Available online at www.sciencedirect.com
ScienceDirect
Theory in Biosciences 125 (2007) 93–121
Theory in Biosciences
www.elsevier.de/thbio

Geometric robustness theory and biological networks

Nihat Ay[a,b,*], David C. Krakauer[b]

[a]*Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[b]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Received 16 June 2006; accepted 19 June 2006

Abstract

We provide a geometric framework for investigating the robustness of information flows over biological networks. We use information measures to quantify the impact of knockout perturbations on simple networks. Robustness has two components, a measure of the causal contribution of a node or nodes, and a measure of the change or exclusion dependence, of the network following node removal. Causality is measured as statistical contribution of a node to network function, whereas exclusion dependence measures a distance between unperturbed network and reconfigured network function.We explore the role that redundancy plays in increasing robustness, and how redundacy can be exploited through error-correcting codes implemented by networks. We provide examples of the robustness measure when applied to familiar boolean functions such as the AND, OR and XOR functions. We discuss the relationship between robustness measures and related measures of complexity and how robustness always implies a minimal level of complexity.

Keywords: Robustness; Complexity; Networks; Information Geometry; Knockouts

*Corresponding author.
E-mail address: nay@mis.mpg.de (N. Ay).

1431-7613/$ - see front matter © 2006 Published by Elsevier GmbH.
doi:10.1016/j.thbio.2006.06.002

Introduction

A very useful geometric abstraction for biological organizations is provided by networks (Strogatz, 2001; Newman, 2003). In network terms, individual genes, proteins, cells, organisms, or species are represented as nodes, and rules of interaction among nodes are represented by edges. For genes, edges encode the combined processes of transcription, translation and activation or inhibition (Wagner, 1994), whereas for species, the edges would represent predator–prey relationships or patterns of mutualism (Dunne et al., 2002). The rules of interaction can be conceived generally as flows, typically in the form of energy in ecology, or information transport in genetics and neuroscience. It becomes possible within this framework, to ask questions about regularities that are likely to have application at multiple levels of organization.

In previous papers (Krakauer, 2004; Krakauer and Plotkin, 2004), motivated by Chomsky's usage in linguistics (Chosmky, 1981), we have adopted a Principles and Parameters approach to robustness, in which one seeks to establish a set of very general *principles* leading to some form of robustness – considered essentially as an invariance under perturbation – and the *parameters*, or mechanisms, giving rise to the invariance in each particular case. Where we might assume that the principles will generalize to include genes, individuals and species, the parameters are likely to be system dependent.

Some important principles (instantiated by a wide range of mechanisms) of robustness are redundancy, purging and modularity. Some important parameters are component duplication, feedback control and comparmentalization (Krakauer, 2004). Biologists typically study the mechanisms of robustness. In this paper, we aim to demonstrate that some fundamental statistical requirements (principles) must be met by any mechanism effective at promoting functional invariance following perturbation.

In this paper, we introduce an information-geometric approach to robustness (Amari, 1985; Amari and Nagaoka, 2000). Information theory has the virtue of a long history of inquiry into the robustness properties of signal-receiver pairs communicating over noisy, finite bandwidth channels (Cover and Thomas, 2001). Where we depart from standard approaches, is to consider structural perturbations to the model in the form of the knockout or elimination of channels, and consider the set of communication strategies which result in the least departure from unperturbed network function. This deviation is captured, quantitatively in terms of a distance between probability distributions, and makes use of insights from information geometry (Amari, 1985; Amari and Nagaoka, 2000).

One important emphasis in this paper is that a general mathematical definition of robustness need include at least two components: some measure of insensitivity following perturbation (present in all treatments of robustness) and a measure of flow through the system prior to the perturbation (novel to this contribution). To speak of a non-trivial robust system it is necessary to demonstrate that the system performs interesting aggregate behavior in advance of perturbation. By not including the "causal" contribution of nodes, an inert system, or systems of non-interacting

components, is liable to be deemed robust. This suggests an important connection between the notion of robustness and the notion of complexity, a connection that we explore in this paper.

The paper is arranged as follows. We firstly introduce the mathematical formalism for flows over networks defined in information theoretic variables. We then introduce a new probabilistic approach to network perturbation. We introduce a novel definition of robustness in terms of the joint causal contribution of nodes to network function and the functional exclusion dependence on these nodes. For clarity, the mathematics is illustrated throughout the paper with a simple network consisting of multiple inputs and a single output. We then show how this network can be extended to general recurrent network structures. We apply the robustness measure to simple boolean functions, and discuss the construction of robust networks which implement simple error-correcting codes. Finally, we explore the connections of our robustness measure to related robustness studies and the correspondence between robustness and statistical measures of complexity. It is our position that selection for increased robustness is one of the primary drivers for increasing network complexity.

Notation and preliminaries

Stochastic maps

In this section, we introduce a simple network structure which forms the basis of our subsequent analysis. This simple network is intended to capture the structure of a component of a larger genetic regulatory circuit, a component of the nervous system or a trophic network in ecology. All of these systems can be thought about in terms of the flow of energy or information through a network. In the final section of the paper we introduce a way of thinking about the more general setting of recurrent networks. Here, we analyze feed-forward components of the network as shown in Fig. 1. We choose to describe these network flows in terms of a stochastic map T, which describes how a node labeled as 0 generates an output y based on information from an input vector $x_1, \ldots, x_N$. The output node pertains to some property of system function, such as gene activation, postsynaptic firing, or successful predation, whereas the inputs refer to the set of transcription factors, presynaptic activities, or prey species.

In order to simplify our notation, we write Λ for the set $\{1, \ldots, N\}$ of input units and Λ_0 for the set $\Lambda \cup \{0\}$ of all units. The state set of a unit $i \in \Lambda_0$ is denoted by $\boldsymbol{X}_i$. For simplicity, we assume that their corresponding cardinalities, written as $n_i := |\boldsymbol{X}_i|$, are finite. The formal description of the transformation T is given by a Markov transition matrix

$$T : \boldsymbol{X}_\Lambda \times \boldsymbol{X}_0 \to [0, 1], \quad (x, y) \mapsto T(y|x),$$

where T is the function performed by the network, and the input set is given by $\boldsymbol{X}_\Lambda := \boldsymbol{X}_1 \times \cdots \times \boldsymbol{X}_N$. The value $T(y|x)$ is the conditional probability of generating

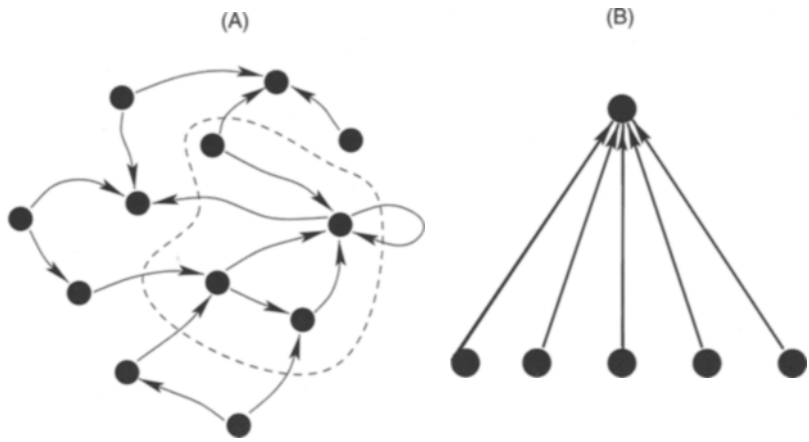

Fig. 1. A schematic representation of a general network structure in which nodes represent genes, cells, or individuals, and edges flows of information or energy between nodes. We isolate a feed-forward component of the network as shown in (B) to include those nodes circled in (A). In this simple structure, a single node receives inputs from a number of nodes in its parent set. The receiving node integrates the inputs and produces an output. For example, in this figure, a transformation T has $N = 5$ inputs.

the output y given the input $x = (x_1, \ldots, x_N)$. This implies that for every $x \in \boldsymbol{X}_\Lambda$,

$$\sum_{y \in \boldsymbol{X}_0} T(y|x) = 1, \tag{1}$$

to conserve total probability.

Given a deterministic map $f : \boldsymbol{X}_\Lambda \to \boldsymbol{X}_0$, we can always consider it as a special case of a Markov transition matrix T^f as follows:

$$T^f(y|x) := \begin{cases} 1 & \text{if } y = f(x), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Example 1. Consider the simple network structure of Fig. 1B. Nodes are assumed to have the two states "0 = not active" and "1 = active", corresponding to the presence or absence of a gene, an active neuron or a species.[1] Thus, one has $\boldsymbol{X}_i = \{0, 1\} \subset \mathbb{R}$ for all units $i \in \Lambda_0$. The system parameters are the *edge/channel weights* w_i, $i \in \Lambda$, which describe the strength of interaction among the individual input nodes i and the output node 0, and a *threshold value* θ for the output node which controls its sensitivity to the input. In the case of genetics, the weight corresponds to the binding strength of a transcription factor and the threshold the minimum activation concentration. In neuroscience the weight describes the product of the

[1] In further examples of the paper we will also use "−1" and "+1" as states, whereby we interpret a "0"-state as being knocked out.

density of postsynaptic receptors and neurostransmitter, and the threshold controls the sensitivity. In ecology, the weight would describe the efficiency of energy transduction and the threshold the minimum energy conversion required for individual survival. In all cases we assume that given an input vector $x = (x_i)_{i\in\Lambda} \in \{0,1\}^N$ in a first step, the output node assumes a value given by the function

$$h(x) = \sum_{i\in\Lambda} w_i x_i - \theta,$$

and then, in a second step, it generates the output 1 with the probability

$$T^\beta(1|x) = \frac{1}{1+\mathrm{e}^{-\beta\, h(x)}}.$$

The normalization property (1) then implies that the output 0 is generated with probability $1 - T(1|x)$. Here, the *inverse temperature* β controls the stochasticity of the map T^β. It is easy to see that in the generic case where all points $\{0,1\}^N$ are contained either in the positive halfspace $\mathscr{H}^+ = \{x \in \{0,1\}^N : h(x)>0\}$ or in the negative halfspace $\mathscr{H}^- = \{x \in \{0,1\}^N : h(x)<0\}$, the transformation T^β converges for $\beta \to \infty$ to a deterministic transformation T^f according to (2).

Whereas in Example 1 and in most further examples of this paper we restrict ourselves to direct pairwise interactions, the framework is sufficiently general to apply to higher-order interactions, such as we observe in cascades and networks of interactions.

Some information-theoretic definitions

As we stated in the Introduction, this paper adopts principles from information theory, to construct a general theory of robustness in biological networks. In this section, we review some definitions of basic quantities from information theory that we shall need later in developing our definition of robustness. Those interested in a more detailed exposition are directed to (Cover and Thomas, 2001).

Given an arbitrary subset $S \subseteq \Lambda_0$, we write $\boldsymbol{X}_S$ rather the more cumbersome notation $\prod_{i\in S} \boldsymbol{X}_i$, and we consider the natural projection

$$X_S : \boldsymbol{X}_{\Lambda_0} \to \boldsymbol{X}_S, \quad x = (x_i)_{i\in\Lambda} \mapsto x_S = (x_i)_{i\in S}.$$

With an input distribution p on $\boldsymbol{X}_\Lambda$ and a stochastic map T from $\boldsymbol{X}_\Lambda$ to $\boldsymbol{X}_0$ we have the joint probability vector

$$P(X_\Lambda = x, X_0 = y) = p(x)T(y|x), \quad x \in \boldsymbol{X}_\Lambda,\ y \in \boldsymbol{X}_0. \tag{3}$$

The projection X_S becomes a random variable with respect to P. Now consider three subsets $A, B, C \subseteq \Lambda_0$. The *entropy* of X_C is then defined as

$$H_P(X_C) = -\sum_{z\in\boldsymbol{X}_C} P(X_C = z)\ln(P(X_C = z)).$$

This quantity is a natural measure of the uncertainty that one has about the outcome of X_C. Once we know the outcome, this uncertainty is then reduced to zero. This justifies the interpretation of $H_P(X_C)$ as the *information gain* after knowing the

outcome of X_C. Now, having information about the outcome of the second variable X_B reduces the uncertainty about X_C. More precisely, the *conditional entropy* of X_C given X_B is defined as

$$H_P(X_C|X_B) = -\sum_{y\in X_B, z\in X_C} P(X_B = y,\ X_C = z)\ln(P(X_C = z|X_B = y)),$$

and we have $H_P(X_C)\geqslant H_P(X_C|X_B)$. Using these entropy terms, the *mutual information* of X_C and X_B is then defined as the uncertainty of X_C minus the uncertainty of X_C given X_B:

$$I_P(X_C : X_B) = H_P(X_C) - H_P(X_C|X_B).$$

The *conditional mutual information* of X_C and X_B given X_A is defined in a similar way:

$$I_P(X_C : X_B|X_A) = H_P(X_C|X_A) - H_P(X_C|X_A, X_B).$$

In the following, we will simplify the notation by writing these quantities without explicitly mentioning P, which is always assumed to be given by an input distribution p and a transformation T according to (3).

Source exclusion

The geometry of source exclusion

Generally, the robustness of a system refers to approximate invariance of function with respect to perturbation, where the robustness is based on structural reorganization of the system initiated by the perturbation. In our context, we are interested in the robustness of a transformation T. We consider perturbations consisting in the exclusion of some input sources/nodes, so that the input comes from a subset $A \subseteq \Lambda$ of the original input nodes. The effect of a knockout on the output node 0 is to induce a switch to a new transformation T_A, which we refer to as a different mechanistic modality. In other words, the transformation T is replaced by a transformation T_A which describes the value of the output node after exclusion of $\Lambda\backslash A$.

Obviously, the transformation T_A has to be an element of the following set:

$$\begin{aligned}\mathcal{T}_A &:= \text{set of transitions that do not depend on the input from} \Lambda\backslash A \\ &= \{T : T(y|x_A, x_{\Lambda\backslash A}) = T(y|x_A, x'_{\Lambda\backslash A}) \text{ for all } x_A, x_{\Lambda\backslash A}, x'_{\Lambda\backslash A}, y\}.\end{aligned}$$

In order to quantify the effect of the exclusion, we consider the following distance-like measure, which is known variously as *information divergence*, *relative entropy*, and *Kullback–Leibler* (*KL*) *distance* of a transition T_1 from a transition T_2 with respect to an input distribution p:

$$D_p(T_1\|T_2) := \sum_{x\in X_\Lambda,\ y\in X_0} p(x)T_1(y|x)\ln\frac{T_1(y|x)}{T_2(y|x)}.$$

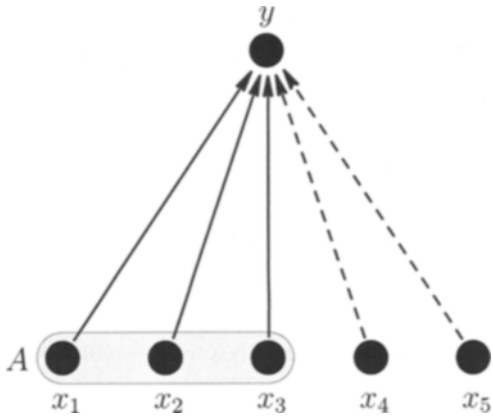

Fig. 2. An example of a knockout procedure: we illustrate the exclusion of the nodes forming the set {4, 5} from the complete set {1, 2, 3, 4, 5} of input nodes and we ask what impact this has on the value of y.

We have $D_p(T_1 \| T_2) \geqslant 0$, and $D_p(T_1 \| T_2) = 0$ if and only if $T_1(\cdot|x) = T_2(\cdot|x)$ for all x with $p(x) > 0$. Now, the effect of the replacement of T by T_A is naturally measured by the distance $D_p(T \| T_A)$. This distance offers a way of comparing probability distributions and here, Markov transitions (Figs. 2 and 3).

Exclusion dependence and invariance

Taking all possible exclusions into account, we have to consider a family $(T_A)_{A \subseteq \Lambda}$ of transformations that describe the different mechanistic modalities for which T_Λ coincides with T. Modeling the expectation of exclusion by a probability measure $r(A)$, $A \subseteq \Lambda$, that is $r(A) \geqslant 0$ for all A, and $\sum_{A \subseteq \Lambda} r(A) = 1$, we define the (*expected*) *exclusion dependence* as

$$C(r, p, (T_A)_{A \subseteq \Lambda}) := \sum_{A \subseteq \Lambda} r(A)\, D_p(T_\Lambda \| T_A). \tag{4}$$

We say that $(T_A)_{A \subseteq \Lambda}$ is (r,p)-*exclusion independent* if the corresponding exclusion dependence vanishes. A trivial example of exclusion independence is given by a family of transformations that do not depend on any input source. More precisely, with an output probability measure q on X_0, we can define a family of transformations by

$$T_A(y|x) := q(y), \quad A \subseteq \Lambda. \tag{5}$$

Here, T_A is equal to T_Λ for all A, and therefore the exclusion dependence vanishes. This example highlights a very important aspect of robustness that we shall return to in later sections. In particular, that exclusion independence is a necessary component of robustness by leading to invariance, but is not sufficient, as systems of non-interacting nodes are those least effected by knockout. We shall see later that

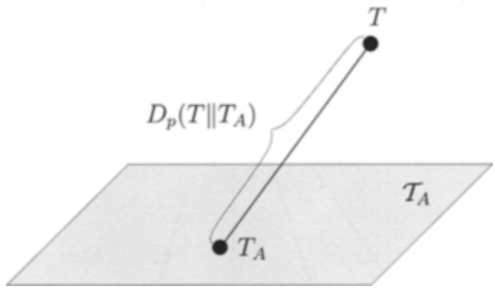

Fig. 3. The effect of the replacement of T by T_A following knockout. The plane $\mathcal{T}_A$ represents the space of transformations that depend only on the input from A. The point T_A is a particular Markov transition situated at a distance $D_p(T\|T_A)$ from the complete transformation T. This distance quantifies the effect size.

robustness requires that a system maintains some information flow while maximizing exclusion independence.

One way of thinking about non-trivial cases of exclusion dependence, is that a low exclusion dependence arises through switching from one transformation T_A to another transformation T_B while keeping the system behavior invariant. This is illustrated in Fig. 4.

The following theorem makes this invariance property more precise:

Theorem 1. *Let $T_A \in \mathcal{T}_A$, $A \subseteq \Lambda$, a family of transformations. Then*

$$\sum_{A,B\subseteq\Lambda,\ x\in X_\Lambda,\ y\in X_0} r(A)\,r(B)\,p(x)\,|T_A(y|x) - T_B(y|x)| \leqslant 2\cdot\sqrt{2C(r,p,(T_A)_{A\subseteq\Lambda})}. \tag{6}$$

In particular, we have complete invariance if $(T_A)_{A\subseteq\Lambda}$ is (r,p)-exclusion independent.

Proof.

$$\sum_{y\in X_0}|T_A(y|x) - T_B(y|x)| \leqslant \sum_{y\in X_0}|T_A(y|x) - T_\Lambda(y|x)| + \sum_{y\in X_0}|T_B(y|x) - T_\Lambda(y|x)|$$
$$\leqslant \sqrt{2D_p(T_A(\cdot|x)\|T_A(\cdot|x))} + \sqrt{2D_p(T_A(\cdot|x)\|T_B(\cdot|x))}. \tag{7}$$

With Jensen's inequality, this leads to

$$\sum_{x\in X_\Lambda} p(x)\sum_{y\in X_0}|T_A(y|x) - T_B(y|x)| \leqslant \sqrt{2D_p(T_\Lambda\|T_A)} + \sqrt{2D_p(T_\Lambda\|T_B)},$$

and

$$\sum_{A,B\subseteq\Lambda} r(A)r(B)\sum_{x\in X_\Lambda} p(x)\sum_{y\in X_0}|T_A(y|x) - T_B(y|x)|$$
$$\leqslant 2\cdot\sqrt{2\sum_{A\subseteq\Lambda} r(A)\,D_p(T_\Lambda\|T_A)}. \qquad \square$$

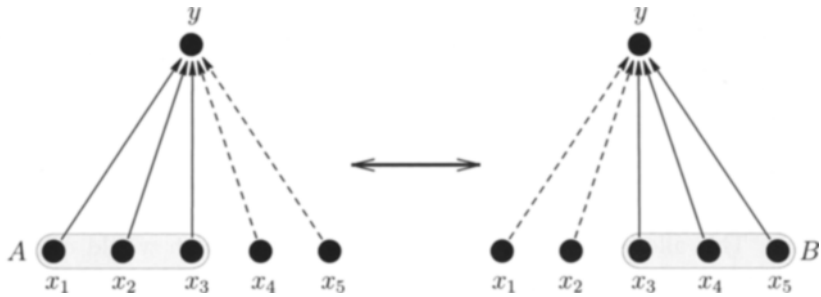

Fig. 4. Switching between the input sets $\{1,2,3\}$ and $\{3,4,5\}$ while keeping the output y the same. In this case, the network is able to extract the same information from $\{1,2,3\}$ as $\{3,4,5\}$ by switching to a new transformation.

The left-hand term of (6) refers to the mean L^1- distance between the mechanistic modalities or transformations, T_A and T_B, averaging over all subsets of Λ and all inputs x. The right-hand side represents the exclusion dependence defined by (4). This implies that systems with low exclusion dependence will lead to mechanistic modalities with similar behavior.

The role of adaptation and mechanistic constraints

After the exclusion of some input sources the output unit has to use a different transformation T'_A, which is in many cases not directly related to the global transformation T so that we have a high dependence of T on the exclusion of $\Lambda\backslash A$. On the other hand, low exclusion dependence is the property that underlies the invariance of function. If the system has the internal flexibility or variability that allows for an adaptation of T'_A so as to move closer to the global transformation T, then it will modify parameters in such a way that the KL distance is as small as possible. We illustrate this important aspect of adaptation within the context of Example 1. Excluding some of the input sources so that the only available sources derive from the subset $A \subseteq \Lambda$ means that the energy function $h(x) = \sum_{i\in\Lambda} w_i x_i - \theta$ has to be replaced by $h_A(x) = \sum_{i\in A} w_i x_i - \theta$. This leads to the transformation given by

$$T'_A(1|x) = \frac{1}{1 + e^{-\beta h_A(x)}},$$

and in general, it is not true that $\inf_{T'\in\mathcal{T}_A} D_p(T\|T') = D_p(T\|T'_A)$. In other words, the parameters w_i, $i \in A$, have to be adjusted in order to minimize the distance $D_p(T\|T')$. The transition T'_A describes the function of the network after removal of the source channels $\Lambda\backslash A$ and before the adaptation process leads to a reconfiguration T_A of the transformation T'_A. As a result of mechanistic constraints it is generally not possible to find a transformation T_A that satisfies

$$D_p(T\|T_A) = D_p(T\|\mathcal{T}_A) := \inf_{T'\in\mathcal{T}_A} D_p(T\|T'). \tag{8}$$

In this example, these constraints take the form of the exclusively pairwise interactions among the input units and the output unit. In order to arrive at T_A with the minimality property (8) it will in practice often be necessary to include higher-order interactions, such as those among triplets and above. To work with a purely probabilistic notion of exclusion dependence, we assume that the mechanistic flexibility or variability of the system is always sufficient to achieve this minimality property. In other words we start from the optimally adapted configuration given constraints. This allows us to make precise statements which would otherwise be confounded by sub-optimal switching. If T_A satisfies (8) then it has a nice interpretation as a projection onto $\mathcal{T}_A$, which is shown in Fig. 5. This projection is uniquely defined for all inputs x with $p(x)>0$.

One can easily prove that a projection T_A satisfies

$$T_A(y|x) = P(X_0 = y|X_A = x_A) = \frac{\sum_{\substack{x' \in x_\Lambda \\ x'_A = x_A}} p(x')T(y|x')}{\sum_{\substack{x' \in x_\Lambda \\ x'_A = x_A}} p(x')}, \tag{9}$$

whenever $P(X_A = x_A)>0$.

With the family $(T_A)_{A\subseteq\Lambda}$ given by (9) we define the exclusion dependence of T with respect to r and p as follows:

$$C(r,p,T) := \sum_{A\subseteq\Lambda} r(A) \inf_{T'\in\mathcal{T}_A} D_p(T\|T') = C(r,p,(T_A)_{A\subseteq\Lambda}). \tag{10}$$

Thus, the exclusion dependence of a transformation T gives us information about the effect of source exclusion, in those cases where the system has the mechanistic flexibility or variability to discover the best behavior. In this sense our definition does

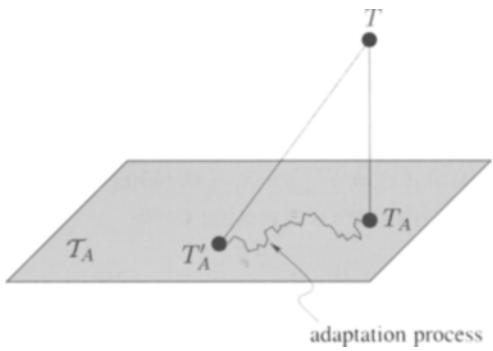

Fig. 5. Following perturbation, the network performs transformation T'_A which depends only on input units A. The network then adapts in order to restore towards its previous function T. Typically this restoration will not return the system to T as: (1) information has been lost from the inputs and this restricts outputs to fall into the set $\mathcal{T}_A$ (2) the parameters or edge values of the network impose constraints on the variability of the system to adapt within $\mathcal{T}_A$. Assuming that mechanistic constraints do not limit adaptation, then the best the network can do is return to T_A. The distance between T_A and T we refer to as the exclusion dependence.

not include the concrete information about the underlying mechanisms of adaptation, and therefore it represents a probabilistic quantity.

Example 2. We consider two binary input units with the states -1 and $+1$. In this example, we want to calculate the exclusion dependence of the boolean functions XOR, NOR, AND, and OR in their ± 1-representation:

XOR	−1	+1
−1	−1	+1
+1	+1	−1

NOR	−1	+1
−1	+1	−1
+1	−1	−1

AND	−1	+1
−1	−1	−1
+1	−1	+1

OR	−1	+1
−1	−1	+1
+1	+1	+1

We assume that the input distribution p_β on the configuration set $\{-1,+1\}^2$ is parametrized by a parameter $\beta \in \mathbb{R}$ as follows:

$$p_\beta(x_1,x_2) := \frac{e^{\beta x_1 x_2}}{2(e^{\beta}+e^{-\beta})}.$$

We can control the correlation of the inputs with the parameter β, and we have completely correlated function when,

$$\lim_{\beta\to\infty} p_\beta = \tfrac{1}{2}(\delta_{(-1,-1)}+\delta_{(+1,+1)}), \qquad \lim_{\beta\to-\infty} p_\beta = \tfrac{1}{2}(\delta_{(-1,+1)}+\delta_{(+1,-1)}).$$

For $\beta = 0$, the inputs are independent. Finally, we have to specify the knockout distribution. To this end we consider a probability $\alpha \in [0,1]$ for an input node to remaining an information source after knockout. With an independence assumption this gives us

$$r(\emptyset) = (1-\alpha)^2, \quad r(\{1\}) = r(\{2\}) = \alpha(1-\alpha), \quad r(\{1,2\}) = \alpha^2.$$

The exclusion dependence of the boolean functions is shown in Fig. 6. It depends on the two parameters α and β. It is in all cases maximal for $\beta = 0$, which means independence of the inputs. Furthermore, for $\alpha = 1$, which means no knockout, we have vanishing exclusion dependence. Fig. 6 also shows the sections given by fixed value $\alpha = \frac{1}{2}$.

Exclusion independence with error-correcting codes

In this section, we consider distances in terms of the information-theoretic measures introduced in earlier sections. We do this to make the redundancy property of the input distribution more transparent. Rewriting the distance of a transformation T from a projection T_A with respect to p as

$$D_p(T\|T_A) = I(X_0 : X_{A\setminus A}|X_A) \tag{11}$$

implies the following version of $C(r,p,T)$:

$$C(r,p,T) = \sum_{A\subseteq\Lambda} r(A)\, I(X_0 : X_{A\setminus A}|X_A).$$

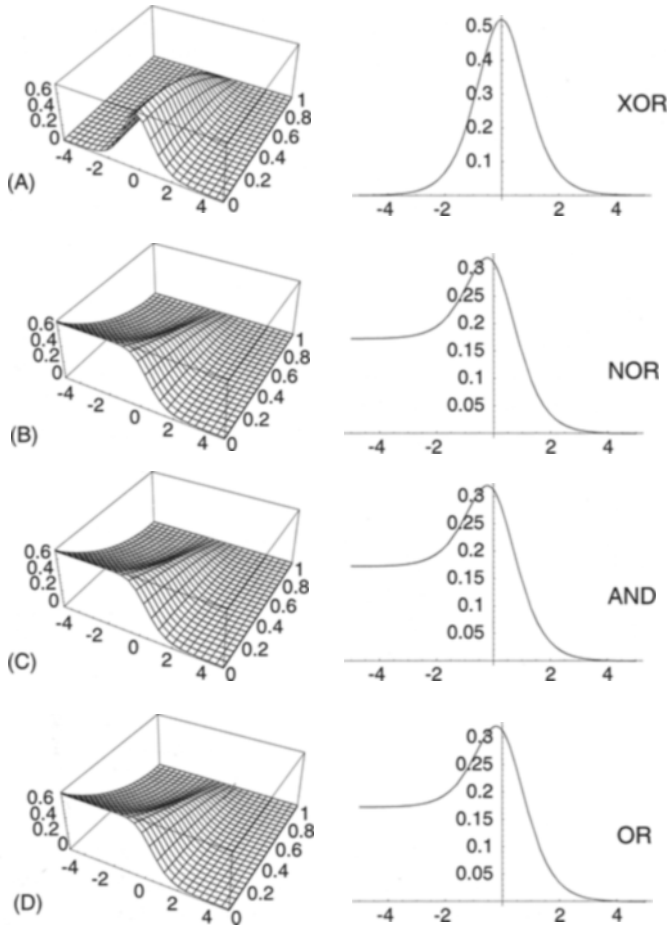

Fig. 6. Exclusion dependence quantified for four boolean operations and a range of knockout probabilities given by α, $0 \leqslant \alpha \leqslant 1$ (left panel), and the corresponding sections for $\alpha = \frac{1}{2}$ (right panel).

Although the exclusion dependence explicitely involves the specification of the transformation T, there is an interesting upper bound that just depends on the input distribution. We can see this by decomposing the mutual information term.

Proposition 1.

$$C(r,p,T) \leqslant \sum_{A \subseteq A} r(A)\, H(X_{A \setminus A} | X_A). \tag{12}$$

Proof. We give an upper bound of the conditional mutual informations that appear in (11):

$$I(X_0 : X_{A \setminus A} | X_A) = H(X_{A \setminus A} | X_A) - H(X_{A \setminus A} | X_0, X_A) \leqslant H(X_{A \setminus A} | X_A). \qquad \square$$

The upper bound in estimate (12) controls the redundancy of the input distribution. The lower it is, the more redundant is the input distribution. In the most redundant case where

$$\sum_{A\subseteq\Lambda} r(A)H(X_{\Lambda\setminus A}|X_A) = 0,$$

there are maps $f_A : \boldsymbol{X}_A \to \boldsymbol{X}_{\Lambda\setminus A}$, $A \subseteq \Lambda$, so that for all A with $r(A)>0$ and all x with $p(x)>0$, the following holds:

$$x_{\Lambda\setminus A} = f_A(x_A).$$

Estimate (12) provides a method of controlling the exclusion dependence by varying the redundancy properties of the input distribution. One nice application of this method comes from the theory of error-correcting codes. The Hamming distance of two input strings $x = (x_i)_{i\in\Lambda}$ and $x' = (x'_i)_{i\in\Lambda}$ is defined as $d(x,x') = |\{i \in \Lambda : x_i \neq x'_i\}|$. Now consider a non-empty set $\mathscr{C} \subseteq \boldsymbol{X}_\Lambda$ and the corresponding minimal distance

$$d(\mathscr{C}) := \min\{d(x,x') : x, x' \in \mathscr{C},\ x \neq x'\}.$$

With a maximal number $t \in \mathbb{N}$ of excluded input units satisfying $2t+1 \leqslant d(\mathscr{C})$, and with δ_x denoting the Dirac measure concentrated in x, we define

$$p := \frac{1}{|\mathscr{C}|}\sum_{x\in\mathscr{C}} \delta_x, \tag{13}$$

and

$$r(A)>0 \quad \text{if } |A| \geqslant N - t.$$

Now assume that we have the information $x_A = (x_i)_{i\in A}$ from an input $x = (x_i)_{i\in\Lambda} \in \mathscr{C}$, where $|A| \geqslant N - t$ elements. For the reason that $2t+1 \leqslant d(\mathscr{C})$, there is no other $x' \in \mathscr{C}$ with $x'_A = x_A$ than x, and x can be recovered from the fragmentary information x_A. More precisely, $x_A = x'_A$ implies the following contradiction:

$$d(\mathscr{C}) \leqslant d(x,x') \leqslant N - |A| \leqslant N - (N-t) = t < 2t+1 \leqslant d(\mathscr{C}).$$

Thus, we have $H(X_{\Lambda\setminus A}|X_A) = 0$ for all A with $r(A)>0$. According to inequality (12), this implies

$$C(r,p,T) = 0.$$

In this case every transformation T is (r,p)-exclusion independent. Now we assume that p is given by a code $\mathscr{C}$ according to (13) with a minimal distance $d(\mathscr{C})$ that allows recovery of exclusions of at most t input variables, that is $2t+1 \leqslant d(\mathscr{C})$. It is not hard to see that for two subsets A and B of Λ with at least $N-t$ elements, and for all $x \in \boldsymbol{X}_0$ with $p(x)>0$:

$$T_A(y|x) = T_B(y|x). \tag{14}$$

In other words, the information from A can be replaced by the information from B if A and B are sufficiently large, and we have complete invariance in the sense of Theorem 1. This statement goes beyond asserting that A and B are mere duplicates, and relates to statistical correlations between A and B (Fig. 7).

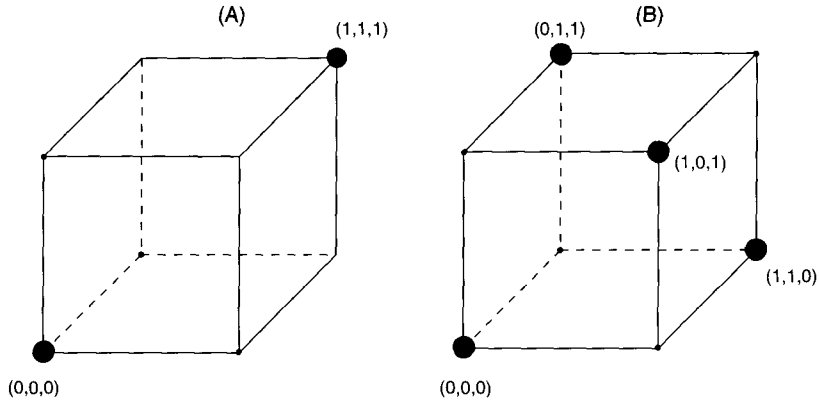

Fig. 7. One way in which networks can respond robustly to perturbations is to implement error-detecting codes. In these examples we present two codes $\mathscr{C}_1$ and $\mathscr{C}_2$ in $\{0,1\}^3$. The code $\mathscr{C}_1$ in (A) has two code words with $d(\mathscr{C}_1) = 3$, and the code $\mathscr{C}_2$ in (B) has four code words with $d(\mathscr{C}_2) = 2$. The first code is 1-bit error correcting and the second is 1-bit error detecting. Error correction in A is possible as the correct sequence can always be recovered by tracing the shortest path (1-bit) back to the appropriate code word.

Information flow and robustness

A definition of information flow

In this section we provide a formal definition of robustness. As we shall see, the concept of exclusion dependence is closely related to the concept of robustness. Intuitively, a low exclusion dependence of a transformation T contributes to the robustness of T. On the other hand, we have seen the example (5) of trivial exclusion independence where the transformation is not effected by exclusion of a subsystem simply because it does not use any information from this subsystem. In order to have a measure for robustness, we have to relate the dependence of the system behavior from the exclusion of a subsystem to the contribution of that subsystem. In order to quantify the contribution of the subsystem, we follow Pearl's approach to causality (Pearl, 2000). Within this approach, the notion of intervention is essential. One has to make a distinction between observing that a variable X is in state x and forcing X to be in state x, which is related to experimental intervention. Let us apply this to the following situation: we have two input variables X and Y, and one output variable Z. This output is generated according to a transition matrix $T(z|x,y)$. With respect to the input distribution p, X and Y may be correlated. In order to put this situation into Pearl's framework of a directed acyclic graph (DAG), we consider an additional variable W that gives rise to the statistical dependence of X and Y. We get the diagram shown in Fig. 8.

With respect to the graph of Fig. 8A the joint distribution of the vector (W, X, Y, Z) can be written as

$$P(W = w, X = x, Y = y, Z = z) = P(W = w)P(X = x|W = w)P(Y = y|W = w)T(z|x, y).$$

It is easy to see that, with respect to this distribution,

$$P(Z = z|X = x) = \frac{\sum_y p(x, y)T(z|x, y)}{\sum_y p(x, y)}. \tag{15}$$

It is exactly the projection of T with respect to p that we had before (compare with (9)), and it is the conditional probability of observing $Z = z$ after having observed $X = x$. In order to describe information flows and causal effects, in (15) we have to replace the observation $X = x$ by the intervention method of *doing* $X = x$, written as $\mathrm{do}(X = x)$. Following Pearl, the corresponding joint distribution of the remaining variables W, Y, Z

$$P(W = w, Y = y, Z = z|\mathrm{do}(X = x)) = P(W = w)P(Y = y|W = w)T(z|x, y).$$

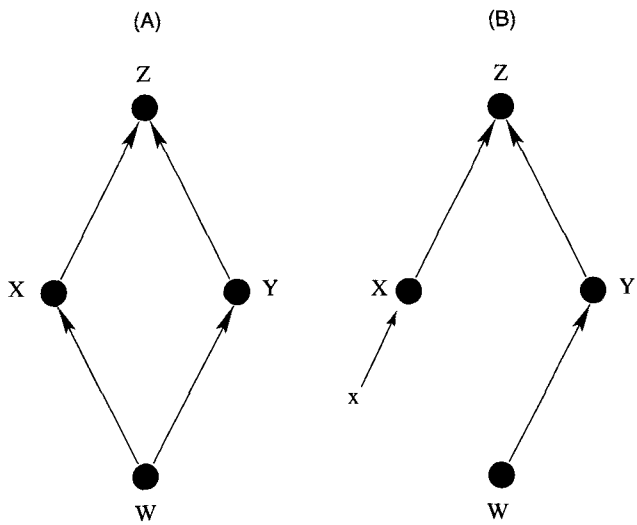

Fig. 8. Simple example of statistical causality presented in terms of two directed acyclic graphs, where we wish to determine the contribution of variable Y to variable Z. In (A) correlations between X and Y arise as the result of a shared input from W. In order to capture the contribution of Y to Z we fix the value of $X = x$. This is shown in (B). If Y does not contribute to Z then there is no information flow from Y to Z and this can only be ascertained for a fixed value of X.

Now we compute the conditional probability of observing $Z = z$ after setting $X = x$:

$$P(Z = z|\mathrm{do}(X = x)) = \sum_{w,y} P(W = w)P(Y = y|W = w)T(z|x, y)$$
$$= \sum_{y} p(y)T(z|x, y). \tag{16}$$

Note that this is different from formula (15), if X and Y are not independent. In other words, the marginalization with respect to the do-intervention does not use any correlation between X and Y. This implies that Z does not depend on Y, if and only if $P(Z = z|\mathrm{do}(X = x)) = T(z|x, y)$ for all x, y, z. Thus, it is natural to measure the contribution of Y to Z by the deviation of the transition T from the corresponding interventionally generated conditional distribution $P(Z = \cdot|\mathrm{do}(X = \cdot))$. Through analogy with exclusion dependence, we quantify the contribution of Y to the output Z by the KL distance of the transition kernel from its marginal (16):

$$\mathrm{IF}(Y \to Z, p, T) := \sum_{x,y,z} p(x, y)T(z|x, y) \ln \frac{T(z|x, y)}{\sum_{y'} p(y')\, T(z|x, y')}. \tag{17}$$

It is not hard to see that

$$\mathrm{IF}(Y \to Z, p, T) \leqslant H(Y).$$

In other words, if the entropy of Y is low, then we cannot expect a high contribution of that subsystem to the function. But on the other hand, high entropy of Y does not necessarily imply high information flow from Y to Z. If T does not use any information from Y, that is $T(z|x, y) = T(z|x, y')$ for all $x \in \boldsymbol{X}$, $y, y' \in \boldsymbol{Y}$, and $z \in \boldsymbol{Z}$, then we have

$$\sum_{y'} p(y')\, T(z|x, y') = \sum_{y'} p(y')\, T(z|x) = T(z|x) = T(z|x, y).$$

Therefore, in this case we have $\mathrm{IF}(Y \to Z, p, T) = 0$.

Example 3. Consider the state spaces

$$\boldsymbol{X} = \boldsymbol{Y} = \boldsymbol{Z} = \{-1, +1\},$$

and the input distribution p and the family $T^{\gamma,\delta}$, $\gamma.\delta \in \mathbb{R}$, of transition kernels defined as

$$p(x, y) := \frac{\mathrm{e}^{xy}}{2(\mathrm{e}^{1} + \mathrm{e}^{-1})}, \quad T^{\gamma,\delta}(z|x, y) := \frac{\mathrm{e}^{\gamma zx + \delta zy}}{\mathrm{e}^{\gamma x + \delta y} + \mathrm{e}^{-\gamma x - \delta y}}. \tag{18}$$

The plot in Fig. 9B shows the shape of the information flow $\mathrm{IF}(Y \to Z, p, T^{\gamma,\delta})$ according to our formula (17):

From Fig. 9B we see that the maximal information flow from Y to Z is in the region where $|\delta| > |\gamma|$. This shape shows that in order to have any contribution of Y, the absolute value of its coupling δ with Z has to be greater than the absolute value of the coupling γ of X with Z.

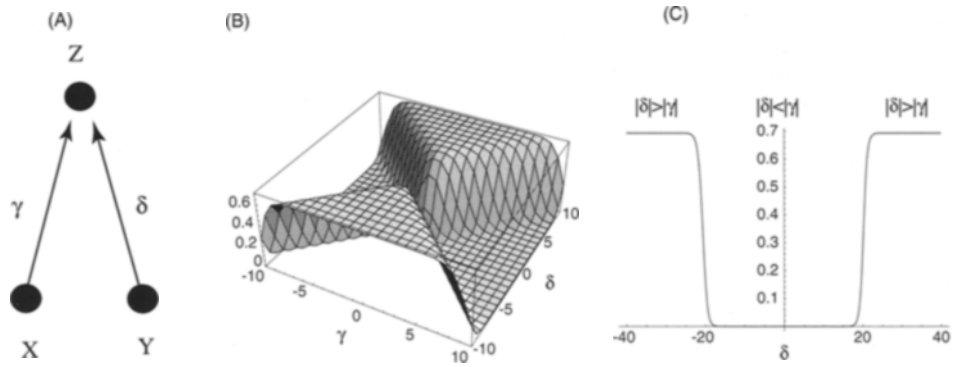

Fig. 9. An example of statistical causality in a network. In (A) we depict a simple model with two parameters, γ and δ which serve as the edge weights. The two inputs are assumed to be correlated as in (18). (B) shows the contribution of Y to the values of Z for a range of parameter values as given by (17). This is plotted for a single value of $\gamma = 20$ in (C), which shows how there is no causal contribution from Y to Z until $|\delta|$ has exceeded $|\gamma|$.

A definition of robustness

Now we are ready to define robustness: first, we apply our robustness idea to the situation of three variables X, Y, Z, which we have discussed in the context of information flow. The variable Z is computed from the pair (X, Y), and we exclude Y. How robust is this computation against this exclusion? Of course, the system should be able to compensate this exclusion, that is, the exclusion dependence should be low. But considering this as the only source for robustness would imply that we have high robustness in the case where Z does not depend on Y at all. In order to avoid this, we have to relate the exclusion dependence to the contribution of Y within the computation of Z. Therefore, we define the *robustness* against the exclusion of Y as the information flow $Y \to Z$ minus the exclusion dependence. This leads to the following formula:

$$R(X,p,T) := \sum_{x,y,z} p(x,y)T(z|x,y) \ln \frac{\sum_y p(x,y)T(z|x,y)}{p(x)\sum_y p(y)T(z|x,y)}. \tag{19}$$

This measures the amount of statistical dependence between x and y that is used for computing z in order to compensate the exclusion of y. The robustness vanishes if for all x and all z

$$\sum_y p(x,y)T(z|x,y) = p(x)\sum_y p(y)T(z|x,y)$$

or equivalently

$$\sum_y T(z|x,y)(p(x,y) - p(x)p(y)) = 0 \tag{20}$$

There are two extreme cases where this equality holds. The first case is when there is no statistical dependence between x and y that can be used for compensation. Then

$p(x,y) = p(x)p(y)$, and equality (20) holds. The other extreme case is when there is statistical dependence, but this dependence is not used by T. In this case $T(z|x,y) = T(z|x,y')$ for all y, y'. Now we analyze the situation where the robustness becomes maximal.

Theorem 2. *The robustness measure is convex in T, and we have the following upper bound*:

$$R(X,p,T) \leqslant \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} = I(X:Y). \tag{21}$$

Let $f : \boldsymbol{X} \times \boldsymbol{Y} \to \boldsymbol{Z}$ be a (deterministic) function that satisfies the following partial injectivity condition: For all x we have

$$f(x,y) = f(x,y') \quad \Rightarrow \quad y = y'.$$

Then the upper bound of (23) *is attained, that is $R(X,p,T^f) = I(X:Y)$. (Here T^f is defined according to* (2))

Proof. Using the joint convexity of the KL distance, we see that $R(X,p,T)$ is convex in T.

We obtain estimate (23) by using the log sum inequality:

$$\begin{aligned} R(X,p,T) &= \sum_{x,y,z} p(x,y)T(z|x,y) \ln \frac{\sum_y p(x,y)T(z|x,y)}{p(x)\sum_y p(y)T(z|x,y)} \\ &\leqslant \sum_{x,y,z} p(x,y)\, T(z|x,y) \ln \frac{p(x,y)T(z|x,y)}{p(x)p(y)T(z|x,y)} \\ &\quad \text{(log sum inequality)} \\ &= \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

Now let f satisfy the partial injectivity condition as defined in the theorem. Then

$$\begin{aligned} R(X,p,T^f) &= \sum_{x,y,z} p(x,y)\delta_{f(x,y)}(z) \ln \frac{\sum_y p(x,y)\delta_{f(x,y)}(z)}{p(x)\sum_y p(y)\delta_{f(x,y)}(z)} \\ &= \sum_{x,y,z} p(x,y)\delta_{f(x,y)}(z) \ln \frac{p(x,y)\delta_{f(x,y)}(z)}{p(x)p(y)\delta_{f(x,y)}(z)} = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

Here we used the fact that the sums $\sum_y p(x,y)\delta_{f(x,y)}(z)$ and $\sum_y p(y)\delta_{f(x,y)}(z)$ consist of one summand at most. □

Example 4. Consider a first variable X_{past} which represents knowledge of the past and a second variable X_{future} which represents the knowledge of the future. Having access to the future allows for the performance of the perfect "prediction" which is given by the map $f : (x_{\text{past}}, x_{\text{future}}) \mapsto x_{\text{future}}$. On the other hand, prediction within a non-anticipating situation can be interpreted as performing this projection f after

knockout of the future variable. For the reason that f satisfies the partial injectivity condition of Theorem 2, the robustness of the prediction against knockout of the future coincides with the mutual information $I(X_{\text{past}} : X_{\text{future}})$ between the past and the future. This is a fundamental quantity which underlies several complexity measures like excess entropy (Crutchfield and Packard, 1983), effective measure complexity (Grassberger, 1986), and predictive information (Bialek et al., 2001). The relation between these measures and multi-information has been discussed in Erb and Ay (2003).

Now come back to the situation of N input variables $(X_i)_{i\in\Lambda}$, $\Lambda = \{1,\dots,N\}$, and one output variable X_0. The exclusion is described by a probability measure $r(A)$, $A \subseteq \Lambda$. This implies the following mean robustness:

$$R(r,p,T) := \sum_{A\subseteq\Lambda} r(A)R(X_A,p,T). \tag{22}$$

As direct implication of Theorem 2, we have the following upper bound:

$$R(r,p,T) \leqslant \sum_{A\subseteq\Lambda} r(A)I(X_A : X_{\Lambda\setminus A}). \tag{23}$$

This upper bound is interesting for the reason that it relates our robustness measure to a complexity measure introduced by Tononi et al. (1994), which we refer to as *TSE-Complexity*. It is defined as follows:

$$C_{\text{TSE}} := \sum_{k=1}^{N-1}\left(\frac{1}{\binom{N}{k}}\sum_{\substack{A\subseteq\Lambda\\|A|=k}} H_p(X_A) - \frac{k}{N}H_p(X_\Lambda)\right).$$

There are several choices for the distribution r to get such a relation. With $\lceil x\rceil$ denoting the smallest integer $\geqslant x$ we define

$$r^{(1)}(A) := \frac{2}{N(N+1)}\cdot\frac{|A|}{\binom{N}{|A|}},$$

$$r^{(2)}(A) := \begin{cases} \frac{1}{\lceil\frac{N}{2}\rceil}\cdot\frac{1}{\binom{N}{|A|}} & \text{if } |A|\leqslant\lceil\frac{N}{2}\rceil,\\ 0 & \text{otherwise,}\end{cases}$$

$$r^{(3)}(A) := \begin{cases} \frac{1}{N-\lceil\frac{N}{2}\rceil+1}\cdot\frac{1}{\binom{N}{|A|}} & \text{if } |A|\geqslant\lceil\frac{N}{2}\rceil,\\ 0 & \text{otherwise.}\end{cases}$$

With these definitions we have the following equalities relating the upper bound in (23) to the TSE-complexity:

$$\sum_{A\subseteq\Lambda} r^{(i)}(A)\,I(X_A : X_{\Lambda\setminus A}) = c^{(i)}\cdot C_{\text{TSE}}, \quad i = 1,2,3$$

with

$$c^{(i)} := \begin{cases} \dfrac{2}{N+1} & \text{for } i = 1, \\ \dfrac{1}{\lceil \frac{N}{2} \rceil} & \text{for } i = 2, \\ \dfrac{1}{N - \lceil \frac{N}{2} \rceil + 1} & \text{for } i = 3. \end{cases}$$

Since robustness is frequently favored by selection mechanisms in noisy contexts, evolutionary processes are likely to lead to the emergence of complex networks indirectly. We return to this point in discussion.

Case study: duplication and robustness

Perhaps the most simple way of creating robustness in informational networks is through a physical duplication of components. Here, we explore duplication in the context of our mathematical framework.

Let $T : \boldsymbol{X} \times \boldsymbol{X}_0 \to [0, 1]$ be a stochastic map, and let p be a probability distribution on $\boldsymbol{X}$. In order to have several copies of this map, we consider the N-fold cartesian product $\boldsymbol{X}^N$, and we define the input probability distribution

$$\widetilde{p} := \sum_{x \in \boldsymbol{X}} p(x)\delta_{(x,\ldots,x)}.$$

With a probability α to leave an edge as an input source, we set

$$r(A) := \alpha^{|A|}(1-\alpha)^{N-|A|}.$$

We define the extension of the map T to the set of N inputs by choosing one input node with probability $1/N$ and then applying the map T to that node. This leads to

$$\widetilde{T}(y|x_1, \ldots, x_N) := \frac{1}{N}(T(y|x_1) + T(y|x_2) + \cdots + T(y|x_N)).$$

Duplication implies a duplication of both nodes and of edges. Mechanistically this corresponds to duplicating an information source and its channel. Above we have assumed that identical sources are selected randomly, with a probability $1/N$. Mechanistically this states that the target cannot discriminate among inputs and selects randomly. In genetics two identical genes produce identical protein products, and all else being equal, these are just as likely to enter into a reaction pathway. However, duplication is often associated with mutations, in which case either the node, the channel, or both can be modified. To capture this case we would have to replace $1/N$ and related terms with the channel weights. We shall only consider the simple case here.

Now we want to study the robustness properties with respect to the number N of copies and the probability α by specifying T as the identity map on the set $\{-1, +1\}$

with uniform distribution $p(+1) = p(-1) = \frac{1}{2}$. The output node just copies the input: $x \mapsto x$. First let us consider the exclusion dependence: if the information from at least one node is given, then the output can be computed, because the information from all the other nodes is completely redundant with the given information. This means that the only contribution to our exclusion dependence measure comes from the situation where no input is given. This happens with probability $(1-\alpha)^N$, and we get the exclusion dependence $(1-\alpha)^N \ln(2)$. Now, we calculate the contribution of a subset $\{1,\ldots,N\}\backslash A$ which is excluded with probability $\alpha^{|A|}(1-\alpha)^{N-|A|}$. Given a configuration x_A on the subset A, elementary calculations lead to

$$\sum_{x_{A\backslash A}\in X_{A\backslash A}} p(x)\, T(y|x_A, x_{A\backslash A}) = \frac{1}{N}\sum_{j\in A} T(y|x_j) + \frac{N-|A|}{2N},$$

which implies the following contribution of $A\backslash A$:

$$-\ln\left(\frac{|A|}{N} + \frac{N-|A|}{2N}\right).$$

Taking the mean with respect to the probability distribution r determined by α leads to the following mean information flow:

$$-\sum_{k=0}^{N}\binom{N}{k}\alpha^k(1-\alpha)^{N-k}\ln\left(\frac{N+k}{2N}\right).$$

Following our concept of robustness, we have to take this information flow minus the exclusion dependence as the measure of robustness

$$-\sum_{k=0}^{N}\binom{N}{k}\alpha^k(1-\alpha)^{N-k}\ln\left(\frac{N+k}{2N}\right) - (1-\alpha)^N\ln(2).$$

Fig. 10 illustrates some important points about the role of simple duplication in promoting robustness. Duplication of nodes always reduces their individual causal contribution to network function. This is because the function is now averaged over a larger number of identical inputs. Each node makes a small contribution to the total network behavior. At the same time duplication reduces the exclusion dependence, allowing the network to function when nodes are removed. The total robustness initially increases with duplications, but then declines gradually as the contribution of duplicates falls. The greater the probability of removing nodes, the greater the benefits of duplication for network robustness.

Sketch of an extension to networks

In this section, we extend our notion of robustness to the general setting of networks. We consider a finite set V of interacting units. The interaction of the units is qualitatively modeled by directed edges, which are ordered pairs $(u,v) \in V \times V$. The set of edges is denoted by E. Given a unit v, $\mathrm{pa}(v) := \{u \in V : (u,v) \in E\}$ is the set

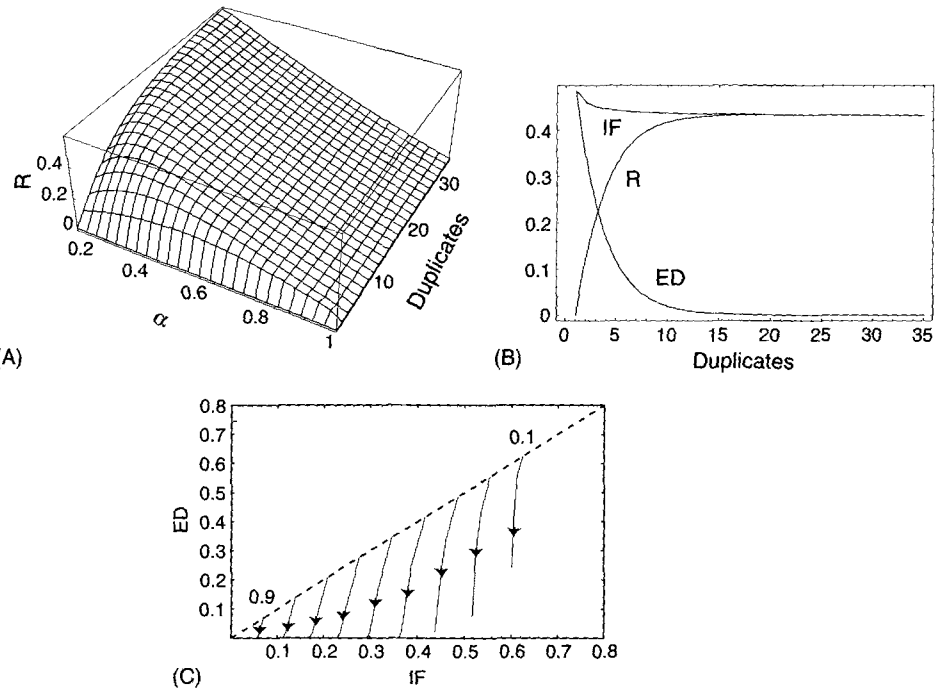

Fig. 10. Contribution of information flow and exclusion dependence on network robustness. Plots measure robustness as a function of duplicated units and unit knockout probabilities. In (A) the robustness surface is maximum for multiple duplicates and low knockout probability. In (B) we fix α at 0.3 and plot the information flow, exclusion dependence and the robustness independently. In (C) each downward curve with terminating arrow represents a different knockout probability, the direction of the arrow indicates increasing numbers of duplicated units.

of units that provide direct information to v. With state sets X_v, $v \in V$, we consider a family of transformations given by Markov transition kernels

$$T^v : X_{\mathrm{pa}(v)} \times X_v \to [0,1], \quad (x,y) \mapsto T^v(y|x).$$

For each node v, $\mathrm{pa}(v)$ plays the role of the set $\Lambda = \{1,\ldots,N\}$ as shown in Fig. 1. The kernel T^v corresponds to a transformation T of a simple network. A combination of such simple networks allows for formalizing the general situation of a recurrent network. This gives us the global dynamics $T : X_V \times X_V \to [0,1]$ defined by

$$T(y|x) := \prod_{v \in V} T^v(y|x_{\mathrm{pa}(v)}).$$

The aspects of a recurrent network that are not captured by the simple network study are given by the fact that the global dynamics generates the probability distribution p. Therefore, p depends on the network properties and is not given from outside as an input distribution (Fig. 11).

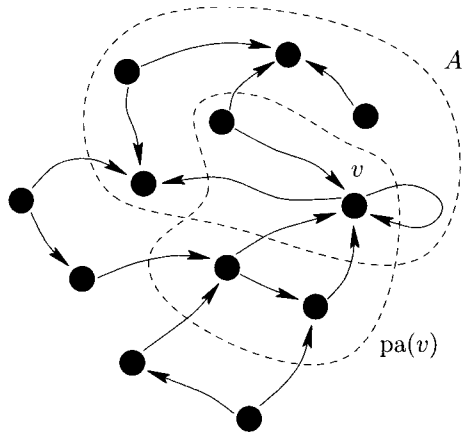

Fig. 11. The generalized, recurrent network. The simple feedforward architecture described in this paper can be extracted by recognizing the subsystem A of V, a node v and its parent set $\mathrm{pa}(v)$. The parent set are the input nodes and v the output node.

Now we assume that nodes are knocked out according to a probability distribution, and we denote the probability for the remaining subsystem $A \subseteq V$ by $r(A)$. Given an element $v \in A$, its parent set $\mathrm{pa}(v)$ can be decomposed as the disjoint union of the subset that is contained in A and the subset that is not contained in A:

$$\mathrm{pa}(v) = (\mathrm{pa}(v) \cap A) \cup (\mathrm{pa}(v) \backslash A).$$

The robustness of v is given by the contribution of the parent set of v that is not contained in A minus the exclusion dependence after its knockout. Note that both quantities clearly depend on the spatial probability distribution p, which need not be stationary with respect to T. Using the robustness formula (19), we consider the following average as the robustness of A:

$$\frac{1}{|A|} \sum_{v \in V} [\text{Robustness of } v \text{ against exclusion of } \mathrm{pa}(v) \backslash A].$$

Finally, we take the mean with respect to r in order to calculate the total robustness of the network:

$$\sum_{A \subseteq V, v \in A} \frac{r(A)}{|A|} \cdot [\text{Robustness of } v \text{ against exclusion of } \mathrm{pa}(v) \backslash A]$$

$$\leqslant \sum_{A \subseteq V} \frac{r(A)}{|A|} \sum_{v \in A} I(X_{\mathrm{pa}(v) \cap A} : X_{\mathrm{pa}(v) \backslash A}) \leqslant \sum_{A \subseteq V} r(A) I(X_A : X_{V \backslash A}).$$

This upper bound coincides with the bound in (23) so that we can make the same statements on the relation between robustness and complexity within this extension to networks. The main message here is that one cannot expect any robustness of a dynamical system if there are no spatial correlations. If the dynamics of the network, which is given by T, is generating a distribution p, it has to display spatial

redundancies in order to make temporal robustness possible. This result is a direct consequence of our previous study of simple networks and demonstrates the generality and applicability of our approach to the setting of recurrent networks.

Robust properties of flows through biological networks

In this contribution, we have presented an informational and geometric framework for thinking about robustness employing information-theoretic measures defined over networks. Network structures can be described at a range of different scales of biological organization. A study of the properties of robust networks can provide general insights into robustness mechanisms in biology and is an important part of an ongoing research program in systems biology (Wagner, 2000; Albert et al., 2000; Strogatz, 2001). One value of our approach is that it makes explicit what is meant by robustness, and provides numerical means for calculating robustness values. It should be emphasized that our formalism is statistical and not mechanistic, and hence we do not commit to any one robustness mechanism. This gives the approach a fair degree of generality, but also limits its predictive utility.

Thinking about network robustness in terms of functional invariance following knockout perturbation, raises questions about the generic properties of robust networks. Previous studies on this topic can be divided into those stressing static, topological and distributional properties of networks, and those emphasizing dynamic properties arising from flows through pathways and networks. The best known of the first group are large, scale-free networks in which connectivity, as measured by the degree moment of the distribution, is given by a power law. This ensures that random node elimination will leave the majority of edges as they were. They have the deficiency, however, of being vulnerable to targeted elimination of highly connected nodes (Albert et al., 2000; Callaway et al., 2000). Of the second group, there are a number of models considering the impact of redundancy on robustness, in particular pleiotropic genetic interactions, in which a single gene can adopt two or more functions (de Visser et al., ; Wagner, 2000). These papers tend to emphasize those forces required to maintain redundancy over generational time. A third class of robustness study considers in some detail mechanisms ensuring invariance such as feedback control and of repair processes acting on either DNA or proteins (Tyson et al., 2001; Shmulevich et al., 2003; Barkai and Leibler, 1997).

Our models do not deal with complex networks with feedback structure and dynamics. One way in which feedback can be included is by considering the sum over all linear decompositions of a recurrent network into feedforward architectures (see section "Sketch of an extension to networks" above). However, the feed-forward network makes clear several important insights that are expected to generalize to extended cases:

- Robustness quantifies a network's ability to reconfigure, exploiting redundant activity in connected nodes to promote invariance.

- Robustness measures both information flow, or causality, and exclusion dependence: networks comprised of completely uncorrelated nodes, in which there is no flow and hence no causal relationships, cannot be by definition robust.
- Redundancies arise through correlations among nodes and need not imply identity among nodes (simple duplication leading to perfect correlation).
- Error-correcting codes represent one general mechanism for exploiting redundancies in the distribution of node activity to achieve robustness.
- Robust networks need include some mechanism of adaptation, whereby networks are able to reconfigure in order to arrive at alternative means of using correlated information to function.
- Network structures that produce robustness also increase statistical measures of network complexity. This is because robustness increases through correlated activity among nodes as does the ability of a network to perform computations.
- Robustness is likely to come first and complexity second, as persistence is a primary requirement for evolution.

Robustness and error correction

One intuitive notion of robustness is functional invariance following perturbation of a biosystem. The perturbation could be anything from the removal of a gene, to the eradication of a species. Function is assessed by monitoring the impact on phenotypes or ecosystems. Typically in developmental genetics, knockouts that leave phenotypes unchanged, are interpreted as functionally redundant and suggest some robustness mechanism (Krakauer, 2003). The same argument has been made for animal social systems, in which critical social networks remain statistically unchanged following individual removal while others reconfigure (Flack et al., 2006). In many cases different structures can produce the same function. Following perturbation, when a new structure reproduces the function prior to perturbation without adaptive reconfiguration, we speak of neutral networks (Schuster et al., 1994) or neutral sets. If the new structure produces the same function by means of reconfiguration, then the system manifests adaptation. In both cases, performing the same function following perturbation relies on redundancies in the network. Adaptive reconfiguration and neutrality differ with respect to how redundancies are structured and exploited.

The best understood structuring of redundancies to achieve functional invariance are error-correcting codes (Peterson and Weldon, 1972). These codes are used when a signal is transmitted once, and a correction implemented immediately upon detection. The greater the rate of error, the more redundancy required to compensate for error. Error-correcting codes typically arrange a signal in a geometric configuration in order to assign parity check bits to bits in the signal in a consistent manner. For example, in rectangular codes each bit of a signal is indexed by two check bits, and the value of the check bit indicates which bit in the original signal is modified. This way messages can be restored. An argument could be made that biological robustness is produced by mechanisms implementing network error

correction. Error correction during DNA repair (Friedberg, 1985) involves the methylation state of one DNA strand providing information to repair enzymes about correct nucleotide sequences prior to replication error. This is a repair network with error correction conferred by a redundant base pair and a methyl group serving a parity function.

Mechanisms of adaptation

Error-correcting codes minimize exclusion dependence in such a way that networks do not require reconfiguration in order to function. The idea behind this is that there is one sequence which uniquely leads to the function of interest, and hence mechanisms need be in place that restore this sequence. Alternatively, there could be many configurations equally able to perform a function, in which case error correction involves reconfiguring networks to alternative networks which map to the same function. Selection acting on the space of networks could achieve this degeneracy in several ways.

One way is to create neutral sets. Neutral sets include all networks mapping to the same function and the perturbations transforming one network into another within the set. The best studied of these sets, are neutral networks in which members of a neutral set are related to one another according to the Hamming distance metric. Mutations are most probable among sequences differing by a single base pair and edges connect genomes differing by a single-point mutation. In the case of RNA, network neutrality is the outcome of the physics of folding: different sequences fold into the same, minimum free energy structure (Schuster et al., 1994). An RNA neutral network describes all sequences connected by single-point mutations sharing a common minimum free energy folded form.

Another possibility requires more degrees of freedom and an explicit mechanism by which nodes can reorganize their connections. Genetic networks have been observed to modify their topology on the basis of the removal of key genes (Wagner, 2005). At the individual level individuals are capable of learning to switch to alternative sources of energy or information when a favored resource is removed. A canonical ecological example is prey switching and its ability to stabilize predator–prey dynamics in ecology (van Baalen et al., 2001).

Role and persistence of duplication

Robustness arises out of correlations among units, and these correlations need to be constructed. In evolutionary time, these correlations arise as the outcome of a mutation selection process operating on populations and even speciation events. During ontogeny, correlations arise through learning mechanisms. In each case, robust correlations can arise through direct benefits experienced by individuals independent of exclusion dependence, or become targeted directly as a mechanism for reducing exclusion dependence at a higher level.

We explored the robustness of networks comprising many identical duplicated elements. Duplication does not modify the function of a network, it only modulates

sensitivity to knockout. It has been pointed out that this indirect benefit of redundancy, weakens the ability of stabilizing selection to preserve duplicates in a population (Krakauer and Nowak, 1999). However, since duplicated components will initially contribute randomly to network function, each duplicate has a non-zero influence on function. This allows selection to operate directly on the causal contribution of duplicates. We have shown that this contribution drops exponentially with duplicate number, and hence exclusion-dependent perturbations, knocking out nodes, will be required to maintain large copy numbers. However, a degree of exclusion independence will arise through incidental selection on the direct contribution of duplicates. This can be made more intuitive by contrasting the role of duplicates in biology with duplicates in engineered systems. In biology duplicates are employed regularly through a local stochastic selection process (e.g. enzyme kinetics), whereas in engineered systems, backups are only deployed upon failure of the primary system. Hence in biology there is an ongoing contribution to function from duplicates which can increase the efficiency of the natural selection process or learning mechanisms.

Robustness and complexity

We have derived a measure of robustness to quantify the impact of network perturbations on the ability of networks to perform a function. The measure quantifies the distance between the wild-type network and an optimally reconfigured, perturbed network. We observe that the robustness has an upper bound describing a network complexity measure (Tononi et al., 1994). This complexity measure allows for calculation of the extent of integration and segregation in complex biological networks. The complexity measure is derived by considering the averaged, summed mutual information between all bi-partitions of a network. If any two bipartitions are independent, then the mutual information will be zero. If the bipartitions behave randomly or identically, then the complexity measure is low. When the networks possess independent components which are capable of global interaction the complexity measure is correspondingly high.

Integration and segregation when present jointly, lead to high complexity. Highly integrated activity of nodes in a network corresponds to high redundancy. Segregation of nodes reduces redundancy and allows for greater variation in activity. Most information processing problems require a mixture of both properties as function need to be decomposed (segregated) and then integrated in order that an appropriate action be initiated. Important nodes have the feature that they make important contributions to a network, and this implies segregation. These nodes also need to share information with other nodes and this implies a reduced exclusion dependence.

What is the relationship of neural network complexity to network robustness? In the complexity measure random bi-partitions are used to assess the extent of communication among brain regions; whereas in the robustness measure, bi-partitions reflect knockouts of large sets of components and their remaining complements. In both cases information flow is required. In the cognitive case,

information flow is assumed to reflect associative power, whereas in the robustness case information flow is required in order that the remaining nodes respond to perturbation. These results show mathematically that characteristics of biocomplexity can derive from selection minimizing the deleterious effects on function following the removal of important nodes. This phenomenon has been referred to as *robust overdesign* (Krakauer and Plotkin, 2002) and suggests that some amount of biological complexity and diversity relates to overcoming instabilities engendered by perturbations.

Summary

We have provided a geometric analysis of robustness for biological networks based on the concept of information flow. Robustness has two components: the first quantifies the contribution of nodes to network function *in situ* and the second the consequences for function of node removal-exclusion dependence or their contribution *ex situ*. Nodes in robust networks contribute both unique features to network function and correlated features promoting redundancy. Networks exploit redundancy directly through error correction and neutrality or through adaptive reconfiguration in order to minimize exclusion dependence. We have found that maximizing robustness also leads to an increase in network complexity – robustness serves as a lower bound on complexity. The robustness measure has been applied to boolean functions as test cases and provides a general framework for analyzing biological network data.

References

Albert, R., Jeong, H., Barabasi, A.L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382.

Amari, S., 1985. Differential-Geometric Methods in Statistics, Lecture Notes in Statistics, vol. 28. Springer, Heidelberg.

Amari, S., Nagaoka, H., 2000. Methods of Information Geometry, AMS Translations of Mathematical Monographs, vol. 191. Oxford University Press, Oxford.

Barkai, N., Leibler, S., 1997. Robustness in simple biochemical networks. Nature 376, 307–312.

Bialek, W., Nemenman, I., Tishby, N., 2001. Predictability, Complexity, and Learning. Neural Comput. 13, 2409–2463.

Callaway, D.S., Newman, J.E.J., Strogatz, S.H., Watts, D.J., 2000. Network robustness and fragility: percolation on random graphs. Phys. Rev. Lett. 85, 5468–5471.

Chosmky, N., 1981. Principles and parameters in syntactic theory. In: Hornstein, N., Lightfoot, D. (Eds.), Explanations in Linguistics. Longman, London.

Cover, T.M., Thomas, J.A., 2001. Elements of Information Theory. Wiley, New York.

Crutchfield, J.P., Packard, N.H., 1983. Symbolic dynamics of noisy chaos. Physica D 7, 201–223.

de Visser, J.A.G.M., Hermisson, J., Wagner, G.P., Meyers, L.A., Bagheri-Chaichian, H., Blanchard, J.L., Chao, L., Cheverud, J.M., Elena, S.F., Fontana, W., Gibson, G., Hansen, T.F., Krakauer, D., Lewontin, R.C., Ofria, C., Rice, S.H., von Dassow, G., Wagner, A., Whitlock, M.C., 2003. Evolution and detection of genetic robustness. Evolution 57, 1959–1972.

Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Networks structure and biodiversity loss in food webs: robustness increases with connectance. Ecol. Lett. 5, 558.

Erb, I., Ay, N., 2003. Multi-information in the thermodynamic limit. J. Stat. Phys. 115, 949–976.

Flack, J., Girvan, M., de Waal, F., Krakauer, D.C., 2006. Policing stabilizes construction of social niches in primates. Nature, 439, 426–429.
Friedberg, E.C., 1985. DNA Repair. W.H. Freeman, New York.
Grassberger, P., 1986. Toward a quantitative theory of self-generated complexity. Int. J. Theor. Phys. 25 (9), 907–938.
Krakauer, D.C., 2003. Genetic Redundancy, Evolution and Comparative Genomics. Encyclopedia of the human genome. Nature Publishing Group. MacMillan Publishers, New York.
Krakauer, D.C., 2004. Robustness in biological systems: a provisional taxonomy. In: T.S. Dreisboeck, J. Yasha Kresh. (Eds.), Complex Systems Science in Biomedicine, Kluwer Academic Press, Dordrecht.
Krakauer, D.C., Nowak, M.A., 1999. Evolutionary preservation of redundant duplicated genes. Semin. Cell. Dev. Biol. 10, 555–559.
Krakauer, D.C., Plotkin, J.B., 2002. Redundancy, antiredundancy, and the robustness of genomes. Proc. Natl. Acad. Sci. USA 99, 1405–1409.
Krakauer, D.C., Plotkin, J.B., 2004. Principles and parameters of molecular robustness. In: Jen, E. (Ed.), Robust Design: A Repertoire for Biology, Ecology and Engineering. Oxford University Press, Oxford, pp. 115–133.
Newman, M., 2003. The structure and function of complex networks. SIAM Rev. 45, 167–256.
Pearl, J., 2000. In: Causality. Cambridge University Press, Cambridge.
Peterson, W.W., Weldon, E.J., 1972. Error Correcting Codes. MIT Press, Cambridge, MA.
Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L., 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc. Roy. Soc. (London) B, 255, 279–284 (1994).
Shmulevich, I., Lahdesmaki, H., Dougherty, E.R., Zhang, W., 2003. The role of certain postclasses in Boolean network models of genetic networks. Proc. Natl. Acad. Sci. USA 100, 10734–10739.
Strogatz, S.H., 2001. Exploring complex networks. Nature 410, 268–276.
Tononi, G., Sporns, O., Edelman, G.M., 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. USA 91, 5033–5037.
Tyson, J.J., Chen, K., Novak, B., 2001. Network dynamics and cell physiology. Nat. Rev. Mol. Bio. 2, 908–916.
van Baalen, M., Krivan, V., van Rijn, P.C.J., Sabelis, M.W., 2001. Alternative food, switching predators, and the persistence of predator–prey systems. Am. Nat. 157.
Wagner, A., 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. Proc. Natl. Acad. Sci. USA 91, 4387–4391.
Wagner, A., 2000. Robustness against mutations in genetic networks of yeast. Nat. Genet. 24, 355–361.
Wagner, A., 2005. Robustness and Evolvability in Living Systems. Princeton University Press, Princeton, NJ.