NC ms 2802

Fair Attribution of Functional Contribution in Artificial and Biological Networks

Alon Keinan[1], Ben Sandbank[1], Claus C. Hilgetag[2],

Isaac Meilijson[3], Eytan Ruppin[1,4]

[1]School of Computer Sciences, Tel-Aviv University, Tel-Aviv, Israel

{*keinanak,sandban,ruppin*}*@post.tau.ac.il* (tel: +972-3-6406528)

[2]School of Engineering and Science, International University Bremen, Bremen, Germany

c.hilgetag@iu-bremen.de (tel: +49-421-2003542)

[3]School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

isaco@post.tau.ac.il (tel: +972-3-6408826)

[4]School of Medicine, Tel-Aviv University, Tel-Aviv, Israel

Abstract

This paper presents the *Multi-perturbation Shapley value Analysis (MSA)*, an axiomatic, scalable and rigorous method for deducing causal function localization from multiple perturbations data. The MSA, based on fundamental concepts from game theory, accurately quantifies the contributions of network elements and their interactions, overcoming several shortcomings of previous function localization approaches. Its successful operation is demonstrated in both the analysis of a neurophysiological model and of reversible deactivation data. The MSA has a wide range of potential applications, including the analysis of reversible deactivation experiments, neuronal laser ablations and transcranial magnetic stimulation "virtual lesions", as well as in providing insight on the inner workings of computational models of neurophysiological systems.

1 Introduction

How is neural information processing to be understood? One of the fundamental challenges is to identify the individual roles of the network's elements, be they single neurons, neuronal assemblies or cortical regions, depending on the scale on which the system is analyzed. Localization of specific functions in the nervous system is conventionally done by recording neural activity during cognition and behavior, mainly using electrical recordings and functional neuroimaging techniques, and correlating the activations of the neural elements with different behavioral and functional observables. However, this correlation does not necessarily identify causality (Kosslyn, 1999). To allow the correct identification of the elements that are responsible for a given function, lesion studies have been traditionally employed in neuroscience, in which the functional performance is measured after different elements of the system are disabled. Most of the lesion investigations in neuroscience have been *single lesion* studies, in which only one element is disabled at a time (e.g. Squire, 1992; Farah, 1996). Such single lesions are very limited in their ability to reveal the significance of interacting elements, such as elements with functional overlap (Aharonov, Segev, Meilijson, & Ruppin, 2003).

Acknowledging that single lesions are insufficient for localizing functions in neural systems, Aharonov et al. (2003) have presented the Functional Contribution Analysis (FCA). The FCA analyzes a data set composed of numerous multiple lesions that are inflicted upon a neural system, along with

its performance scores in a given function. In each multiple lesion experiment composing the lesion data set, several elements are lesioned concurrently and the system's level of performance is measured. The FCA uses these data to yield a prediction of the system's performance level when a new, untested, multiple lesion damage is imposed on it. It further yields a quantification of the elements' contributions to the function studied as a set of values minimizing the prediction error (a *contribution* of an element to a function should measure its importance to the successful performance of the function).

The definition of the elements' contributions to each function within the FCA framework is an operative one, based on minimizing the prediction error within a predefined prediction model. *There is no inherent notion of correctness of the contributions found by this method.* In particular, there are incidents where several different contributions assignments to the elements yield the same minimum prediction error. In such cases the FCA algorithm may reach different solutions, providing accurate predictions in all cases, but yielding quite different contributions. Furthermore, since the contributions are calculated in an embedded manner with the prediction component, the statistical significance of the results is unknown. In large systems with numerous insignificant elements, testing the statistical significance of the elements' contributions is essential for locating the significant ones and focusing on them for the rest of the analysis. Devoid of this capacity, the FCA approach has serious limitations on the size of networks it can analyze.

This paper presents a new framework, the Multi-perturbation Shapley

value Analysis (MSA), addressing the same challenge of defining and calculating the contributions of network elements from a data set of multiple lesions or other types of perturbations and their corresponding performance scores. A set of multiple perturbation experiments is viewed as a *coalitional game*, borrowing relevant concepts from the field of game theory. Specifically, we define the desired set of contributions to be the *Shapley value* (Shapley, 1953), which stands for the *unique* fair division of the game's worth (the network's performance score when all elements are intact) among the different players (the network elements). While in traditional game theory the Shapley value is a theoretical tool which assumes full knowledge of the game, we have developed and studied methods to compute it approximately with high accuracy and efficiency from a relatively small set of multiple perturbation experiments. The MSA framework also quantifies the interactions between groups of elements, allowing for higher order descriptions of the network, further developing previous work on high-dimensional FCA (Segev, Aharonov, Meilijson, & Ruppin, 2003). The MSA has the ability to utilize a large spectrum of more powerful predictors, compared with the FCA, and to obtain a higher level of prediction accuracy. Lastly, it incorporates statistical tests of significance for removing unimportant elements, allowing for the analysis of large networks.

Besides biological experiments, the MSA can be utilized for the analysis of neurophysiological models. In recent years such models have been developed in large numbers and in different contexts, providing valuable insight

on the modeled systems. The availability of a computer simulation model for a system, however, does not imply that the inner-workings of that model are known. Analyzing such models with the MSA can extend their usefulness by allowing for a deeper understanding of their operation. We demonstrate the workings of the MSA for the analysis of the building block of a model of lamprey swimming controller (Ekeberg, 1993), successfully uncovering the neuronal mechanism underlying its oscillatory activity as well as the redundancies inherent in it. Additionally, the contributions of specific synapses to different characteristics of the oscillation, such as frequency and amplitude, are inferred and analyzed.

To test the applicability of our approach to the analysis of real biological data, we applied the MSA to reversible cooling deactivation experiments in cats. Our aim was to establish the contributions of parietal cortical and superior collicular sites to the brain function of spatial attention to auditory stimuli, as tested in an orienting paradigm (Lomber, Payne, & Cornwell, 2001). As we shall show, the MSA correctly identifies the elements' contributions and interactions, verifying quantitatively previous qualitative findings.

The remainder of this paper is organized as follows: Section 2 describes the concept of coalitional games and how it is utilized in the MSA framework for calculating the contribution of network elements, as well as the interactions between them. Section 3 demonstrates the application of the MSA to the neurophysiological model and section 4 presents an MSA of reversible deactivation experiments. Our results and possible future applications of the

MSA are discussed in section 5. The appendix illustrates the workings of the MSA on a toy problem, comparing it with single lesion analysis and the FCA.

2 The Multi-Perturbation Shapley Value Analysis

2.1 Theoretical Foundations

Given a system (network) consisting of many elements, we wish to ascribe to each element its contribution in carrying out the studied function. In single lesion analysis, an element's contribution is determined by comparing the system's performance in the intact state and when the element is perturbed. Such single lesions are very limited in their ability to reveal the significance of interacting elements. One obvious example is provided by two elements that exhibit a high degree of functional overlap, that is, *redundancy*: lesioning either element alone will not reveal its significance. Another classical example is that of the *"paradoxical" lesioning effect* (Sprague, 1966; Kapur, 1996). In this paradigmatic case, lesioning an element is harmful, but lesioning it when another specific element is lesioned is beneficial for performing a particular function. In such cases, and more generally in cases of *compound processing*, where the contribution of an element depends on the state of other elements, single lesion analysis is likely to be misleading, resulting in

erroneous conclusions.

The Multi-perturbation Shapley value Analysis (MSA) presented in this paper aims at quantifying the contribution of system's elements, while overcoming the inherent shortcomings of the single lesion approaches. The starting point of the MSA is a data set of a series of *multi-perturbation* experiments studying the system's performance in a certain function. In each such experiment, a different subset of the system's elements are perturbed concomitantly (denoting a *perturbation configuration*) and the system's performance following the perturbation in the function studied is measured. Given this data set, our main goal is to assign values that capture the elements contribution (importance) to the task in a fair and accurate manner. Note that this assignment is nothing but the classical functional localization goal in neuroscience, but now recast in a formal, multi-perturbation framework.

The basic observation underlying the solution presented in this paper to meet this goal is that the multi-perturbation setup is essentially equivalent to a coalitional game. That is, the system elements can be viewed as "players" in a game. The set of all elements which are left intact in a perturbation configuration can be viewed as a "coalition" of players. The performance of the system following the perturbation can then be viewed as the "worth" of that coalition of players in the game. Within such a framework, an intuitive notion of a player's importance (or contribution) should capture the worth of coalitions containing it (i.e. the system's performance when the corresponding element is intact), relative to the worth of coalitions which do

not (i.e. relative to the system's performance when this element, perhaps among others, is perturbed). This intuitive equivalence, presented formally below, enables us to harness the pertaining game theoretical tools to solve the problem of function localization in biological systems.

Using the terminology of game theory, let a *coalitional game* be defined by a pair (N, v), where $N = \{1, \ldots, n\}$ 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, such that $v(\phi) = 0$.[1] In the context of multi-perturbations, N denotes the set of all elements, and for each $S \subseteq N$, $v(S)$ denotes the performance measured under the perturbation configuration in which all the elements in S are intact and the rest are perturbed.

A *payoff profile* of a coalitional game is the assignment of a payoff to each of the players. A *value* is a function that assigns a unique payoff profile to a coalitional game. It is *efficient* if the sum of the components of the payoff profile assigned is $v(N)$. That is, an efficient value divides the overall game's worth (the networks' performance when all elements are intact) between the different players (the network elements). A value that captures the importance of the different players may serve as a basis for quantifying, in the context of multi-perturbations, the contributions of the system's elements.

The definite value in game theory and economics for this type of coalitional game is the *Shapley value* (Shapley, 1953), defined as follows. Let the

[1]This type of game is most commonly referred to as a *coalitional game with transferable payoff*.

marginal importance of player i to a coalition S, with $i \notin S$, be

$$\Delta_i(S) = v(S \cup \{i\}) - v(S). \tag{1}$$

Then, the Shapley value is defined by the payoff

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

of each player $i \in N$, where $\mathcal{R}$ is the set of all $n!$ orderings of N and $S_i(R)$ is the set of players preceding i in the ordering R. The Shapley value can be interpreted as follows: Suppose that all the players are arranged in some order, all orders being equally likely. Then $\gamma_i(N, v)$ is the expected marginal importance of player i to the set of players who precede him. The Shapley value is efficient since the sum of the marginal importance of all players is $v(N)$ in any ordering.

An alternative view of the Shapley value is based on the notion of balanced contributions. For each coalition S, the *subgame* (S, v^S) of (N, v) is defined to be the game in which $v^S(T) = v(T)$ for any $T \subseteq S$. A value Ψ satisfies the *balanced contributions property* if for every coalitional game (N, v) and for every $i, j \in N$

$$\Psi_i(N, v) - \Psi_i(N \setminus \{j\}, v^{N\setminus\{j\}}) = \Psi_j(N, v) - \Psi_j(N \setminus \{i\}, v^{N\setminus\{i\}}), \tag{3}$$

meaning that the change in the value of player i when player j is excluded from the game is equal to the change in the value of player j when player i is excluded. This property implies that *objections* made by any player to any other regarding the division are exactly balanced by the *counterobjections*.

The unique efficient value that satisfies the balanced contributions property is the Shapley value (Myerson, 1977, 1980).

The Shapley value also has an axiomatic foundation. Let player i be a *null player* in v if $\Delta_i(S) = 0$ for every coalition S ($i \notin S$). Players i and j are *interchangeable* in v if $\Delta_i(S) = \Delta_j(S)$ for every coalition S that contains neither i nor j. Using these basic definitions, *the Shapley value is the only efficient value that satisfies the three following axioms, further pointing to its uniqueness* (Shapley, 1953):

Axiom 1 *(Symmetry) If i and j are interchangeable in game v then $\Psi_i(v) = \Psi_j(v)$.*

Intuitively, this axiom states that the value should not be affected by a mere change in the players' "names".

Axiom 2 *(Null player property) If i is a null player in game v then $\Psi_i(v) = 0$.*

This axiom sets the baseline of the value to be zero for a player whose marginal importance is always zero.

Axiom 3 *(Additivity) For any two games v and w on a set N of players, $\Psi_i(v+w) = \Psi_i(v) + \Psi_i(w)$ for all $i \in N$, where $v+w$ is the game defined by $(v+w)(S) = v(S) + w(S)$.*

This last axiom constrains the value to be consistent in the space of all games.

In the fifty years since its construction, the Shapley value as a unique fair solution has been successfully used in many fields. Probably the most important application is in cost allocation, where the cost of providing a service should be shared among the different receivers of that service. This application was first suggested by Shubik (1962), and the theory was later developed by many authors (e.g. Roth (1979) and Billera, Heath, and Raanan (1978)). This use of the Shapley value has received recent attention in the context of sharing the cost of multicast routing (Feigenbaum, Papadimitriou, & Shenker, 2001). In epidemiology, the Shapley value has been utilized as a mean to quantify the population impact of exposure factors on a disease load (Gefeller, Land, & Eide, 1998). Other fields where the Shapley value has been used include, among others, politics (starting from the *strategic voting* framework introduced by Shapley and Shubik (1954)), international environmental problems and economic theory (see Shubik (1985) for discussion and references).

The MSA, given a data set of multi-perturbations, uses the Shapley value as the unique fair division of the network's performance between the different elements, assigning to each element its contribution, as its average importance to the function in question[2]. The higher an element's contribution

[2]Since $v(\phi) = 0$ does not necessarily hold in practice, as it depends on the performance measurement definition, Shapley value efficiency transcribes to the property according to which the sum of the contributions assigned to all the elements equals $v(N) - v(\phi)$.

according to the Shapley value, the larger is the part it causally plays in the successful performance of the function[3]. This set of contributions is a unique solution, in contrast to the multiplicity of possible contributions assignments that may plague an error minimization approach like the FCA.

In the context of multi-perturbations, Axiom 1 of the Shapley value formulation entails that if two elements have the same importance in all perturbation configurations, their contributions will be identical. Axiom 2 assures that an element that has no effect in any perturbation configuration will be assigned a zero contribution. Axiom 3 indicates that if two separate functions are performed by the network, such that the overall performance of the network in all multi-perturbation configurations is defined to be equal to the sum of the performances in the two functions, then the total contribution assigned to each element for both functions will be equal to the sum of its individual contributions to each of the two functions.

It should be noted that other game-theoretical values can be used instead of the Shapley value within the MSA framework. Specifically, Banzhaf (1965) has suggested an analogue of the Shapley value and Dubey, Neyman, and Weber (1981) have later generalized the Shapley value to a whole family of semi-values, all satisfying the three axioms but without the efficiency property, a natural requirement for describing fair divisions.

[3]Since no limitations are enforced on the shape of v, a negative contribution is possible, indicating that the element hinders, on the average, the function's performance.

Once a game is defined, its Shapley value is uniquely determined. Yet, employing different perturbation methods may obviously result in different values of v and as a consequence, different Shapley values. Aharonov et al. (2003) and Keinan, Meilijson, and Ruppin (2003) discuss different perturbation methods within the framework of neurally-driven evolved autonomous agents and the effects that they may have on the elements' contributions found.

The workings of the MSA and the intuitive sense of its "fairness" are demonstrated in a simple, toy-like example in Appendix A, for the interested reader.

2.2 Methods

2.2.1 Full Information Calculation: The Original Shapley Value

In an ideal scenario, in which the full set of 2^n perturbation configurations along with the performance measurement for each is given, the Shapley value may be straightforwardly calculated using equation (2), where the summation runs over all $n!$ orderings of N. Equivalently, the Shapley value can be computed as a summation over all 2^n configurations, properly weighted by the number of possible orderings of the elements,

$$\gamma_i(N, v) = \frac{1}{n!} \sum_{S \subseteq N \setminus \{i\}} \Delta_i(S) \cdot |S|! \cdot (n - |S| - 1)!. \tag{4}$$

Substituting according to equation (1) results in

$$\gamma_i(N,v) = \frac{1}{n!} \sum_{S\subseteq N, i\in S} v(S)\cdot(|S|-1)!\cdot(n-|S|)! - \frac{1}{n!} \sum_{S\subseteq N, i\notin S} v(S)\cdot(|S|)!\cdot(n-|S|-1)!, \quad (5)$$

where each configuration S contributes a summand to either one of the two sums, depending on whether element i is perturbed or intact in S. Thus, the Shapley value calculation consists of going through all perturbation configurations and calculating for each element the two sums in the above equation.

2.2.2 Predicted and Estimated Shapley Value

Obviously, the full set of all perturbation configurations required for the calculation of the Shapley value is often not available. In such cases, one may train a predictor, using the given subset of the configurations, to predict the performance levels of new, unseen perturbation configurations. Given such a predictor, the outcomes of all multi-perturbation experiments may be extracted and a *predicted Shapley value* can be calculated on this data according to equation (5). The functional uncoupling (as opposed to the FCA) between the predictor component and the contributions calculation *provides the MSA with the freedom to use any predictor.*

When the space of multi-perturbations is too large to enumerate all configurations in a tractable manner, the MSA can sample orderings and calculate an unbiased estimator of each element's contribution as its average marginal

importance (equation (1)) over all sampled orderings. Additionally, an estimator of the standard deviation of this Shapley value estimator can be obtained, allowing the construction of confidence intervals for the contribution of each of the elements, as well as the testing of statistical hypotheses on whether the contribution of a certain element equals a given value (e.g., zero or $1/n$).

The multi-perturbation experiments that should be performed in the estimated Shapley value method are dictated by the sampled orderings. But, in many biological applications a data set consisting of multi-perturbation experiments is pre-given and should be analyzed. In such cases, the MSA employs an estimated predicted Shapley computation: A performance predictor is trained using the given set of perturbation configurations and serves as an oracle supplying performance predictions for any perturbation configuration as dictated by the sampling.

Based on the estimation method, the MSA provides another method for handling networks with a large number of elements. The *two-phase MSA procedure* is motivated by the observation that often only a small fraction of the possibly large number of elements significantly contributes to the specific function being analyzed. Based on a small sample, the method's first phase finds those elements with significant contributions. Then, the second phase focuses on finding the accurate contributions of those significant elements. This phase may use the same small sample from the first phase, while focusing on a much smaller perturbations space, thus allowing for faster training and

for an increased scalability.

The different MSA prediction and estimation variants, intended for use in various neuroscience applications, have been tested in the theoretical modeling framework of neurally-driven evolved autonomous agents, where their accuracy, efficiency and scalability were established. A detailed account of these results and of the above estimation and two-phase methods used in obtaining them will be published elsewhere. This paper presents analyses that utilize the full information Shapley value calculation (section 3 and appendix A) and the predicted Shapley value (section 4), demonstrating the applicability of the concepts underlying the MSA.

2.3 Two-Dimensional MSA

The Shapley value serves as a summary of the game, indicating the average marginal importance of an element over all possible elements orderings. For complex networks, where the importance of an element strongly depends on the state (perturbed or intact) of other elements, a higher order description may be necessary in order to capture sets of elements with significant interactions. For example, when two elements exhibit a high degree of functional overlap, that is, *redundancy*, it is necessary to capture this interaction, aside from the average importance of each element. Such *high-dimensional analysis* provides further insights into the network's functional organization.

We focus on the description of two-dimensional interactions. A natural definition of the interaction between a pair of elements is as follows: Let

$\gamma_{i,\bar{j}} = \gamma_i(N \setminus \{j\}, v^{N\setminus\{j\}})$ be the Shapley value of element i in the subgame of all elements without element j. Intuitively, this is the average marginal importance of element i when element j is perturbed. Let us now define the coalitional game (M, v^M), where $M = N \setminus \{i,j\} \cup \{(i,j)\}$ ((i,j) is a new compound element) and $v^M(S)$, for $S \subseteq M$, is defined by

$$v^M(S) = \begin{cases} v(S) & : \quad (i,j) \notin S \\ v(S \setminus \{(i,j)\} \cup \{i,j\}) & : \quad (i,j) \in S \end{cases} \tag{6}$$

where v is the characteristic function of the original game with elements N. Then, $\gamma_{i,j} = \gamma_{(i,j)}(M, v^M)$, the Shapley value of element (i,j) in this game, is the average marginal importance of elements i and j when jointly added to a configuration. The two-dimensional interaction between element i and element j, $j \neq i$, is then defined as

$$I_{i,j} = \gamma_{i,j} - \gamma_{i,\bar{j}} - \gamma_{j,\bar{i}} \tag{7}$$

which quantifies *how much the average marginal importance of the two elements together is larger (or smaller) than the sum of the average marginal importance of each of them when the other one is perturbed.* Intuitively, this symmetric definition ($I_{i,j} = I_{j,i}$) states how much "the whole is greater than the sum of its parts" (*synergism*), where the whole is the pair of elements. In cases where the whole is smaller than the sum of its parts, that is, when the two elements exhibit functional overlap, the interaction is negative (*antagonism*). This two-dimensional interaction definition coincides with the Shapley interaction index which is a more general measure for the interaction among any group of players (Grabisch & Roubens, 1999).

The MSA can classify the type of interaction between each pair even further: By definition, $\gamma_{i,\bar{j}}$ is the average marginal importance of element i when element j is perturbed. Based on equation (7), $\gamma_{i,\bar{j}} + I_{i,j}$ is the average marginal importance of element i when element j is intact. When both $\gamma_{i,\bar{j}}$ and $\gamma_{i,\bar{j}} + I_{i,j}$ are positive, element i's contribution is positive, irrespective of whether element j is perturbed or intact. When both are negative, element i hinders the performance, irrespective of the state of element j. In cases where the two measures have inverted signs, we define the contribution of element i as *j-modulated*. The interaction is defined as *positive modulated* when $\gamma_{i,\bar{j}}$ is negative, while $\gamma_{i,\bar{j}} + I_{i,j}$ is positive, causing a "paradoxical" effect. We define the interaction as *negative modulated* when the former is positive while the latter is negative. The interaction of j with respect to i may be categorized in a similar way, yielding a full description of the type of interaction between the pair. Classical "paradoxical" lesioning effects, for instance, of the kind reported in the neuroscience literature (Sprague, 1966; Kapur, 1996) are defined when both elements exhibit positive modulation with respect to one another. As evident, the rigorous definition of the type of interaction presented in this section relies on an average interaction over all perturbation configurations. Thus, it does not necessarily coincide with the type of interaction found by using only single perturbations and a double perturbation of the pair, as conventionally described in the neuroscience literature.

3 MSA of a Lamprey Segment

Computer simulation of neuronal models is becoming an increasingly important tool in neuroscience. Modeling enables exploring the relation between the properties of single elements and the emergent properties of the system as a whole. Neural network models are usually highly complex networks composed of many interacting neural elements. In many cases it is possible to determine an architecture and find parameter values which provide a good fit between the behavior of the model and that of the real physiological system. However, it is usually very difficult to assess each element's function within the model system, that is, to understand *how* the elements interact with each other to create the emergent properties that the system displays as a whole. In fact, there is a wide gap between the large body of work invested in constructing neural models, and the paucity of efforts to analyze and understand the workings of these models in a *rigorous manner*.

Neuronal models may be easily manipulated, allowing to measure their behavior under different perturbations, which makes them amenable to analysis. Using the MSA it is straightforward to compute the contribution of each element to each function the model system simulates and the interactions between different elements in each of the different functions. Insights from such an analysis may be utilized for further study of the system being modeled. In the remainder of this section we demonstrate the application of the MSA to the analysis of a fish swimming model. Section 3.1 briefly

introduces the model and section 3.2 presents the results obtained by the MSA.

3.1 The Oscillatory Segment Model

The lamprey is one of the earliest and most primitive vertebrates. Much like an eel, it swims by propagating an undulation along its body, from head to tail. The neural network located in its spinal cord which controls the motion has been extensively studied (see, for example, Brodin, Grillner, and Rovainen (1985); Buchanan and Grillner (1987); Grillner, Wallen, and Brodin (1991); Grillner et al. (1995); Buchanan (1999); Parker and Grillner (2000); Buchanan (2001). For a recent review see Buchanan (2002)). Physiological experiments on isolated spinal cords have shown that this network is a central pattern generator (CPG), as it is able to generate patterns of oscillations without oscillatory input from either the brain or sensory feedback. Because small parts of the spinal cord (up to 2 segments) can be isolated and still produce oscillations when subjected to an excitatory bath, the CPG is thought of as an interconnection of local (segmental) oscillators. The complete controller is formed of approximately 100 identical copies of segmental oscillators, linked together to form a chain.

The CPG controlling the lamprey's swimming movement has been modeled at several levels of abstraction. The model analyzed in this section is a connectionist model of the swimming controller developed by Ekeberg (1993). This model demonstrates that the inter-neuron connectivity is in it-

self sufficient for generating much of the lamprey's range of motions, without resorting to complicated neuronal mechanisms. Following the real lamprey, the model is essentially a chain of identical segments, each of which capable of independently producing oscillations. Being the basic building block of the model, the properties of the individual segment determine to a large extent the overall behavior of the model. We therefore focus our analysis on the individual segment.

Figure 1A shows a schematic diagram of the neural network of a segmental oscillator. It is a symmetrical network, composed of 4 neurons on each side. The neurons are modeled as non-spiking leaky integrators, where each unit can be regarded as a representative of a large population of functionally similar (spiking) neurons, its output representing the mean firing rate of the population. There are four types of neurons: The motoneurons (MN) provide the output to the muscles; the excitatory interneurons (EIN) maintain the ipsilateral activity by exciting all of the neurons on their side of the segment; the contralateral inhibitory neurons (CCIN) suppress the contralateral activity; and the lateral inhibitory neurons (LIN) suppress the ipsilateral CCIN. When this network is initialized to an asymmetric state (with all neurons on one side excited) and a constant external bias is applied to all neurons (interpreted within the model as input from the lamprey's brain stem), the network exhibits a stable pattern of oscillations with the left and right motoneurons out of phase, as shown in figure 1B. When a hundred such segments are linked in a long chain as in the full model, these local

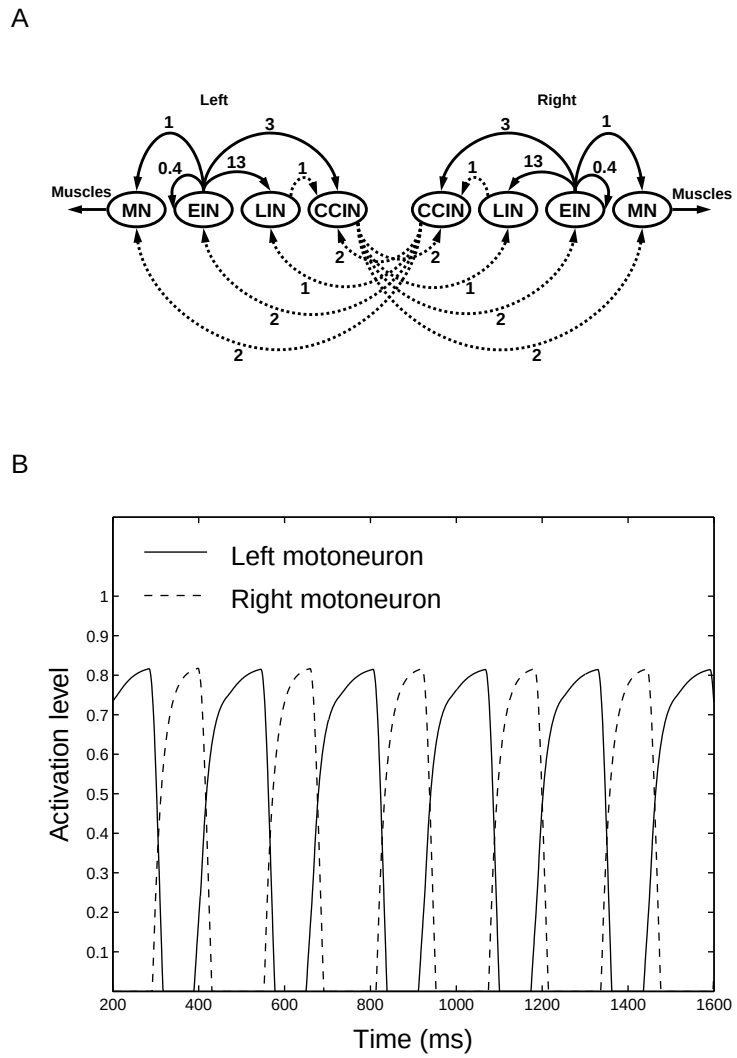


Figure 1: *The lamprey's segmental oscillator.* A. The neural network controlling the segmental oscillator. Solid lines indicate excitatory synapses, dashed lines are inhibitory synapses. The number next to each synapse denotes the synaptic weight. The connections leading from the motoneurons to the muscles are not included in the model and are shown for illustration purposes only. B. The activation levels of the left- (solid line) and right- (dashed line) motoneurons, after the network has been initialized to an asymmetric state with all left neurons maximally active. The figure focuses on the behavior of the system after it stabilizes, starting from the 200th millisecond.

oscillations give rise to coordinated waves of activity which travel along the chain (corresponding to the lamprey's spinal cord), producing the lamprey's swimming motion.

3.2 Segmental Analysis

In order to understand how the connectivity between the neurons gives rise to the behavior of the segmental network as a whole, the analysis was conducted on the level of the synapses. Because the neural network is completely left-right symmetrical, and its task is also inherently symmetrical, the basic neural element in our analysis is a symmetric pair of synapses. That is, each perturbation configuration defines a set of "synapses" to be perturbed, where each perturbation of a "synapse" denotes the simultaneous perturbation of both the left and right corresponding synapses. As a consequence, the MSA quantifies the importance of each such pair of synapses. Throughout this section, the shorter term "synapse" is used to refer to a synaptic pair.

3.2.1 The Oscillation Generating Mechanism

The first experiment aims to discover which synapses are important in producing the pattern of oscillations in the motoneurons, and hence for generating the motion that propels the lamprey. From viewing the structure of the full network it is not clear how these oscillations are generated and maintained. Hence, we utilize the MSA to uncover this mechanism: All possible multi-perturbation configurations are inflicted upon the model, test-

ing whether the motoneurons exhibit stable oscillations. The perturbation method used is clamping the activation level of a synapse to its mean.[4] This allows the MSA to quantify the extent to which the oscillations mediated by a given synapse contribute to the existence of oscillations in the motoneurons, without affecting the level of bias of each neuron. Viewing each model neuron as a representative of a large population of spiking neurons, this perturbation method can be shown to be equivalent to using stochastic lesioning (Aharonov et al., 2003) on the level of the individual synapses connecting different neuronal populations.[5]

Figure 2A presents the contributions (Shapley value) of the synapses, based on the full information available. Evidently, out of the nine synapses in the model, only five contribute to the generation of oscillations, while the others have a vanishing contribution. Figure 2B shows the "oscillatory backbone" of the network, including only those 5 synapses. Viewing just these synapses, it is possible to understand the mechanism generating the oscillations: Essentially, when the left CCIN becomes active, it inhibits the right EIN. The reduced activity of the EIN causes a reduction in the activity of the right LIN. This, in turn, causes increased activation in right CCIN,

[4]The activation level of a synapse is used here to refer to the activation level of the presynaptic neuron, as mediated by the synapse to the postsynaptic neuron.

[5]Stochastic lesioning is performed by randomizing the firing pattern of the perturbed element without changing its mean activation level.

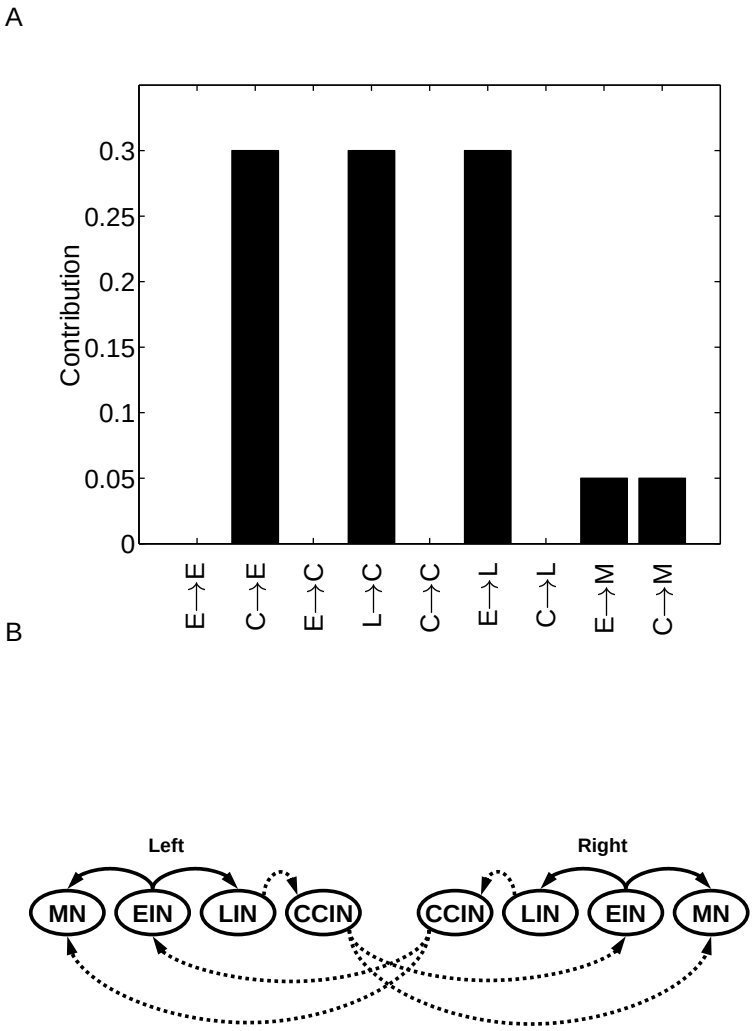


Figure 2: *Results of the MSA on the task of motoneuronal oscillations.* A. The contributions of the 9 synapses in the network. Synapses are presented in the form presynaptic neuron→postsynaptic neuron. Due to space considerations only the first letter of each neuron is displayed (e.g. E stands for the EIN). B. The "oscillatory-backbone" of the network, composed of the 5 synapses with non-zero contributions. As before, solid lines indicate excitatory synapses and dashed lines indicate inhibitory synapses.

which inhibits the left EIN, and so on. The constant bias to the network is what causes the EIN's activation level to swing back up when it becomes too low, thus maintaining the process. The two remaining synapses, CCIN→MN and EIN→MN mediate the oscillations to the motoneurons.

The MSA also quantifies the interaction between pairs of synapses. Specifically, it reveals a functional overlap between the CCIN→MN and EIN→MN synapses, expressed by a strongly negative interaction ($I_{EIN\to MN,CCIN\to MN} = -0.25$). The positive contribution of each of the two synapses when the other is perturbed ($\gamma_{EIN\to MN,\overline{CCIN\to MN}} = \gamma_{CCIN\to MN,\overline{EIN\to MN}} = 0.25$) and the zero contribution of each of them when the other is intact indicate that perturbing one of the synapses nullifies the oscillations in the MNs if and only if the other is perturbed. In other words, the two synapses are totally redundant with respect to each other. In contradistinction, every other pair of synapses in the backbone exhibit a positive interaction (synergism) indicating that the two synapses comprising it cooperate and rely on each other in order to perform this function. Thus, the MSA successfully uncovers the synapses generating the oscillations, including those that are backed up by other synapses, in which case it also identifies the type of interaction between them.

It should be noted that this description of the dynamics underlying the network's oscillatory activity could not have been obtained by examining the structure of the original network, as one might postulate that a different set of synapses is important for this task. For example, the CCIN→EIN and

EIN→LIN synapses might be replaced by a single CCIN→LIN synapse to the same effect. It should also be emphasized that using only single perturbations to analyze the model would not have revealed the importance of the CCIN→MN and EIN→MN synapses, since if only one of them is perturbed, the oscillations are still preserved as they are mediated to the motoneurons by the remaining synapse. Using multi-perturbation experiments, however, overcomes the intrinsic limitations of single-perturbation experiments and fully reveals the workings of the network.

3.2.2 Oscillation Characteristics

The results of the previous section suggest that only 5 synapses generate the oscillations in the network, while the others may be perturbed without affecting their existence. This, however, does not mean that the other synapses play no role in determining other aspects of the system's behavior. It is also possible that the 5 synapses participating in the "oscillatory-backbone" of the network contribute to other functions as well. To examine this issue further, we perform a finer analysis which identifies and quantifies the synaptic contributions to four characteristics of the network's oscillatory behavior. The first three characteristics pertain to the difference between the left and the right motoneuron's activity level during a time span of 3 seconds. This difference is in itself oscillatory in the intact network, signifying in each moment the extent to which the left muscle innervated by the segment's motoneuron is more stimulated than the right one. The three characteristics we measure

are the amplitude and frequency of this oscillation, and the standard deviation of the difference in activity around its mean. The fourth characteristic examined is the degree to which the left and right motoneurons are out of phase with each other.

In each multi-perturbation configuration the network's performance is measured in each of these four characteristics, interpreted as functions within the MSA framework, and the contribution of each of the elements to each function is quantified separately by calculating the corresponding Shapley value using full information. This type of analysis requires a more delicate type of perturbation, such that the oscillations in the network are preserved under all perturbation configurations[6]. To this end, we use a perturbation method which only slightly shifts a perturbed synapse's activation level towards its mean. Thus, oscillations mediated by the perturbed synapse are somewhat attenuated, but not wholly negated. If the magnitude of the shift is small enough, the network retains its oscillatory behavior under all perturbation configurations. The selection of this perturbation method can be justified if we recall that a single neuron stands for a large population of spiking neurons, and hence the perturbation's overall impact is equivalent to applying stochastic informational lesioning (Keinan et al., 2003) to the

[6]Another reason for using more delicate perturbations is that they allow to evaluate the performance of the system closer to its true, intact behavior, and thus gain a more precise sense of the contribution of each synapse (see Keinan et al. (2003) for more on this approach to perturbation).

individual synapses linking different populations[7].

Several observations can immediately be made on the results of the analysis, shown in Figure 3. First, synapses that contribute nothing to the oscillation-generating function, such as the CCIN→LIN synapse, do make significant contributions to other tasks. The opposite is also true, as synapses in the oscillation-generating backbone make vanishing contributions to other functions. Second, a given synapse may have a positive contribution to one function, while effectively hindering another. For example, the CCIN→CCIN synapse has a strong positive contribution to the left-right firing anti-phase, but a negative contribution to the frequency of the oscillations. This suggests that the network is the result of compromising a number of contradicting demands. Lastly, different functions are localized to a different degree. On the one hand, the oscillation amplitude function is very localized, having a single synapse with a contribution of almost 0.9. On the other hand, the frequency function is more distributed, as four synapses have contributions greater than 0.1 in absolute value.

An analysis of the distribution of contributions to each of the functions may provide further understanding of the network's behavior. As an example we discuss the results for the amplitude function. Evidently, the CCIN→MN

[7]Informational lesioning views a lesioned element as a noisy channel and enables the application of different levels of perturbation, corresponding to different magnitudes of noise in the channel, while maintaining the element's mean activation level.

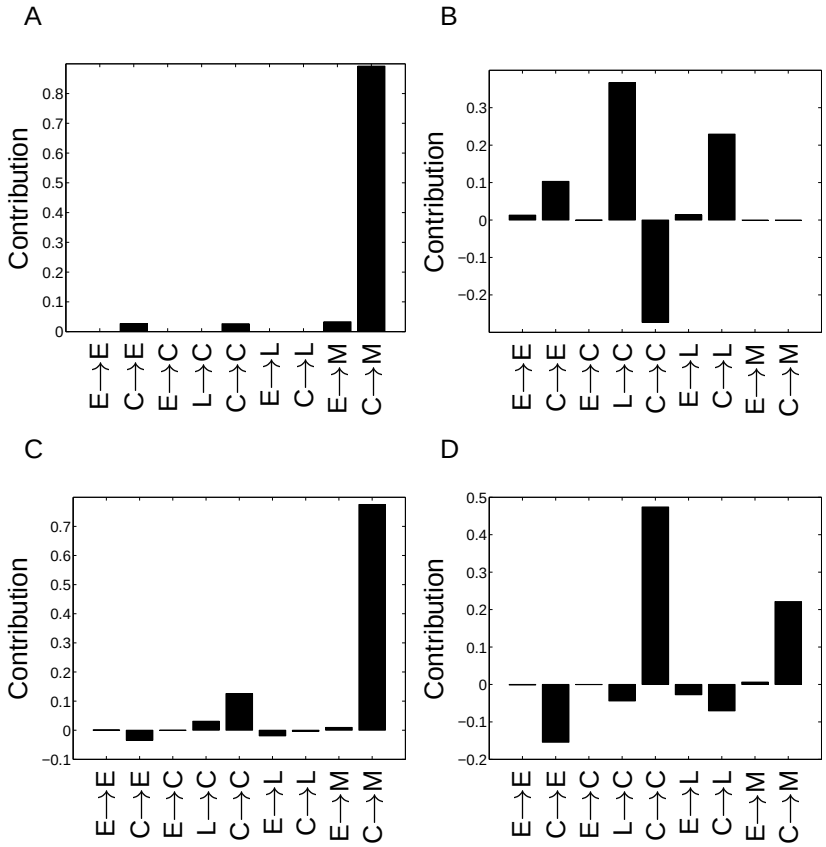


Figure 3: *Results of the MSA with different oscillation-characteristics taken as functions (see main text).* Shown are the contributions to the oscillation's amplitude (A), frequency (B), the standard deviation of the difference between the activity of the left and right motoneurons (C), and to the degree to which the left and right motoneurons' firing is in anti-phase (D). The results are normalized such that the sum of the absolute values of the contributions in each function equals 1.

synapse is the only one with a significant contribution to the amplitude of the MN's oscillations (Figure 3A). Interestingly, while the results of section 3.2.1 have shown that the CCIN→MN and the EIN→MN synapses are totally redundant with respect to one another in generating and maintaining the oscillations, the results of the current analysis reveal that the CCIN→MN synapse influences the activity of the MN to a far greater degree than the EIN→MN. This synapse is indeed sufficient for preserving the oscillations in the MN when the CCIN→MN synapse is clamped to its mean activation value, but these would be of much lower amplitude than when the CCIN→MN synapse is intact. Another conclusion that can be drawn from these results is that the amplitude of the oscillations of the CCIN is rather robust to perturbations to synapses. If it were not true and perturbing some synapse had a large impact on the amplitude of the CCIN's oscillations, it would also have affected the amplitude of the MN, and hence that synapse would have had a large contribution to this function. Because the MSA reveals no such contributions, the conclusion follows.

The contributions to the frequency function (Figure 3B) demonstrate how the MSA infers the true importance of neural elements, which in this simple network and for this function can be at least partially deduced through an examination of the network's structure. We focus here on the two synapses that have non-vanishing contributions to the frequency function and that do not belong to the "oscillatory-backbone", namely CCIN→CCIN and CCIN→LIN. The CCIN→CCIN's strong negative contribution suggests that on the aver-

age, when it is perturbed, the frequency of the MN's oscillations is increased. This is due to the fact that as the left side of the segment becomes more excited the CCIN→CCIN synapse inhibits the activity of the right CCIN. As this CCIN becomes more inhibited, its inhibitory influence on the overall activity of the left side decreases, causing a further increase in the activity of the left side. Therefore, this synapse indeed serves to prolong the bursts of activity, reducing the frequency of the oscillations. In contrast, the CCIN→LIN synapse has a positive contribution to the MN's frequency of oscillations, which means that, on the average, when it is perturbed, the frequency is decreased. The dynamics underlying this effect are as follows: As the left side of the segment becomes more excited, the left CCIN inhibits the right LIN, reducing the inhibition effect on the right CCIN. As the right CCIN becomes more active, it serves to terminate the left side burst. As mentioned in section 3.2.1, this synapse could in principle participate in the oscillatory backbone together with (or instead of) the pair of synapses CCIN→EIN and EIN→LIN. But while the CCIN→LIN synapse is not necessary nor sufficient for generating oscillations in the network, this finer analysis reveals that it is important in modulating the frequency of those oscillations. Clearly, the MSA provides interesting insights to the identification of neural function, even in this relatively simple example.

In summary, this section presents and analyzes the basic building block of a model for the neural network responsible for lamprey swimming. Using the MSA it is possible to understand how the oscillations are generated and

maintained in the network, which could not be inferred from the network's structure alone. The MSA succeeds where single perturbation analysis fails and detects the importance of synapses that are completely backed up by other synapses, in which case it also correctly identifies the type of interaction between them and quantifies it. Lastly, using the MSA it is possible to find the contribution of each synapse to various characteristics of the oscillations, revealing different levels of functional localization and providing further insight into the workings of the network.

4 MSA of Reversible Deactivation Experiments

To test the applicability of our approach to the analysis of biological "wetware" network data, we applied the MSA to data from reversible cooling deactivation experiments in the cat. Specifically, we investigated the brain localization of spatial attention to auditory stimuli (based on the orienting paradigm described in Lomber, Payne, Hilgetag, & Rushmore, 2002). Spatial attention is an essential brain function in many species, including humans, that is underlying several other aspects of sensory perception, cognition, and behavior. While attentional mechanisms proceed efficiently, automatically and inconspicuously in the intact brain, perturbation of these mechanism can lead to dramatic behavioral impairment. So-called neglect patients, for instance, have great difficulties, or even fail, to reorient their attention to spatial locations after suffering specific unilateral brain lesions,

with resulting severe deficits of sensory (e.g., visual, auditory) perception and cognition (Vallar, 1998). From the perspective of systems neuroscience, attentional mechanisms are particularly interesting, because this function is known to be widely distributed in the brain. Moreover, lesions in the attentional network have resulted in "paradoxical" effects (Kapur, 1996), in which the deactivation of some elements results in a better-than-normal performance (Hilgetag, Theoret, & Pascual-Leone, 2001) or reversed behavioral deficits resulting from earlier lesions (e.g., Sprague, 1966). Such effects challenge traditional approaches for lesion analysis and provide an ideal testbed for novel formal analysis approaches, such as the MSA.

In the reversible deactivation experiments analyzed here, auditory stimulus detection and orienting responses in intact and reversibly lesioned cats were tested in a semi-circular arena, in which small speakers were positioned bilaterally from midline (0) to 90 degrees eccentricity, at 15 degree intervals. The animals were first trained to stand at the center of the apparatus, facing and attending to the 0 degrees position speaker presenting a white noise hiss. Their subsequent task was to detect and orient toward noise coming from one of the peripheral speakers. They were rewarded with moist (high incentive) or dry (low incentive) cat food depending on whether they correctly approached the peripheral stimulus or the default central position, respectively. After the animals attained a stable near-perfect baseline performance, cryoloops were surgically implanted over parietal cortical (pMS) and collicular (SC) target structures, following an established standard procedure (Lomber, Payne, &

Horel, 1999). Because these regions are found in both halves of the brain, altogether four candidate target sites (pMS_L, pMS_R, SC_L and SC_R) were deactivated, one or two of them in each experiment. Once baseline performance levels were reestablished, the animals were tested during reversible cooling deactivation of unilateral and bilateral cortical and collicular sites (for further details see Lomber et al. (2001)). The technique allows for selective deactivation of just the superficial cortical or collicular layers, by keeping the cooling temperature to a level at which the deeper layers are still warm and active. Deactivation of the deeper layers, on the other hand, also requires deactivation of the superficial ones, due to the placement of the cryoloops on top of the cortical or collicular tissue. Nineteen single and multi-lesion experiments were performed (mostly as control experiments for the deactivation of visual structures, but also in their own right, e.g., Lomber et al., 2001) and another 14 lesion configurations were deduced by assuming mirror-symmetric effects resulting from lesions of the two hemispheres, yielding data for a total of 33 single and multi-lesion experiments, out of the 256 possible configurations.

Figure 4A shows the predicted Shapley value (section 2.2.2) of the different regions involved in the experiments, using Projection Pursuit Regression (PPR) (Friedman & Stuetzle, 1981) for prediction, trained using the 33 configurations. It is evident that only regions SC_R-deep and SC_L-deep play a role in determining the auditory attentional performance. Both have a contribution equal to half the overall predicted performance of the system (0.2). Due to the outlined limitation of the experimental procedure, when a

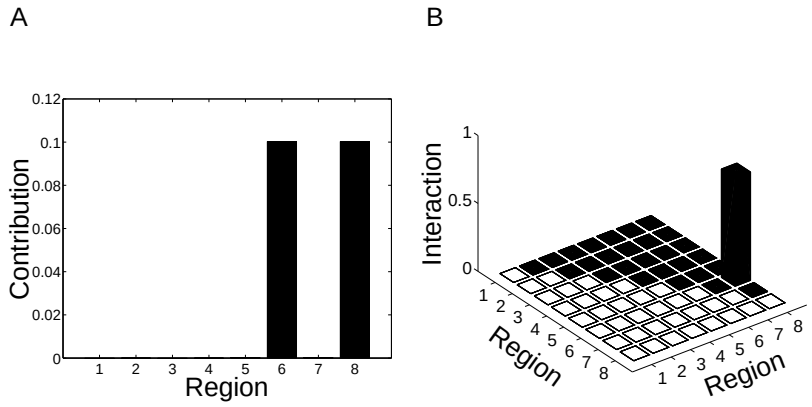


Figure 4: *One- and two-dimensional MSA of reversible deactivation experiments.* The eight regions correspond to the pMS and the SC of both sides, with separation between deep and superficial layers in each. A. Predicted Shapley value of the eight regions. Regions 6 and 8 represent SC_R-deep and SC_L-deep, respectively. B. The symmetric interaction $I_{i,j}$ between each pair of regions ($i < j$).

deep component of the SC is lesioned, the superficial one is lesioned as well (regions 5 and 7 in Figure 4). Nevertheless, the MSA successfully reveals that only the deep SC regions are the ones of significance, which concurs with previous interpretation of collicular deactivation results (Lomber et al., 2001). Interestingly, the FCA using the same experiments assigns different contributions in different runs. In most runs, it uncovers the role played by the deep SC regions. Alas, it usually also assigns non-vanishing contributions to the superficial SC regions. Furthermore, in some runs it assigns significant contributions to the pMS regions. This testifies to the disadvantages of an operative approach such as the FCA, that are overcome by the MSA by offering a unique fair solution for the contributions.

We further performed a two-dimensional MSA to quantify the interactions between each pair of regions, finding only one significant interaction, between SC_R-deep and SC_L-deep, $I_{6,8} = 0.8$ (Figure 4B). Furthermore, observing the negative contribution of each of the two regions when the other one is lesioned ($\gamma_{6,\bar{8}} = \gamma_{8,\bar{6}} = -0.3$) and the positive contribution when the other one is intact ($\gamma_{6,\bar{8}} + I_{6,8} = \gamma_{8,\bar{6}} + I_{8,6} = 0.5$), the MSA concludes that each of the two regions is positively modulated by the other, uncovering the "paradoxical" type of interaction assumed to take place in this system (Hilgetag, Kotter, & Young, 1999; Hilgetag, Lomber, & Payne, 2000; Lomber et al., 2001). This analysis testifies to the usefulness of the MSA in deducing the functionally important regions as well as their significant interactions.

The prediction made by the MSA to lesioning configurations that were

not performed in the experiments is that lesioning either one of the deep SC regions will result in large deficit in (contralateral) spatial attention to auditory stimuli, while lesioning both will result in a much smaller deficit. Lesioning any subgroup of the other regions involved in these experiments will not influence the performance in this task. In particular, it is predicted that as long as the deep SC regions are intact, the animal will exhibit full orientation performance, even when some or all of the superficial collicular layers, the superficial parietal cortical layers and the deeper parietal cortical layers are lesioned.

5 Discussion

Over the last two years we have developed and described a new framework for quantitative causal function localization via multi-perturbation experiments (Aharonov et al., 2003; Segev et al., 2003; Keinan et al., 2003). This paper presents an important new step within this project, replacing an earlier ad-hoc error minimization approach by the MSA – an axiomatic framework based on a rigorous definition of all elements' contributions via the Shapley value. The latter is a fundamental concept borrowed from game theory, which has been classically used for providing fair solutions to cost allocation problems. The MSA accurately approximates the Shapley value in a scalable manner, making it a more accurate and efficient method for function localization than its predecessor, the FCA. The prediction and estimation variants of

the MSA are specifically geared toward experimental biological applications, in which only a limited number of multi-perturbation experiments can be performed.

As demonstrated in this paper, the MSA can provide new insights to the workings of a classical neural model, that of a segment of the lamprey, revealing a "skeletal" subnetwork that is primarily responsible for its CPG activity. We also showed that the MSA framework is capable of dealing with behavioral data from experimental deactivation studies of the cat brain, quantitatively identifying the main interaction underlying a "paradoxical" lesioning phenomenon observed previously in studies of spatial attention (Hilgetag et al., 1999, 2000; Lomber et al., 2001). The presented applications, however, involve relatively simple networks of limited size. The more advanced estimation and prediction methods will be needed in order to meet the challenge of function localization in more complex systems.

As pointed out previously, and as evident from effects such as "paradoxical" lesion phenomena and functional backup, single lesion approaches do not suffice to portray the correct function localization in a network. Why then has the great majority of lesioning studies in neuroscience up until now relied on single lesioning solely? First, it has been very difficult (and in many systems, practically impossible) to reliably create and test multiple lesions. Moreover, single lesion studies have been perceived as already being fairly successful in providing insights to the workings of neural systems. Naturally, the lack of a more rigorous multi-perturbation analysis has made the test-

ing/validation of this perception practically impossible, and it may well have been the case that the lack of an analysis method has made such experiments seem futile. We hope that at least this last obstacle has been remedied by the introduction of the MSA.

But the question remains; have we reached the stage where we can now perform multi-perturbation experiments and corresponding analyses of biological networks? We believe that the answer is affirmative, and newly introduced experimental tools hold great promise, both in neuroscience and for the analysis of genetic and metabolic networks. In neuroscience, there is now an exciting new prospect of carrying out experimental perturbation studies in human subjects using Transcranial Magnetic Stimulation (TMS). This technique allows to induce "virtual lesions" in normal subjects performing various cognitive and perceptual tasks (Pascual-Leone, Wasserman, Davey, & Rothwell, 2002; Rafal, 2001). The methodology can be utilized to co-deactivate doublets of brain sites (Hilgetag et al., 2003) and potentially even triplets. Additionally, recent retrospective lesion studies of stroke patients have reconstructed patients' lesions and analyzed the resulting multi-lesion data using statistical tools (e.g. Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000; Bates et al., 2003). Such data may be more rigorously analyzed by the MSA, processing the multi-lesion data to capture the contributions of and the significant high-dimensional interactions between regions or voxels. Going beyond the realm of neuroscience to biology in general, the recent discovery of RNA interference (RNAi) (Hammond, Caudy, & Hannon, 2001;

Couzin, 2002) has made the possibility of multiple concomitant gene knock-outs a reality. Using RNAi vectors it is now possible to temporarily block the transcription of specified genes for a certain duration and measure the performance of various cellular and metabolic indices of the cell, including the expression levels of other genes. As with TMS, RNAi is limited at this stage to just a few elements that are knocked out concomitantly, but this is just the beginning. Multi-perturbation studies are a necessity, and they are hence bound to take place, starting in the very near future. The MSA framework presented in this paper is a harbinger of this new kind of studies, offering a novel and rigorous way of making sense out of them.

Appendix A: A Toy Problem

In this appendix we analyze a very simple system as a test case for the MSA, in the purpose of illustrating its operation in a concrete, "minimal" example. After introducing the system, we present the Shapley value obtained and explain its intuitive correctness in this case. We then compare the results of the MSA with both single lesion analysis and the FCA of this system.

The System

Let us define a system of elements $\{e_1, \ldots, e_n\}$, where the lifetime of element e_i is exponentially and independently distributed with parameter λ_i (expectation of $1/\lambda_i$). We define the performance of the system as the expected

time where at least one of the elements remains functioning, that is, the expectation of the maximum of the individual lifetimes. For clarity, we focus on the case $n = 3$, but the results presented throughout this section hold for any number of elements.

We calculate the performance for each perturbation configuration inflicted upon the system, where a perturbed element is simply removed from the system. Obviously, $v(\phi) = 0$ and $v(\{e_i\}) = \frac{1}{\lambda_i}$, for $i \in \{1, 2, 3\}$, since there is only one element in the system in the latter and none in the former. For $i \neq j$, $i, j \in \{1, 2, 3\}$, $v(\{e_i, e_j\})$ is the expectation of a random variable with a cumulative distribution function which is the product of the two cumulative distribution functions: $(1 - e^{-\lambda_i x}) \cdot (1 - e^{-\lambda_j x})$. Calculating the expectation using the cumulative distribution function yields

$$v(\{e_i, e_j\}) = \frac{1}{\lambda_i} + \frac{1}{\lambda_j} - \frac{1}{\lambda_i + \lambda_j}. \tag{8}$$

Similarly,

$$v(\{e_1, e_2, e_3\}) = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} - \frac{1}{\lambda_1 + \lambda_2} - \frac{1}{\lambda_1 + \lambda_3} - \frac{1}{\lambda_2 + \lambda_3} + \frac{1}{\lambda_1 + \lambda_2 + \lambda_3} \tag{9}$$

is the performance of the system in its intact state.

The formulas of v may also be viewed intuitively based on the inclusion-exclusion formula. For instance, equation (9) is illustrated by the following equation using Venn diagrams:

where each Venn diagram corresponds to a term in equation (9), in the same

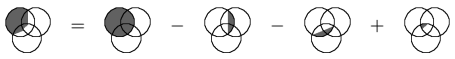
order. A region in a Venn diagram in the illustration is the expected time where all the elements corresponding to including groups are functioning. Thus, an intersection of several groups is the expectation of the minimum of the corresponding distributions. In this case of exponential distributions, the minimum is also exponentially distributed with a parameter equal to the sum of the parameters of the different distributions. Thus, the resulted expectation of each such distribution (equation (9)).

The Shapley Value

Knowing the performance of each coalition of elements, the Shapley value is obtained using equation (5). The contribution of element e_1 according to the Shapley value is

$$\gamma_1(N, v) = \frac{1}{\lambda_1} - \frac{1}{2} \cdot \frac{1}{\lambda_1 + \lambda_2} - \frac{1}{2} \cdot \frac{1}{\lambda_1 + \lambda_3} + \frac{1}{3} \cdot \frac{1}{\lambda_1 + \lambda_2 + \lambda_3}, \qquad (10)$$

and similarly for the other elements. Illustrating the meaning of the resulted contribution of e_1 using Venn diagrams, we get



,

in the same order as the terms in equation (10). As seen from the left-hand side Venn diagram, element e_1 is accredited for a third of the time when it is functioning with both elements e_2 and e_3 (the rest is divided equally between the contributions of elements e_2 and e_3), for half of the time when it is functioning with either e_2 or e_3 (the other half is contributed to the other

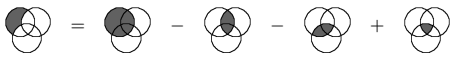
element) and for the whole time when it is functioning alone. That is, the Shapley value divides the intact performance of the system (equation (9)) to the different elements such that each term is divided equally to all elements composing it, denoting a fair division of the system performance to the different elements.

Single Lesion Analysis and FCA

The single lesion approach consists of perturbing one element within each experiment and measuring the decrease in performance. Using the same notation of the MSA, the contribution assigned to element i using single lesion analysis is proportional to $v(N)-v(N\setminus\{e_i\})$. For the test case system, this equals (for $i=1$)

$$\sigma_1(N,v)=\frac{1}{\lambda_1}-\frac{1}{\lambda_1+\lambda_2}-\frac{1}{\lambda_1+\lambda_3}+\frac{1}{\lambda_1+\lambda_2+\lambda_3}, \tag{11}$$

and similarly for the other elements. Illustrating σ_1 using Venn diagrams, we obtain



,

in the same order as the terms in equation (11). The left-hand side Venn diagram illustrates that *with single lesion analysis, each element is only accredited for the expected time when it is functioning alone, without considering its previous contribution while other elements were still functioning.* Thus, the Shapley value is much more informative in capturing the true con-

tribution of the elements in comparison with using only single perturbation experiments.

Figure 5 compares the Shapley value with the single lesion and FCA contributions, for the case where $n = 4$ and $\lambda_i = 1/i$, for $i = 1, \ldots, 4$.[8] Evidently, the FCA contributions and the Shapley value differ significantly for 3 out of the 4 elements, even when considering the large standard deviations of the former. The FCA contributions resemble the contributions assigned by the single lesion analysis, testifying that the FCA fails, too, in capturing a fair attribution of contributions in this case.

Acknowledgments

We acknowledge the valuable contributions and suggestions made by Ranit Aharonov, Shay Cohen, Alon Kaufman and Ehud Lehrer. This research has been supported by the Adams Super Center for Brain Studies in Tel Aviv University and by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities.

[8] A concrete example must be used since no general formulas for the contributions exist in an FCA analysis.

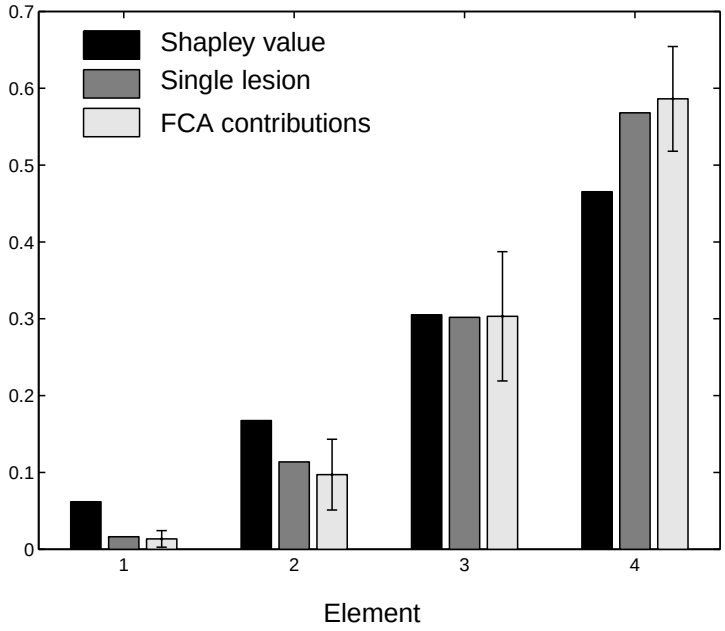


Figure 5: *Comparison between the FCA, single lesion analysis (dark gray bars) and the MSA (black bars) on the test case.* The FCA contributions (light gray bars) are mean and standard deviations across 10 FCA runs. Both the Shapley value and the FCA are based on the full set of all 2^4 perturbation configurations. The Shapley value, the single lesion contributions and the FCA contributions within each of the 10 runs are normalized such that their sum equals 1.

References

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. R. (2000). A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *Journal of Neuroscience*, *20*(7), 2683–2690.

Aharonov, R., Segev, L., Meilijson, I., & Ruppin, E. (2003). Localization of function via lesion analysis. *Neural Computation*, *15*(4), 885–913.

Banzhaf, J. F. (1965). Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, *19*, 317–343.

Bates, E., Wilson, S. M., Saygin, A., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience*, *6*(5), 448–450.

Billera, L. J., Heath, D., & Raanan, J. (1978). Internal telephone billing rates—a novel application of non-atomic game theory. *Operations Research*, *26*, 956–965.

Brodin, L., Grillner, S., & Rovainen, C. M. (1985). N-methyl-d-aspartate (nmda), kainate and quisqualate receptors and the generation of fictive locomotion in the lamprey spinal cord. *Brain Research*, *325*(1–2), 302–306.

Buchanan, J. T. (1999). Commissural interneurons in rhythm generation and intersegmental coupling in the lamprey spinal cord. *Journal of Neurophysiology*, *81*(5), 2037–2045.

Buchanan, J. T. (2001). Contributions of identifiable neurons and neuron classes to lamprey vertebrate neurobiology. *Progress in Neurobiology*, *63*(4), 441–466.

Buchanan, J. T. (2002). Spinal cord of lamprey: Generation of locomotor patterns. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks.* MIT Press / Bradford Books.

Buchanan, J. T., & Grillner, S. (1987). Newly identified 'glutamate interneurons' and their role in locomotion in the lamprey spinal cord. *Science*, *236*(4799), 312–314.

Couzin, J. (2002). Small RNAs make big splash. *Science*, *298*, 2296–2297.

Dubey, P., Neyman, A., & Weber, R. J. (1981). Value theory without efficiency. *Mathematics of Operations Research*, *6*(1), 122–128.

Ekeberg, O. (1993). A combined neuronal and mechanical model of fish swimming. *Biological Cybernetics*, *69*(5–6), 363–374.

Farah, M. J. (1996). Is face recognition 'special'? evidence from neuropsychology. *Behavioral Brain Research*, *76*, 181–189.

Feigenbaum, J., Papadimitriou, C. H., & Shenker, S. (2001). Sharing the cost of multicast transmissions. *Journal of Computer and System Sciences*, *63*(1), 21–41.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, *76*(376), 817–823.

Gefeller, O., Land, M., & Eide, G. E. (1998). Averaging attributable fractions in the multifactorial situation: Assumptions and interpretation. *J Clin*

Epidemiol, *51*(5), 437–441.

Grabisch, M., & Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, *28*, 547–565.

Grillner, S., Deliagina, T., Ekeberg, O., Manira, A. E., Hill, R. H., Lansner, A., Orlovsky, G. N., & Wallen, P. (1995). Neural networks that coordinate locomotion and body orientation in lamprey. *Trends in Neurosciences*, *18*(6), 270–279.

Grillner, S., Wallen, P., & Brodin, L. (1991). Neuronal network generating locomotor behavior in lamprey: Circuitry, transmitters, membrane properties, and simulation. *Annual Review Neuroscience*, *14*, 169–199.

Hammond, S. M., Caudy, A. A., & Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Gen.*, *2*(2), 110-119.

Hilgetag, C. C., Kotter, R., Theoret, H., Classen, J., Wolters, A., & Pascual-Leone, A. (2003). A bilateral competitive network for visual spatial attention in humans. *Neurocomputing*, *52–54*, 793–798.

Hilgetag, C. C., Kotter, R., & Young, M. P. (1999). Inter-hemispheric competition of sub-cortical structures is a crucial mechanism in paradoxical lesion effects and spatial neglect. *Prog Brain Res*, *121*, 121–141.

Hilgetag, C. C., Lomber, S. G., & Payne, B. R. (2000). Neural mechanisms of spatial attention in the cat. *Neurocomputing*, *38*, 1281–1287.

Hilgetag, C. C., Theoret, H., & Pascual-Leone, A. (2001). Enhanced visual

spatial attention ipsilateral to rTMS-induced virtual lesions of human parietal cortex. *Nature Neuroscience*, *4*(9), 953-957.

Kapur, N. (1996). Paradoxical functional facilitation in brain-behaviour research. A critical review. *Brain*, *119*, 1775–1790.

Keinan, A., Meilijson, I., & Ruppin, E. (2003). Controlled analysis of neurocontrollers with informational lesioning. *Philosophical Transactions of the Royal Society of London: Series A*, *361*(1811), 2123–2144.

Kosslyn, S. M. (1999). If neuroimaging is the answer, what is the question? *Philosophical Transactions of the Royal Society of London: Series B*, *354*, 1283–1294.

Lomber, S. G., Payne, B. R., & Cornwell, P. (2001). Role of the superior colliculus in analyses of space: Superficial and intermediate layer contributions to visual orienting, auditory orienting, and visuospatial discriminations during unilateral and bilateral deactivations. *J Comp Neurol*, *441*, 44–57.

Lomber, S. G., Payne, B. R., Hilgetag, C. C., & Rushmore, R. J. (2002). Restoration of visual orienting into a cortically blind hemifield by reversible deactivation of posterior parietal cortex or the superior colliculus. *Exp Brain Res*, *142*, 463–474.

Lomber, S. G., Payne, B. R., & Horel, J. A. (1999). The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function. *J Neurosci Methods*, *86*, 179–194.

Myerson, R. B. (1977). Graphs and cooperation in games. *Mathematics of Operations Research*, *2*, 225–229.

Myerson, R. B. (1980). Conference structures and fair allocation rules. *International Journal of Game Theory*, *9*, 169–182.

Parker, D., & Grillner, S. (2000). Neuronal mechanisms of synaptic and network plasticity in the lamprey spinal cord. *Progress in Brain Research*, *125*, 381–398.

Pascual-Leone, A., Wasserman, E., Davey, N., & Rothwell, J. (2002). *Handbook of transcranial magnetic stimulation.* Oxford University Press.

Rafal, R. (2001). Virtual neurology. *Nature Neuroscience*, *4*(9), 862–864.

Roth, A. E. (1979). *Axiomatic models of bargaining.* Berlin: Springer Verlag.

Segev, L., Aharonov, R., Meilijson, I., & Ruppin, E. (2003). High-dimensional analysis of evolutionary autonomous agents. *Artificial Life*, *9*(1), 1–20.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. II, pp. 307–317). Princeton: Princeton University Press.

Shapley, L. S., & Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, *48*(3), 787–792.

Shubik, M. (1962). Incentives, decentralized control, the assignment of joint costs and internal pricing. *Management Science*, *8*, 325–343.

Shubik, M. (1985). *Game theory in the social sciences.* Cambridge, MA:

MIT Press.

Sprague, J. M. (1966). Interaction of cortex and Superior Colliculus in mediation of visually guided behavior in the cat. *Science*, *153*, 1544–1547.

Squire, L. R. (1992). Memory and the Hippocampus: A synthesis of findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.

Vallar, G. (1998). Spatial hemineglect in humans. *Trends in Cognitive Sciences*, *2*, 87–97.