Localization of Function Via Lesion Analysis

Ranit Aharonov[1] Lior Segev[2] Isaac Meilijson[3]

Eytan Ruppin[2]

[1]Center for Neural Computation

The Hebrew University, Jerusalem, Israel

ranit@alice.nc.huji.ac.il

[2]School of Computer Sciences

Tel-Aviv University, Tel-Aviv, Israel

{*liorseg*, *ruppin*}@*post.tau.ac.il*

[3]School of Mathematical Sciences

Tel-Aviv University, Tel-Aviv, Israel

isaco@post.tau.ac.il

August 1, 2002

Abstract

This paper presents a general approach for employing lesion analysis to address the fundamental challenge of localizing functions in a neural system. We describe the Functional Contribution Analysis (FCA) which assigns contribution values to the elements of the network such that the ability to predict the network's performance in response to multi-lesions is maximized. The approach is thoroughly examined on neurocontroller networks of evolved autonomous agents. The FCA portrays a stable set of neuronal contributions and accurate multi-lesion predictions, which are significantly better than those obtained based on the classical single lesion approach. It is also utilized for a detailed synaptic analysis of the neurocontroller connectivity network, delineating its main functional backbone. The FCA provides a quantitative way of measuring how the network functions are localized and distributed among its elements. Our results question the adequacy of the classical single lesion analysis traditionally used in neuroscience, and show that using lesioning experiments to decipher even simple neuronal systems requires a more rigorous multi-lesion analysis.

1 Introduction

One of the fundamental challenges in understanding a nervous system is to characterize how its various functions are localized and distributed among the system elements. These elements may be single neurons, neuronal assemblies or even cortical areas, depending on the scale on which one chooses to analyze the system. Even simple nervous systems are capable of performing multiple and unrelated tasks, often in parallel. Each task recruits some of the elements of the system, and often the same element participates in several tasks. This poses a difficult challenge when one attempts to identify the roles of the network elements, and to assess their contributions to the different tasks.

The localization of specific tasks in a nervous system is conventionally done in Neuroscience either by recording the activity of the system elements during behavior, or by measuring the deficit in performance after lesioning specific elements. Obviously inferring significance from recording unit activity during behavior is problematic because correlation of neuronal activity with performance does not necessarily identify causality. For example, it is possible that some areas raise their activity while a given task is performed not because they significantly contribute to the processing of that task, but rather because they are activated by other regions that do play an important role in this task processing. To try and uncover the units and regions whose activation really underlies a specific function or behavior, neuroscientists have traditionally adopted a system analysis approach, where the normal operating modes of a system are studied by lesioning and perturbing its processing in many possible ways. These abnormal manipulations are aimed to provide one with deeper insights into the system's normal, regular, operation.

Albeit, the vast majority of lesion studies in neuroscience have been *single lesion* studies, in which only one network component is abolished at a time (e.g. [Squire, 1992; Farah, 1996]). Such single lesions are very limited in their ability to reveal the significance of elements which interact in complex ways. A straightforward conceptual example demonstrating this is when two elements have a high degree of redundancy with respect to the processing of a function to which they equally contribute. Then, lesioning either element alone will not reveal its true significance, since no reduction in the performance of that function will occur. The problematic and limited value of single lesion analysis has already been widely noted in the neuroscience literature [Sprague, 1966; Farah, 1990; Sitton, Mozer and Farah, 2000; Young, Hilgetag and Scannell, 2000]. One classical example discussed is that of the *paradoxical lesioning* effect [Sprague, 1966]. In this paradigmatic case, lesioning area A alone is harmful but lesioning area A given that area B is lesioned is beneficial, hence the apparent "paradox". Importantly, it demonstrates that looking at a single lesion alone may be misleading, as the beneficial influence of an area depends on the general state of the system. Last but not least, current lesioning studies in neuroscience have yielded mostly qualitative measures, lacking the ability to precisely quantify the contribution of a unit to the performance of the system and, perhaps even more importantly, to predict the effect of new, multiple-site lesions.

Addressing these challenges, this paper presents a Functional Contribution Analysis (FCA). The FCA framework gives a rigorous, operative definition for the neurons' contributions to the system's performance in various tasks, and an algorithm to measure them via multi-lesion analysis. Acknowledging that single lesions are insufficient for localizing functions in neural systems, the FCA framework assumes an existing data set of multiple lesions to a neural system and their corresponding system performance scores in a given set of tasks. The FCA uses these data to calculate the elements' contributions to the various tasks, yielding an accurate prediction of the performance of a neural system when a new, untested, multiple lesion damage is inflicted upon it. Thus, the FCA harnesses the full power of the lesion approach both to learn about the localization of functions in the intact, unperturbed network, and to predict its response to damage.

Although still infrequent, multiple lesion studies are now being performed in an increasing frequency in neuroscientific studies of reversible inactivation [Lomber and Payne, 2001]. However, these data are still of fairly limited magnitude. Hence, the development and study of the FCA method was primarily done in a theoretical modeling framework, that of *neurally-driven evolved autonomous agents (EAAs)* [Ruppin, 2002]. EAA agents are software programs living in a simulated virtual environment that autonomously perform typical animat tasks like gathering food, navigating, evading predators, and seeking prey and mating partners. Each agent is controlled by an artificial neural network "brain". This network receives and processes sensory inputs from the surrounding environment and governs the agent's behavior via the activation of the motors controlling its actions. It is developed via genetic algorithms that apply some of the essential ingredients of inheritance and selection to a population of agents that undergo evolution. In recent years, much progress has been made in finding ways to evolve EAAs which successfully cope with diverse behavioral tasks [Floreano and Mondada, 1996; Scheier, Pfeifer and Kunyioshi, 1998; Kodjabachian and Meyer, 1998; Floreano and Mondada, 1998; Gomez and Miikkulainen, 1997], employing networks composed of several dozen neurons and hundreds of synapses (see [Meyer and Guillot, 1994; Kodjabachian and Meyer, 1998; Yao, 1999; Guillot and Meyer, 2000; Guillot and Meyer, 2001] for recent reviews).

EAAs are a very promising model system for studying neural processing and developing methods for its analysis (see [Ruppin, 2002] for a detailed discussion of this issue). They are *less biased* than conventional neural networks used in neuroscience modeling as their architecture is typically not pre-designed, but emergent. Their potential is further substantiated by their *feasibility:* the small size of the evolved networks, the simplicity of their environments and behavioral tasks, coupled with the full information available regarding the model dynamics, form conditions that help make the analysis of the network's dynamics an amenable task. Furthermore, numerous EAA studies have already shown that networks emerging in EAA systems can manifest interesting biological-like characteristics and provide additional computational insights [Cangelosi and Parisi, 1997; Ijspeert and amd D. Willshaw, 1999; Aharonov-Barki, Beker and Ruppin, 2001].

Identifying the importance of elements of a system to varying tasks is often used as a basis for claims concerning the degree of localization versus distribution of computation in that system (e.g. [Wu, Cohen and Falk, 1994]). The distributed representation approach hypothesizes that computation emerges from the joint interaction between numerous simple elements [McClelland, Rumelhart and the PDP Research Group, 1986; Hopfield, 1982]. The local representation hypothesis suggests that activity in single neurons represents specific and well-defined concepts (the grandmother cell notion) or performs specific computations (see [Barlow, Bartholemew, Bremner and Brunk, 1972]). This fundamental question of distributed versus localized computation in nervous systems has attracted ample attention (spanning from Lashley's seminal paper [Lashley, 1929] to today's ongoing controversies [Downing, Jiang, Shuman and Kanwisher, 2001; Haxby, Gobbini, Furey, Ishai, Schouten and Pietrini, 2001; Cohen and Tong, 2001]. However, there seems to be a lack of a precise and rigorous definition for these terms [Thorpe, 1995]. The ability of the FCA to quantify the contribution of elements to tasks yields, as a "by-product", a precise determination of the level of distribution of processing in the network.

This paper presents a new tool for localizing functions in biological and artificial neural networks, and introduces a quantitative approach for studying the long debated issue of local versus distributed computation in the brain. It capitalizes on the availability of full information that makes EAA models a good test-bed for studying neural processing. The paper is organized as follows: section 2 describes the functional contribution approach, and provides the necessary definitions. In section 3 we apply the FCA to evolved neurocontrollers, demonstrating its capabilities to unravel the underlying neuronal contributions. We examine the implications of the type of lesioning method employed, and show the FCA's superiority over the classical neuroscience single lesion approach. Section 4 studies the localization of different tasks with the FCA, and section 5 shows that the FCA can be utilized for a detailed synaptic analysis of the underlying network connections. Our results and their implications are discussed in section 6.

2 The Functional Contribution Approach

The FCA aims to compute the localization of a task from a series of multi-lesion experiments. In each such experiment, a lesioning configuration specifies which of the agent's neurocontroller units (be they neurons, synapses, or any higher-order module) is lesioned. Given this lesion, the agent is run in the environment in which it was evolved, and its performance (i.e., its fitness or any other well defined performance measure) is measured. The FCA uses this data to compute a *contribution vector* $\mathbf{c} = (c_1, ..., c_N)$, where c_i is the contribution of element i to the task in question, and N is the number of elements in the network. The goal of the FCA is to assign the contribution values which provide the best prediction of the agent's performance in terms of Mean Squared Error (MSE) under all possible multi-site lesions.

2.1 Definitions

Suppose that a multi-lesion experiment is performed where a set of elements in the network is lesioned and the neurocontroller network then performs a certain task. The result of this experiment is described by the pair $\{\mathbf{m}, p_m\}$ where the lesion configuration vector $\mathbf{m}$ has $m_i = 0$ if element i was lesioned and 1 if it was left intact, and

p_m is the corresponding performance of the lesioned network, divided by the baseline performance of the fully intact network. For a system of N elements there are 2^N such possible multi-lesion experiments, i.e., 2^N possible lesioning configurations $\mathbf{m}$.

The underlying idea of our definition is that the contributions of the units are those values which allow the most accurate prediction of performance following lesions of any degree to the system. Within the FCA framework, given a configuration $\mathbf{m}$ of lesioned and intact units, the predicted performance $\tilde{p}_\mathbf{m}$ of the network is the sum of the contribution values of the intact units $(\mathbf{m} \cdot \mathbf{c})$, as evaluated by a non-decreasing function f,

$$\tilde{p}_\mathbf{m} = f(\mathbf{m} \cdot \mathbf{c}) \ . \tag{1}$$

The function f is a non-decreasing piecewise polynomial. It is non-decreasing to reflect the notion that beneficial elements (those whose lesioning results in performance deterioration) should have positive contribution values, and that negative values indicate elements that hinder performance.

Given these definitions, we seek to find the pair $\{\mathbf{c}, f\}$, which minimizes the mean squared prediction error

$$E = \frac{1}{2^N} \sum_{\{\mathbf{m}\}} (\tilde{p}_\mathbf{m} - p_\mathbf{m})^2 \ . \tag{2}$$

The vector $\mathbf{c}$ which minimizes this error is the *contribution vector* for the task tested, and the corresponding f is its adjoint *performance prediction function*. Since multiplying $\mathbf{c}$ and dividing f by the same scalar maintains the prediction, we arbitrarily normalize $\mathbf{c}$ such that $\sum_{i=1}^{N} |c_i| = 1$. Thus, the optimal contribution vector $\mathbf{c}$ and its accompanying performance prediction function f minimize the MSE of predicted versus actual performance, over all possible lesioning configurations.

The FCA is related to two known modeling and prediction methods, that of Generalized Linear Modeling and that of Projection Pursuit Regression. The Generalized Linear Model (GLM, see [McCullagh and Nelder, 1989]) approach can be succinctly defined by the relation $g(\tilde{p}_{\mathbf{m}}) = \mathbf{m} \cdot \mathbf{c}$, where g is the transfer function akin to the performance prediction function f in the FCA formulation (if $g = f^{-1}$ the above equation reduces to Eq. 1). In GLM, the contribution vector $\mathbf{c}$ can be computed via a closed formula, and hence results in a much faster and deterministic computation than that performed within the FCA (see section 2.2). However, GLM cannot be used to predict performance since in general g is not reversible; many lesioning configurations can lead to the same performance (e.g., complete failure of the agent). Projection Pursuit Regression (PPR, see [Huber, 1985]) has a formulation very similar to the FCA. PPR approximates the response surface g by a sum of ridge functions $g(\mathbf{x}) \sim \sum_{j=1}^{k} f_j(\mathbf{c}_j \cdot \mathbf{x})$, which, together with the projection directions $\mathbf{c}_j$ are found in an iterative, greedy manner, minimizing the distance from a target random variable. The FCA is hence a special case of PPR in which $k = 1$ and f is non-decreasing. Using only one ridge function and a non-decreasing f enables one to preserve the intuitive meaning of $\mathbf{c}$ as the contribution values vector.

2.2 The FCA Algorithm

The contribution vector $\mathbf{c}$ and the prediction function $\mathbf{f}$ for a task are computed using a training set of lesioning configurations and their corresponding performance levels in the task at hand. The FCA is a gradient descent search algorithm which iteratively updates f and $\mathbf{c}$ until it reaches a local minimum of the training error E'

$$E' = \frac{1}{n} \sum_{\mathbf{m} \in M} [f(\mathbf{m} \cdot \mathbf{c}) - p_{\mathbf{m}}]^2, \tag{3}$$

where M is the training set and n is its size. The steps of the FCA are:

1. **Choose** a random initial normalized contribution vector $\mathbf{c}$ for the task, and compute f as in step 4.

2. **Compute c.** Using the current f compute new values of $\mathbf{c}$ by performing a gradient descent on the error E' (Eq. 3), searching for $\mathbf{c}$ values that minimize E' using the current f.

3. **Re-normalize c**, such that $\sum_{i=1}^{N} |c_i| = 1$.

4. **Compute f.** Given the current $\mathbf{c}$, perform isotonic regression on the pairs $\{\mathbf{m} \cdot \mathbf{c}, p_{\mathbf{m}}\}$ in the training set. Use a smoothing spline on the result of the regression to obtain a new f.

   Steps 2 through 4 are repeated a fixed number of times.

In [Segev, Aharonov, Meilijson and Ruppin, 2002] an algorithm for the efficient selection of the training set in FCA is presented. This Adaptive Lesion Selection algorithm is based on active learning [Cohn, Ghahramani and Jordan, 1995; Engelbrecht and Cloete, 1999], and selects the next lesioning configuration based on the information contained in the current set of configurations. It is shown to considerably reduce the required training set size. For simplicity, throughout most of the paper, the training set used is the full configuration set. However, utilizing the adaptive algorithm (or even random selection), a relatively small training set suffices to achieve low MSE on the test set (see, e.g., sections 3.3 and 5).

2.3 Experimental Protocol

The FCA algorithm is stochastic, and as such is dependent on both the initial conditions and the random choices made in the gradient descent step. To reduce stochasticity, each of the MSE values reported in this paper is a mean of 10 FCA runs. A single run consists of 10 *trials*, each of which is initialized with a different random contribution vector. The result of the trial with the lowest MSE on the *training* set is chosen for that run. We note, however, that the sensitivity to initial conditions is rather small in most cases. For each initial condition, the FCA executes a fixed number I of iterations of updating f and $\mathbf{c}$ (see section 2.2). Throughout the paper, $I = 150$, except for the analysis of agent `S22` where $I = 300$. All FCA results presented are mean and standard deviation of 10 FCA runs (i.e. a total of 10×10 initial condition choices). The MSE values are normalized by dividing them by the variance of the agent's performances $p_{\mathbf{m}}$ in the test set. Thus, a normalized MSE of q means that the FCA explains a fraction $R^2 = 1 - q$ of the variance. That is, if one would predict for any configuration a performance level equal to the mean performance of the agent, then the normalized MSE would equal 1, corresponding to $R^2 = 0$.

3 FCA of Evolutionary Neurocontrollers

Using evolutionary simulations, Aharonov et. al. [Aharonov-Barki, Beker and Ruppin, 2001] have previously developed autonomous agents controlled by fully recurrent artificial neural networks. High performance levels were attained by agents performing simple life-like tasks of foraging and navigation. In [Aharonov-Barki, Beker and Ruppin, 2001], classical neuroscience methods were used to analyze the evolved neurocontrollers. Here the FCA is developed and studied in this environment and is applied to the analysis of EAA neurocontrollers, demonstrating several aspects of the approach.

3.1 The EAA Environment

The agents live in a virtual grid arena of size 30×30 cells surrounded by walls (Figure 1, after [Aharonov-Barki, Beker and Ruppin, 2001]). "Poison" is randomly scattered all over the arena (consuming this resource results in a negative reward).

"Food", the consumption of which results in a positive reward, is randomly scattered in a restricted 10×10 "food zone" in a corner of the arena. The agent's goal is to find and eat as many food items as possible during its life, while ignoring the poison items. The agents are equipped with a set of sensors, motors, and a fully-recurrent neurocontroller. The neurocontroller is coded in the genome and evolved; the sensors and motors are given and constant.

The initial population consists of 100 agents equipped with random neurocontrollers. Each agent is evaluated in its own environment, which is initialized with 250 poison items and 30 food items. The life cycle of an agent (an *epoch*) is 150 time steps, in each of which one motor action takes place. At the beginning of an epoch the agent is introduced to the environment at a random location and orientation. At the end of its life-cycle each agent is assigned a general performance score (fitness) calculated as the number of food items minus the number of poison items it consumes, divided by 30, so that the maximum possible fitness is 1. Since the agent's lifetime is short, this is practically an unattainable score. Simulations last for a large number of generations, ranging between 5000 and 30000. All the details concerning the simulations and genetic methods are given in [Aharonov-Barki, Beker and Ruppin, 2001].

The agent's neurocontroller consists of 10 to 22 internal recurrently connected McCulloch-Pitts binary neurons (the number of neurons is fixed within a given simulation experiment). In addition to the internal neurons, the agent has five input neurons connected to the five basic sensors, collectively called the somatosensor (see Figure 1, after [Aharonov-Barki, Beker and Ruppin, 2001]). Four of these sensors encode the presence of a resource (food or poison, without distinction between the two), a wall, or an empty cell in the cell the agent occupies and in the three cells directly in front of it. The fifth sensor is a "smell" sensor which can differentiate between food and poison underneath the agent, but gives a random reading if the agent is in an

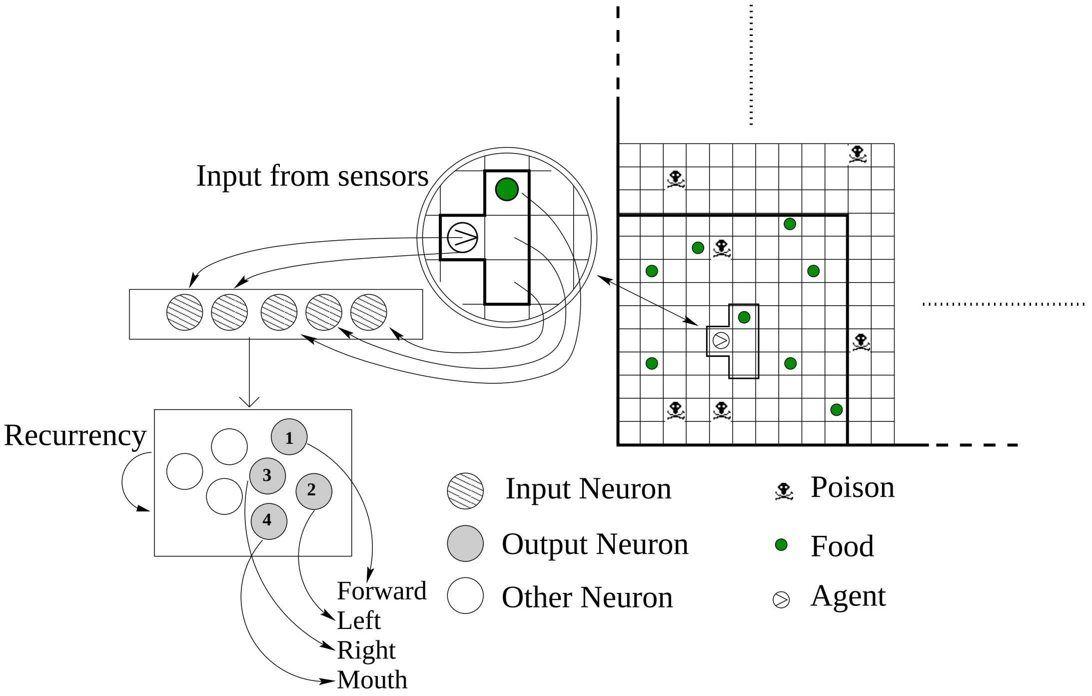

Figure 1: *An outline of the grid arena* and the agent's neurocontroller. For illustration purposes the borders of the food zone are marked, but they are invisible to the agent. The agent is marked by a small arrow on the grid, whose direction indicates its orientation. The T-shape marking in front of the agent denote grid cells which it senses. The output neurons control four motors: move forward, turn left or right, and open mouth. Eating can take place only when the agent stands still and its mouth is open. Output neurons and interneurons are fully connected to each other.

empty cell. In some simulations agents were equipped with an extra position sensor returning the location ((x, y) coordinates) of the agent in the arena. Four out of the internal neurons are motor neurons. They initiate movement forward, a turn left or right, and control the state of the mouth – open or closed (see caption). In each step a sensory reading occurs, network activity is then updated (network updating is synchronous), and a motor action is taken according to the resulting activity in the designated output neurons. Importantly, eating only takes place if the agent is neither moving nor turning. Thus, eating is a costly process requiring a time step with no other movement, in a lifetime of limited time steps. The task is especially difficult when the position sensor is not provided, because only partial and local sensory information about the environment is available to the agent. Thus navigating to the foodzone must rely only on the local environment and not on global position cues. As shown previously ([Aharonov-Barki, Beker and Ruppin, 2001]), algorithms employing no memory are quite poor in this task.

In [Aharonov-Barki, Beker and Ruppin, 2001] two types of agents were studied: an SP-type agent possessing both a somatosensor and a position sensor, and an S-type agent possessing a somatosensor only. For both types of agents the most successful strategy relied upon a switch between two behavioral modes – *exploration* and *grazing*. The exploration mode consists of moving in straight lines, ignoring resources in the sensory field which are not directly under or in front of the agent, and turning at walls. The grazing mode, on the other hand, consists of turning to resources to the right or left in order to examine them, turning at walls, and maintaining the agent's location on the grid in a relatively small, restricted region. In both modes the agent consumes food items that it finds. The exploration mode is mostly observed when the agent is out of the food zone, allowing it to explore the environment and find the food zone. Inside the food zone, however, the agents almost always display grazing behavior, which results in efficient consumption of food.

Examining the networks of successful agents equipped with a position sensor (SP-type) has revealed an important common feature [Aharonov-Barki, Beker and Ruppin, 2001]: certain interneurons have a position-sensitive response. One of these neurons typically fires outside the food zone and remains quiescent inside the food zone and in its close vicinity. Clamping this *"position-sensitive cell"* and measuring the behavioral mode of the agent has revealed that it acts as a *command neuron*. When this command neuron is active the agent assumes an exploration behavior, *regardless* of where it is in the environment, whereas clamping it to a silent state results in grazing behavior. Thus, the place-dependent activity map of the command neuron serves as an accurate predictor of the agent's behavior; where the map values are low the agent will graze, and where they are high it will explore. In the following sections we apply the FCA to one such agent, whose network consists of 10 neurons (`SP10`).

Similar to the SP case, also in agents possessing only the somatosensor, *and lacking a position sensor* (S-type), at least one neuron whose activity is place-dependent was identified. We study such agents in subsequent sections, one whose network consists of 22 neurons (`S22`), and another with a 10 neuron network (`S10`). These command neurons emerge even though no explicit information about location is available to S-type agents. Moreover, their activity is not locked to the absolute position in the arena, but rather to the location of the food zone, which may vary at different experiments [Aharonov-Barki, Beker and Ruppin, 2001]. Aharonov et. al. have previously demonstrated that this computational ability is based on a spontaneously emerging stochastic memory mechanism. The "knowledge" of location is actually a form of memory of the time elapsed since the last eating event [Aharonov-Barki, Beker and Ruppin, 2001]. As will become clear from the analysis below, this memory mechanism relies on feedback from the motor neurons, supplying information about eating events of the agent.

As discussed above, the FCA approach requires a well defined quantitative measure for performance on each task analyzed. We consider three tasks throughout the work. First, the general survival task of maximizing fitness as defined above. Further, we consider two subtasks which correspond to the two emergent behaviors of successful agents, exploration and grazing. Exploration performance, p^e, is measured by randomly placing the agent in the arena and measuring the number of steps t elapsed before the agent reaches the food zone for the first time. If the agent fails to reach the food zone in less than 1000 steps, t is set to 1000. The exploration performance is computed by

$$p^e = \frac{T-t}{T} - \frac{\pi_b/S}{t/T}, \tag{4}$$

where π_b is the number of poison items consumed before reaching the food zone, $S = 30$ is the total number of food items in the arena, and $T = 150$ is the original number of steps in an epoch. The first term is concerned with the speed of reaching the food zone, and becomes negative for $t > T$. The second term penalizes poison consumption which should be avoided in any behavioral mode, and is normalized to be consistent with the general performance normalization. Grazing performance, p^g, is measured by placing the agent in the arena for T time steps. Denote by s the total number of food sources eaten, and by π_a the number of poison items consumed after reaching the food zone for the first time, then,

$$p^g = \frac{(s-\pi_a)/S}{(T-t)/T}, \tag{5}$$

where t is measured as before. The numerator is consistent with the general performance normalization, and the denominator normalizes by the time spent inside the food zone.

3.2 Selecting The Lesioning Method

An important determinant of a lesioning analysis is the manner in which the lesioning itself is performed. We have studied two main lesioning methods, *biological lesioning* and *stochastic lesioning*. Biological lesioning refers to the classical method employed in most neuroscience lesioning experiments, where the lesioned component is ablated or cooled, i.e., completely silenced. In our model, accordingly, in biological lesioning a lesioned neuron is deactivated completely by cutting all its output connections to the rest of the network. In contrast, Stochastic lesioning is performed by making the firing pattern of the lesioned component random rather than by completely silencing its output. That is, at every time step a lesioned neuron fires with probability equal to its overall mean firing rate in normal behavior, independent of its input field. This ensures that the lesioning does not effect the mean field of other neurons, affecting only the information content received from the lesioned neuron[1]. In both methods, when lesioning motor neurons, we do not alter the activity transmitted to the motors themselves. This enables us to isolate the role of the motor units in the computation of the recurrent controller networks (i.e. their contribution to other neurons), without completely immobilizing the agent.

Before advancing to study the FCA in depth as described further on in this paper, we ran a few experiments to gauge the accuracy of FCA analysis obtained with both lesioning methods. Figure 2A compares the test MSE obtained by the FCA on the controller of agent `S10` in the general, grazing and exploration tasks discussed in section 3.1, once employing biological lesioning (black bars) and once using stochastic lesioning (light bars). As is evident, the MSE obtained from biological lesioning is significantly poorer than that obtained when using stochastic lesioning. Indeed, as panels B and C demonstrate, the contribution values of the neurons computed

[1]Similarly, in the synaptic analysis, a lesioned synapse transmits a stochastic version of the pre-synaptic activity.

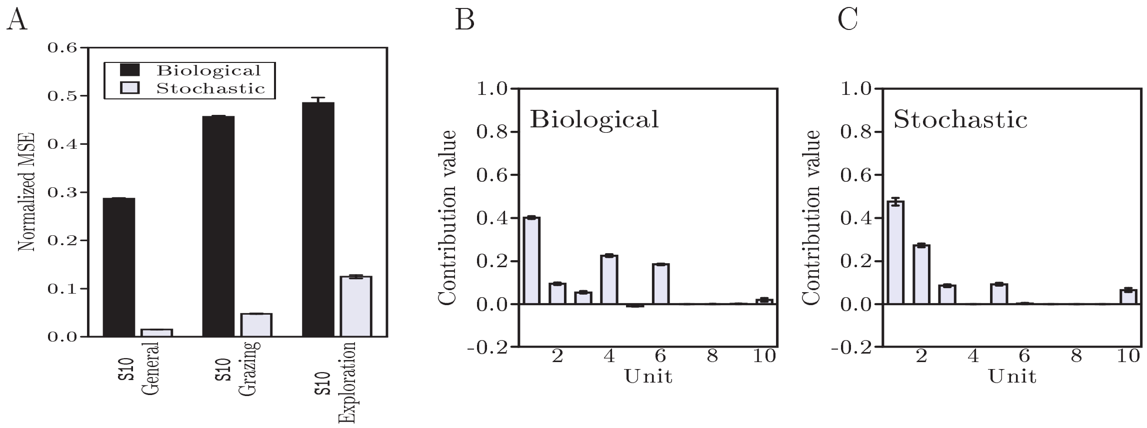

Figure 2: *Comparison of biological and stochastic lesioning* A: MSE obtained by the biological versus stochastic lesioning for agent `S10` in the general, grazing and exploration tasks. B,C: Contribution values obtained using biological lesioning (B) and stochastic lesioning (C) for agent `S10` in the general fitness task.

by both lesioning methods differ, showing that the poorer prediction capability of biological lesioning stems from a real difference in the significance assigned to the system elements.

These results testify to the crucial importance of selecting the lesioning method. Specifically, they demonstrate that the classical ablation (biological) lesioning conventionally used in animal lesioning experiments might be questionable. A possible explanation for this phenomenon is that when a neuron is silenced it transfers other unlesioned neurons into a regime in which they cannot "fire" as their mean field is far below threshold. Thus, the effective part of the network which is lesioned is much larger than what was intended, and considered "lesioned" by the FCA. This results in a situation where lesions percolate to other network elements, disabling and isolating specific elements, and creating dispersed effective lesions. Stochastic lesioning also spreads the damage from lesioned neurons to neighboring neurons, but to a much lesser extent. As a result, biological lesioning is much more difficult to analyze, because the difference between the *direct* lesioning configuration used by the FCA and the *effective* one is much larger, causing FCA-like algorithms to be erratic and less accurate.

Biological lesioning is hence not fit for accurate lesion analysis (at least, in the FCA framework, which is fairly general and employs multi-unit lesions) even in very simple neurocontrollers. In view of these results, we have chosen to employ stochastic lesioning in the rest of the paper. This method is definitely superior for the analysis of animat agents. We shall return to discuss these two lesioning methods from a neuroscience perspective in section 6.

3.3 FCA of Agents' Performance

We consider first the performance obtained in the general task, i.e., maximizing fitness by consuming food and avoiding poison. We study the three agents described in section 3.1, the two S-type agents (`S10` and `S22`) and the SP-type agent (`SP10`). We apply the FCA to these neurocontrollers by performing lesioning experiments and testing the performance levels of the agents. For each agent, a small training set was chosen using the adaptive algorithm (section 2.2) or a random selection, and the FCA was applied 10 times (section 3.2). The result of the FCA is the contribution vector $\mathbf{c}$ and the performance prediction function f. The prediction capability of the FCA is measured by computing the MSE on a test set consisting of all 2^{10} configurations in `S10` and `SP10`, and of 30,000 random configurations in `S22`[2]. The normalized MSE on these test sets were all below 0.07, demonstrating that the FCA succeeded in finding a pair $\{\mathbf{c}, f\}$, which brings the error in Eq. 2 to a very low value even when trained on a very limited subset of the 2^N configuration space.

Figures 3A, 3C and 3D depict the contribution vectors computed for the three agents. Each sub-figure depicts the mean and standard deviation of the contributions of the different neurons of the agent, over the 10 runs of the FCA. Importantly, the

[2]Since obtaining the full 2^{22} configuration set is impossible, we chose a large enough set such that almost any randomly chosen set of that size results in practically the same error.

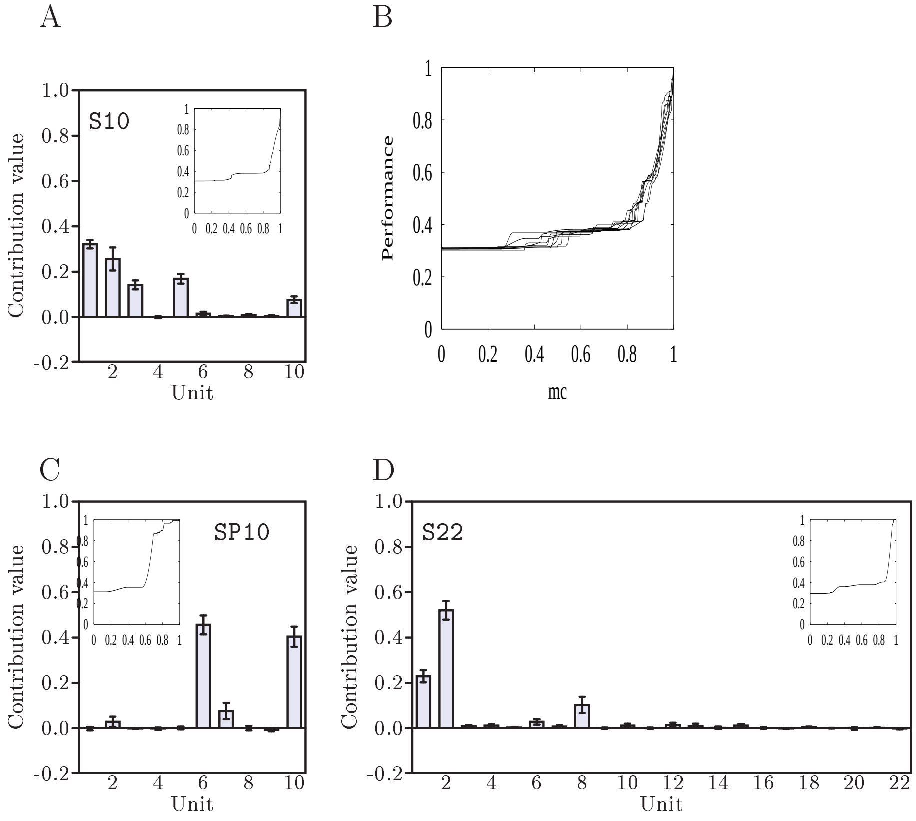

Figure 3: *Contribution values of the agents' neurons.* The mean and standard deviation of the contribution values computed by the FCA are plotted in each panel (vertical axis). A. Agent `S10`. FCA trained on 45 lesioning examples obtained by the Adaptive Lesioning algorithm. B. The corresponding ten performance prediction functions. The x-axis is $m \cdot c$, y-axis is the predicted performance $f(m \cdot c)$ (Eq. 1). C. Agent `SP10`. FCA trained on 45 lesioning examples obtained by the Adaptive Lesioning algorithm. D. Agent `S22`. FCA trained on 99 random lesioning examples. The inset in panels A,C and D depicts one performance prediction function. Neurons 1-4 (on the horizontal axis) are the motor neurons (see Fig. 1).

FCA converges to similar minima of the error in all runs, yielding consistent measures of the contributions. As expected from previous receptive field analysis of the agents [Aharonov-Barki, Beker and Ruppin, 2001], we find that the command neurons (neuron 5 in `S10`, neuron 6 in `SP10`, and neuron 8 in `S22`) receive significant and positive contributions in all three agents. Interestingly, in the two S-type agents, some of the motor neurons (neurons 1 through 4) receive high contribution values, testifying to the importance of the recurrency from the output units[3]. This is consistent with the notion that the feedback from the motor units is required for signaling eating events, enabling the agents to switch to and remain in grazing mode. In `SP10`, where location information is available to the agent, this recurrent connection from the motor neurons is not significant. The inset in each panel plots the performance prediction function obtained by the FCA. All ten performance prediction functions look similar (as can be seen for the case of `S10`, in Figure 3B), so one was chosen for clarity of exposition. The similar form of the prediction functions again testifies to the convergence of the FCA to similar minima of the prediction error.

3.4 Comparison to the Single Lesion Approach

An important aspect of the FCA is the integration of information obtained from multiple lesions, going beyond the limited information available in single lesions alone. Here, we compare the FCA approach to the single lesion approach using the evolved agents. To study the significance of using multiple lesions, we compute the null hypothesis, i.e. we calculate the contribution values relying only on the performance levels obtained when single units are lesioned. The contribution of neuron i is taken to be the decrease in the performance due to lesioning that neuron alone, normalized as before[4]. The black bars in Figure 4(A) depict the contributions computed from single lesions for the general performance of agent `S10`. The light bars describe the contributions computed by the FCA, using all possible lesioning configurations. As evident, the contributions assigned by the FCA differ significantly from those obtained with knowledge of single lesion results.

[3] Recall that under lesioning, the motor actions governed by an output neuron are kept intact, and only the information transfer to the rest of the network is disturbed.

[4] $c_i = (1 - p_i)/(\sum_{i=1}^{N} |1 - p_i|)$, where p_i denotes the performance when neuron i is lesioned alone.

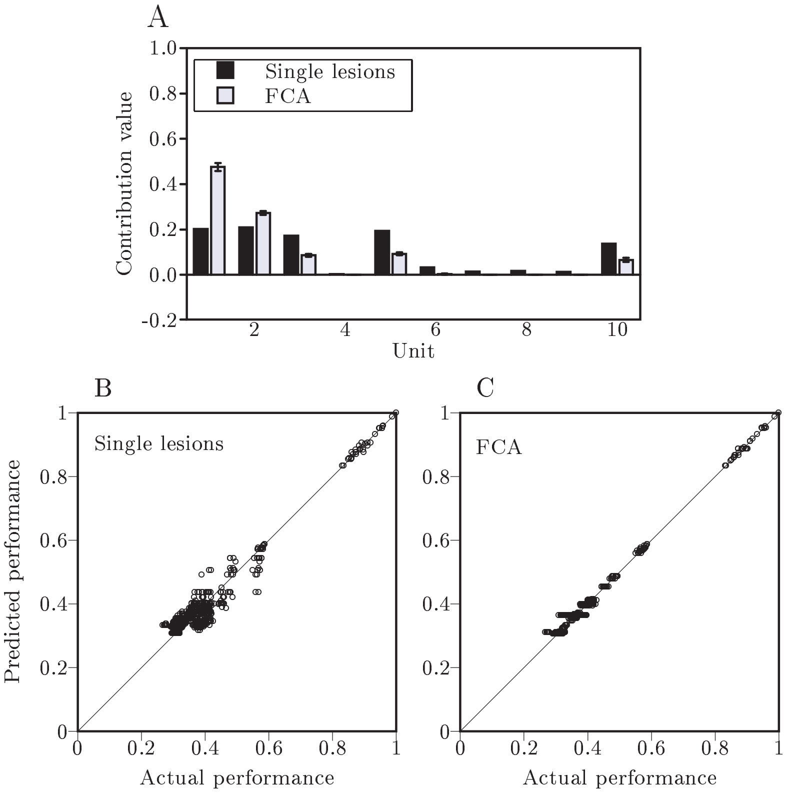

Figure 4: *Comparison between the single lesion approach and the FCA on agent* `S10`. A: Contribution values obtained by the two methods. B: Predicted versus actual performance on the test set using single lesions information. C: Predicted versus actual performance on the test set using the FCA.

Do the contribution values computed by the FCA using multiple lesions describe the system more accurately than those inferred from single lesions? The obvious criterion is the prediction ability. In principle, the single lesion approach implies a linear prediction function, which would lead to very inaccurate performance predictions. Thus, to obtain a fair comparison, we take the contribution vector found by the single lesion analysis, and optimally fit its corresponding prediction function to the test set. This ensures that any difference in prediction capabilities stems from a difference in the contributions themselves. Panels B and C of Figure 4 compare the predicted versus actual performance, using single lesions (B) and the FCA (C). Clearly, single lesions do not reveal the true contributions of the neurons, testifying to the existence of some form of interaction between the effects of the different units on the network function.

Figure 5 compares the test MSE obtained for several agents and tasks using single lesion information (black bars) and using the FCA derived contributions (light bars). These results demonstrate again that in some agents and tasks complex interactions exist, yielding suboptimal contributions when considering only single lesions. The information from multiple lesions in such networks is indeed essential for revealing the true contributions of the network elements. The more complex, possibly non-linear interactions are modeled well by the FCA, because although being a linear model, it is generalized by the non-linear prediction function[5]. However, obviously many forms of interactions cannot be modeled with the basic FCA. A classical example of such an interaction was given by Sprague in the neuroscience literature [Sprague, 1966], showing that deficits in orienting towards a stimulus, resulting from large cortical visual lesions, can be reversed through additional removal of contralateral areas. This has been known as the Sprague effect, or paradoxical lesioning [Hilgetag, Lomber and Payne, 2000; Lomber and Payne, 2001]. To address these challenges we have generalized the FCA to include high-dimensional context-dependent interactions. This non-linear generalization is based on the usage of *compound elements*, which are specific combinations of a few simple elements. A preliminary discussion of the high-dimensional FCA appears in [Segev, Aharonov, Meilijson and Ruppin, 2002], and a more thorough treatment will be given in a separate paper. Note that for agent `SP10`, single lesion information suffices to achieve very accurate contributions, testifying to the simpler functional organization of the network[6].

[5]A simple example of such non-linear interactions is a two unit system which functions perfectly unless both units are lesioned. The basic FCA models such a case by assigning both units a contribution value of 0.5, with $f(0) = 0$ and $f(0.5) = f(1) = 1$.

[6]However, it still may be that localization of a specific task in this agent requires an FCA analysis.

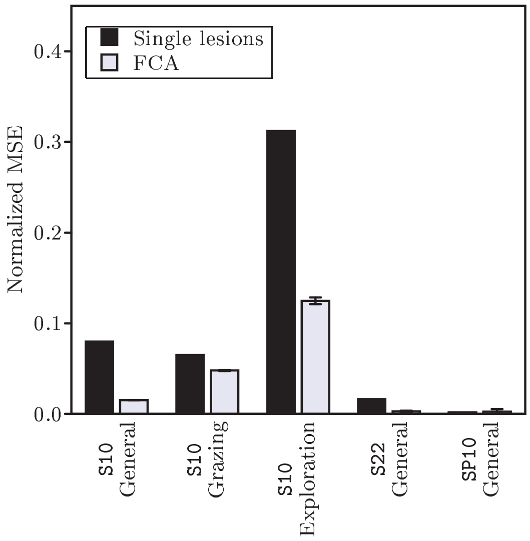

Figure 5: *Comparison of MSE obtained by the single lesion approach and the FCA.* Train and test sets are the full 2^{10} configuration set for `S10` and `SP10`. For `S22`, the training set is a 5,000 configuration set, and the test set consists of 20,000 configurations. All predictions used an optimally fitted prediction function to the test set (see text). In most tasks the standard deviation of the MSE is too small to be seen.

4 Localization of Several Tasks

In the previous section we have discussed the results of applying the FCA to the overall performance measure of the agent, its fitness. We showed how the FCA can be used to find the contribution of each neuron to the task of maximizing the fitness. However, as discussed in section 3.1, the EAAs perform two additional subtasks – grazing and exploration. In this section, we show how the FCA can be used to localize these tasks in parallel.

Consider an agent with a neurocontroller network of interconnected neurons that performs a set of K different functional tasks. Addressing the question of which elements contribute to which tasks, it is natural to think in terms of a contribution matrix, where C_{ik} is the contribution of element i to task k, as shown in Figure 6. The elements studied may be neurons, synapses or any higher-order module.

As before, the data analyzed for computing the contribution matrix is gathered by

Element	Task 1	Task 2	$\cdots$	Task K	
1	C_{11}	C_{12}	$\cdots$	C_{1K}	
2	C_{21}	C_{22}	$\cdots$	C_{2K}	$\rightarrow S_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
N	C_{N1}	C_{N2}	$\cdots$	C_{NK}	
	$\downarrow$ L_1				

Figure 6: *The Contribution Matrix.* The network consists of N elements and performs K different functional tasks. The entry C_{ik} is the contribution of element i to task k.

inflicting a series of multiple lesions onto the agent's neurocontroller network. Under each lesion, the resulting performance of the agent in different tasks is measured. Given this data, the FCA finds the contribution values C_{ik} that provide the best performance prediction on average for all possible multi-site lesions (the kth column of the matrix is the contribution vector of task k as described in section 2). Following the spirit of [Lashley, 1929], the localization L_k of task k can now be defined as a deviation from equipotentiality along column k (e.g., L_1 in Figure 6), and similarly, S_i, the specialization of neuron i is the deviation from equipotentiality along row i of the matrix (e.g., S_2 in Figure 6).

More formally, L_k is the standard deviation of column k of the contribution matrix divided by the maximal possible standard deviation,

$$L_k = \frac{\mathrm{std}(C_{\cdot k})}{\sqrt{(N-1)/N^2}}. \tag{6}$$

Note that L_k is in the range $[0, 1]$ where $L_k = 0$ indicates full distribution and $L_k = 1$ indicates localization of the task to one neuron alone. Similarly, if neuron i is highly specialized for a certain task, $C_{i\cdot}$ will deviate strongly from a uniform distribution, and thus we define S_i, the *specialization* of neuron i, as

$$S_i = \begin{cases} 2 \cdot \mathrm{std}(|C_{i\cdot}|) & \text{if the number of tasks, } K \text{, is even} \\ \\ \frac{2 \cdot \mathrm{std}(|C_{i\cdot}|)}{\sqrt{(K^2-1)/K^2}} & \text{otherwise.} \end{cases} \tag{7}$$

Note that S_i uses the absolute value of the contributions to reflect the intuition that the specialization of the unit is determined more by the magnitude of its contribution than by its sign. Again, $S_i = 1$ indicates maximal specialization of neuron i. We note, however, that in principle other measures can be defined. The important point is that the contribution matrix enables one to define such quantitative measures to capture qualitative notions.

We concentrate on one agent (`S10`), and consider the two tasks which correspond to the two emergent behaviors of successful agents, exploration and grazing as defined in section 3.1. The FCA is performed separately for each task yielding a contribution vector and a prediction function for the task. Figure 7 depicts the contribution values computed by the FCA for the two tasks. As evident, the localization of the two tasks within the network differs quite considerably. Notably, neuron 10 which is an interneuron, contributes positively to grazing behavior, but hinders exploration.

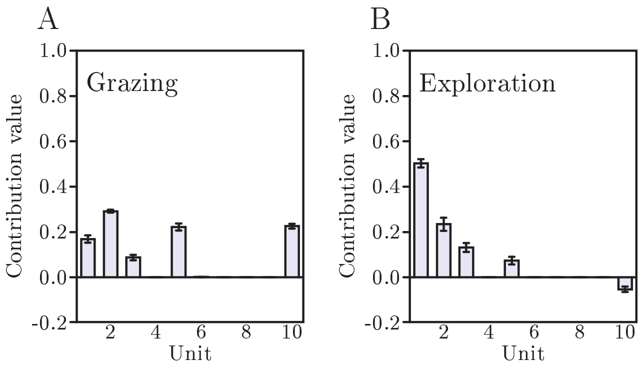

Figure 7: *Contribution values for Grazing and Exploration.* A: Grazing task. B: Exploration task.

Given these data, we can now construct the contribution matrix of the agent, depicted in Table 1, along with the measures of localization and specialization computed from it. Since the size of the evolved neurocontrollers was set arbitrarily, there are a number of neurons which do not participate in any task the agent performs, and hence the *effective size of the network*, considering only neurons with non-vanishing contributions, is smaller. This should be taken into account when considering task localization in the network. Thus, the effective localization $\tilde{L}_k$ of the different tasks performed by the agent (Table 1), is computed only on the non-vanishing neurons (neurons 1,2,3,5 and 10 in this agent), after normalizing the sum of their absolute values of contributions to 1. In this agent, exploration is more localized than grazing. Note that even though interneuron 10 has contribution values of similar absolute magnitude as neuron 5, its specialization is double that of the latter, as it contributes positively to grazing behavior while actually hindering exploration.

N	Grazing	Exploration	S_i
1	**0.1696**	**0.5034**	**0.3338**
2	**0.2916**	**0.2344**	**0.0571**
3	**0.0878**	**0.1324**	**0.0445**
4	0.0000	0.0000	0.0000
5	**0.2228**	**0.0740**	**0.1488**
6	0.0015	0.0006	0.0009
7	-0.0001	0.0001	0.0002
8	0.0007	0.0001	0.0006
9	0.0001	0.0001	0.0001
10	**0.2257**	**-0.0547**	**0.2804**

L_k	0.3684	0.5323
$\tilde{L}_k$	0.1698	0.4691

Table 1: *Localization and specialization in agent* S10. Each entry to the table is the contribution of the corresponding neuron to each of the two subtasks – grazing and exploration. The internal part of the table is thus the contribution matrix (see Fig. 6) of agent S10. The localization, L_k, effective localization, $\tilde{L}_k$, and specialization, S_i, are displayed for each task and neuron.

5 Synaptic Analysis

The FCA is not limited to a specific level of analysis, and thus the same network can be studied concomitantly on the neuronal and synaptic level. In this section we apply the FCA to agent `S10` analyzed above, to determine the contribution values of its synapses, and hence, the underlying structure of its synaptic network. The neuronal analysis presented above revealed that five out of the ten neurons are significant. Thus, as a first step we consider a reduced network consisting of only the significant neurons and their synapses. Recall, however, that insignificance of a motor neuron testifies only to the insignificance of its outgoing synapses, and not to the insignificance of its activity insofar as it controls the motor. Thus, the reduced network of agent `S10` contains the six indispensable neurons (the four motor neurons, neuron 5 and neuron 10), and *all* the synapses connecting them (excluding the outgoing synapses from neuron 4, which has a vanishing contribution value). The reduced network has a total of 30 synapses[7]. The agent driven by this reduced network still attains a very high level of performance, 0.36, which is 90% of the performance obtained by the original agent, driven by the fully connected intact neurocontroller. The next step is to use the FCA to find the backbone of the network, i.e., the significant synapses within the reduced neurocontroller.

Applying the FCA for this synaptic analysis, with a training set consisting of 2000 randomly chosen configurations, the normalized MSE on a randomly chosen test set of 20000 configurations is 0.05. The training set is extremely small relative to the full configuration space (2^{30} configurations). Nevertheless, the FCA managed to reach a fairly low test error, testifying to the potential scalability of the method. Figure 8A depicts the synaptic contribution values obtained by the FCA. The right

[7]We consider here only the internal network and not the synapses coming from input neurons, as lesioning those synapses causes a degree of damage which basically causes a total malfunctioning of the agent.

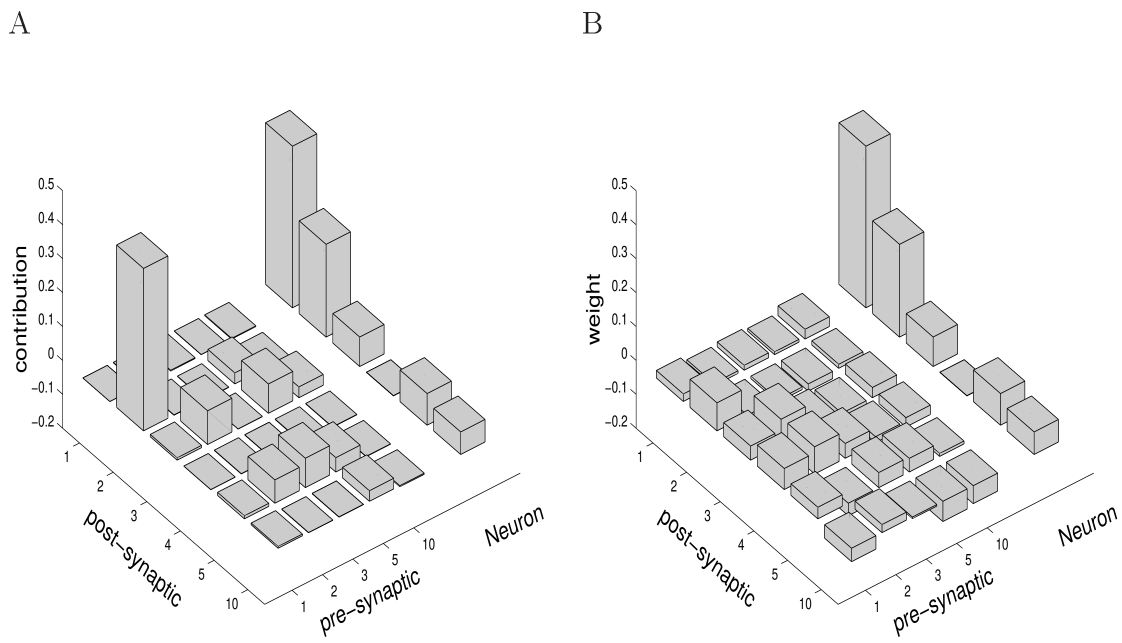

Figure 8: *Comparison of synaptic contributions computed by the FCA and the synaptic weights.* A: The FCA computed contributions of the *synapses* and *neurons* (right row) of the reduced network. B: The weights of the reduced network synapses, normalized such that the sum of the absolute values of these weights is 1. Neurons 1, 2, 3, 5 and 10 are recurrently connected. Neuron 4 is the motor neuron (controlling the mouth) whose outgoing synapses are insignificant.

part of the figure (marked "Neuron") depicts the contribution values of the neurons of the agent `S10`, as computed by the FCA on the neuronal level. Figure 8B depicts the weight values of the agent's neurocontroller, normalized such that the sum of the absolute values of the synaptic weights is 1, also accompanied by the neuronal contributions computed by the FCA. Evidently, the network structure emerging from considering the synaptic contributions (panel A) differs quite significantly from that inferred directly from the synaptic weights (panel B) (correlation of 0.31).

Is the synaptic significance obtained by the FCA more accurate than that derived from the raw synaptic weights themselves? To answer this question we lesion the synapses incrementally and measure the deterioration of the agent's performance. Figure 9 depicts the performance of the agent as a function of the number of lesioned synapses for three methods of incremental lesioning starting from the intact reduced network of the 30 internal synapses. In the first lesioning experiment we use the synaptic contribution values computed by the FCA and lesion by ascending order of significance. The graph depicts the mean and standard deviation using contribution values from ten different FCA runs. In the second lesioning experiment we use the absolute weight as a measure of importance and lesion by ascending order of the absolute value of the weights. In the third experiment we lesion incrementally by choosing the next synapse to be lesioned randomly. The graph depicts the mean and standard deviation of 100 such choices.

As evident, the FCA is much better at identifying the significant synapses of the network. For example, leaving only 15 synapse in the entire internal network, i.e. 15% of the total internal network and 50% of the reduced network, salvages 80% of the original performance level. In contrast, using the weights as a significance measure to choose the 15 important synapses leaves only 50% of the original performance[8].

[8]Using the actual weight value and not the absolute value achieves even lower performance on all levels of synaptic lesioning.

Not surprising, choosing 15 synapses randomly completely destroys the agent's performance. In summary, the FCA proves very reliable in identifying the significant network backbone and can be used to select synapses for pruning a neural network.

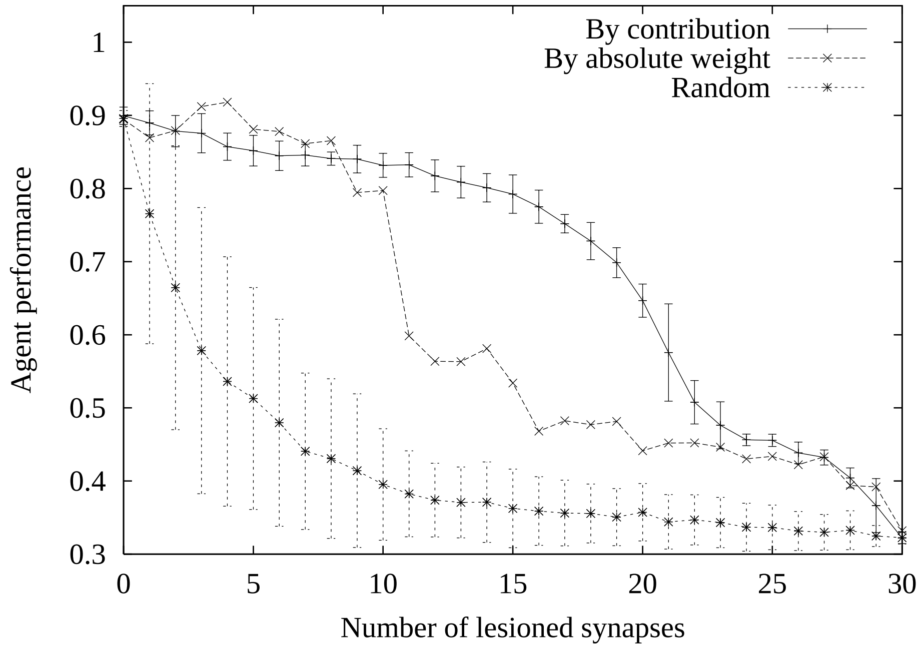

Figure 9: *Agent performance as a function of lesioning degree.* Performance of agent `S10` as a function of the number of synapses lesioned, for three different methods of incremental synaptic pruning. The performance is given relative to the performance of the original fully intact network. All methods start from the reduced network consisting of 30 synapses, i.e. 70 internal synapses lesioned. *By contribution*: For each of ten FCA runs, synapses are incrementally lesioned by ascending order of their contribution value. Line depicts mean and standard deviation of measured performance levels over the ten runs. *By absolute weight*: Performance levels following lesioning by ascending order of absolute synaptic weights. *Random*: Mean and standard deviation of performance levels over 100 random orders of synaptic lesions.

Using the FCA computed synaptic contributions, a minimal effective connectivity can be inferred, to simplify the analysis of the network. This is especially important for fully recurrent networks where understanding the structure becomes a daunting task. Looking at Fig 8A, one observes that there are nine synapses with relatively large contribution values, compared to the rest. Figure 10A depicts the network

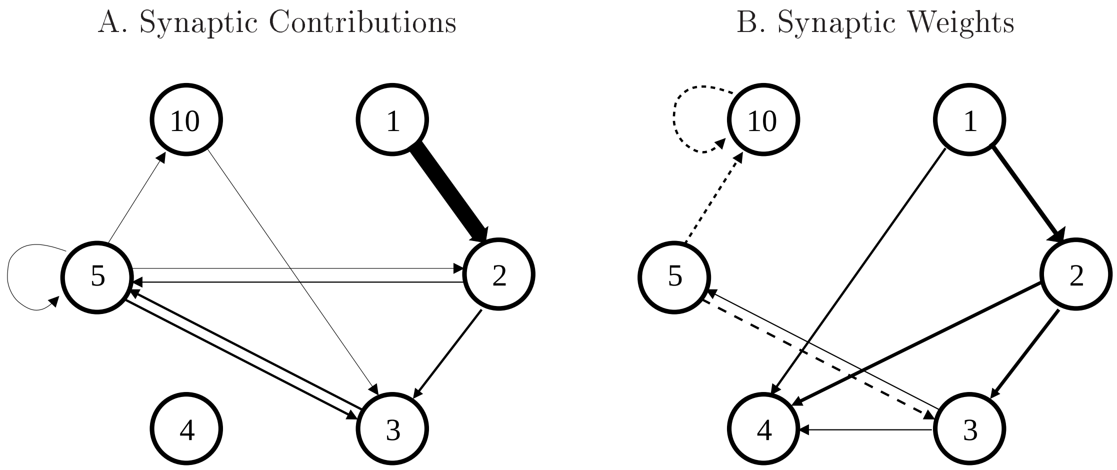

Figure 10: *Network backbone*. A. The nine synapses of the internal network with the largest contribution values. B. Similarly for the largest synaptic weights (by absolute value). Dashed lines denote negative values. The thickness of the lines scale with their absolute value, and are normalized to have the same sum in each panel.

formed by these nine largest synaptic contributions, and correspondingly Fig 10B depicts the nine synapses with largest weight values (in absolute value). Although part of the resulting backbone is similar, there are two important differences. First, although the synaptic weights imply that the connections into neuron 4 are important, the FCA correctly reveals that this is not the case. Second, the FCA analysis clearly shows the central role of the command neuron (neuron 5), a feature of the network much less apparent when considering the synaptic strengths. The functional role of the command neuron in these networks has been studied in detail in [Aharonov-Barki, Beker and Ruppin, 2001]. Examining the network backbone of Figure 10A one can postulate that the synapse from neuron 1 (forward step) to neuron 2 (left turn) is responsible for the move and turn behavior observed in grazing mode. Moreover, the synapses from the turn motor neurons (2 and 3) to the command neuron (neuron 5) is clearly important, as expected from previous analysis.

As discussed above the network functional connectivity emerging from the FCA analysis is more accurate in terms of being able to predict the network performance.

Note, that although only nine synapses receive a high contribution value (Figure 8), the performance of the network with only these synapses intact is still rather low (Figure 9). This is because other synapses are important in certain *combinations*, improving the performance if they are intact[9]. However, the nine high contribution synapses are strongly beneficial in many lesioning configurations, and are hence picked up by the basic FCA.

6 Discussion

This paper presents a functional contribution analysis that addresses the fundamental challenge of localization of function in neural networks. The FCA is based on the assignment of contribution values to the basic elements of the network, such that the ability to predict the network's performance in response to multi-lesions is maximized. The algorithm is thoroughly examined in an EAA environment. The results shown testify that:

- In the recurrent yet simple EAA neurocontrollers studied here, the basic version of the FCA suffices to portray a stable set of contributions and yield fairly accurate multi-lesion predictions. These are significantly better than those obtained with a single lesion analysis method. The FCA may be efficiently trained on a relatively small subset of all possible multi-lesion configurations. It can be used to localize several subtasks that the agent performs, and compute the corresponding localization and specialization indices.

- FCA analysis on the synaptic level testifies to the potential scalability of the approach for larger systems. It also presents an important method for synaptic pruning and visualizing the effective "backbone" of the network architecture.

[9]High-dimensional FCA is necessary to reveal these combinations.

These results should be viewed as an encouraging but first step in the attempt of understanding neural information processing in neural systems. Several important issues and limitations of the current FCA arise subsequently to this work, and deserve considerable attention in the future. First and foremost, the accurate analysis of complex networks requires high-dimensional descriptions involving compound elements that are composed of several basic units (neurons or synapses) that interact together. Such high-dimensional analysis is essential, for example, for an accurate account of subnetworks displaying the "paradoxical lesioning" effect. Second, the FCA makes one think about the "correct" or "best" way to perform the lesions in the network. Our results indicate that the conventional way of lesioning currently employed in neuroscience, that is, completely silencing the lesioned unit(s), is essentially inadequate for general multi-lesion studies. In this study we have focused on stochastic lesioning, but is it the best one possible? Lastly, the FCA provides important and useful information about function localization and the true effective network size, but it does not otherwise advance our understanding of how neural information processing takes place. The way in which information is represented, encoded and manipulated on the neural, assembly and network levels, remains an open question.

The FCA has potential significance beyond the scope of EAAs. Multi-lesion analysis algorithms like the FCA will become an essential tool in neuroscience for the analysis of reversible inactivation experiments, combining reversible neural cooling deactivation with behavioral testing of animals. These methods alleviate many of the problematic aspects of the classical single lesioning technique (ablation), enabling the acquisition of reliable data from multiple lesions of different configurations (for a review see [Lomber, 1999]). If the underlying network of brain regions studied is simple enough, then perhaps the "biological" ablation method of lesioning currently employed in these experiments will suffice. However, our study suggests that if these networks are more complex, ablation lesions may result in excessive perturbations of the system and fail to reveal its normal operation (the problematic nature of analyzing lesioning data due to this excessive lesioning has also been discussed in [Young, Hilgetag and Scannell, 2000]). If that will turn out to be the case, then other lesioning methods that cause smaller perturbations (much alike stochastic lesioning) should be employed.

FCA-like algorithms could also be used for the analysis of transcranial magnetic stimulation studies which aim to induce multiple transient lesions and study their cognitive effects (see [Walsh and Cowey, 2000] for a review). They could also possibly be used for studying the functional efficacy of synaptic projections on different components of the neuronal dendritic tree, in parallel to recent studies examining this question with other methods [London, Shribman, Hausser, Larkum and Segev, 2002]. In both cases, stochastic lesioning may prove to be the preferred lesioning method employed both experimentally and in the FCA analysis. Another potentially interesting application is the analysis of functional imaging data. One direction is to assess the contributions of each element to others, i.e., extending previous network's effective connectivity studies employing linear models (e.g., [Friston, Frith and Frackowiak, 1993]) with high-dimensional FCA. The second direction involves the application of the FCA to the meta-analysis of functional imaging data. This will require development and study of a continuous version of the current binary FCA representations, where results of various activation studies of similar tasks are viewed as different "multiple-lesion" probes of the task in hand.

Finally, the FCA studied in this paper has focused on studying each task separately, and then assembling this information and presenting it in a contribution matrix. The latter, however, could serve as an object of further elaborate analysis beyond the computation of localization/specialization indices. This analysis could involve a singular value decomposition of the contribution matrix, reducing it to a lower dimension and revealing possible hidden similarities in the localization of different tasks, in a manner analogous to the Latent Semantic Analysis used in Information Retrieval applications [Deerwester, Dumais, Landauer, Fumas and Harshman, 1990].

In summary, the FCA provides a rigorous method for determining function localization, yielding useful insights into the organization of both animat and animate nervous systems.

Acknowledgments

We acknowledge the valuable contributions and suggestions made by Hanoch Gutfreund, Nira Dyn and Matan Ninio, and the technical help provided by Oran Singer. This research has been supported by the Adams Super Center for Brain Studies in Tel Aviv University, and by the Israel Science Foundation. Ranit Aharonov is supported by the Horowitz foundation.

References

Aharonov-Barki, R., Beker, T. and Ruppin, E. (2001). Emergence of memory-driven command neurons in evolved artificial agents. *Neural Computation,* **13**, 691–716.

Barlow, R., Bartholemew, D., Bremner, J. and Brunk, H. (1972). *Statistical inference under order restrictions.* New York: John Wiley.

Cangelosi, A. and Parisi, D. (1997). A neural network model of Caenorhabditis Elegans: The circuit of touch sensitivity. *Neural Processing Letters,* **6**, 91–98.

Cohen, J. and Tong, F. (2001). The face of controversy. *Science,* **293**(5539), 2405–2407.

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1995). Active Learning with

Statistical Models. In G. Tesauro, D. Touretzky and T. Leen (Eds.), *Advances in Neural Information Processing Systems*, Volume 7, pp. 705–712. The MIT Press.

Deerwester, S., Dumais, S., Landauer, T., Fumas, G. and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the society for information science,* **41**(6), 391–407.

Downing, P. E., Jiang, Y., Shuman, M. and Kanwisher, N. (2001). A Cortical area selective for visual processing of the human body. *Science,* **293**(5539), 2470–2473.

Engelbrecht, A. and Cloete, I. (1999). Incremental Learning using Sensitivity Analysis. In *IEEE IJCNN*, Washington DC.

Farah, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press.

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

Floreano, D. and Mondada, F. (1996). Evolution of homing navigation in a real mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics - Part B,* **26**(3), 396–407.

Floreano, D. and Mondada, F. (1998). Evolutionary neurocontrollers for autonomous mobile robots. *Neural Networks,* **11**(7-8), 1461–1478.

Friston, K., Frith, C. and Frackowiak, R. (1993). Time-dependent changes in effective connectivity measured with PET. *Human Brain Mapping,* **1**, 69–79.

Gomez, F. and Miikkulainen, R. (1997). Incremental evolution of complex general behavior. *Adaptive Behavior,* **5**(3/4), 317–342.

Guillot, A. and Meyer, J. (2001). The animat contribution to cognitive systems research. *Journal of Cognitive Systems Research,* **2**, 157–165.

Guillot, A. and Meyer, J.-A. (2000). From SAB94 to SAB2000: What's new, Animat? In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press.

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J. and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science,* **293**(5539), 2425–2430.

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA,* **79**, 2554–2558.

Huber, P. J. (1985, June). Projection Pursuit. *The Annals of Statistics,* **13**(2), 435–475.

Ijspeert, A. and amd D. Willshaw, J. H. (1999). Evolving swimming controllers for a simulated lamprey with inspiration from neurobiology. *Adaptive Bahvior,* **7**, 151–172.

Kodjabachian, J. and Meyer, J. (1998). Evolution and development of neural controllers for locomotion, gradient-following and obstacle-avoidance in artificial insects. *IEEE Transactions on Neural Networks,* **9**(5), 796–812.

Lashley, K. S. (1929). *Brain mechanisms in intelligence.* Chicago: University of Chicago Press.

Lomber, S. and Payne, B. (2001). Task-specific reversal of visual hemineglect following bilateral reversible deactivation of posterior parietal cortex: A comparison with deactivation of the superior colliculus. *Visual Neuroscience,* **18**(3), 487–499.

Lomber, S. G. (1999). The advantages and limitations of permanent or reversible deactivation techniques in the assesment of neural function. *J. of Neuroscience Methods,* **86**, 109–117.

London, M., Shribman, A., Hausser, M., Larkum, M. and Segev, I. (2002). The information efficacy of a synapse. *Nature Neuroscience,* **5**(4), 333–340.

McClelland, J. L., Rumelhart, D. and the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, volume 2: Psychological and biological models.* Massachusetts: MIT Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall.

Meyer, J.-A. and Guillot, A. (1994). From SAB90 to SAB94: Four years of animat research. In D. Cliff, P. Husbands, J.-A. Meyer and S. Wilson (Eds.), *Proceedings of the Third International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press.

Ruppin, E. (2002). Evolutionary autonomous agents: A neuroscience perspective. *Nature Reviews Neuroscience,* **3**, 132–141.

Scheier, C., Pfeifer, R. and Kunyioshi, Y. (1998). Embedded neural networks: Exploiting constraints. *Neural Networks,* **11**(7-8), 1551–1569.

Segev, L., Aharonov, R., Meilijson, I. and Ruppin, E. (2002). Localization of Function in Neurocontrollers. In *Proceedings of the International Conference on the Simulation of Adaptive Behavior (SAB2002), Edinburgh.*

Sitton, M., Mozer, M. and Farah, M. J. (2000). Superadditive effects of multiple lesions in a connectionist architecture: Implications for the neuropsychology of Optic aphasia. *Psychological Review,* **107**, 709–734.

Sprague, J. (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.

Thorpe, S. (1995). Localized versus distributed representations. In M. A. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks.* Massachusetts: MIT Press.

Walsh, V. and Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience,* **1**, 73–79.

Wu, J., Cohen, L. B. and Falk, C. X. (1994). Neuronal activity during different behaviors in Aplysia: A distributed organization? *Science,* **263**, 820–822.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE,* **87**(9), 1423–1447.

Young, M., Hilgetag, C. and Scannell, J. (2000). On imputing function to structure from the behavioural effects of brain lesions. *Phil. Trans. R. Soc. Lond. B,* **355**, 147–161.