RESEARCH ARTICLE

Disease evolution in reaction networks: Implications for a diagnostic problem

Abolfazl Ramezanpour[1,2], **Alireza Mashaghi**[1]*

1 Medical Systems Biophysics and Bioengineering, Leiden Academic Centre for Drug Research, Faculty of Science, Leiden University, Leiden, The Netherlands, **2** Physics Department, College of Sciences, Shiraz University, Shiraz, Iran

* a.mashaghi.tabari@lacdr.leidenuniv.nl

Abstract

We study the time evolution of symptoms (signs) with some defects in the dynamics of a reaction network as a (microscopic) model for the progress of disease phenotypes. To this end, we take a large population of reaction networks and follow the stochastic dynamics of the system to see how the development of defects affects the macroscopic states of the signs probability distribution. We start from some plausible definitions for the healthy and disease states along with a dynamical model for the emergence of diseases by a reverse simulated annealing algorithm. The healthy state is defined as a state of maximum objective function, which here is the sum of mutual information between a subset of signal variables and the subset of assigned response variables. A disease phenotype is defined with two parameters controlling the rate of mutations in reactions and the rate of accepting mutations that reduce the objective function. The model can provide the time dependence of the sign probabilities given a disease phenotype. This allows us to obtain the accuracy of diagnosis as a function of time by using a probabilistic model of signs and diseases. The trade-off between the diagnosis accuracy (increasing in time) and the objective function (decreasing in time) can be used to suggest an optimal time for medical intervention. Our model would be useful in particular for a dynamical (history-based) diagnostic problem, to estimate the likelihood of a disease hypothesis given the temporal evolution of the signs.

Author summary

Here, we use concepts from statistical physics and reaction network dynamics to introduce a measure to quantify the tradeoff between the accuracy of diagnosis and an early diagnosis. This measure is used to suggest an optimal time for medical intervention depending on the number of observed signs (medical tests). We present a stochastic model using a reverse simulated annealing algorithm for numerical simulation of disease evolution. The model can provide the time dependence of the sign probabilities given a disease phenotype. This in turn allows us to anticipate the accuracy of diagnosis as a function of time by using a probabilistic model of signs and diseases.

Introduction

Early diagnosis considerably reduces the human and financial cost of disease treatments. However, an early diagnosis requires an accurate characterization of disease states, understanding the mechanisms of disease development (dynamics), and the way a disease influences the other ones (disease interactions) [1–3]. This, in turn, allows us to construct more accurate diagnostic models and algorithms to uncover a hidden disease pattern in the early stages of its progress [4–8]. This study aims to clarify these concepts within a chemical reaction network as a microscopic model for the time evolution of molecular concentrations (the system signs) [9].

Defining a disease state and differentiating diseases based on medical signs/symptoms is not always trivial. From a statistical point of view, however, it makes sense to define the microscopic variables as the signs and define a disease state as a macroscopic (Gibbs) state of the sign probability distribution [10, 11]. A disease state then may appear by (see Fig 1): (i) changing smoothly the sign values with no phase transition, e.g., in ageing, (ii) a discontinuous phase transition, e.g., when the stress exceeds a critical value [12], (iii) a continuous phase transition which can be classified depending on its critical behaviour [13]. In recent studies, we constructed probabilistic models of signs and diseases and introduced a diagnostic algorithm that is based on the simulation of the diagnostic process [14–16]. The main finding was a two-stage diagnostic strategy, which starts by suggesting at each step one medical test and observing the outcome of that medical test (sign). Then, for a critical number of observations, the model undergoes a phase transition to an ordered phase, where it would be safe to suggest a sequence of medical tests at once, relying on the model predictions. It is very helpful here to have an accurate probabilistic model that captures the relevant statistical correlations of the sign and disease variables. A microscopic model would then be needed to construct such a probabilistic model for simulation of the diagnostic process.

The problem of disease development can be studied at different spatiotemporal resolutions. For example, the aim of molecular pathology is to understand cellular and molecular mechanisms that underlie the diseases [17]. At the level of cell population, one can study dynamics of tumor cells and their interactions with immune system with the aim of controlling the disease process [18]. Here the methods of ecological and resource-consumer theory can be used to study disease dynamics and the host-pathogen interactions [19]. At a larger scale, a complex system approach can be employed to investigate e.g. the neural dynamics of neurological disorders [20]. On the other hand, clinical data acquisition and monitoring of disease dynamics
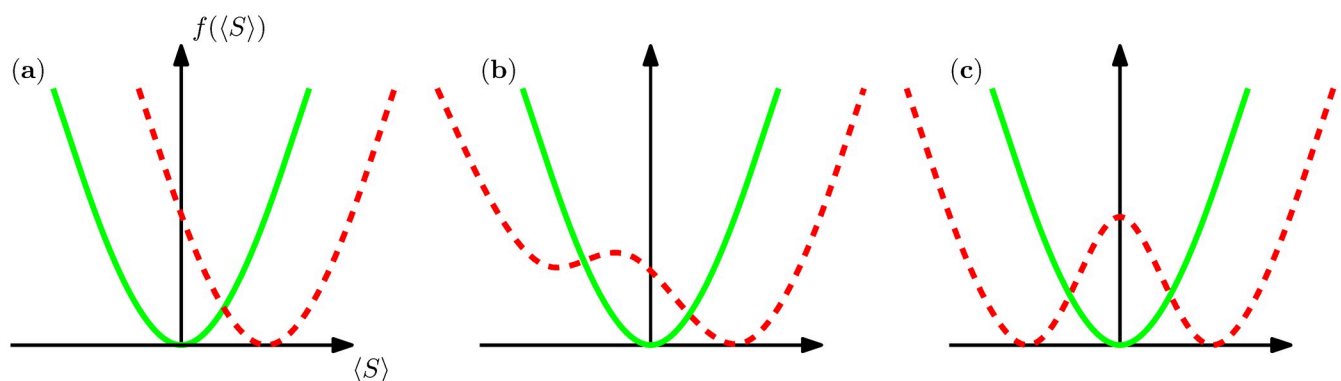

Fig 1. A schematic view of possible scenarios for disease development in the space of the sign probability distribution $P(\mathbf{S})$. The average value of a binary sign variable $S_i = \pm 1$ is shown by $\langle S \rangle$. The minimums of the free energy $f(\langle S \rangle)$, associated with the Gibbs measure $P(\mathbf{S})$, represent the possible macroscopic states of the system. (a) A healthy state changes smoothly without any phase transition. On the other hand, a disease state can emerge by a discontinuous (b) or continuous (c) phase transition. The healthy and disease cases are represented by the solid and dashed lines, respectively.

https://doi.org/10.1371/journal.pcbi.1007889.g001

are essential for understanding of disease development [21, 22]. The methods of complex dynamic systems and machine learning can be employed here to analyse the data and construct reasonable dynamical models.

In this paper, we are going to study the time evolution of disease states in a simple model of biochemical reactions. Unfortunately, the concepts of signs, symptoms, and diseases are not always well defined with clear boundaries. This will expectedly change in the future when high resolution molecular picture of diseases are available, thanks to advances in multi-omics approaches. The use of a reaction network as a specific microscopic model provides clear definitions for the above concepts. We take the number of molecules as the microscopic or sign/symptom variables. Here the symptoms are equivalent in meaning to the signs and are used interchangeably. The value of an observed sign (as a medical test) reveals the number or concentration of a biochemical species. The reactions here play the role of interactions between the microscopic variables and the reaction rates give the strength of these interactions. On the other hand, we define the diseases or disease phenotypes as the macroscopic or emergent behaviours displayed by the sign variables. A defect in the reaction network is defined below as a specific deviation in the reaction rates of the healthy network. One may consider a defect as a simple and well defined disease, but we stress that in general a disease is defined by the collective contributions of the signs with (possibly) multiple defects.

We introduce an effective dynamics for temporal evolution of diseases, which is inspired from thermal annealing of physical systems to reach a low temperature (ordered) state of the system. We apply it to a specific model of microscopic signs (molecular concentrations in a reaction network) to study disease development and its consequences for disease diagnosis in different stages of the process. A reaction network provides us with a testbed to model disease developments and disease-disease interactions, and follow the response of the sign variables (here molecular concentrations) as the time pass [23–25]. Specifically, we start from a healthy reaction network which maximizes the mutual information between a subset of signal and response variables as the objective function of the system [26–28]. The optimization step can be done by a local optimization algorithm like the simulated annealing algorithm [29]. Then, we introduce some defects (mutations) in the system and run the simulated annealing algorithm in the reverse direction, which on average decreases the objective function of the system. The system dynamics here is controlled by the rates of introducing the defects and the rate of accepting a decrease in the objective function. Note that by maximizing the mutual information in the healthy state, the system could be placed in a critical region close to a phase transition point. And, the reverse annealing algorithm can induce a phase transition from the healthy state (say an ordered phase) to a disease state (disordered phase). Another aim of this study is to provide a microscopic model of time-dependent signs and diseases which can be used for diagnosis from the history or dynamics of the observed signs, and which in addition allows us to model disease-disease interactions more explicitly.

As a first step, in this study we use the synthetic data generated by the above dynamical model to investigate the time dependence of the diagnostic performance with a simple probabilistic model of signs and diseases. We see how the accuracy of diagnosis with such a diagnostic model improves with time as the reaction network deviates from the healthy state of maximum objective function. This information would be necessary for quantification of the tradeoff between the accuracy of diagnosis and the level of disease progression at the time of diagnosis. This, in particular, can be used to suggest an optimal intervention or screening time for specific diseases and diagnostic models.

The paper is structured as follows. In Sec The model and definitions we define the model and give the main equations. In Sec Modelling disease evolution we present the stochastic

model of disease evolution, and the results of numerical simulations for a small reaction network. The concluding remarks are given in Sec Discussion.

Results

The model and definitions

We consider a system of N interacting species (molecules) with M chemical reactions [9, 30]. Fig 2 displays a graph representation of some elementary reaction networks. A reaction network can be represented by a number of reaction pathways as the fundamental building blocks (basis) of the reaction network [26]. We use the integers $0 \leq X_i(t) \leq X_{max}$ to show the number of molecules for species $i = 1, \ldots, N$ at time t. Here X_{max} is the maximum number of molecules that is allowed by the biological system. The whole set of molecular numbers are denoted by vector $\mathbf{X}(t) = \{X_1(t), \ldots, X_N(t)\}$. To be specific, let us assume that our reaction network is a signal transduction network, where the signals are encoded in the temporal concentration of some signal species. The set of signal variables is denoted by $\mathcal{S}$. To each signal species $i \in \mathcal{S}$
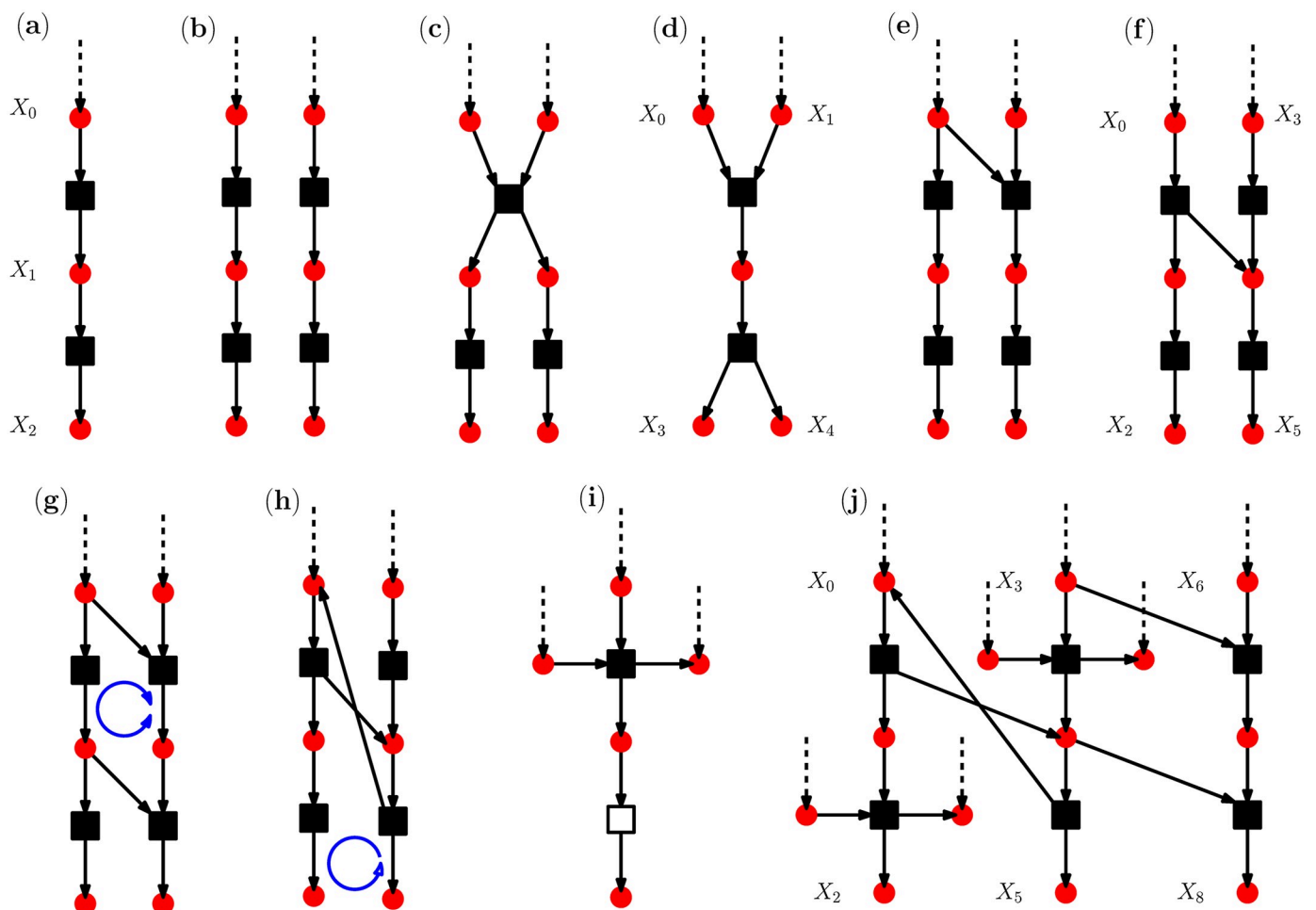

Fig 2. Illustration of the reaction networks. The solid circles and squares display the species and the reactions, respectively. A reaction could be reversible (full square) or irreversible (empty square). A dashed arrow shows that the species concentration is driven externally. (a) a single pathway, (b) two non-interacting pathways, (c)-(d) two pathways interacting through common reactions and species, (e)-(f) two pathways interacting through a link, (g)-(h) two pathways interacting through a coherent (h) or incoherent (g) cycle, (i) a pathway with an irreversible reaction which also interacts with two externally driven species, (j) a larger reaction network of interacting pathways with $M = 6$ reactions and $N = 13$ species.

we assign the subset of associated response species $\mathcal{R}(i)$. The signal generated by species i is transmitted through the network to species $j \in \mathcal{R}(i)$. This means that a change in the concentration of signal molecules is expected to significantly affect the activation level of the response molecules.

In general, a reaction $r = 1, \ldots, M$ is represented by the stoichiometric coefficients $v_r^-(i)$ and $v_r^+(i)$

$$\sum_{i=1}^{N} v_r^-(i)X_i \rightarrow \sum_{j=1}^{N} v_r^+(j)X_j. \tag{1}$$

The coefficients $v_r^-(i), v_r^+(i)$ take only non-negative integer values. The reactions happen stochastically with a probability that is determined by a propensity function $\eta_r(\mathbf{X})$. From the stochastic simulation of this process one can extract the probability distribution of the number of species. Let $\Pr(X_i, X_j)$ be the joint probability distribution of variables X_i and X_j in a stochastic process governed by the above reactions. The mutual information of the two variables, which measures any statistical dependency of the variables, is given by

$$\mathrm{MI}(i,j) = \sum_{X_i}\sum_{X_j}\Pr(X_i, X_j) \log \frac{\Pr(X_i, X_j)}{\Pr(X_i)\Pr(X_j)}. \tag{2}$$

We need the above measure later to quantify the degree of statistical correlations between the signal and response variables.

The average value of a single variable is denoted by

$$\langle X_i \rangle = \sum_{X_i} X_i \Pr(X_i). \tag{3}$$

The activation level of species i can be represented by a coarse-grained variable $x_i \in \{-1, 0, +1\}$, where

$$x_i = -1 \quad \text{if } \langle X_i \rangle < X_i^* - \delta X_i$$
$$x_i = +1 \quad \text{if } \langle X_i \rangle > X_i^* + \delta X_i$$

otherwise $x_i = 0$. Later, we use these variables as the signs in a probabilistic model of disease diagnosis. Here X_i^* defines the threshold value for variable i and δX_i denotes the standard deviation in X_i. We shall set $X_i^* = X_{max}/2$. The threshold value here is defined according to the number of coarse-grained levels, but in practice its precise value would depend on the biological details of the system.

The linear correlation coefficient of two variables is defined as follows

$$C(i,j) = \frac{\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle}{\sqrt{\langle (X_i - \langle X_i \rangle)^2 \rangle \langle (X_j - \langle X_j \rangle)^2 \rangle}}. \tag{4}$$

Note that in general the mutual information $\mathrm{MI}(i,j)$ gives more information about the statistical dependence between the variables than the above linear correlations. Nevertheless, it is easy to see from the sign of $C(i,j)$ that the two variables are positively or negatively correlated. The above definitions give the basic statistical measures that are used in the following to characterize the statistical state of the system.

Evolution by the stochastic chemical kinetics.   The system evolves in time by M (reversible or irreversible) chemical reactions, where $X_i(t)$ gives the number of species i at time t. Each chemical reaction $r = 1, \ldots, M$ is identified with the state-change vectors (stoichiometric

coefficients) v_r^-, v_r^+ and the propensity function $\eta_r(\mathbf{X})$, which determines the reaction flux and in general depends on the X_i. The state-change vectors give the changes in the number of molecules; i.e., after reaction r the X_i change to $X_i + (v_r^+(i) - v_r^-(i))$. The propensity function $\eta_r(\mathbf{X})dt$ gives the probability that reaction r happens in the time interval $(t, t + dt)$ given the $\mathbf{X}(t) = \{X_1(t), \ldots, X_N(t)\}$. Note that each reaction happens with the above probability independently of the other reactions as in a Poisson process [31]. However, a single specie may be involved in several reactions affecting the dynamics of the associated reactions. The propensity function depends on the number of molecules in the left hand side of the reaction $L_r = \{i : v_r^-(i) > 0\}$. The propensity function can take different forms depending on the nature of reaction. For instance, within the mass-action kinetics [9]:

$$\eta_r(\mathbf{X}) = \kappa_r \prod_{i \in L_r} \frac{X_i!}{v_r^-(i)!(X_i - v_r^-(i))!}, \tag{5}$$

where the reaction rate κ_r scales with the system volume Ω as $1/\Omega^{|L_r|-1}$. It is assumed that the molecules are confined in a bounded region of space denoted by Ω. The scaling of the reaction rates with the volume is to ensure that X_i/Ω is well defined in the thermodynamic limit $\Omega \to \infty$.

The average values of stochastic variables $X_i(t)$ satisfy the following equation

$$\frac{d\langle X_i(t) \rangle}{dt} = \sum_r (v_r^+(i) - v_r^-(i)) \langle \eta_r(\mathbf{X}) \rangle. \tag{6}$$

We see that even solving for the averages $\langle X_i(t) \rangle$ is difficult due to the coupling of the variables in the right hand side of the equation. In the following, however, we resort to a numerical simulation of the above process. By simulating a large population of identical reaction networks we also obtain the joint probability distributions which are needed for computation of the mutual informations.

One can use the Stochastic Simulation Algorithm (or Gillespie algorithm) to simulate the time evolution of a reaction network [32]. Let us assume that a reaction happened at time t. Then, we need to know the time to the next reaction (τ) and the index of the next reaction (r). Given that the system is currently in state $\mathbf{X}(t)$, the joint probability distribution of two random variables $\rho(\tau, r)$ is:

$$\rho(\tau, r) = \eta_r(\mathbf{X})e^{-\eta(\mathbf{X})\tau}, \qquad \eta(\mathbf{X}) \equiv \sum_r \eta_r(\mathbf{X}). \tag{7}$$

Note that no reaction should happen in the time interval τ. That is why we need the exponential factor with the total propensity function $\eta(\mathbf{X})$. The Gillespie algorithm then goes as follows [32]:

- compute the η_r and η given the $\mathbf{X}(t)$

- extract τ and r from the probability distribution $\rho(\tau, r)$:

  - draw two random numbers u_1, u_2 from the uniform distribution in $(0, 1)$

  - then $\tau = \frac{1}{\eta} \log(1/u_1)$ and r is the smallest integer satisfying $\sum_{r'=1}^{r} \eta_{r'} > u_2 \eta$

- change $t \to t + \tau$ and $\mathbf{X} \to \mathbf{X} + (v_r^+ - v_r^-)$

Note that we should always check to have $0 \leq X_i \leq X_{max}$. That means that a reaction should have enough reactants to happen, and we keep the $X_i \leq X_{max}$.

This is a simple but very time consuming algorithm for simulation of large reaction networks. The reader can refer to other references in [32] and [33, 34] for more efficient and sophisticated algorithms.

Modelling disease evolution

In this section, we introduce a simple dynamical model for the time evolution of defects in the reaction network. We shall assume that the number of species and the pattern of interactions by the reactions are fixed. Moreover, we assume that the healthy state is an optimal state for an appropriately defined objective function, which depends on the nature of functions we expect from the system. For instance, if the main task of a system is to remember a number of patterns, then a good objective function could be the number or quality of the stored patterns. A reaction network of interacting pathways can be considered as a signal transaction and processing system, among the other functions [26–28]. To be specific, in the following we assume that the healthy state is defined by the state of maximum mutual information between the signal variables and the associated responses. A more suitable measure is the transfer entropy from the signal to response variables. This is a directed measure of information transmission (or causal relation) which is expected from the signal transduction part of the reaction network in a cell. Mutual information is however a symmetric measure of statistical dependence which is closely related to the transfer entropy and at the same time it is computationally easier to estimate; because here one needs two-point joint probabilities whereas for computation of transfer entropy one needs three-point conditional probabilities.

Given the structure of the reaction network, we first use an approximate optimization algorithm (e.g. a zero-temperature Monte Carlo) to find the reactions rates that maximize the following objective function:

$$\mathcal{E} \equiv \sum_{i \in \mathcal{S}} \left(\frac{1}{|\mathcal{R}(i)|} \sum_{j \in \mathcal{R}(i)} \mathrm{MI}(i,j) \right). \tag{8}$$

Note that the mutual information are obtained from the stochastic simulation of the species dynamics which depends on the reaction rates. Therefore, one can search in the space of the reaction rates to find the optimal parameters that maximize the mutual information and so the objective function.

We start from the stationary state of the optimal (healthy) state and denote the associated objective function value by $\mathcal{E}_{old}$. We also attribute an elementary defect $\hat{d}_r$ to each reaction which means that the associated reaction rate is deviating from the healthy value, e.g. due to mutations in the related genes. A general defect pattern is obtained by a combination (or a set) of these elementary defects $\mathbf{D} = \sum_{r=1}^{M} D_r \hat{d}_r$. The number of nonzero coefficients $D_r = 0, 1$ gives the number of present elementary defects in $\mathbf{D}$, which is denoted by $|\mathbf{D}|$ (cardinality of set $\mathbf{D}$). We call $\mathbf{D}$ a defect pattern because in general it can be a superposition of multiple elementary defects. Then, the time evolution of a defect (or pathologic variation) pattern $\mathbf{D}$ is modelled in the following way: For $t = 1, \ldots, t_{max}$ do:

- each reaction rate changes with (mutation) probability $\alpha_r(t|\mathbf{D})$ from κ_r to $\kappa_r \pm \delta\kappa$

- run the stochastic simulation (Gillespie) algorithm to compute the new objective function $\mathcal{E}_{new}$

- accept the changes in the reaction rates if the objective function increases, otherwise, accept the changes with probability $\exp(\beta(t|\mathbf{D})\Delta\mathcal{E})$. If the changes are accepted set $\mathcal{E}_{old} = \mathcal{E}_{new}$

The mutation probabilities $\alpha_r(t|\mathbf{D})$ are expected to be negligible at the beginning (for a healthy state) and increase with time step t. This probability is greater than zero only for the elementary defects that are presented by $\mathbf{D}$. For instance, in case that only elementary defect D_a is present, the mutation probability is denoted by $\alpha_r(t|D_a)$ which is nonzero only for $r = a$. On the other hand, the parameter $\beta(t|\mathbf{D})$ in the acceptance probabilities is expected to be very large for a healthy state and decrease with t depending on $\mathbf{D}$. Here $\beta(t|\mathbf{D}) \geq 0$ plays the role of an inverse temperature in a reverse simulated annealing algorithm [29]. While the probabilities $\alpha_r(t|\mathbf{D})$ control the rate of local changes (mutations) in the reaction network, the global parameter $\beta(t|\mathbf{D})$ determines the system susceptibility to such local mutations (as the immune system). Note that in the standard simulated annealing the temperature decreases slowly to reach an optimal state of the system, which maximizes the objective function (minus the energy function). Here, however, we are deviating from the optimal state by increasing a temperature-like parameter $1/\beta$. That is why we call the above process a reverse annealing.

Let $\tau_r(\mathbf{D})$ denote the time scale in which $\alpha_r(t|\mathbf{D})$ changes from zero to one. Similarly we define the time scale $\tau_\beta(\mathbf{D})$, for changing the probability of accepting a decrease in the objective function. The rates $1/\tau_r(\mathbf{D})$ could in general be nonzero for an arbitrary subset of the reactions.

Interactions between two evolving defects $\mathbf{D}_a$ and $\mathbf{D}_b$ may change the rate of mutations and acceptance probabilities, for instance, by

$$\frac{1}{\tau_r(\mathbf{D}_a + \mathbf{D}_b)} = \frac{1}{\tau_r(\mathbf{D}_a)} + \frac{1}{\tau_r(\mathbf{D}_b)} + \frac{\lambda_r^{ab}}{\tau_r(\mathbf{D}_a * \mathbf{D}_b)}, \tag{9}$$

$$\frac{1}{\tau_\beta(\mathbf{D}_a + \mathbf{D}_b)} = \frac{1}{\tau_\beta(\mathbf{D}_a)} + \frac{1}{\tau_\beta(\mathbf{D}_b)} + \frac{\lambda_\beta^{ab}}{\tau_\beta(\mathbf{D}_a * \mathbf{D}_b)}. \tag{10}$$

The additional rates $1/\tau_r(\mathbf{D}_a * \mathbf{D}_b)$, $1/\tau_\beta(\mathbf{D}_a * \mathbf{D}_b)$ and couplings $\lambda_r^{ab}, \lambda_\beta^{ab}$ show how much the two defects alter the reactions that are not directly affected by the single defects. Here, we use the fact that the rate of two independent Poisson processes with rates $1/\tau_1$ and $1/\tau_2$ is given by $1/\tau = 1/\tau_1 + 1/\tau_2$. Then, a deviation from this case, due to an interaction between the two processes, is represented by an additional term λ_{12}/τ_{12}. Note that $\mathbf{D}_a * \mathbf{D}_b$ is not a simple multiplication; it is just a notation we introduced to represent the interaction between the two defects.

Evaluation of diagnosis accuracy. Given a defect pattern $\mathbf{D}$, we can compute the molecular numbers $X_i(t)$ and the associated mutual information $MI(i, j)$ and correlations $C(i, j)$ at each time step t of the disease evolution algorithm described above. Here, we are going to map the coarse-grained variables x_i to binary sign variables $\mathbf{S} = \{S_i = \pm 1: i = 1, \ldots, N\}$ in a diagnostic problem. The aim here is to find out the underling defects $\mathbf{D}$, assuming that we have observed the values of $N_O \leq N$ signs at time t. We can compute the conditional defect probabilities from a probabilistic model $\Pr(\mathbf{D}, \mathbf{S})$ of the sign and defect variables given the observed signs. Then, the most probable defect pattern is taken for the diagnosis. A computationally simple model (call it the D1S1 model) is obtained by assuming that the elementary defects are marginally independent of each other $\Pr(\mathbf{D}) = \prod_a \Pr(D_a)$, and the signs are conditionally independent of each other $\Pr(\mathbf{S}|\mathbf{D}) = \prod_i \Pr(S_i)$ [5–7].

In the following, we shall use the above model for diagnosis, assuming that the number of present defects $|\mathbf{D}|$ is one. This simplification is to avoid the unnecessary complications of the diagnosis problem and focus on the temporal evolution of diagnosis accuracy. To construct the D1S1 model (i.e. to set the model parameters), we need the conditional probabilities $\Pr_t(S_i|\text{nodefect})$ and $\Pr_t(S_i|D_r)$, in the healthy state (with no defect) and in presence of only one elementary defect (D_r) [14]. This information is obtained (for all the D_r) from the disease evolution algorithm as describe above.

To evaluate the model predictions, first we run the disease evolution algorithm for a given elementary defect D_r. This results to the associated sign values $S_i(t)$ (obtained from the molecular numbers) for different times t. Then, it is assumed that we know only the value of N_O signs (the observed signs) which is used by the diagnostic model (D1S1). From this model we infer the most probable defect and take it as the model prediction, which can be compared with the true one D_r. In this way, we obtain the accuracy of diagnosis by the D1S1 model, conditioned on the number of observed signs:

$$AC(t : N_O) = \frac{P_{true} + N_{true}}{P_{total} + N_{total}}. \qquad (11)$$

This is the ratio of true positive (P_{true}) and true negative (N_{true}) to the total number of positive (P_{total}) and negative (N_{total}) cases. Obviously $0 \leq AC(t: N_O) \leq 1$ and $AC(t: N_O) = 1/2$ for a completely random diagnosis. Moreover, the accuracy is expected to increase with the observation time at which the N_O signs are measured; the larger time t the more specific are the observed signs regarding the involved defects.

Now, imagine that we are to find the best time for observing N_O signs assuming that $|\mathbf{D}| = 1$. In one hand, we have the objective function $\mathcal{E}(t)$ which is a decreasing function of time and shows the disease progress. On the other hand, we have the diagnosis accuracy $AC(t: N_O)$ which improves with the disease progress in time. An optimal intervention time $t^*(N_O)$ then can be obtained by maximizing a weighted sum of the two functions:

$$\mathcal{L}(t : N_O) \equiv \lambda \frac{\mathcal{E}(t)}{\mathcal{E}(0)} + (1 - \lambda) \frac{\left(AC(t : N_O) - \frac{1}{2}\right)}{1/2}, \qquad (12)$$

using the normalized quantities (divided by the maximum values) with $0 \leq \lambda \leq 1$. The value of λ determines the importance that we associate with each term. In practice, the λ value is gradually adjusted in a learning process to find the optimal value, which could depend on the number of present defects $|\mathbf{D}|$ and the nature of diseases. Fig 3 displays schematically the expected behaviour of the above quantities. In practice, we do not have access to the future values of the objective function and the unobserved signs to compute the diagnosis accuracy. But, we can use the microscopic model of disease evolution (here the reaction network) in addition to the probabilistic diagnostic model (here the D1S1 model) to simulate the above process and find an estimation of the optimal intervention time. This highlights the importance of simulation-based methods in a diagnostic problem.

Note that here for simplicity we assume that the number of involving (present) elementary defects is one ($|\mathbf{D}| = 1$). In addition, more accurate (and sophisticated) diagnostic models can be used by considering the possibility of direct defect-defect and sign-sign interactions in the probabilistic model. Such models could be very useful to deal with the case that multiple interacting defects are at work [14]. Finally, the stochastic dynamics of the reaction network and the defects in the disease evolution algorithm provides an ensemble of sign histories which are consistent with a given defect pattern $\mathbf{D}$. This statistical information would be very useful for having a more accurate diagnosis considering the effect of such disorders on the system dynamics.

Numerical simulations. In this section we present the results of numerical simulation for the reaction networks of Fig 2. For example, consider the reaction network of Fig 2(j) with three interacting pathways, $N = 13$ species, and $M = 6$ reversible reactions. We assume that each reaction pathway has its own signal and response species; the selection of these variables in general depends on the specific function of the reaction network which is considered. Here, the signal variables and the associated responses are: $\mathcal{S} = \{X_0, X_3, X_6\}, \mathcal{R}(X_0) = X_2, \mathcal{R}(X_3) = X_5, \mathcal{R}(X_6) = X_8$.
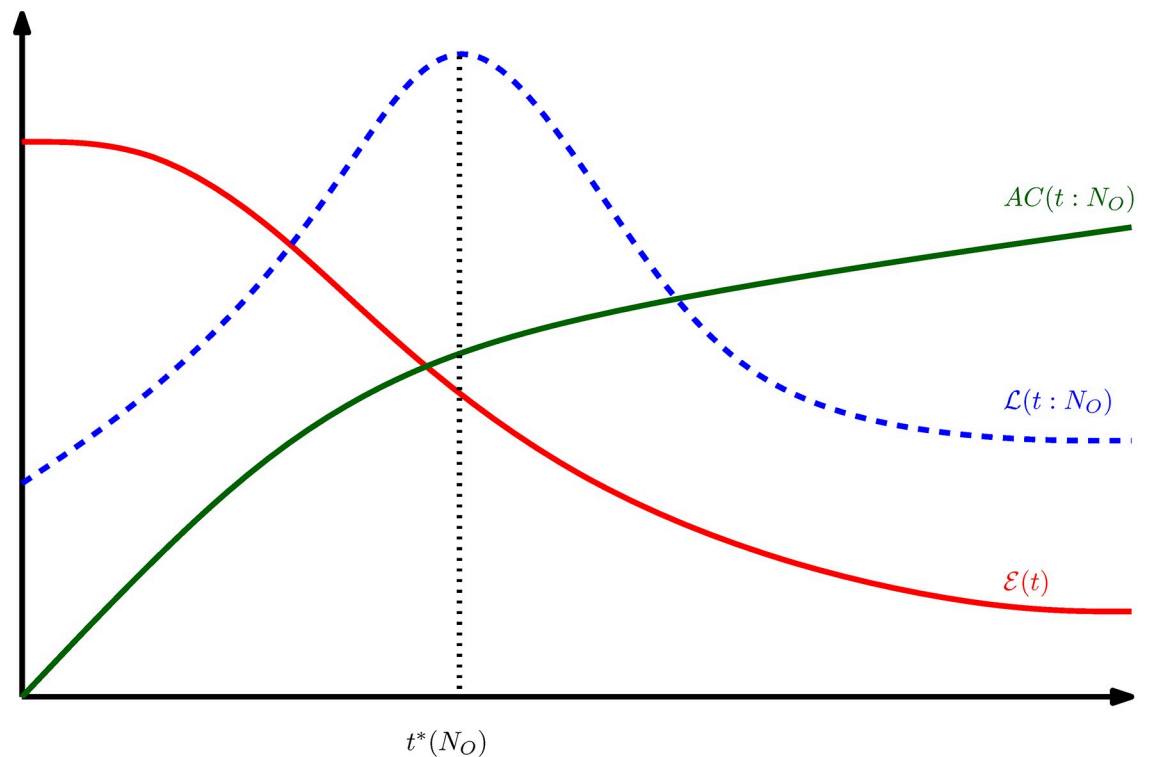
Fig 3. A schematic behaviour of disease progress and diagnosis accuracy with time. An optimal intervention time $t^*(N_O)$ can be defined by maximizing another measure $\mathcal{L}(t : N_O)$ constructed by the objective function $\mathcal{E}(t)$ plus the diagnosis accuracy $AC(t: N_O)$, given the number of observed signs N_O, and the number of present diseases $|\mathbf{D}|$.

And, the objective function is

$$\mathcal{E} = \mathrm{MI}(X_0, X_2) + \mathrm{MI}(X_3, X_5) + \mathrm{MI}(X_6, X_8). \tag{13}$$

We use a zero-temperature Monte Carlo algorithm to find a local maximum of the objective function by changing only the reaction rates, with the constraint that $\kappa_r \in (0, 2)$; this range of values was chosen such that the models can display different non-trivial phases in a reasonable computation time.

Fig 4 shows the optimal features of some selected reaction networks (from Fig 2) obtained in this way. The figure displays the coarse-grained activities x_i, the mutual information $MI(i, j)$, and the correlation coefficients $C(i, j)$. Note that the coefficients $C(i, j)$ display only linear correlations between X_i and X_j. Nevertheless, it is easier to say that the two variables are positively or negatively correlated by looking at the sign of $C(i, j)$ than the value of mutual information. We observe in the figure that negative correlations appear in networks of interacting pathways due to the presence of feedback loops; a single pathway like the one on panel (a), displays merely positive correlations as expected. Moreover, due to the same interactions in the reversible network of panel (e), we see that by maximizing the above objective function we also obtain considerable cross mutual informations, e.g., for $\mathrm{MI}(X_0, X_5)$ and $\mathrm{MI}(X_3, X_8)$.

The optimized reaction network gives the initial condition for the time evolution of the model with defects $\mathbf{D}$ as follows. At each step t, we change the κ_r to $\kappa_r \pm \delta\kappa$ with probability $\alpha(t) = \min(1, t/\tau_\alpha)$ for the reactions that are affected by $\mathbf{D}$. Here $\delta\kappa = 0.05$, $\tau_\alpha = 100, 400$, and still we have the constraint that $\kappa_r \in (0, 2)$. The other parameter decreases with time step t as $\beta(t) = max(0, \beta_0(1 - t/\tau_\beta))$, where $\beta_0 = 100$ and $\tau_\beta = 100, 400$. The mutual informations after
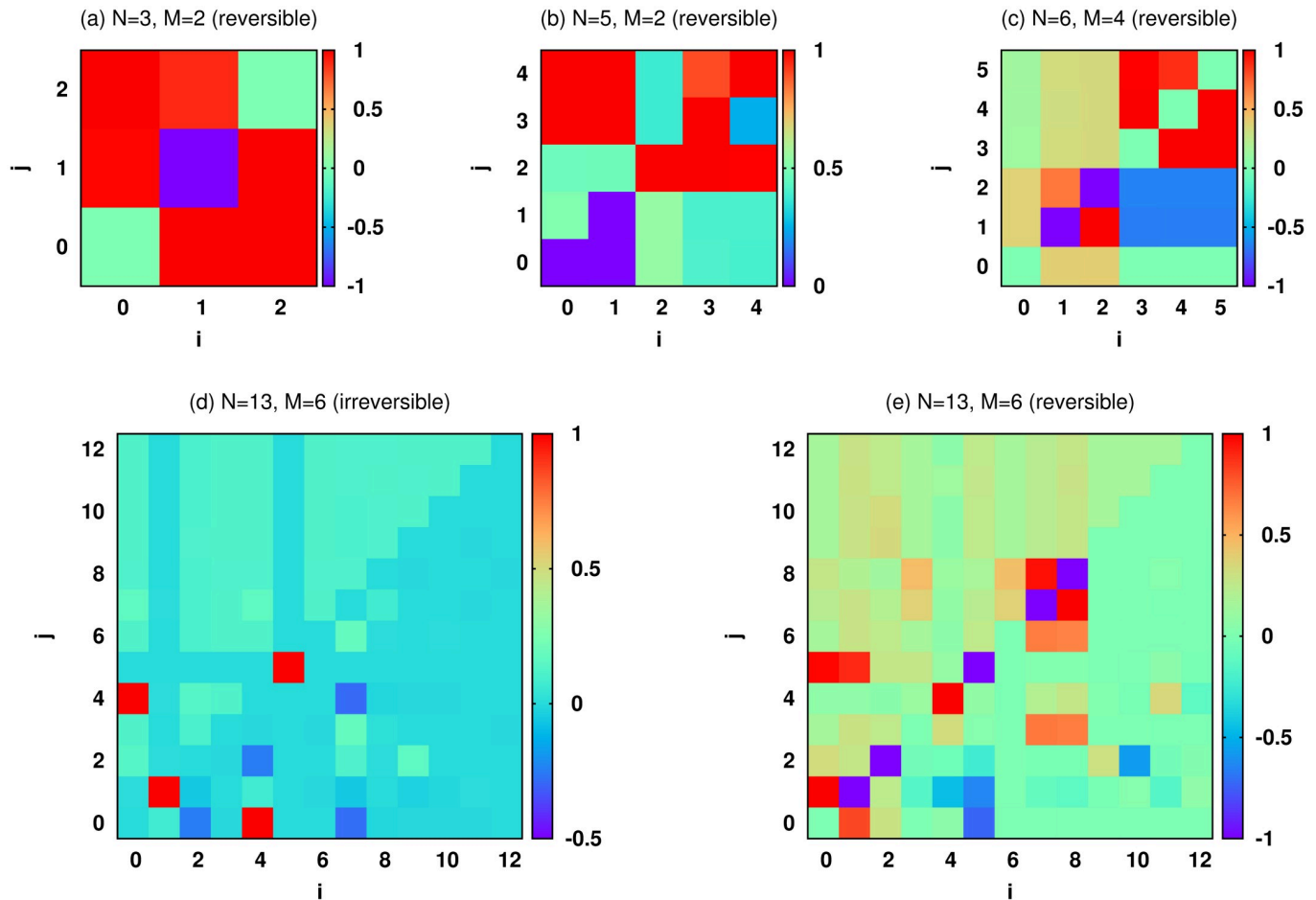
Fig 4. Statistical characterization of the optimal (healthy) reaction networks. The matrix displays mutual informations $\mathrm{MI}(i, j)$ (the upper triangle of the matrix), correlations coefficients $C(i, j)$ (the lower triangle of the matrix), and concentration values x_i (the diagonal elements) for the reaction networks of Fig 2 (a), 2(d), 2(f) and 2(j): (a) the single chain ($\mathcal{S} = X_0, \mathcal{R}(X_0) = X_2$), (b) the two-chain network with two common reactions ($\mathcal{S} = \{X_0, X_1\}, \mathcal{R}(X_0) = X_3, \mathcal{R}(X_1) = X_4$), (c) the two-chain network with a connecting link ($\mathcal{S} = \{X_0, X_3\}, \mathcal{R}(X_0) = X_2, \mathcal{R}(X_3) = X_5$), (d) and (e) the three-chain network with reversible and irreversible reactions ($\mathcal{S} = \{X_0, X_3, X_6\}, \mathcal{R}(X_0) = X_2, \mathcal{R}(X_3) = X_5, \mathcal{R}(X_6) = X_8$). The $\mathrm{MI}(i, j)$ and $C(i, j)$ are scaled to have maximum value one.

https://doi.org/10.1371/journal.pcbi.1007889.g004

each update of the reaction rates are estimated by the stochastic evolution of the reaction network for Δt_{eq} = 200 iterations. Then, we run the algorithm for another Δt_{av} = 200 iterations to extract the necessary statistical information. Note that all parameters here are dimensionless and should be scaled to be related to the real quantities. The parameters like the reaction rates and the time scales are chosen to display the typical behaviour of the system in different regimes. To be specific, in the following we focus on the temporal evolution of diseases in the larger reaction network of Fig 2(j).

Figs 5 and 6 display the average objective function (mutual information) and distance from the healthy state we observe for some defect patterns **D**. Here the average is taken over at least 100 realizations of the stochastic disease dynamics. The numerical results for the behaviour of single dynamical realizations are given in S1 Text. As the figures show, besides the number of present defects which is displayed in panels (b),(e), it is the relative difference of the two time scales τ_α and τ_β that determines the qualitative behaviour of the system. For instance, in panels (c),(f) of the figures we observe that the transition from the healthy state is usually sharper when the rate of mutations $1/\tau_\alpha$ is smaller than the rate of accepting the mutations; it makes
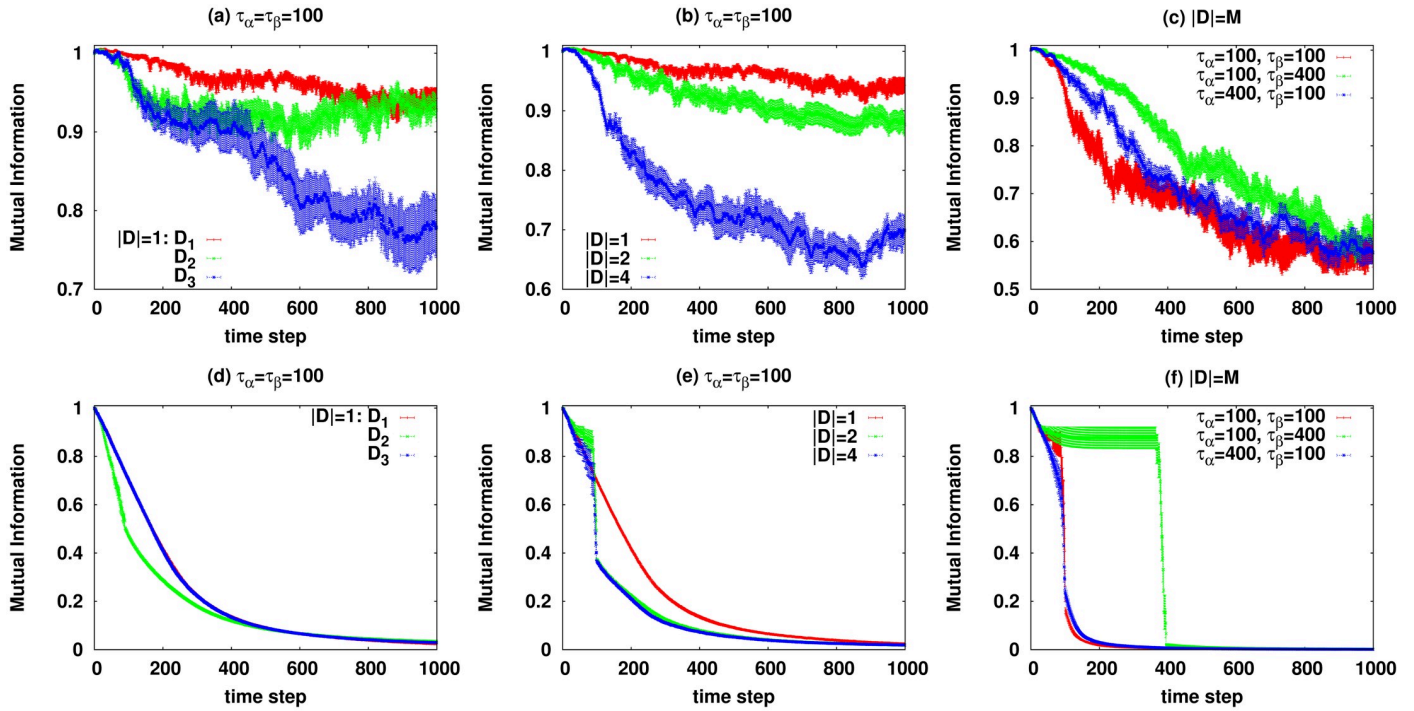
Fig 5. The average objective function (scaled to have maximum one) in presence of various defects. The data are for the reaction network of Fig 2(j) with $N = 13$ species and $M = 6$ reactions. Top panels show the results with reversible reactions for: (a) evolution with one defect ($|\mathbf{D}| = 1$) for $\tau_\alpha = \tau_\beta = 100$, (b) evolution with one, two, and four defects ($|\mathbf{D}| = 1, 2, 4$) for $\tau_\alpha = \tau_\beta = 100$, (c) evolution with all defects ($|\mathbf{D}| = M$) for different $\tau_\alpha = 100, 400$ and $\tau_\beta = 100, 400$. Bottom panels (d),(e),(f) show the results with irreversible reactions.

https://doi.org/10.1371/journal.pcbi.1007889.g005

sense as in this case the probability of accepting the changes is already considerable when the chance of happening the mutations becomes large. On the other hand, a two-stage behaviour is observed in panels (c),(f) when the rate of mutations is significantly larger than $1/\tau_\beta$. Here, there is a period where the mutations happen frequently but the rate of accepting the mutations is small and they are mostly rejected. Fig 6 shows the distance of the system state at time step t from the initial state $d(t) = \sum_{i=1}^{N} |c_i(t) - c_i(0)|/N$, where $c_i(t) = \langle X_i(t) \rangle / X_{max}$. The figure shows the times that the system spends around a microscopic state and how the distance changes in presence of various defects.

Now, we consider the temporal behaviour of the diagnosis accuracy in the above reaction network with reversible reactions. The disease evolution algorithm is run for various elementary defects to obtain the binary sign variables $S_i(t) = \pm 1$ from the molecular numbers $X_i(t)$. Let us assume that the value of N_O signs are observed at time step t. The aim then is to find out the elementary defect that resulted in the observed values. Fig 7 shows the average accuracy of the predictions made by the D1S1 model of Ref [14], assuming that $|\mathbf{D}| = 1$ and $\tau_\alpha = \tau_\beta = 100$. To construct the probabilistic diagnostic model we need the conditional probabilities $\Pr_t(S_i|\text{nodefect})$, and $\Pr_t(S_i|D_r)$ at time step t. This information is obtained from numerical simulation of the dynamics of the reaction network in the healthy state (no defect), and in presence of only one defect (D_r), respectively. The reported accuracy is averaged over different elementary defects and dynamical realizations.

The accuracy of diagnosis is expected to improve with the evolution time of the defects; obviously, there is more statistical information about the defects in the data that are extracted at larger times and this enhances the ability to distinguish between different defects. We
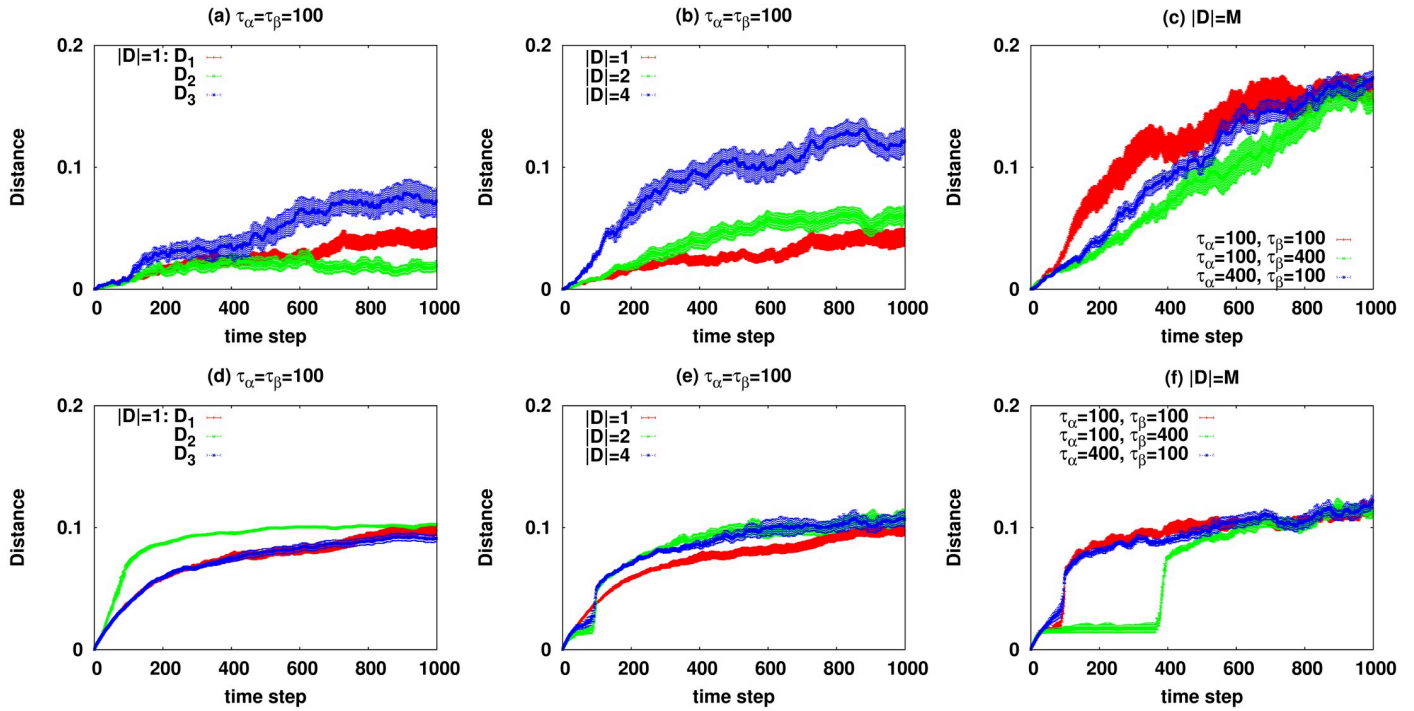
Fig 6. The average distance $d(t) = \sum_i |c_i(t) - c_i(0)|/N$ of the concentration values $c_i(t) = \langle X_i(t) \rangle / X_{max}$ at different time steps. The data are for the reaction network of Fig 2(j) with $N = 13$ species and $M = 6$ reactions. Top panels show the results with reversible reactions for: (a) evolution with one defect ($|\mathbf{D}| = 1$) for $\tau_\alpha = \tau_\beta = 100$, (b) evolution with one, two, and four defects ($|\mathbf{D}| = 1, 2, 4$) for $\tau_\alpha = \tau_\beta = 100$, (c) evolution with all defects ($|\mathbf{D}| = M$) for different $\tau_\alpha = 100, 400$ and $\tau_\beta = 100, 400$. Bottom panels (d),(e),(f) show the results with irreversible reactions.
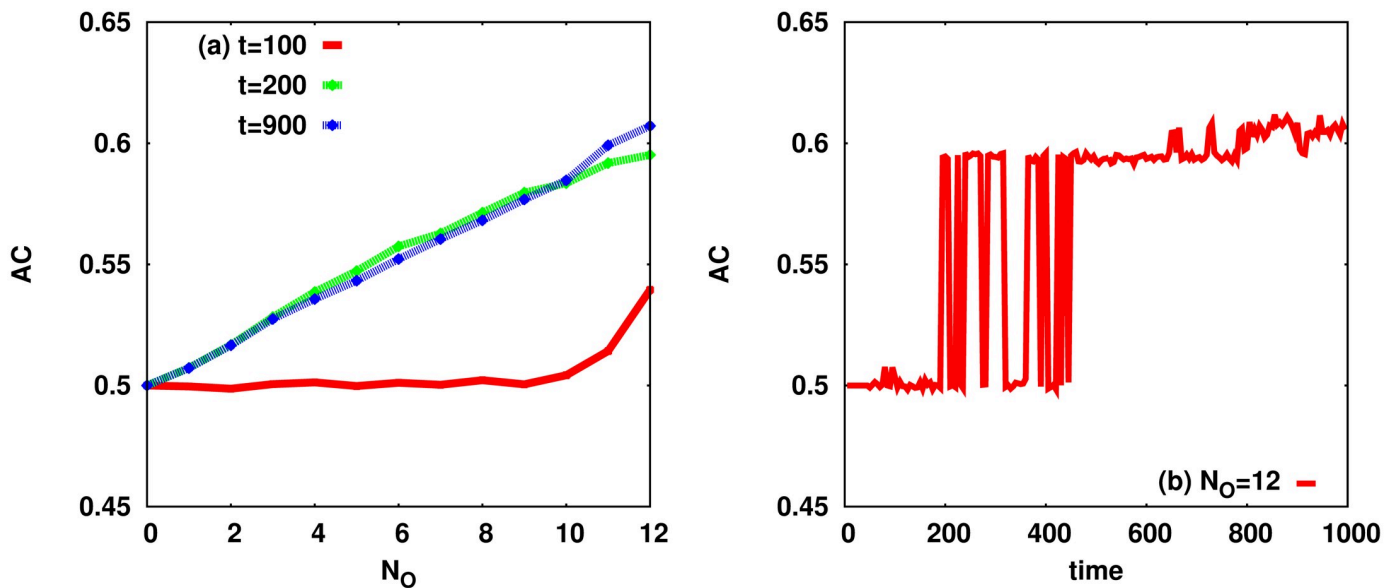
Fig 7. Accuracy of predictions with the probabilistic diagnostic model of Ref [14]. The accuracy (true positive + true negative)/(total positive + total negative) at different time steps t with $\tau_\alpha = \tau_\beta = 100$: (a) Using the D1S1 model in the presence of one defect ($|\mathbf{D}| = 1$) and (b) Time dependence of the accuracy in presence of one defect ($|\mathbf{D}| = 1$) for $N_O = 12$.

observe in Fig 7(a) that this improvement (with respect to random predictions) in the accuracy is very sharp; it is nearly zero for times less than 200 and then it takes a nearly constant value (for given N_O) at larger times. The figure (panel (b)) also shows how the accuracy (for a given $N_O = 12$) changes with time when the number of present defects $|\mathbf{D}| = 1$. Here it is clearer to see the discontinuous behaviour of the average accuracy; in particular, there is a transition time interval (200, 400) where the system displays two macroscopic states with different values for the diagnostic accuracy. When the accuracy exhibits such a behaviour, the optimal intervention time for diagnosis, considering the decreasing objective function, is just after the transition.

Note that here we are using only the statistical information that are obtained at a given time step t. A more accurate diagnosis should take into account the whole history of the sign variables, which is the subject of our future studies.

Discussion

In summary, we presented a stochastic model using the reverse annealing algorithm for the simulation of disease evolution in time. The dynamical model depends on two parameters, which control the rate of introducing mutations (defects) and the rate of accepting a decrease in the objective function. The relative strength of these parameters determines the overall behaviour of the system. For instance, the transition to the disease state is sharper when the rate of generating a mutation is lower than the rate of accepting the mutation. Moreover, we used a probabilistic diagnostic model to estimate the accuracy of diagnosis as the system performance degrades in time in the presence of some defects. The results show how much the diagnostic accuracy improves by the elapsed time of a disorder. This allows us to quantify the tradeoff between the accuracy of diagnosis and the degree of disease progression.

A microscopic model of disease evolution could be useful for a diagnostic problem which is to consider the history (dynamics) of the sign variables $\mathbf{S}(t_i \to t_f) = \{\mathbf{S}(t_i), \ldots, \mathbf{S}(t_f)\}$. For instance, given the time dependence of the molecular concentrations $X_i(t)$ in a chemical reaction network, then a relevant problem is to reconstruct the time evolution of the model parameters $\alpha_r(t), \beta(t)$. Obviously, a diagnosis that relies on the likelihoods of defects for a given history $\mathbf{S}(t_i \to t_f)$, would be more accurate than a diagnosis that is solely based on the current sign values.

The model presented in this study is of course a toy model, which does not recapitulate fully the biological complexity of disease characterization and development. However, we believe that it could be a good starting model to capture some of the essential ingredients of disease progression and the clinical diagnostic problem. The main objectives behind this work are:

- to provide a microscopic model that generates synthetic data which are needed for construction of deeper probabilistic models of signs and diseases.

- to present a dynamical model of disease evolution which (I) can be used to model explicitly disease-disease interactions and (II) can be used directly in a diagnostic problem that is based on the dynamics or history of the observed signs. That is to infer the effective parameters α_r and β to uncover the underlying defects behind the dynamics.

As a first application, we used the proposed simple diagnostic model to provide insights into the optimal intervention time (considering the accuracy of diagnosis and the level of disease development). The second application can involve, for example, a well reconstructed part of the cell reaction network with a history of observed molecular concentrations, to localize the position of diverging reactions in the network from the inferred parameters α_r. On the

other hand, a global measure of the network deviation from the healthy state is provided by the parameter β which somehow quantifies the overall level of destructive noises in the reaction network.

It would be interesting to apply the methodology of this study to the Hopfield model of associative neural networks [35]. In addition to connections with the study of mental disorders, the Hopefield model is also related to simple models of the immune system [36, 37]. Finally, the study can also be done for a personal reaction network which has been reconstructed from a single-person biomedical data [38–40]. This, in turn, would allow for a more precise and personalized diagnosis.

Methods

The results of numerical simulations are obtained for the reaction networks of Fig 2. The reaction networks are chosen in a way that mimic the biological structures. The number of molecules for each species is restricted to $0 \leq X_i(t) < X_{max} = 1000$. The value of X_{max} is chosen to be of the order of the molecular numbers in real biological systems. The initial number of molecules is chosen randomly and uniformly in $(0, X_{max})$. For the externally driven species, which are indicated by dashed arrows in the figure, we assume that the number of molecules at any time obeys a uniform probability distribution. It means that the number of these species is not determined by the system dynamics but these species still affect the reactions in which they play the role of reactants. To compute the probabilities and mutual informations we do the numerical simulation in parallel for a population of $\mathcal{N}_{pop} = 10^5$ independent and identical reaction networks.

We use a zero-temperature Monte Carlo algorithm to find a local maximum of the objective function. We start from random initial values for the κ_r. At each step, new reaction rates are suggested around the current values. Then, the mutual informations are estimated from the stochastic dynamics of the reaction network for $\Delta t_{eq}^0 = 4000$ iterations to reach a stationary state. We run the algorithm for another $\Delta t_{av}^0 = 1000$ iterations to extract the statistical information needed for computation of the objective function; at each iteration, we allow for M reactions to happen according to the dynamical rules of the stochastic evolution (Gillespie) algorithm. The suggested changes in the reaction rates are accepted only if the objective function increases.

Supporting information

S1 Text. Supporting information are provided in file S1_Text.
(PDF)

Author Contributions

Conceptualization: Alireza Mashaghi.

Investigation: Abolfazl Ramezanpour.

Methodology: Abolfazl Ramezanpour.

Resources: Alireza Mashaghi.

Supervision: Alireza Mashaghi.

Writing – original draft: Abolfazl Ramezanpour, Alireza Mashaghi.

Writing – review & editing: Abolfazl Ramezanpour, Alireza Mashaghi.

References

1. Tierney LM, Papadakis MA, McPhee SJ, editors. Current Medical Diagnosis and Treatment, 1997. McGraw-Hill/Appleton & Lange; 1996.

2. Liu W, Wu A, Pellegrini M, Wang X. Integrative analysis of human protein, function and disease networks. Scientific reports. 2015 Sep 24; 5:14344. https://doi.org/10.1038/srep14344 PMID: 26399914

3. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015 Feb 20; 347(6224):1257601. https://doi.org/10.1126/science.1257601 PMID: 25700523

4. Spielgelharter DJ. Probabilistic expert systems in medicine. Stat Sci. (1987) 2:3–44.

5. Heckerman D. A tractable inference algorithm for diagnosing multiple diseases. InMachine Intelligence and Pattern Recognition 1990 Jan 1 (Vol. 10, pp. 163–171). North-Holland.

6. Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. Methods of information in Medicine. 1991; 30(04):241–55. PMID: 1762578

7. Heckerman DE, Shortliffe EH. From certainty factors to belief networks. Artificial Intelligence in Medicine. 1992 Feb 1; 4(1):35–52.

8. Nikovski D. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. IEEE Transactions on Knowledge and Data Engineering. 2000 Jul; 12(4):509–16.

9. Goutsias J, Jenkinson G. Markovian dynamics on complex reaction networks. Physics reports. 2013 Aug 10; 529(2):199–264.

10. Mezard M, Montanari A. Information, physics, and computation. Oxford University Press; 2009 Jan 22.

11. Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012 Sep 7.

12. Cramer AO, van Borkulo CD, Giltay EJ, van der Maas HL, Kendler KS, Scheffer M, Borsboom D. Major depression as a complex dynamic system. PloS one. 2016 Dec 8; 11(12):e0167490. https://doi.org/10.1371/journal.pone.0167490 PMID: 27930698

13. Goldenfeld N. Lectures on phase transitions and the renormalization group. CRC Press; 2018 Mar 8.

14. Ramezanpour A, Mashaghi A. Toward First Principle Medical Diagnostics: On the Importance of Disease-Disease and Sign-Sign Interactions. Frontiers in Physics. 2017 Jul 25; 5:32.

15. Ramezanpour A, Mashaghi A. Uncovering hidden disease patterns by simulating clinical diagnostic processes. Scientific reports. 2018 Feb 5; 8(1):1–3.

16. Mashaghi A, Ramezanpour A. Statistical physics of medical diagnostics: Study of a probabilistic model. Physical Review E. 2018 Mar 16; 97(3):032118. PMID: 29776109

17. Mitchell MG. Molecular Pathology and the Dynamics of Disease. Academic Press; 2018 Apr 16.

18. Asachenkov A, Marchuk G, Mohler R, Zuev S. Disease dynamics. Springer Science & Business Media; 1993 Dec 23.

19. Smith VH, Holt RD. Resource competition and within-host disease dynamics. Trends in ecology & evolution. 1996 Sep 1; 11(9):386–9.

20. Shaw CA. Neural dynamics of neurological disease. John Wiley & Sons, Incorporated; 2017 Feb 3.

21. Goldstein B, McNames J, McDonald BA, Ellenby M, Lai S, Sun Z, Krieger D, Sclabassi RJ. Physiologic data acquisition system and database for the study of disease dynamics in the intensive care unit. Critical care medicine. 2003 Feb 1; 31(2):433–41. PMID: 12576948

22. Sjölinder H, Jonsson AB. Imaging of disease dynamics during meningococcal sepsis. PLoS One. 2007; 2(2). https://doi.org/10.1371/journal.pone.0000241 PMID: 17311106

23. Bhalla US, Iyengar R. Emergent properties of networks of biological signaling pathways. Science. 1999 Jan 15; 283(5400):381–7. PMID: 9888852

24. de Ronde W, Tostevin F, Ten Wolde PR. Multiplexing biochemical signals. Physical review letters. 2011 Jul 19; 107(4):048101. PMID: 21867046

25. Van Wijk R, Tans SJ, Ten Wolde PR, Mashaghi A. Non-monotonic dynamics and crosstalk in signaling pathways and their implications for pharmacology. Scientific reports. 2015 Jun 18; 5(1):1–3.

26. Schilling CH, Palsson BO. The underlying pathway structure of biochemical reaction networks. Proceedings of the National Academy of Sciences. 1998 Apr 14; 95(8):4193–8.

27. Kremling A, Jahreis K, Lengeler JW, Gilles ED. The organization of metabolic reaction networks: a signal-oriented approach to cellular models. Metabolic Engineering. 2000 Jul 1; 2(3):190–200. PMID: 11056061

28. Tyson JJ, Novák B. Functional motifs in biochemical reaction networks. Annual review of physical chemistry. 2010 May 5; 61:219–40. https://doi.org/10.1146/annurev.physchem.012809.103457 PMID: 20055671

29. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. science. 1983 May 13; 220 (4598):671–80. PMID: 17813860

30. Temkin ON, Zeigarnik AV, Bonchev DG. Chemical reaction networks: a graph-theoretical approach. CRC Press; 1996 Aug 5.

31. Van Kampen NG. Stochastic processes in physics and chemistry. Elsevier; 1992 Nov 20.

32. Gillespie DT. Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. 2007 May 5; 58:35–55. PMID: 17037977

33. Gupta A, Briat C, Khammash M. A scalable computational framework for establishing long-term behavior of stochastic reaction networks. PLoS computational biology. 2014 Jun; 10(6). https://doi.org/10.1371/journal.pcbi.1003669 PMID: 24968191

34. Hepp B, Gupta A, Khammash M. Adaptive hybrid simulations for multiscale stochastic reaction networks. The Journal of chemical physics. 2015 Jan 21; 142(3):034118. PMID: 25612700

35. Amit DJ. Modeling Brain Function (Cambridge. Cambridge Univ. Press [2] PA Getting (1988), Neural Control of Rhythmic Movements in Vertebrates, eds, AH Cohen, S. Rossignol, 8L S. Grillner, New York: John Wiley. 1989;85:101-27.

36. Agliari E, Barra A, Guerra F, Moauro F. A thermodynamic perspective of immune capabilities. Journal of theoretical biology. 2011 Oct 21; 287:48–63. PMID: 21824481

37. Agliari E, Annibale A, Barra A, Coolen AC, Tantari D. Immune networks: multitasking capabilities near saturation. Journal of Physics A: Mathematical and Theoretical. 2013 Sep 27; 46(41):415003.

38. Papin JA, Hunter T, Palsson BO, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. Nature reviews Molecular cell biology. 2005 Feb; 6(2):99–111. PMID: 15654321

39. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences. 2007 Feb 6; 104(6):1777–82.

40. Fröhlich F, Kaltenbacher B, Theis FJ, Hasenauer J. Scalable parameter estimation for genome-scale biochemical reaction networks. PLoS computational biology. 2017 Jan; 13(1).