

Multi-Knockout Genetic Network Analysis: The Rad6 Example

Alon Kaufman

Center of Neural Computation,
Hebrew University, Jerusalem, Israel
akaufman@alice.nc.huji.ac.il

Martin Kupiec

Department of Molecular Microbiology and Biotechnology,
Tel Aviv University, Tel Aviv, Israel
martin@post.tau.ac.il

Eytan Ruppin

School of Computer Science and School of Medicine,
Tel Aviv University, Tel Aviv, Israel
ruppin@post.tau.ac.il

Abstract

*A novel and rigorous Multi-perturbation Shapley Value Analysis (MSA) method has been recently presented [12]. The method addresses the challenge of defining and calculating the functional causal contributions of elements of a biological system. This paper presents the first study applying MSA to the analysis of gene knockout data. The MSA identifies the importance of genes in the Rad6 DNA repair pathway of the yeast *S. cerevisiae*, quantifying their contributions and characterizing their functional interactions. Incorporating additional biological knowledge, a new functional description of the Rad6 pathway is provided, predicting the existence of additional DNA polymerase and RFC-like complexes. The MSA is the first method for rigorously analyzing multi-knockout experiments, which are likely to soon become a standard and necessary tool for analyzing complex biological systems.*

1. Introduction

The success of genome sequencing projects has allowed biologists to identify almost all genes responsible for producing the biological complexity of several model organisms. The next important task is to quan-

tify their importance in various functions [5] and understand the interactions among the genes [1]. This paper utilizes a novel conceptual and mathematical method for causal system identification in biological systems, based on the analysis of multiple knockout experiments. Our study is based on the Multi-perturbation Shapley value Analysis (MSA) method, recently developed and applied to the analysis of artificial and biological neural systems [11, 12]. In this paper we apply the MSA to the analysis of genetic knockout experiments of the Rad6 DNA repair system in the yeast *S. cerevisiae*. The MSA assigns *Contribution Values (CVs)* to the genes, denoting their relative importance to the investigated function. It further identifies the types of interactions between these genes and the contribution of important groups of genes. By incorporating additional, known biological data, the knockout experiments data can further be utilized to obtain a functional inference description of the Rad6 pathway.

Localization of function (system identification) in genetic networks is conventionally addressed by high-throughput expression profiling, mainly using DNA microarray techniques. These strategies have yielded large amounts of useful information [17], however, such correlation methods do not necessarily identify causality. Previous studies have indeed shown that there may be, at times, a weak correlation between the expression of different genes and their role in various cellular func-

tions [19, 7]. This may occur because, in general, the *causal* identification of the elements that are responsible for a given function actually requires perturbation studies [14]. Toward this end, gene knockout studies have been traditionally employed, in which functional performance is measured after deletion or mutations of different genes in the network [19]. However, in practice, deleting a single gene in an organism often has little phenotypic effect, due to the existence of gene duplicates and alternative metabolic pathways [9]. That is, in complex networks such as metabolic or gene transcription networks, the contribution of an element depends on the state of other elements. Single perturbation analysis is hence likely to be misleading, resulting in erroneous conclusions, and the need for multi-perturbation studies has become clear [9].

Acknowledging that single perturbations are insufficient for faithful localization of function and system identification, Keinan *et al.* [12] have recently developed the MSA, a systematic approach for the analysis of multi-perturbation experiments in neural systems. In the genetic network analysis setup on which we focus in this paper, we apply the MSA method to analyze a set of multiple gene knockout experiments. In each such individual experiment, a few genes are knocked out together, and the resulting performance level of the investigated task is recorded. Given these data, the MSA outputs a set of CVs which denote the importance of each gene, as well as a quantification of the interactions between the genes. We then describe the task studied in a quantitative way leading to a functional inference description. The MSA is based on an axiomatic approach borrowed from Game Theory, and yields a unique and fair attribution of contributions among the investigated elements.

The potential of the MSA and its extensions for the analysis of genetic networks is demonstrated via the multi-knockout investigation of the Rad6 DNA repair system in the yeast *Saccharomyces cerevisiae*. To this end, the remainder of this paper is organized as follows: Section 2 provides a brief description of the multiple perturbation analysis method and some of the different algorithmic variants it encompasses. Section 3 presents the Rad6 DNA repair system in the yeast, and section 4 describes the MSA study of this system. In section 5 we describe how additional biological data can be incorporated to yield a functional inference description. Our conclusions and future applications are briefly discussed in section 6.

2. Multi-perturbation Analysis

Given a system of a number of elements, we wish to ascribe to each element its contribution (importance in terms of causal responsibility) in carrying out a certain function. To achieve this goal, assume one can measure the system's performance for this function (*e.g.*, the ability to survive UV irradiation), and that one can introduce multiple perturbations to the system before measuring the performance. Accordingly, the data for a multi-perturbation analysis is a series of such perturbation experiments, which in theory can consist of 2^n experiments for a system of n genes. In each such experiment, a different subset of the system's elements are perturbed concomitantly (denoting a perturbation configuration) and the system's performance function is measured. Given this data the Multi-perturbation Shapley value Analysis (MSA) method assigns a contribution value to each of the elements, and identifies the important interactions between the system's elements. We provide below a brief description of the MSA method. A detailed and extensive description of the MSA can be found in [12].

2.1. The Basic One-Dimensional MSA

The MSA is based on the observation that a set of multi-perturbation experiments can be viewed as a coalitional game, borrowing the latter concept and analytical approach from Game Theory [16]. A *coalitional game* is defined by a pair (N, v) , where $N = \{1, \dots, n\}$ is the set of all *players* and $v(S)$, for every $S \subseteq N$, is a real number associating a worth with the *coalition* S . In the context of genetic multi-knockouts, N denotes the set of all genes investigated, and for each $S \subseteq N$, $v(S)$ denotes the performance function measured under the multi-knockout experiment in which all genes in S are intact and the rest are knocked-out. Let the *marginal contribution* of gene i to a perturbation configuration S , with $i \notin S$, be

$$(1) \quad \Delta_i(S) = v(S \cup \{i\}) - v(S),$$

the contribution value of each gene $i \in N$ is then defined as the *Shapley value* [16]:

$$(2) \quad \gamma_i(N, v) = \frac{1}{n!} \sum_{R \in \mathcal{R}} \Delta_i(S_i(R))$$

where \mathcal{R} is the set of all $n!$ orderings of N and $S_i(R)$ is the set of genes preceding i in the ordering R . It should be noted that the Shapley Value is a representative of a broader family of semi-values, differing in the way experiments are weighted. The enumeration

over the set of $n!$ orderings \mathcal{R} , as defined by the Shapley Value, essentially enumerates over all sets of 2^n possible multi-knockout configurations, weighting them such that multi-knockouts of a size k (k , the number of knocked-out genes) receive equal weighting as multi-knockouts of size l ($l \neq k$). The Shapley value is efficient, that is, it divides the overall system's worth (the performance gap $v(N) - v(\emptyset)$) between the different genes. It is the only efficient value satisfying three basic assumptions (null players contribution, symmetry and additivity across games) which all apply naturally in biological systems [12]. The contribution of a gene, γ_i , measures its importance, that is, the part it plays in the successful performance of the function studied.

As described above in Eqs.(1) and (2), the original Shapley value requires full knowledge of the behavior of the game (system) over all possible coalitions (multi-knockout configurations). When applying the Shapley value concept to biological systems this scenario is of course unrealistic. One encounters two main problems: 1) *The missing data problem*, where only a partial set of the perturbation configurations may be available even in relatively small systems. 2) *The scalability problem*, *i.e.*, determining the CVs of the elements in a large system where a full Shapley value computation in accordance with Eq.(2) is simply intractable.

To address the scalability problem, the MSA uses estimation methods to compute the CVs approximately from a relatively small set of experiments (see [12]). In this paper, we focus on the analysis of a small size genetic system which does not require the use of estimation methods. Yet, even in our system we encounter the missing data problem, where only a partial set of all possible multi-knockout experiments are accessible. We hence train a predictor on the available dataset to predict the performance scores of the missing, unseen, multi-knockout configurations. A standard "leave-one-out" procedure is used during training to obtain a predictor with a low generalization error. Given such a predictor, the CVs may now be accurately approximated using Eqs.(1) and (2).

2.2. The Two-Dimensional MSA Interactions

The one-dimensional CV denotes for the average marginal contribution of an element to a given function. To capture the dependency of an element's importance on the state (perturbed or intact) of other elements, a higher order description is needed (for example, a pair of genes that only in combination are synthetically lethal [18]). The Shapley interaction index [8] defines the interaction magnitude in all possible dimensions. We focus here on the description of two-dimensional inter-

actions, which are used in our genetic data analysis. The two-dimensional interaction between a pair of genes i and j , describes how much does the contribution of gene i depend on the state of gene j . To calculate this interaction we divide the experimental dataset into two subsets, those where gene j is intact and those where gene j is knocked-out. Let $\gamma_{i,j}$, be the Shapley value of gene i in the experimental setup where gene j is always *intact*, and $\gamma_{i,\bar{j}}$ be the Shapley value of gene i in the experimental setup where gene j is always *knocked-out*. Intuitively, these values represent the average marginal contribution of gene i when gene j is intact or knocked-out, accordingly. The two-dimensional interaction between the genes i and j , which quantifies *how much the state of j influences the average marginal contribution of gene i* , is defined in two equivalent ways by [8, 12]:

$$(3) \quad I_{i,j} = \gamma_{i,j} - \gamma_{i,\bar{j}}$$

and

$$(4) \quad I_{i,j} = \gamma_{(i,j)} - \gamma_{i,\bar{j}} - \gamma_{\bar{j},i},$$

where $\gamma_{(i,j)}$ is the Shapley value of a combined element (i, j) . That is, a new compound element that is considered intact when both genes i and j are intact, and knocked-out when i and j are both knocked-out. This symmetric interaction measure ($I_{i,j} = I_{j,i}$), as phrased in Eq.(4), actually states how much "the whole (i, j) is greater than the sum of its parts" (if $I_{i,j}$ is positive, i contributes more when j is intact and vice versa).

Based on this definition, we now categorize the two-dimensional interactions into a number of different types, giving rigorous definitions conforming with the interaction terminology classically used in genetics [3]. For each pair of genes we compare the average effect of knocking-out each gene separately to the average affect of knocking-out both genes simultaneously, averaging over all possible knockout experiments. This comparison yields the following interaction types: **Additive:** If the interaction's magnitude, $I_{i,j}$, is zero, then the interaction of the pair of genes is completely additive, and each gene's contribution is totally independent of the other ($\gamma_{i,j} = \gamma_{i,\bar{j}}$). **Synergism:** When the average effect of perturbing two genes simultaneously is greater than the sum of the individual average effects. A synergistic interaction reflects some functional overlap between the genes, *i.e.*, the double mutant on average is more harmful than expected from the separate single mutants. In the opposite case, when the effect of perturbing both genes is less than the sum of the individual effects, we further define specific non-symmetrical relations, *Epistasis* and *Suppression*, based on the interaction nomenclature of Brendel *et al.* [3]. An interac-

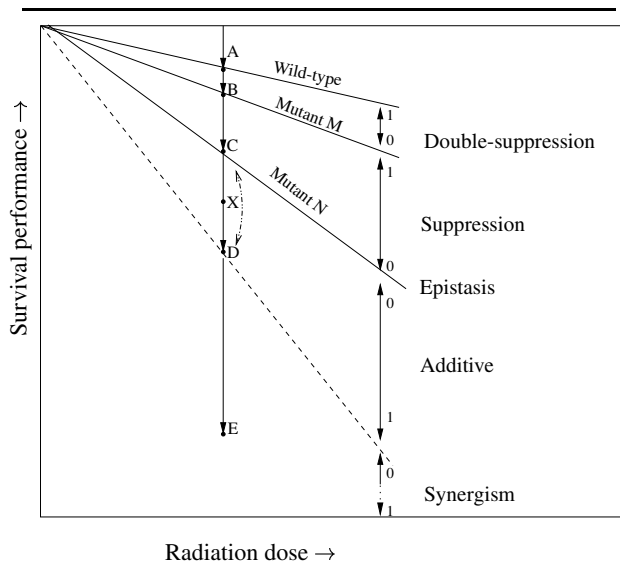


Figure 1. Schematic definition of the basic types of the two-dimensional interactions and relations. The points B and C represent the average phenotypic performance of a single mutant (mutant M and mutant N respectively). A double mutant MN may lie anywhere within the graph determining the interaction classification. For example, if the double mutant MN falls on point X in the figure, the interaction is classified as an additive interaction. The magnitude of an additive interaction can vary from point C (0) up to a full additive interaction (1) represented by point D ($AD=AB+AC$).

tion will be defined **Epistasis** if $\gamma_{i,\bar{j}}$ has no contribution and $\gamma_{i,j}$ has a positive contribution, meaning the intactness of j is essential for i 's contribution. If $\gamma_{i,j}$ is positive and $\gamma_{i,\bar{j}}$ is negative the relation is defined as **Suppression** since the negative contribution of i (when j is knocked-out) is suppressed when j is intact. Figure 1 summarizes these two-dimensional interaction and relation types.

2.3. Functional Inference

This subsection outlines a different way of analyzing multi-knockout data, by describing the investigated performance function as a sum of the "importance" of all possible subsets of genes in the system studied. For example, if we are investigating a simple system with only three genes, the desired description will include eight

terms, including the performance of the null set (when all genes are knocked-out), the incremental contribution of each gene separately (that is, the contribution of each gene without the contribution of the null group), the incremental contribution of the three possible pairs of genes (the contribution of the pair on top of the contribution of each gene separately), and the incremental contribution of the triplet. More generally, this description will include 2^n terms, for a network of n genes, encompassing all possible multi-perturbation configurations in the network. Formally, the performance function, $v(S)$, is expressed in a unique way as,

$$(5) \quad v(S) = \sum_{T \subseteq S} a(T), \quad \forall S \subseteq N,$$

where S denotes the subset of intact genes in the network. The coefficients $a(T)$ are called *dividends* [8], describing the incremental importance of each subset T to the performance studied. These dividends can be calculated from the multi-perturbation data according to Eq.(6) in which the cardinality of the sets S and T are denoted by corresponding lower cases s, t ($s = |S|, t = |T|$),

$$(6) \quad a(S) = \sum_{T \subseteq S} (-1)^{t-s} v(T), \quad \forall S \subseteq N.$$

The dividends computation begins from the dividend of the null group and each iteration computes the dividend (incremental contribution) of the subsequent subsets (that is, the contributions of the single genes minus the null group, the contributions of the pairs minus the single dividends, the contribution of the triplets minus the dividends of the pairs, and so on...). Focusing only on the large dividends, one can express the expected performance level given the intact/perturbed state of the system as a linear summation of these large dividends. This reduced function is a compact and simplified approximation of the performance function, which emphasizes the pathways involved in the network's function and assigns importance values to each of them. Incorporating other sources of biological knowledge together with this reduced performance function may lead to new functional insights as demonstrated in section 5. Section 5 also provides a detailed description of the functional inference process in conjoint with its application.

3. DNA Post Replication Repair

We investigate the DNA post-replication repair (PRR) system of the yeast *Saccharomyces cerevisiae*. Damaging agents such as UV light can create lesions

on the DNA. As DNA polymerases stall on these lesion sites, single-stranded gaps are created. The PRR acts to convert these single-stranded gaps into large molecular weight DNA [4]. In the yeast, this pathway is dominated by the activity of the Rad6 protein. The Rad6 group of genes can be divided into several sub-pathways that are still poorly understood. A main component in the Rad6 pathway is an alternative DNA polymerase, which presumably replaces the canonical DNA polymerase to perform the repair of the DNA lesions. In this paper we focus on the analysis of a subset of genes from the Rad6 group, and on additional genes required for polymerase switching.

Replicative DNA polymerases are stabilized on the DNA by a “sliding clamp”, the trimeric PCNA complex. Replication factor C (RFC), a hetero-pentameric protein complex, is necessary for loading PCNA onto double-stranded DNA at the primer-template junction. RFC includes a large subunit, Rfc1, and four small subunits (Rfc2-5). Lately, several proteins with similarities to Rfc1 were found to form RFC-like complexes (RLCs). These include Elg1, Rad24 and Ctf18. Naturally, several questions have emerged [13], *e.g.*, how does the function of the three RLCs differ among each other, and why are there three different RLCs. One of the possibilities raised in the literature is that the RLCs may act similarly to RFC, loading PCNA-related complexes that act as clamps for specific DNA polymerases [2]. Since the Rev3 protein, a member of the Rad6 repair group, encodes an alternative DNA polymerase, we decided to analyze the relationship between members of the Rad6 group (Rev3 and Rad18) with the genes encoding the large components of the three RLCs.

4. MSA Analysis of the Rad6 Pathway

4.1. The Data

The components of the Rad6 pathway in yeast have been previously identified. In order to analyze more precisely the function of these genes, we carried out an MSA study of multi-knockout experiments performed recently in the lab of one of the authors (M.K.). We analyzed a series of yeast multi-knockout experiments, testing the ability of the resulting mutants to resolve the single-stranded gaps created after UV irradiation. Performance after each knockout experiment is measured by the relative number of colonies that survive compared to the wild-type yeast (varying from 0 to 1). The genes examined were *ELG1*, *CTF18*, *RAD24* (the three RLCs), *REV3* encoding for a specific DNA polymerase ζ , and *RAD18*, a regulatory gene whose product inter-

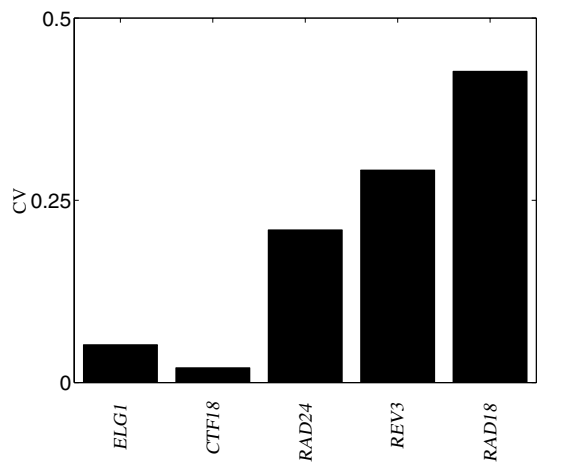


Figure 2. One-dimensional MSA results, showing the CVs of each gene (all contributions sum up to 1).

acts with the Rad6 protein. All these are assumed to play a causal role although they are not the only genes involved in the Rad6 pathway.

The dataset included 21 multi-knockout experiments. The full $2^5 = 32$ multi-knockout set needed for our analysis (Eq.(2)) was obtained using projection pursuit regression as a predictor [6], explaining 83% of the variance of the data.

4.2. One-Dimensional Analysis

Figure 2 shows the CVs of the different genes involved in the experiments, obtained via MSA. The MSA identifies all genes as playing a causal role in the PRR process, assigning precise quantitative contributions for each gene. The most important genes are *RAD18* and *REV3*. All three RLCs play a causal role as well, but their importance markedly differs; *RAD24* is about four times more important than *ELG1* and nine times more important than *CTF18*.

4.3. Two-Dimensional Interactions

We further performed a two-dimensional MSA analysis to quantify and classify the interactions between each pair of genes. Figure 3 shows these interactions and relations. The three established RLC component genes, *ELG1*, *CTF18* and *RAD24* exhibit almost completely additive interactions amongst themselves, suggesting that the role each gene has in PRR is independent from the others (low magnitude of synergism is

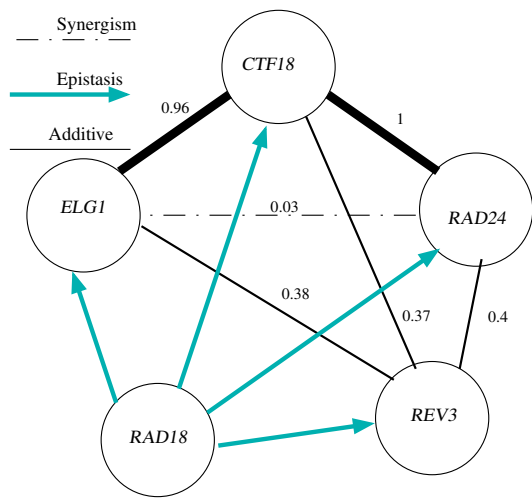


Figure 3. Two-dimensional MSA results. The two-dimensional interactions ($I_{i,j}$) between the different genes, classified according to section 2.2. The numbers above the additive and synergistic interactions edges describe their magnitude as demonstrated in Figure 1 (scaled from 0 to 1). The Epistasis arrows are directed ($i \leftarrow j$), where i 's contribution is dependent on j 's intactness.

close to an additive interaction). All genes have an epistasis relation with *RAD18*, thus, as expected, *RAD18*, a regulatory gene, is crucial for their contribution and an essential gene in the PRR process. The three RLCs have moderate additive interactions with the DNA polymerase *REV3* (~ 0.4), indicating that each of them combined with *REV3* exhibits some extra performance, reinforcing the hypothesis that the RLCs load the DNA polymerase ζ encoded by *REV3* (it should be noted that the two-dimensional interactions are based on average contributions calculated across all possible knockouts to the network, in contradiction to the conventional way of looking at the phenotypic difference between the specific configurations of the two single mutants versus the double mutant solely).

5. Functional Inference With Multi-knockout Data: The Rad6 Example

Calculating the dividends from the 32 multi-knockout configurations data set and their associated performance (section 2.3), one can describe UV sensitivity PRR performance, F , as a linear sum-

mation of significant dividends over the investigated genes. Choosing only dividends with coefficients having an absolute value equal or above 0.05, we obtain ,

$$(7) \quad F = .26 \cdot (d \cdot e) + .2 \cdot (c \cdot e) + .13 \cdot (a \cdot d \cdot e) + .36 \cdot (c \cdot d \cdot e) + .05 \cdot (b \cdot d \cdot e) + .05 \cdot (a \cdot c \cdot e) - .05 \cdot (a \cdot c \cdot d \cdot e)$$

where $a = ELG1$, $b = CTF18$, $c = RAD24$, $d = REV3$ and $e = RAD18$ are Boolean variables, assigned 1 if intact and 0 if knocked-out. F provides a very accurate description of the PRR performance, having a normalized MSE equal to 0.0087, *i.e.*, explaining 99.13% of the variance of the original performance data. The function F is thus a compact analytical representation approximating the PRR function, including only the most important genes and gene-pathways. In order to grasp the underlying properties of such a function, visualization techniques can be of assistance. The Functional Influence Network (FIN), described herewith, is one such visual representation of F . The goal of the FIN construction is to describe the influence each gene has on the performance level F of the PRR pathway. To this end, we would like each gene to be represented by a minimal number of nodes in the network, and the edges connecting nodes should describe the mutual influence of the genes on the function. The FIN is constructed in such a way that the performance level can be predicted given any knock-out configuration. The function F described in Eq.(7), under the constraint of minimal nodes to each element, would induce a complex graph representation, since it includes many clauses with multiple elements in them. Thus, the FIN is constructed in two stages: First, algebraically, we rewrite the function in a decomposed way such that each element appears solely, perhaps in more than one clause. Second, translate the function into a reduced network describing it, the FIN.

The PRR FIN (Figure 4) describes how each gene influences the Rad6 PRR pathway, and has two kinds of edges, non-weighted *connectivity edges* and weighted *influence edges*. The investigated genes are represented as binary nodes, whose state is determined according to the state of the genes, intact or knocked-out. Following is a detailed exposition of the specific construction of the PRR FIN. Based on the FIN and incorporating previous biological knowledge new functional insights can be obtained, as described in the rest of this section. Indeed, one should note that the FIN is not necessarily a unique representation, and its simplicity depends on the

complexity of the interactions between the genes in the network. It is not a general panacea, but in the specific case of the Rad6 pathway, we found it to be of much use.

In the first stage of the FIN construction, algebraic simplification, we search for common variables and rewrite the performance function, minimizing the number of clauses. This is done iteratively until there are no compound elements (elements composed of no more than one literal) left in the clauses. That is, if needed, we add new nodes representing Boolean functions of two or more literals. In our example, the first iteration is

$$(8) \quad F = e \cdot [.26 \cdot d + .2 \cdot c + .13 \cdot a \cdot d + .36 \cdot c \cdot d + .05 \cdot b \cdot d + .05 \cdot a \cdot c - .05 \cdot a \cdot c \cdot d],$$

and, assigning a new variable $\psi = c \cdot d$, we obtain

$$(9) \quad F = e \cdot [d \cdot (.26 + .13 \cdot a + .05 \cdot b) + c \cdot (.2 + .05 \cdot a) + \psi \cdot (.36 - .05 \cdot a)].$$

Since there are no more compound elements within the clauses we turn to the second, FIN construction, stage. Based on Eq.(9), we build the FIN where we start off from a function node, F , connecting it to the variables outside the external level parentheses, assigning the connection weights according to the corresponding coefficients. First we connect F to e , then we connect e to c , d , and ψ (in the FIN ψ will be represented as a Boolean AND node connected by connectivity edges to c and d), with the edges weighted according to Eq.(9), and so on for each clause. The resulting FIN is demonstrated in Figure 4.

Given a perturbation configuration the nodes of the FIN are either intact or knocked-out. Considering the edges only between intact nodes the expected UV sensitivity performance level can be calculated by summing up the weights on the influence (solid lines) which are in the same connected component with the function node, F (the connected component is based on both connectivity edges and influence edges). For example if we consider a mutant where both $REV3$ and $CTF18$ are deleted, the active nodes will be $ELG1$, $RAD24$, $RAD18$, which leaves us with two influence edges in the connected component of F , the edges $ELG1-RAD24$ and $RAD24-RAD18$. In this example the performance computation results in a performance level of 0.25.

The FIN construct yields a number of new insights and suggests answers for a few basic questions concerning the function of the three RLCs. Observe that there are three influence edges connected to F via $RAD18$, showing three main pathways in the FIN, where $RAD18$ is an essential gene in all of them. Incorporation of additional existing biological knowledge can lead to the fol-

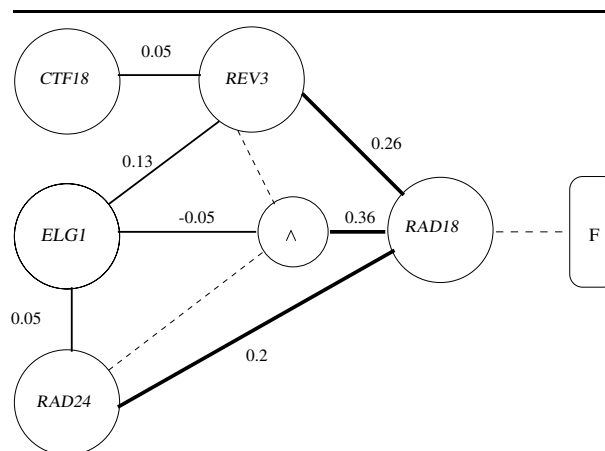


Figure 4. Functional Influence Network. The FIN is constructed based on Eq.(9). The dashed lines are connectivity edges which have no weights assigned to them, the solid lines are the influence edges with a weight assigned to each edge representing functional influence, and used together with the connectivity edges to predict the performance level F if given a multi-knockout configuration.

lowing conclusions:

1) *REV3-RAD18 pathway*: this pathway includes the DNA polymerase ζ encoded by $REV3$. $ELG1$ and $CTF18$ have positive contributions in this pathway suggesting that they possibly play a role loading the DNA polymerase as presumed. However, there are probably some extra genes that play an important role in this pathway in addition to these four genes, since this pathway without any RLC still encompasses 26% of the system's PRR performance! This suggests that there may be some additional DNA polymerase loaders beside those investigated (the edge $REV3-RAD18$ is not dependent on any of the three RLCs investigated, yet, it is known that a DNA polymerase loader is essential for the operation of the DNA polymerase).

2) *RAD24-RAD18 pathway*: As evident, $RAD24$ plays a positive role even without $REV3$. Hence, there is probably another polymerase involved in the PRR process. This pathway is enhanced by $ELG1$, suggesting that both $ELG1$ and $RAD24$ (each with their relative contribution) play a role loading the additional DNA polymerase.

3) *RAD24-REV3-RAD18 pathway*: this pathway combines both $RAD24$ and $REV3$ and is due to their

combined intactness. The pathway includes *RAD24* and *ELG1* as RLCs that possibly load the DNA polymerase ζ encoded by *REV3*.

In summary, the multi-knockout analysis of the Rad6 PRR system, as described in section 4 and this section, shows that each of the RLCs has a different magnitude of contribution to the PRR process, and describes how its contribution depends on other elements. Out of the three RLCs investigated, *RAD24* is the most important one. *RAD18* is a common essential gene in all the three pathways described, taking part in a number of alternative PRR pathways fitting its regulatory role. The analysis suggests that besides the three RLCs there are additional polymerase loading complexes in the yeast. Moreover, DNA polymerase ζ encoded by *REV3* is probably not the only polymerase involved in PRR (possibly there is more than one). Furthermore, the prediction is that this additional DNA polymerase will be independent of *CTF18* (additive interaction), very dependent on *RAD24* and with a moderate dependency on *ELG1*.

6. Discussion

The MSA is a novel method for system identification, based on a rigorous definition of the elements' contributions via the Shapley value. This paper presents the first study of MSA analysis of genetic multi-knockout data, and shows how it can be used to advance our understanding of the underlying genetic pathways. In the specific example shown here, we were capable of addressing pertinent questions raised in the literature. Demonstrating the importance of each RLC, describing when it is required and the pathways via which it exerts its functional influence. The analysis has raised new hypotheses concerning the existence of other components that have a causal role in PRR, yielding specific predictions as to their interactions with the different RLCs.

The MSA and its extensions form the first systematic approach addressing the challenge of analyzing multi-perturbation experiments in a rigorous manner. Although the genetic network described in this paper is indeed small, this will probably be the case of multi-knockout studies in the near future. Due to current technical constraints such multi-knockout experiments will include moderate size networks, which are very suitable for a MSA analysis as presented above. On a longer time scale novel techniques that are being developed, gradually will make large-scale multi-perturbation studies possible. New RNA interference (RNAi) and transposon mutagenesis studies will soon make it possible to undertake systematic genome-wide functional screens that examine the contribution of every gene to a biological process [5]. To date RNAi is limited to just a few elements

whose expression is concomitantly attenuated, but this is just the beginning [19]. A recent published study has already gone beyond a systematic single mutations research, studying double mutants in the yeast using Synthetic Genetic Array (SGA) analysis [18]. Such systematic functional techniques are fundamentally changing how biologists identify the molecular components that derive biological processes. When addressing such large-scale studies the MSA offers variants of estimation to approximate the Shapley value in a scalable manner, allowing for systematic function localization [12]. It should be noted that the basic tenets of our approach continue to hold even if one replaces the Shapley value with other semi-values.

Causal functional genomics methods as the MSA should be integrated in the future with existing global gene expression profiling methods. The information gained from DNA microarray analysis could be used to focus on the relevant target genes that require for further detailed perturbation analysis. Knowing that some set of genes is expressed under some particular conditions can narrow down the gene space investigated, and as previous studies have already shown, can lead to better system identification [10, 15].

Multi-perturbation studies are a necessity, and they are hence bound to take place, starting in the near future. The methods described in this paper are a harbinger of this new kind of studies, offering a novel and rigorous way of making sense out of them.

7. Acknowledgment

We thank Amnon Koren for the execution of the genetic experiments. We acknowledge the valuable contributions and suggestions made by Ya'acov Ritov and Isaac Meilijson, we also thank Doron Betel, Roy Varshavsky, Alon Keinan and Dudi Deutscher for invaluable critical comments during our manuscript preparations.

References

- [1] A. Barabasi and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- [2] S. Ben-Aroya, A. Koren, B. Liefshitz, R. Steinlauf, and M. Kupiec. *ELG1*, a yeast gene required for genome stability, forms a complex related to replication factor c. *Proc. Natl. Acad. Sci. USA*, 100(17):9906–11, 2003.
- [3] M. Brendel and R. Haynes. Interactions among genes controlling sensitivity to radiation and alkylation in yeast. *Mol. Gen. Genet.*, 125:197–216, 1973.

- [4] T. Broomfield, S. Hryciw and W. Xiao. DNA postreplication repair and mutagenesis in *Saccharomyces cerevisiae*. *Mutation Research*, 486(3):167–184, 2001.
- [5] A. Carpentar and D. Sabatini. Systematic genome-wide screens of gene function. *Nature Reviews Genetics*, 5:11–22, 2004.
- [6] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.
- [7] G. Giaever. *et al.*, Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 418:387–391, 2002.
- [8] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565, 1999.
- [9] Z. Gu, L. Steinmetz, X. Gu, C. Scharfe, R. Davis, and W. Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–6, 2003.
- [10] T. Ideker, V. Thorsson, and R. Karp. Discovery of regulatory interactions through perturbation: inference and experimental design. In *Proc. of the Pacif Symp. on Bio-computing*, pages 305–16, 2000.
- [11] A. Keinan, A. Kaufman, N. Sachs, C. C. Hilgetag, and E. Ruppin. Fair localization of function via multi-lesion analysis. *Special issue of the journal of Neuroinformatics, to appear*, 2004. <http://www.cns.tau.ac.il/new/msa.html>.
- [12] A. Keinan, B. Sandbank, C. C. Hilgetag, I. Meilijson, , and E. Ruppin. Fair attribution of functional contribution in artificial and biological networks. *Neural Computation*, 16(9), 2004.
- [13] J. Kim and S. MacNeill. Genome stability: a new member of the RFC family. *Current Biology*, 13(22):R873–R875, 2003.
- [14] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge: University of Cambridge Press, 2000.
- [15] D. Pe’er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(S215-S224), 2001.
- [16] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume II of *Annals of Mathematics Studies* 28, pages 307–317. Princeton University Press, Princeton, 1953.
- [17] L. Steinmetz and R. Davis. Maximizing the potential of functional genomics. *Nature Reviews Genetics*, 5(3):190–201, 2004.
- [18] A. Tong. *et al.*, Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, 2004.
- [19] E. Winzeler. *et al.*, Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285:901–906, 1999.