightarrow R$, the performance function, associates to every set $S \subseteq N$ a number describing the performance level of the system when the set of elements S is intact, $S = \{x \in N | state(x) = 1\}$. For example, in genetic multi-knockout experiments, N denotes the set of all genes, and for each $S \subseteq N$, $F(S)$ denotes the quantitative phenotype measured in the knockout experiment in which all the genes in S are intact and the rest are knocked-out concomitantly. We base our methods on a fundamental result from Game Theory by observing that the multi-perturbation setup is essentially equivalent to a *coalitional game* [2], defined as a set of *players* (N) where for every *coalition* ($S \subseteq N$), $F(S)$ is a real number associating it with a *worth*.

Grabisch *et al.*[6] show that $F(S)$ can be uniquely decomposed into the sum $\sum_{T \subseteq S} a(T)$, where the coeffi-

cients $a(T)$, denoted *dividends*, describe the marginal contribution of each subset T of the set of intact elements S to the studied performance function F. The dividends are calculated based on the performance levels measured in the different multi-perturbation experiments, according to

$$a(S) = \sum_{T \subseteq S} (-1)^{|T|-|S|} F(T), \quad \forall S \subseteq N$$

(where $|S|$ and $|T|$ denote the cardinality of the sets S and T respectively).

Based on the dividends, the performance function F can be represented as a multi-linear polynomial:

$$F(\vec{x}) = \sum_{S \subseteq N} a(S) \cdot \prod_{i \in S} x_i$$

where the vector $\vec{x} \in \{0,1\}^n$ describes the (intact/perturbed) states of the elements in the system. Each summand in the polynomial describes a distinct *functional pathway* with a causal relation to the performance, since its elements must all be intact to influence the value of F. Continuing this terminology, we refer to the polynomial model, $F(\vec{x})$, as the functional influence network (FIN).

2.1 The Compact Functional Network (CFN).

The challenges of analyzing genetic biological data markedly differs from the Game Theory scenario described above since often the full functional description gets very large and unintelligible, containing many "uninteresting" pathways with a very small (but non-zero) influence. To address this problem, Kaufman *et al.* [4] introduced the concept of the CFN, a compact representation which approximates the full functional description. The CFN is in itself a multi-linear polynomial which preserves only the most important summands of the full representation. Figure 1 shows a schematic example of a CFN construction and its visual representation. Given a specific state of the system, the expected performance level of the CFN can be calculated by summing up the dividends of all the intact functional pathways (see [5] for a detailed discussion on the biological relevance of the FIN polynomial representation).

Evidently, the construction of the CFN requires the performance values over all possible multi-knockout experiments. Producing such data is an unrealistic demand in most cases. In addition, the data is typically noisy. Addressing these problems and constructing a CFN based on partial and noisy data, is the end goal of the two previous FIN algorithms [4, 5] as well as the new method discussed in the next section. The basic idea underlying all three methods is the use of machine learning techniques to predict the results of the missing experiments and fill in the

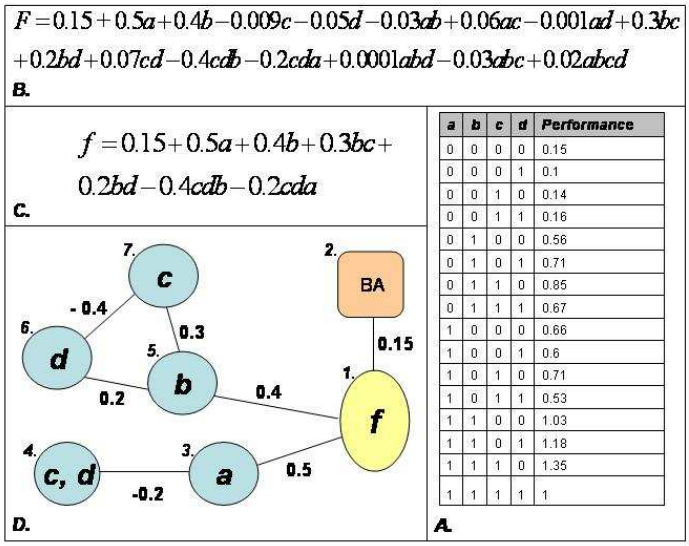

Figure 1: *A simple schematic CFN construction of a 4 element system, taken from [5].* Box A provides the analyzed data set of the 16 (2^4) possible combinations of multi-knockout experiments and their corresponding performance measures. This set yields a unique performance function F, derived by a dividend analysis of the multi-knockout data set, describing the phenotypic behavior of the system (box B). Box C presents the resulting CFN approximating the full functional description. With this compact CFN representation, the approximated performance function f can be visualized in a relatively simple graph (box D). Each of the round-shaped nodes(numbered $3-7$) in the graph corresponds to a set of (possibly only one) elements and is said to be intact if, and only if, all of its corresponding elements are intact. Node 2 corresponds to the empty set, describing the system's basal activity. An additional node (node 1) is the *output node* f. Each simple path which ends at the output node defines a functional pathway (a summand in the CFN) whose elements are those listed on the other nodes along the path. The dividend of each such functional pathway is the weight on its first edge, *e.g.,* the dividend of the functional pathway $(c-d-b-f)$ is the weight on the edge between nodes 7 and 6. Given a knockout experiment where (for example) only a, c and d are intact then the intact functional pathways are $(a-f)$ and $(c,d-a-f)$, and the value of f is $0.5-0.2+0.15$=0.45, the sum of dividends of intact pathways and the basal activity.

missing data. For a formal description and evaluation of the previous algorithms on simulated networks as well as biological data see [5].

3 The Modular Functional Model

Motivated by the empirical observation that many biological networks exhibit a modular organization [7], we introduce the concept of *functional modularity* and utilize it to efficiently construct CFN models. We define a functional module to be a set of elements which act as a single functional unit. A natural way to view a modular system is in a two level hierarchical manner: (i) Inter Module - considering the interactions between entire modules as single units, and (ii) Intra module - considering the interactions between the elements of each module separately.

Formally, using the above notations, assume that the elements in the system can be partitioned into m modules $M = \{mod_1, \ldots, mod_m\}$, where $mod_i \subseteq N$ (note that the intersection between modules can be non-empty). Then the FIN polynomial of the system can be written as

$$F(\vec{x}) = \sum_{S \subseteq M} a(S) \cdot \prod_{mod_i \in S} P_i(\vec{x})$$

where, P_i is the FIN polynomial describing the interactions within module i:

$$P_i(\vec{x}) = \sum_{S \subseteq mod_i, S \neq \emptyset} a(S) \cdot \prod_{j \in S} x_j, \quad \forall i.$$

We construct the polynomials P_i such that $P_i(\vec{x}) = 1$ if all the elements in mod_i are intact, and $P_i(\vec{x}) = 0$ if all are perturbed.

The partitioning of elements into modules is based on a-priori knowledge of the system. For example, in gene or protein networks, we use protein complexes for the partition of the elements. Such data is readily available for a growing number of species. Specifically, in this extended abstract, we use known complexes in yeast as annotated by the Munich Information Center for Protein Sequences (MIPS) [8].

Clearly, partitioning a system into functional modules significantly reduces the amount of computation required to generate the modular functional model. More importantly, however, is that it reduces the number of candidate models when only a part of the required perturbation data is available. In these cases, we would prefer candidate models which reflect a correct modular partition, as deemed by other sources of data.

We demonstrate our modular approach using the DNA post-replication repair (PRR) system of the yeast *Saccharomyces cerevisiae*. This system has previously been the subject of a FIN analysis [4], which led to a number of new biological insights such as the existence of additional, unknown elements, involved in the PRR process.

The experimental data available for our analysis is composed of 21 multi-knockout experiments of 5 genes of interest (RAD18 - a regulatory gene needed for PCNA modification, REV3 which encodes a DNA polymerases and 3 genes that presumably act as clamps for specific DNA polymerases - ELG1, CTF18 and RAD24). The performance measure, F, is the ability of the resulting mutants to resolve single-stranded gaps created after UV irradiation, measured by the relative number of colonies that survive the radiation compared with a wild-type yeast strain.

According to [8], the five elements are partitioned into four modules: the first module containing two clamp loader genes: $mod_1 = \{ELG1, CTF18\}$, and each of the rest contains a single element: $mod_2 = \{RAD24\}$, $mod_3 = \{REV3\}$, $mod_4 = \{RAD18\}$. Let the variables $x_1 \ldots x_5 \in \{0, 1\}$ denote the intact/perturbed states of the genes ELG1, CTF18, RAD24, REV3 and RAD18 respectively. Using the modular formulation and the FINE algorithm [5] for CFN construction, we derived the following functional model:

$F(\vec{x}) = P_4(\vec{x}) \cdot (.24 \cdot P_2(\vec{x}) + .26 \cdot P_3(\vec{x}) \cdot (1 + .09 \cdot P_2(\vec{x})) + .23 \cdot P_1(\vec{x}) \cdot (1 + .19 \cdot P_3(\vec{x})))$

where $P_1(\vec{x}) = .8 \cdot x_1 + .47 \cdot x_2 - .27 \cdot x_1 \cdot x_2$, and for $i \in \{2, 3, 4\}$, $P_i(\vec{x}) = x_j$, such that $x_j \in mod_i$.

A visualization of this modular CFN is given in figure 2. Comparing the modular CFN to the previous FIN analysis of the same data [4] (which does not make any modular assumptions), reveals that both models capture more than 90% of the variance in the experimental data (calculated by comparing the CFN-based predicted performance levels in the given 21 knockout experiments to the actual experimental results, measuring resilience to irradiation). More importantly, both models can be used to draw conclusion as for the functional structure of the system. For example, in both cases we see that RAD18 plays a key role in the functioning of the system, inline with its role as a regulatory gene. In addition, both CFNs reflect the organization of the PRR pathway showing that the clamp loaders operate in functional pathways with REV3, the DNA polymerase, further suggesting that there is probably another DNA polymerase (or perhaps more than one) involved in the PRR process (this since the clamp loaders contribute also without REV3). Interestingly, in contrast to the previous CFN([4]) which excludes the causal role of CTF18, the modular CFN suggests that CTF18 is a weak substitute to ELG1 (redundant in 27% of their functionality) and plays a causal role in the pathway. Indeed, this result of the modular FIN analysis is sup-

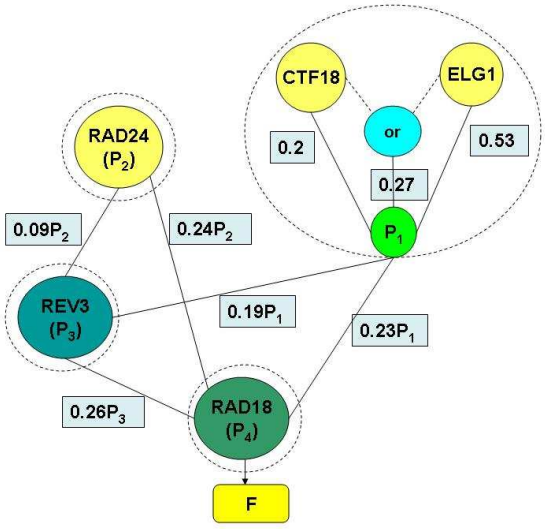

Figure 2: *The modular CFN of the yeast's DNA post-replication repair (PRR) system.* The genes are represented as binary nodes whose state is determined according to the given perturbation experiment, intact or knocked out. Modules are represented as dotted lines, circling the corresponding genes. The output node F corresponds to the performance function - the relative number of colonies that survive UV irradiation. The *or* node connected to $ELG1$ and $CTF18$ gets a value of 1 if at least one of the two is intact, and a value of 0 otherwise. The P_1 node is the output of mod_1. Weights on the edges are dividends values, reflecting the functional influence between their end points.

ported by the finding that CTF18 does play a causal role in the PRR pathway [9]. Finally, we note that the construction of the modular CFN model was based on the results of 18 perturbation experiments (2^4 experiments for the inter module level and 2^2 experiments for the intra module level with 2 experiments overlap between the two sets) whereas the previous CFN construction algorithms required all 32 experiments.

4 Conclusions

Inferring causal relations in a system requires studying the effects of altering the states of its elements. Perturbation studies in biological systems can produce such data, in which functional performance is measured after deletion, mutation or lesioning of the different elements. In complex systems, single element perturbation are not sufficient, and hence multi-perturbation are required. In recent years our lab has developed methods for analyzing such data, uncovering causal dependencies between pairs of elements and inferring the functional pathways.

In this extended abstract we review the functional influence network (FIN) framework and address the challenge of incorporating additional information regarding the system's structure. We define the concept of functional modularity and show how it can be utilized to reduce the amount of computation required to generate a functional model. The significance of the FIN enhancement we present is in the fact that it requires less data and produces an informative and more reliable causal functional description of the network. This enhancement however is only a first step in an ongoing effort to incorporate heterogeneous data into the FIN causal model inference. We are now engaged in developing a framework for functional modeling of cellular systems which considers physical interaction data, such as protein-protein and protein-DNA interactions, in addition to perturbation data.

References

[1] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge: University of Cambridge Press, 2000.

[2] Keinan A, Sandbank B, Hilgetag CC, Meilijson I, and Ruppin E. Fair attribution of functional contribution in artificial and biological networks. *Neural Computation*, 16:1887–1915, 2004.

[3] Cohen S, Dror G, and Ruppin E. Feature selection via coalitional game theory. *Neural Computation, in press.*

[4] Kaufman A, Keinan A, Meilijson I, Kupiec M, and Ruppin E. Quantitative analysis of genetic and neural multi-perturbation experiments. *PLoS Computational Biology*, 1(6):e64, 2005.

[5] Yosef N, Kaufman A, and Ruppin E. Inferring functional pathways from multi-perturbation data. *Bioinformatics*, 14:e539–46, 2006.

[6] Grabisch M, Marichal JL, and Roubens M. Equivalent representations of a set function with applications to game theory and multicriteria decision making. *Mathematics of Operations Research*, **25**(2):157–178, 2000.

[7] Ihmels J et al. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–7, 2002.

[8] HW Mewes et al. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 32:D41–4, 2004.

[9] Ben-Aroya S, Koren A, Liefshitz B, Steinlauf R, and Kupiec M. *ELG1*, a yeast gene required for genome stability, forms a complex related to replication factor C. *Proc. Natl. Acad. Sci. USA*, 100:9906–9911, 2003.