Supporting Online Material

Predictive internal representations underlie anticipatory behavior within microbial genetic networks

Ilias Tagkopoulos[2,3], Yir-Chung Liu[1,3], Saeed Tavazoie[1,3]

[1]Department of Molecular Biology, [2]Department of Electrical Engineering, & [3]The Lewis-Sigler Institute for Integrative Genomics

Princeton University, Princeton, NJ 08544

1. Evolution in Variable Environments (EVE)

1.1 Simulation environment

In our simulation framework (EVE: Evolution in Variable Environments) populations of *in silico* organisms compete and evolve under user-specified dynamic environments. Each organism (cell) contains a biochemical network whose chemical species and their interactions/transformations are modeled according to the "central dogma". Simulations utilized multi-node supercomputer clusters (BlueGene/L and Beowulf running on an average of 500 node workload for over 2 years).

Recent studies (*S1-5*) have made significant progress in simulating biochemical networks and extracting novel insights in the context of focused biological questions (e.g. circadian rhythms, chemotaxis, segmentation, etc.). Extending from these works, the EVE framework integrates many features that improve the biochemical, evolutionary, and ecological realism of our simulations, features that are crucial for simulating microbial regulatory networks in the context of interactions with the environment.

From the biochemical perspective, dynamics are modeled as stochastic events, not ordinary differential equations, a feature that accounts for phenomena arising from small numbers of molecules. Additionally, We take into account the full range of molecular species and their interactions, including genes, mRNAs, siRNAs, miRNAs, proteins, and modified proteins. From the evolutionary perspective, EVE is designed to simulate growth, mutation, and selection in a highly realistic fashion. EVE uses an asynchronous simulation framework, where events such as cell-division, death, mutations etc., occur in a parallel, asynchronous fashion. Another important distinction is the way mutations are modeled, having a range of types and magnitudes that map to specific biological counterparts. Another important dimension of realism is the capacity for gene-duplication and divergence which is a recurring theme in the evolution of complex behavior in many of our experiments.

Another very important advantage of EVE—and essential for addressing our primary thesis here—is the way in which the evolving organisms interact with their environments, sensing and responding to dynamic and temporally structured signals. In this context, our organisms are not evolved to explicitly implement specific functions (e.g. an oscillator), rather they evolve to maximize fitness through learning the probabilistic relationships between signals they can sense and the abundance of resources. For ease of interpretation, our simulations here have focused on sharply structured environments that correspond to familiar logic gates (e.g. delayed dynamic XOR). However, real microbial habits are likely characterized by probabilistic relationships amongst a large number of variables that show rather complex and fuzzy relationships with fitness. This is a fundamental departure from the traditional view of the relationship between regulatory-network structure and function. Our simulation framework provides the capacity to explore this new paradigm.

The details of the EVE simulation environment are provided below.

1.2 Network components

Each organism comprises three types of nodes with distinct functionalities:

- Gene/RNA Nodes
- Protein Nodes
- Modified Protein Nodes

Gene/RNA Nodes capture gene regulation and transcription of RNA (mRNA or siRNA). ***Protein Nodes*** summarize the translation of mRNA to proteins and model post-transcriptional regulation. ***Modified Protein Nodes*** capture the modification (phosphorylation, acetylation, etc) of proteins and other conformational changes that may lead to gain/loss of function. The value of each node is a discrete

variable that corresponds to the number of node molecules present at a particular time. We define as ***node triplet*** any set of a Gene/RNA Node and its corresponding Protein and Modified Protein Nodes.

1.3 Network Topology

The essence of each artificial organism is its internal interaction network. This can be depicted as a directed graph where each vertex is a Gene/RNA, Protein or Modified Protein node, and each edge a regulatory dependency between the two vertices that it connects. Edges are directional and their values can be either negative or positive depending on the type of interaction. A quantitative representation of the resulting graph is the interaction matrix W, where any non-zero value w_{ij} denotes the strength and regulation type (negative – positive) of node *i* by node *j*. Figure S1 depicts a randomly created internal network and its interaction matrix.

1.4 A probabilistic dynamical model

We created a probabilistic dynamical model of intracellular biochemical interactions that regards an organism as a well-stirred and spatially homogeneous reaction chamber. This allows us to efficiently capture stochastic phenomena, for example, noisy gene expression caused by a small number of molecules. However, due to lack of spatial structure, this simulation framework is unsuitable when compartmentalization and spatial localization is important to the function of the network/organism.

The general framework is similar to a dynamic Monte Carlo algorithm (*S6*) with a random selection method where in each time unit the probability of each event (node creation, deletion, mutation, molecule creation and degradation) is calculated and a particular event occurs if its probability is found to be higher than a randomly generated number. Although this approach can be computationally more intensive when compared to first reaction algorithms (*S7*), it requires less memory and is easier to parallelize in multi-processor environments.

1.5 Expression of RNAs, proteins, and modified proteins

The value of each node corresponds to the total number of molecules of a particular type that are present at that given time. The value of each node may fluctuate but always by one molecule at a time when measured in consecutive time points. This fluctuation is a result of molecule creation, transformation or degradation. To calculate the molecule production probability (i.e. the probability that a new molecule will be created) we use two levels of sigmoid functions: One to capture the effect of each individual regulator on the target node, and a second to compute the sum of those effects. More specifically to calculate the molecule production probability of node *i* we first calculate the function:

$$R_i(f_{i1},\dots,f_{in},w_{i1},\dots,w_{in},m_i,s_i) = basal_i + (1-basal_i)\cdot\tanh\left(\frac{\sum_{j=1}^{n}\left(w_{ij}\cdot f_{ij}\left(x_j,\tilde{m}_{ij},\tilde{s}_{ij}\right)\right)-m_i}{s_i}\right)$$

Where the function $f_{ij}\left(x_j,\tilde{m}_{ij},\tilde{s}_{ij}\right)$ corresponds to the regulatory effect of node *j* on node *i*:

$$f_{ij}\left(x_j,\tilde{m}_{ij},\tilde{s}_{ij}\right) = \frac{1}{2}\cdot\left[1+\tanh\left(\frac{x_j-\tilde{m}_{ij}}{\tilde{s}_{ij}}\right)\right]$$

In the previous equations, $\boldsymbol{n}$ is the total number of nodes, $\boldsymbol{basal_i}$ the basal expression parameter, $\boldsymbol{w_{ij}}$ the value of the ith row and jth column in the interaction matrix, $\boldsymbol{x_j}$ the value of node *j*, $\boldsymbol{m_i}$ and $\boldsymbol{s_i}$ the midpoint and slope of the main target-specific sigmoid function, $\tilde{\boldsymbol{m}}_{\boldsymbol{ij}}$ and $\tilde{\mathbf{s}}_{\mathbf{ij}}$ the midpoint and slope of the regulator-specific sigmoid function.

In the case of protein translation and modification, the molecule production probability also depends on the number of substrate molecules (i.e. RNA, proteins) that are present in the environment. Additionally, a realistic model has to capture the saturation effects of the translation and modification machinery (i.e. ribosomes, enzymes) at high molecular concentrations. For this reason R_i is multiplied by yet another function $g(x_i)$ that is also a sigmoid in the case of protein translation and modification:

$$G_i(g_i, R_i) = R_i \cdot g_i(x_i)$$

Where

$$g(x_i) = \begin{cases} \frac{1}{2} \cdot \left[1 + \tanh(\frac{x_i - m}{s})\right] & if\ translation/modification \\ 1 & if\ transcription \end{cases}$$

Finally, the molecule production probability is given by:

$$P_i = \begin{cases} 0 & if\ G_i < 0 \\ G_i & if\ 0 \le G_i \le 1 \\ 1 & if\ G_i > 1 \end{cases}$$

Figure S2 gives a general schematic of the expression model and depicts the effect various values of midpoint and slope have on the sigmoid function $f_{ij}(x_j, \tilde{m}_{ij}, \tilde{s}_{ij})$.

In our experiments, the parameters $\boldsymbol{m}$ and $\boldsymbol{s}$ of function $g(x_i)$ take the heuristic values 10 and 4 respectively. Although RNA molecules are not consumed during translation, each modification event converts one unmodified protein molecule to its modified form and thus each protein modification causes its (unmodified protein) pool to be reduced by one. Proteins that get modified have a low (<0.1) probability of returning to their unmodified state. Although external regulation of the modified to unmodified transitions is a feature in EVE, it was disabled in the experiments we present here.

1.6 Degradation

Degradation of molecules occurs in a somewhat similar manner. The probability that a molecule of node *i* is degraded at any given time unit is proportional to the number of molecules present and is given by:

$$P_{deg_i}(x_i, m_{deg_i}, s_{deg_i}, p_{deg_i}) = \frac{1 + p_{deg_i}}{2} + \frac{1 - p_{deg_i}}{2} \cdot \tanh\left(\frac{x_i - m_{deg_i}}{s_{deg_i}}\right)$$

Where $\boldsymbol{x_i, p_{deg_i}, m_{deg_i}, s_{deg_i}}$ are the number of molecules present, node specific degradation parameter, degradation midpoint and slope respectively ($m_{deg_i} = 50\ and\ s_{deg_i} = 4$ in the experiments presented here).

1.7 Triplet duplication and deletion

A Node triplet can be created or destroyed at any time unit during the course of an experiment. During triplet duplication, a new triplet is created by cloning (duplicating) an existing one. The new triplet initially has exactly the same associations and parameters as the old one, although it may eventually take on its own course during evolution. Similarly, during a triplet deletion event a whole triplet is deleted. It is important to note that these two processes apply on the triplet as a whole (i.e. Gene/RNA, protein and modified protein nodes) and not to a partial set of nodes within the triplet.

1.8 Modeling mutations

Like triplet duplications and deletions, mutations happen at any time point and in any organism. Mutations always target one or more parameters in a particular node. During a mutational event, one of the following node-specific parameter groups is altered:

- Target node midpoint ($\boldsymbol{m_i}$) and slope ($\boldsymbol{s_i}$).
- Interaction weights ($\boldsymbol{w_{ij}}$), regulator specific midpoint ($\boldsymbol{\tilde{m}_{ij}}$) and slope ($\boldsymbol{\tilde{s}_{ij}}$).
- Basal expression parameter ($\boldsymbol{basal_i}$).
- Degradation parameter ($\boldsymbol{p_{deg_i}}$)

The parameters within each group were specifically selected in order to match their biological counterparts. A mutation in the first group captures a sequence or conformational change in the target node that has an effect in its regulation function as a whole. A mutation in the second group models the case where a change in residue(s) or base pair(s) – either in the regulator or target node – results in a change of the interaction affinity between the target and the regulator node. This type of mutation is regulator-target specific since it does not affect other regulatory interactions associated with the target node. Finally a mutation may change the basal expression parameter (e.g. promoter becomes more/less leaky) or the degradation parameter (e.g. point mutations in RNA or protein that affects its stability).

In addition to the type of parameters that a mutation targets, mutations can also be categorized based on the severity of the change. As such, mutations can be grouped in two classes: Mild and Strong mutational events. In mild mutational events, the parameters drift according to a Gaussian distribution around their current value, whereas in strong mutational events they can take on any random value within the valid (initial) range. Strong mutations are relatively rare events in the course of the experiment and they model large impact changes that can result in loss – or even reversal – of function (e.g. an activating transcription factor that becomes a repressor for a specific target gene). Figure S3 demonstrates the variety of possible mutational events accompanying the evolution of a toy oscillator circuit.

1.9 Biochemical parameters and organism attributes

Organism parameters can be classified into three categories and are reported here with the ranges we used during initialization in low mutation rate (LM) experiments. Parameters are sampled from a uniform distribution, unless otherwise stated.

- General parameters:
    - Energy: The energy that the organism has at each time point (initially $8 \cdot 10^5$ energy units).
    - Fitness: The Pearson-correlation between the time series vectors that represent the pattern of response pathway expression and abundance of environmental resources.
    - Maintenance Energy Cost: Energy cost per unit time associated with the maintenance of nodes. The maintenance energy cost for node *i* has a value of $\text{Cost}_i = x_i \cdot M$ energy units, where x_i is the node-specific maintenance cost and M the mutation modifier parameter. For simplicity, in our experiments $x_i = 1$ for all nodes. The mutation modifier per unit time models the cost of DNA repair and other proofreading mechanisms and is the relative ratio between the mutation rates at the beginning of the experiment to its current value. Thus, an organism evolved to have lower mutation rate has a higher mutation modifier parameter M (here M is equal to one since mutation rate per organism is constant).
    - Response pathway cost: Energy cost per unit time associated with maintaining the response pathway (modified protein RP1) active. It is linearly proportional to the number of RP1 molecules present in the organism and it is given by $\text{cost}_{\text{response\_pathway}} = z \cdot P$, where z

is the default cost for one protein per time unit (here z = 10) and P is the number of RP1 molecules.

- Mutation Specific parameters:
  - Triplet Creation Probability: The probability of triplet creation in each time unit (10^{-7} in LM experiments).
  - Triplet Destruction Probability: The probability of triplet destruction in each time unit (10^{-7} in LM experiments).
  - Mild Mutation Probability: The probability of mild mutation in any node per unit time(10^{-6} in LM experiments).
  - Strong Mutation Probability: The probability of strong mutation in any node per unit time ($2 \cdot 10^{-7}$ in LM experiments).
  - Evolvability: The rate of change in the creation, destruction, strong and mild mutation probabilities. This is used only in experiments that allow organisms to vary their mutation probabilities during the course of evolution (here 0.2 in all experiments).
- Kinetic and Network related parameters:
  - Number of Nodes: The total number of nodes in the cell (initially from 0 to 30).
  - Node abundance: Abundance of molecules corresponding to a node (initially from 0 to 30).
  - Connectivity: The fraction of actual to all possible connections in the internal network (initially from 0.01 to 0.20).
  - Basal: A vector that contains the basal levels of transcription, translation and modification for all nodes in the internal network (initially from 0 to 0.5).
  - Degradation: A vector that contains the degradation rates of all node values in the internal network (initially from 0.05 to 1).
  - Interaction matrix: A matrix with all regulatory dependencies between nodes and environmental signals. Distribution of interaction weights initially obeys a power law with $\gamma = 1.5$ while the connection probability of an environmental signal to a node decreases exponentially with the number of nodes (in the same organism) it is already connected to.
  - Target Slope and Midpoint: Two vectors that contain the slope and midpoint of the primary target-specific sigmoid function that is used to model the regulation of a specific node by other nodes and signals (initially slope ranges within [1,3] and midpoint within [-1,1]).
  - Regulator Slope and Midpoint: Two vectors that contain the slope and midpoint of a secondary regulator-specific sigmoid function that is used to model the effect of regulation by a node or signal of another node (initially the regulator slope ranges in [1,4] and the regulator midpoint in [0,10].

1.10 Modeling organism-environment interactions

Organisms compete in an environment of constant population size that may contain any number of signals with arbitrary or predefined characteristics. Various events that have a deleterious or beneficial fitness effect on the organisms occur during the time course of an experiment. The temporal pattern of event occurrence is flexible, ranging from purely stochastic to perfectly periodic. In the present work experiments have a single event type, namely the presence of energy resources that organisms can harvest through the expression of a metabolic pathway that is represented by a modified protein called "response protein 1" (RP1). Environmental signals can be programmed to correlate – partially or fully - with the future occurrence of an event, making it possible for organisms to predict environmental trajectory. To achieve this, the organisms have to "learn" how to extract environmental information from one or more signals. A specific node can couple/decouple its expression to an external signal as a consequence of random mutations. Signal values can range from -1 to 1 and their regulatory effect is directly proportional to that value.

The creation and destruction of any RNA, Protein or Modified Protein has an energy cost. Of special interest is RP1, the modified protein of triplet 1, whose creation cost is ten times more than that of other proteins. In addition to its high creation cost, the response protein has a maintenance cost per unit time that is proportional to the number of its molecules present at any one time. This models well the real metabolic machinery in cells, whose expression and maintenance is energetically costly. The probability of death decreases exponentially with the energy level of the organism and approaches unity as energy levels get close to zero. In order to keep the population size constant, a randomly generated organism is placed in the population after the death of another.

The amount of energy that organisms acquire from the environment per unit time is proportional both to the amplitude of the resource and to the response pathway expression. When an organism's energy level reaches a predefined threshold (in our analysis it is set to twice the initial energy level), the organism undergoes mitosis. During mitosis, the organism is cloned and its energy level is reduced to half. The newly created daughter cell takes the place of the organism with the least energy.

1.11 EVE simulation algorithm

We developed two different algorithms that simulate our evolutionary environment and obtained similar results with both. Although the results presented here are derived from the real-time version of the simulator, both generation-based and real-time implementations are described for the sake of completeness.

1.11.1 Generation based simulator

At the beginning of each experiment, a fixed-sized population of organisms that receive a predefined amount of energy units is created. During a given time duration, organisms undergo mutations while their node values are continuously updated. At the end of the time interval, a new generation of organisms is formed by "sampling with replacement" from the current population. The probability for an organism to be selected for the next generation is directly proportional to the energy that it has at that point. Each time a new generation is formed, the energy of all organisms is initialized. The simulation ends after a predefined number of generations.

1.11.2 Real-time simulator

At the beginning of each experiment, a fixed-sized population of organisms that receive a predefined amount of energy units is created. At each time point during the experiment, organisms can mutate, divide and die in a parallel and asynchronous manner as described in the previous section. Statistics are gathered at periodic time intervals (***epochs***) without this affecting the evolutionary trajectory of the experiment. The experiment ends after a predefined amount of time units have elapsed.

1.12 Auxiliary algorithms

1.12.1 Deconstruction

The deconstruction algorithm was developed to assess the link/node significance and derive the minimal internal network (i.e. the core computational module) of an organism. The algorithm deconstructs networks in either a "parallel" or "serial" mode.

In parallel deconstruction, the significance of each individual link in the internal network of an organism is evaluated within a time window which is at least an order of magnitude larger than the interval between two epochs. During this evaluation process, a link gets knocked out and the fitness of the resulting organism is assessed. Regardless of the link significance, each link is restored in the internal

network at the end of every evaluation. The program terminates once the fitness effect of all links is assessed.

Serial deconstruction operates similarly with the difference that if a link proves to be insignificant (less than 5% fitness change with respect to the initial fitness), it is not restored in the original network, which leads to systematic pruning of non-essential links. To cope with redundant pathways in the network, serial deconstruction performs multiple randomizations (permutations) of link assessment.

While parallel deconstruction is able to find the fitness impact of a single mutation in the network, serial deconstruction enables us to recover redundancies and minimal networks responsible for a particular phenotype. Each link significance is assessed multiple times (3 in our analysis) both in parallel and serial deconstruction, while serial deconstruction creates several (3 in our analysis) permutations of the order by which links get knocked out in order to uncover coexisting minimal networks. The results of parallel and serial deconstruction can be visualized as a fitness matrix (eg. Fig. S13-A), depicting the negative fitness fraction change between the initial and mutated phenotypes.

1.12.2 Epistasis analysis

Epistasis is defined as any non-additive interaction between two or more distinct mutations, such that their combined effect on a phenotype deviates from the sum of their individual effects. Here, we compare the fitness effect of link deletions in an organism's internal network, both when two deletions occur independently or concurrently as a pair. We use a variation of the parallel deconstruction algorithm that knocks out two links simultaneously at each evaluation cycle in order to assess the epistatic potential of link pairs, as manifested in their combined fitness effect (Fig. S13C, S14A).

1.12.3 Reverse engineering

As an alternative approach to network inference, we used the ARACNE algorithm (*S8*) which is based on mutual information analysis of expression data. We conducted time series experiments and gathered the expression signatures of all nodes in any given network. To accurately mimic "microarray" experiments, we collected node expression data under different environmental conditions and stimuli, as for example forcing one of the signals up or down. Figure S15B shows the topology of the reconstructed network that belongs to the organism in Fig. 4, where the reconstruction algorithm discovered 4 out of the 11 links that are essential for the organism to retain its phenotypic behavior. However the reconstructed organism has none of the phenotypic properties of its original counterpart, thus rendering it essentially unfit under the original selection environment. Nevertheless, systematic analysis of reconstructed networks evolved under various mutation rates suggests that network reconstruction can indeed be helpful in identifying a portion of the essential links in the biochemical network, possibly at the expense of a high false-positive rate (Figure S16).

1.13 Additional simulations

All experiments presented here have a fixed population size of N_p= 200 and time duration D = $1.8 \cdot 10^7$ time units. Relative mutation rates are 1:1:10:2 for creation, destruction, mild and strong mutation probabilities respectively. To obtain statistical significance for each environment and mutation rate combination, we ran 64 independent simulations. Based on the selection pressure in the environment, experiments can be classified into the following categories:

- Delayed Gates: Signals and resource are related by OR, AND, XOR, NAND, NOR dynamic logic functions.
- Multi-gates: Signals and resource are interchangeably related by combinations of OR, AND, XOR, NAND, NOR dynamic logic functions.

- Oscillators: Selection pressure to evolve oscillatory expression of RP1 with or without a periodic guiding signal.
- Bi-stable Switches: Selection pressure to evolve bi-stability in environments where two environmental signals operate as ON/OFF pulse switches.
- Duration/variance Locking: Selection pressure to evolve networks that predict the duration of an environmental resource that has fluctuating duration or phase variance.

Figure S4 summarizes the correlation between the environmental signals and the resource presence in each experiment type. Figures S9-S11 depict minimal networks and/or phenotypes of representative organisms that belong in the above categories (evolved oscillatory and multi-gate organisms have large complex networks and thus are not included here).

2. Experimental Methods

2.1 Bacterial strains and growth conditions

E. coli strain MG1655 and its derivates were used in all physiology and evolution experiments. A library of Tn5 transposon mutants (*S9*) with a diversity of ~10^6 was used as the starting point for evolution experiments in order to: 1) increase the diversity of genetic perturbations on which selection acts and 2) allow marking and monitoring of individual mutants within the evolving population. Bacteria were grown in batch or bioreactor vessels in M9 minimal media supplemented with 0.4% glucose.

2.2 Controlled temperature and oxygen perturbations

Physiological perturbations were carried out under a controlled environment in the context of bioreactor (Bioflo 110, New Brunswick Scientific) growth. Thermoelectric sensors and heaters were used to shift temperature profiles between 25^0 C and 37^0 C, and polarographic dissolved oxygen sensors (Mettler Toledo) and nitrogen gas was used to rapidly change oxygen saturation between anaerobic (0% dissolved oxygen) and aerobic (16-21% dissolved oxygen) condition. The cultures were maintained in exponential phase ($O.D._{600}$ 0.2-0.4) through controlled dilution, where fresh media is pumped in and spent media is pumped out at a controlled rate. Prior to the perturbations, cells were maintained in the pre-transition environment for at least eight generations (8-24 hours). All experiments were performed in duplicate, with high reproducibility (Fig. S17). Unless specifically indicated, all perturbations were performed during exponential growth, and cells were harvested at variable intervals and rapidly mixed in an ice-cold ethanol/phenol solution (5% water-saturated phenol in ethanol) in order to stop mRNA degradation. After spinning down the cells at 5000 rpm for 5 minutes at 4^0 C, the media was discarded, and cell pellets were frozen on dry-ice/ethanol for storage at -80^0 C.

2.3 Microarray transcriptional profiling

A hot-phenol procedure was used to extract total RNA. Cell pellets were lysed with 500 µl TE (pH8.0), 50 µl 10% SDS and lysozyme (0.5 mg/ml). Total RNA was extracted sequentially with phenol/chloroform (preheated to 64^0 C), followed by chloroform/isoamyl alcohol. RNA was precipitated with 1/10 volume of 3M NaOAc (pH 5.2) and 2 volumes of ethanol. After incubating overnight at -20^0 C, samples were spun down and pellets were washed with ice cold 70% ethanol (prepared with DEPC- H_2O). RNA was resuspended in water, DNase treated (RQ1 RNase-free DNase/ Promega, WI) and purified using an RNeasy purification kit (Qiagen, CA). Fluorescent cDNA was synthesized from total RNA with 15 µg of total RNA serving as template. Cyanine-labeled Cy3 or Cy5-dUTP (Amersham Bioscience) and pdN6 random hexamers (GE healthcare) were used in reverse transcription reactions utilizing SuperScript II (Invitrogen, CA) for cDNA synthesis. Fluorescent genomic DNA was generated as described in (*S9*). Genomic DNA served as a universal reference for all hybridizations. *E. coli* genomic DNA was fragmented (500 – 2000 bp) using mechanical shearing, and subjected to Cy3 or Cy5-dUTP labeling using the BioPrime DNA labeling system (Invitrogen, CA). Following purifications through CyScribe GFX Purification Kit (Amersham Biosciences), Cy-dye labeled genomic DNA and RNA-derived cDNA probes were combined with a hybridization buffer (5X SSC, 0.1% SDS, 50% formamide, and 10 µg Salmon sperm DNA) and hybridized for ~ 16 hours at 420 C to a DNA microarray containing 95% of all ORFs in the *E. coli* MG1655 genome (*S*9). All microarray experiments were carried out in biological duplicates and showed high reproducibility (Fig. S18).

2.4 Microarray data analysis

Microarray slides were scanned on a GenePix 4000B scanner (Axon Instruments), and fluorescence data for each of the duplicates on each slide were analyzed using GENEPIX PRO 4.0 software. In a quality-

control step, elements with poor spot morphology or exhibiting uneven hybridization caused by dust particles or scratches, were flagged manually and excluded from further analyses. After local background subtraction and global normalization relative to the genomic DNA reference, duplicate measurements on the same array were averaged, yielding a single vector for each hybridization. Seven time-points were assayed for each perturbation, corresponding to 0, 4, 8, 12, 20, 28, and 44 minutes post transition. After relative scaling, biological duplicates from all experiments were averaged, leading to a single set of time-series data for each perturbation. To compare our results to other perturbations, we combined our temperature/oxygen transcriptional responses with previously published studies of UV (*S10*) and osmolarity stress (*S11*) in *E. coli*. For most of the analyses, we focused on changes in gene expression that occurred at various points (*e.g.* 20 minutes) relative to the zero time-point reference. Log (base 2) transformations of fold-changes were utilized in determining the global correlation between various perturbations (**Fig. 5B, 6D**). Sets of differentially expressed genes were defined as showing 1.5 fold or higher increase/decrease in gene expression relative to the time zero reference. Variation of this threshold from 1.33 to 2.00 gave very similar results. The hyper-geometric distribution was used to determine the chance-probability of observing overlaps in sets of differentially regulated genes (**Fig. 5A**). The combined expression dataset (consisting of 2612 genes across 36 conditions/time-points) is available and can be downloaded from our web site.

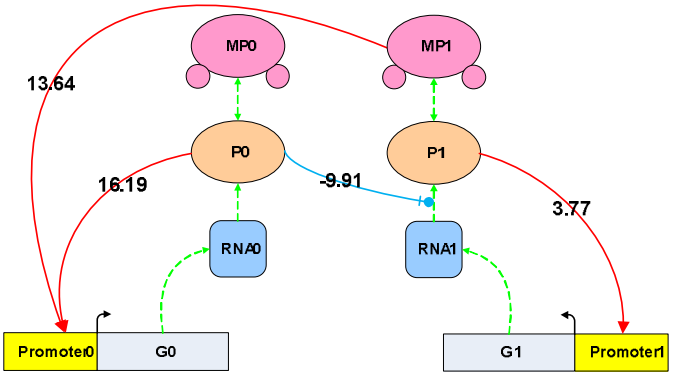
2.5 Experimental evolution under an ecologically incoherent dynamic environment

A diverse library of Tn5 transposon mutants (*S*9) was used as the starting point for evolution of *E. coli* MG1655 for optimal growth under a dynamic environment where temperature and oxygen transitions were positively correlated with a 40 minute time-lag (**Fig. 6A**). In order to avoid selection for periodic behavior, intervals of constant temperature/oxygen were sampled from a Gaussian distribution with a mean of 150 minutes (37C/O2+) and 300 minutes (25C/O2-) and standard deviations of 30 minutes (37C/O2+) and 60 minutes (25C/O2-), respectively. In total, 44 cycles of selection were carried out over the course of 384 hours. The starting population contained $\sim 1 \times 10^{10}$ transposon mutants (diversity of 5×10^5) growing in exponential phase (maintained between $O.D._{600}$ 0.20-0.30 by continuous pumping of fresh media and discarding spent media) in a volume of ~500 ml of M9 media supplemented with 0.4% glucose. Periodic sampling of the culture allowed us to monitor growth-rate and archive representative samples of the population at -80^0 C. The presence of a single randomly inserted transposon in every genome allowed us to monitor population diversity and track high-fitness mutants during the evolution experiment through the application of a gel-based genetic footprinting procedure (Fig. S20). The amplification of transposon insertion sites from a sample of 100 colonies isolated at the end of selection identified a highly abundant mutant that made up 55% of the population. This "evolved" mutant, designated as *aldB*::Tn5 was used in subsequent fitness evaluation and physiology experiments.

2.6 Competition between parental and evolved strains

To establish the fitness advantage of the evolved strain, we carried out competition experiments within the setting of the dynamic environment imposed during the evolution. The evolved and the parental strains were distinguishable through the differential expression of a LacZ marker. This system allowed us to easily quantify the relative fitness advantage of the evolved mutant using colony counting on plates (*S12*). A 60-hour competition experiment established the strong fitness advantage of the evolved strain (Fig. S21).

3. Figures and tables



$$W = \begin{bmatrix} 0 & 16.19 & 0 & 0 & 0 & 13.64 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.77 & 0 \\ 0 & -9.91 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure S1 | A randomly created internal network and corresponding interaction matrix. Red and blue edges denote activation and repression respectively. The absolute value in each edge corresponds to the weight of regulation while a positive/negative sign represents activation/inhibition respectively. Green edges depict creation in the case of mRNA transcription and protein translation, or transformation in the case of protein modification.

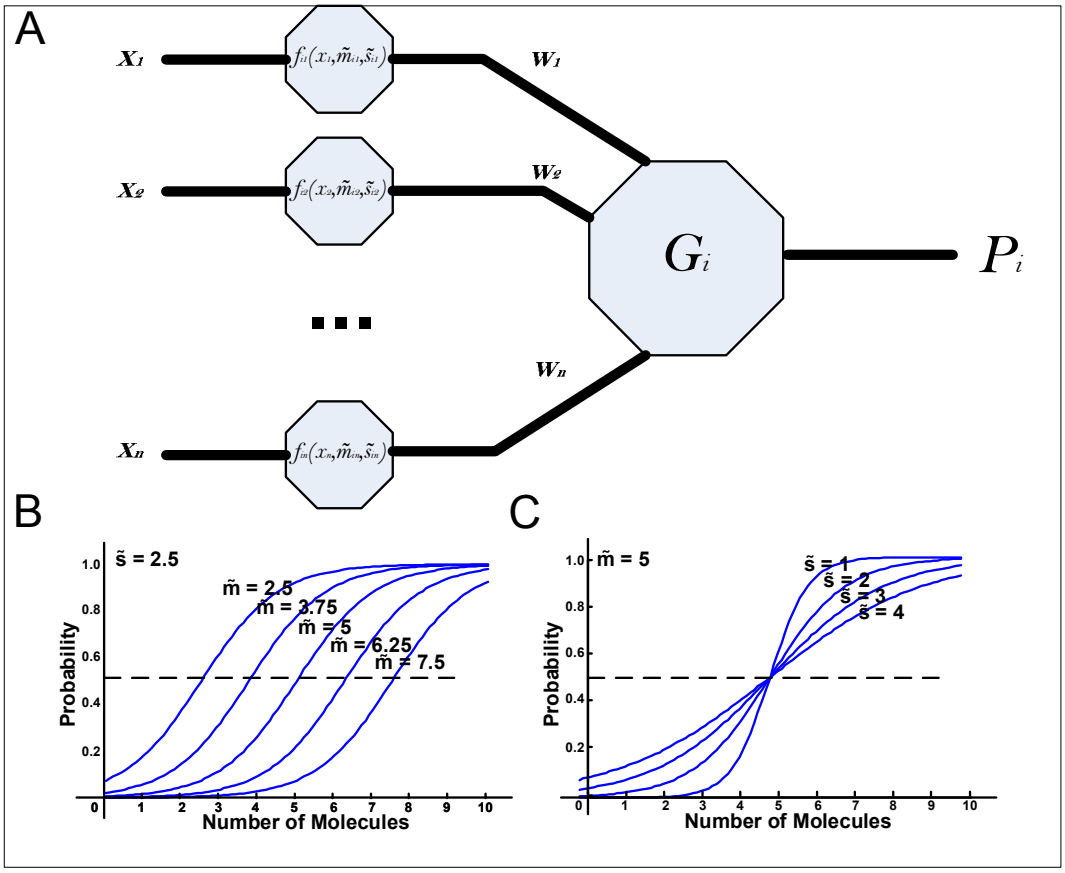


Figure S2 | Expression model and molecule production probability. (A) A two level sigmoid function is used in the expression model. The molecule production probability is represented as a second level sigmoid that has as input the sum of all individual regulatory effects. **(B)** Plot of function $f_{ij}\left(x_j, \tilde{m}_{ij}, \tilde{s}_{ij}\right)$ for different values of the midpoint, while keeping the slope constant ($\tilde{s}_{ij} = 2.5$) **(C)** Plot of function $f_{ij}\left(x_j, \tilde{m}_{ij}, \tilde{s}_{ij}\right)$ for different values of the slope, while keeping the midpoint constant ($\tilde{m}_{ij} = 5$)

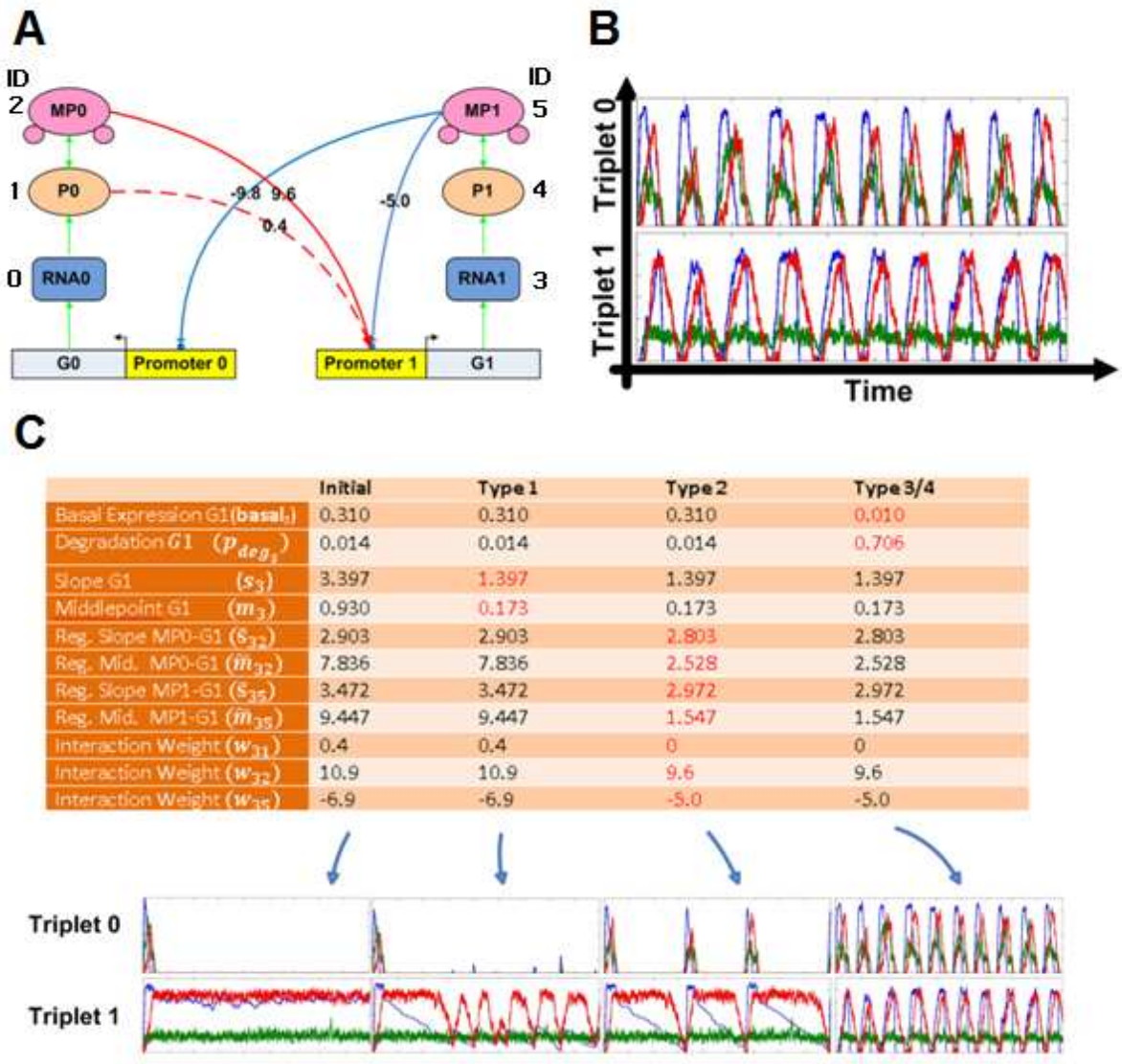


Figure S3 | Impact of mutational events on network parameters. (A) Network topology of a toy oscillator, that was evolved based on an initially user defined, non-functional design. Proteins MP0 and P0 activate (red lines) G1, while MP1 represses (blue line) G0 and G1. The P0-G1 activation (interaction weight w_{31}) is depicted with a dashed line since it is deleted during the evolution. **(B)** Expression profile of both triplets over time. In each plot, blue, green and red waveforms correspond to RNA, protein and modified protein levels respectively. **(C)** Key parameters and their values in the initial state, after type1, type2 and type3/4 mutations. The expression profile of both triplets for each state is shown at the bottom of the figure.

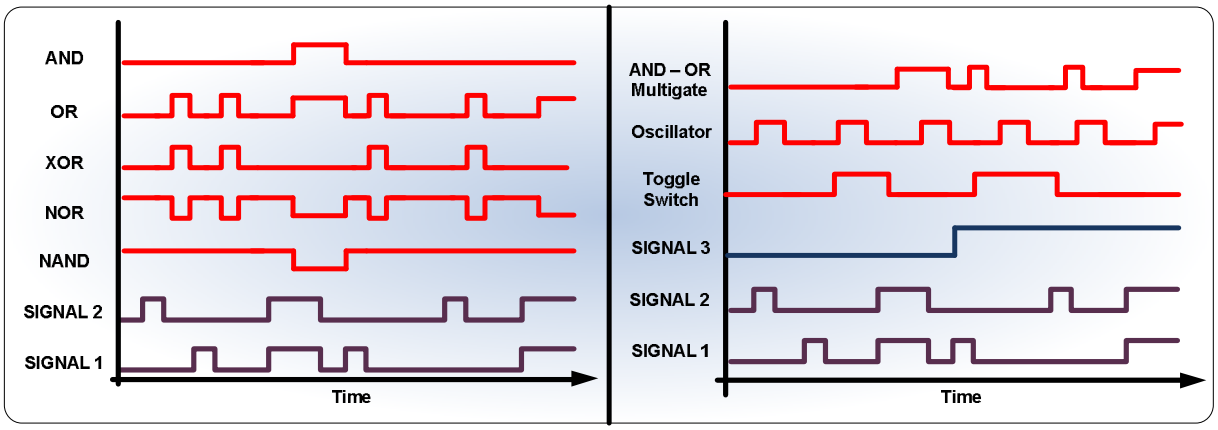


Figure S4 | Correlation of environmental signals with resource abundance for various environments. The resource abundance (red) during a specific signal pattern (black) is given for all resource-signal correlations mentioned in the text. Depending on the selection pressure there can be none, one or more signals present in the environment: oscillators have zero or one, toggle switches and gates have two signals while multi-gates have three. In a toggle switch, Signals 1 and 2 serve as ON and OFF switches respectively. In a multi-gate the third signal indicates a switching event, for example a transition from an AND to an OR logic.

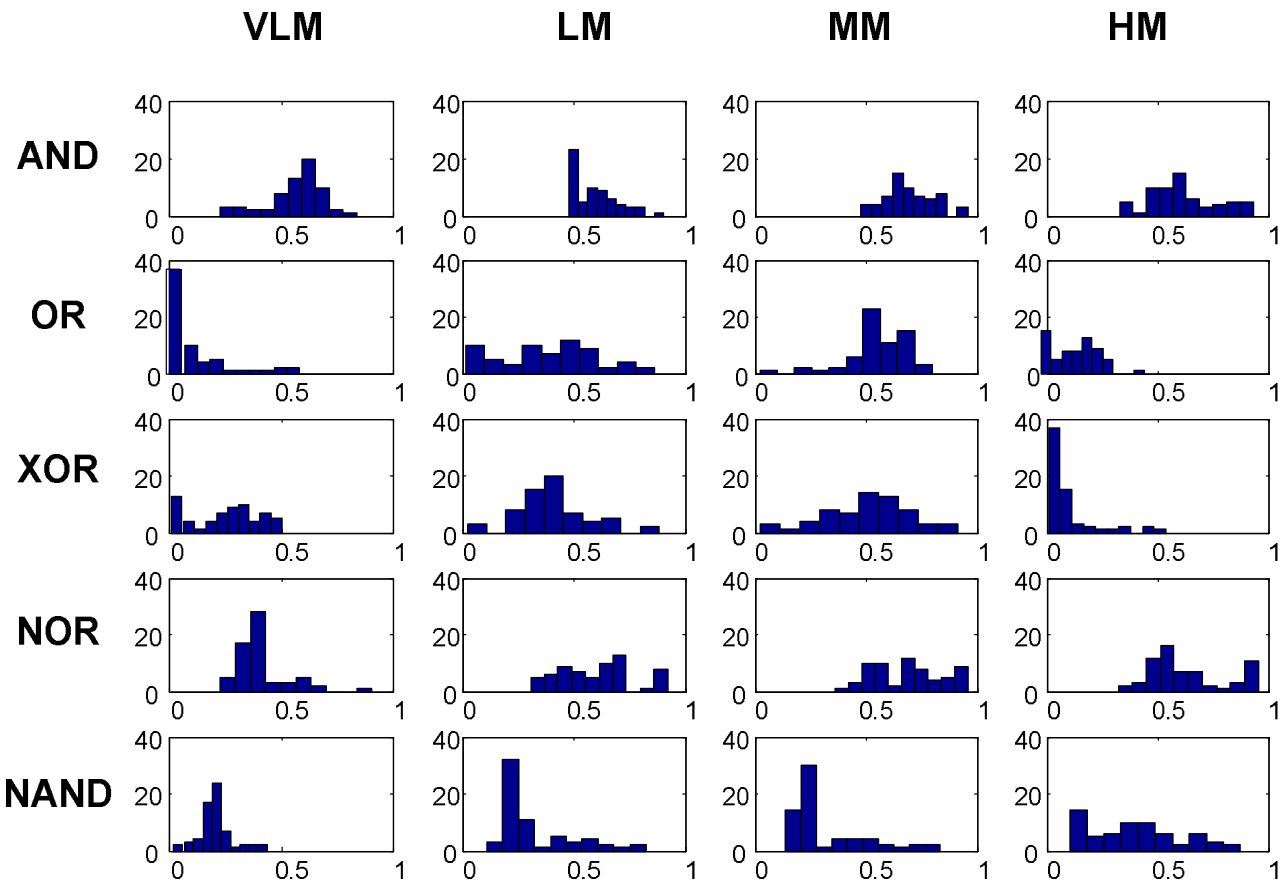


Figure S5 | Fitness (Pearson-correlation) Histogram for all delayed logic gate experiments. Each subplot depicts the histogram of Pearson-correlation between response protein expression and resource abundance of the fittest organism in the final populations (64 final populations per mutation rate/environment) of Supplementary Table 1. Each row corresponds to a gate type (AND, OR, XOR, NOR, NAND) and each column to a mutation rate (VLM, LM, MM, HM). For each subplot, x-axis and y-axis represent Pearson Correlation and number of experiments respectively.

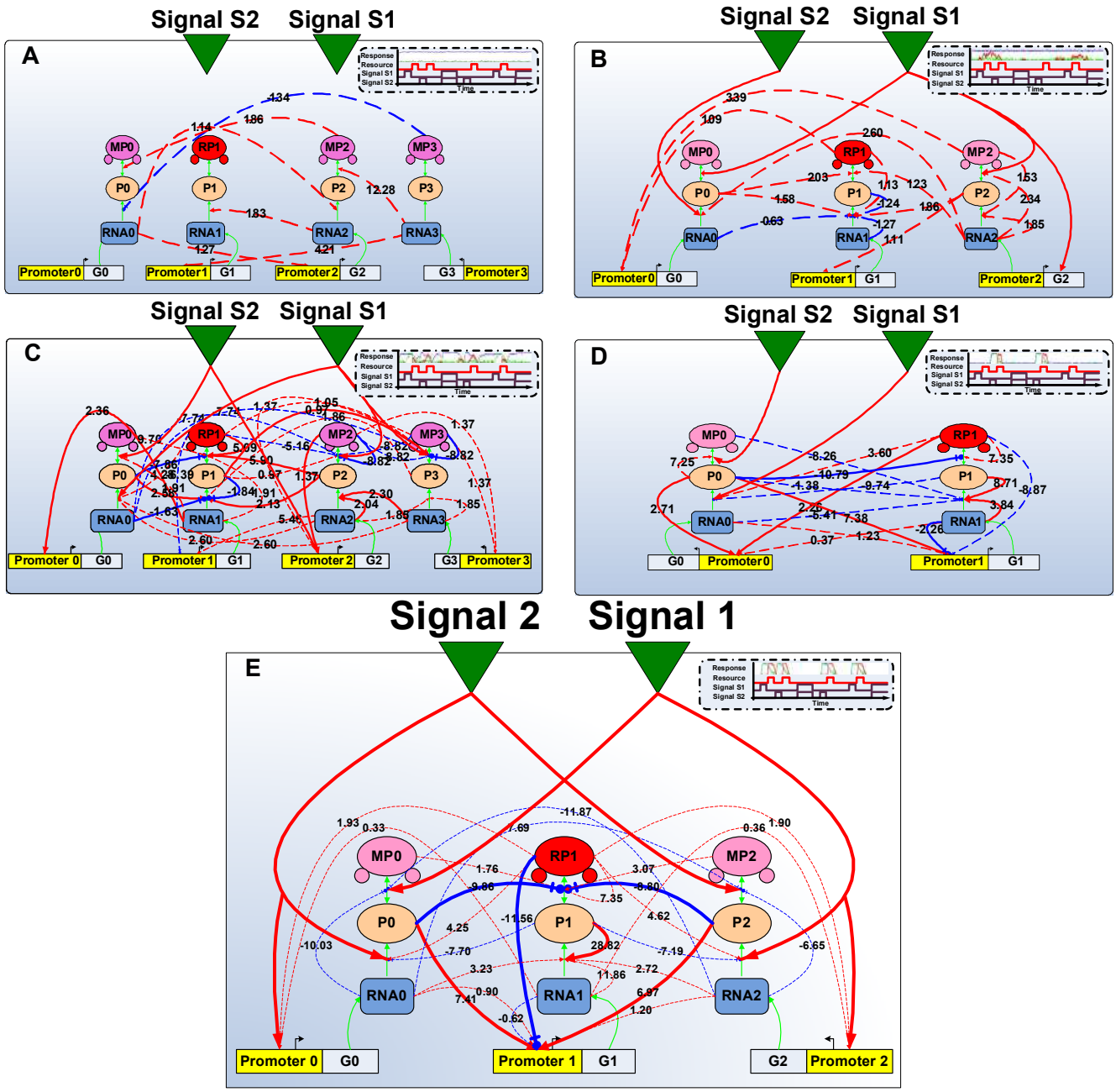


Figure S6 | Network topology of the fittest cell during the evolution of a delayed XOR. The networks portrayed here correspond to the fittest cell at epochs **(1-5)** of [Fig. 3]. Activation and inhibition are represented by red and blue arrows respectively, while solid and dashed lines indicate essential and non-essential links. The phenotypic behavior of each organism is depicted in the upper right corner. **(A, B)** The initially random, low fitness network evolves to a network that partially infers the resource fluctuation in the environment by coupling, although poorly, to signal S1. **(C)** Later, organisms evolve to predict all resource peaks, although their behavior is noisy and not matched well with resource abundance. **(D)** Once a stable phenotype emerges, it takes over the population despite the fact that organisms with higher fitness were present in the population. **(E)** A triplet duplication (triplet 0 to triplet 2) and subsequent optimizing mutations give rise to the final fit organism.

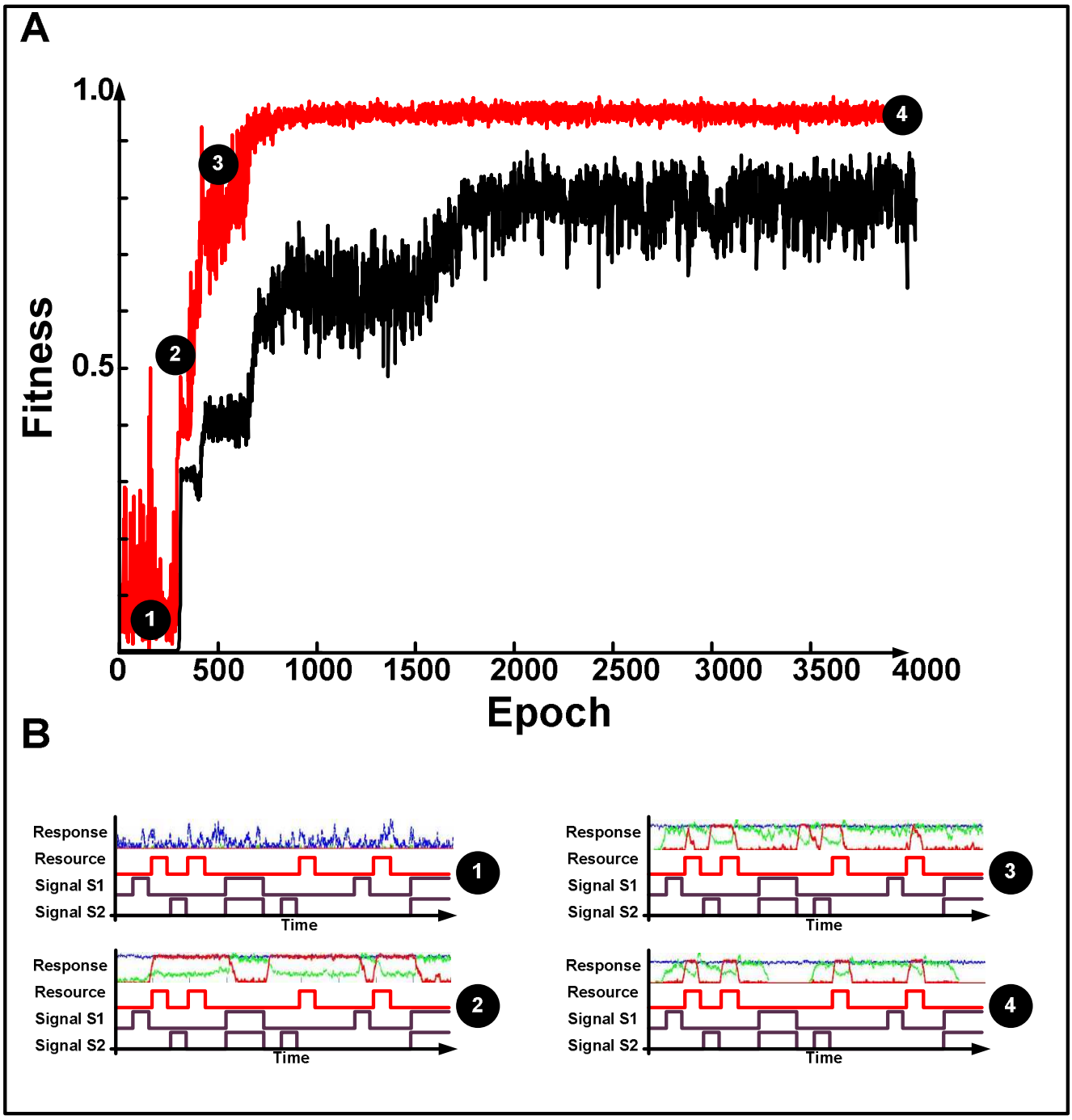


Figure S7 | Fitness trajectory in a delayed XOR experiment. The evolution of a separate XOR experiment under the same conditions as that of [Fig. 3] is depicted. **(A)** Fitness trajectory of a delayed XOR experiment where the presence of any of the two signals alone is necessary and sufficient condition for future resource abundance. Red and black lines correspond to the highest and mean fitness in the population at each epoch respectively. Fitness is defined as the Pearson-correlation between resource abundance and response protein expression. **(B)** The phenotypic behavior of the fittest organism at different points along the evolutionary trajectory. Each subplot consists of four rows: the first row depicts the abundance profiles of the RNA (blue), protein (green), and modified protein (red) of the response pathway. The second, third and fourth rows correspond to the resource abundance, and environmental signals S1 and S2 respectively.

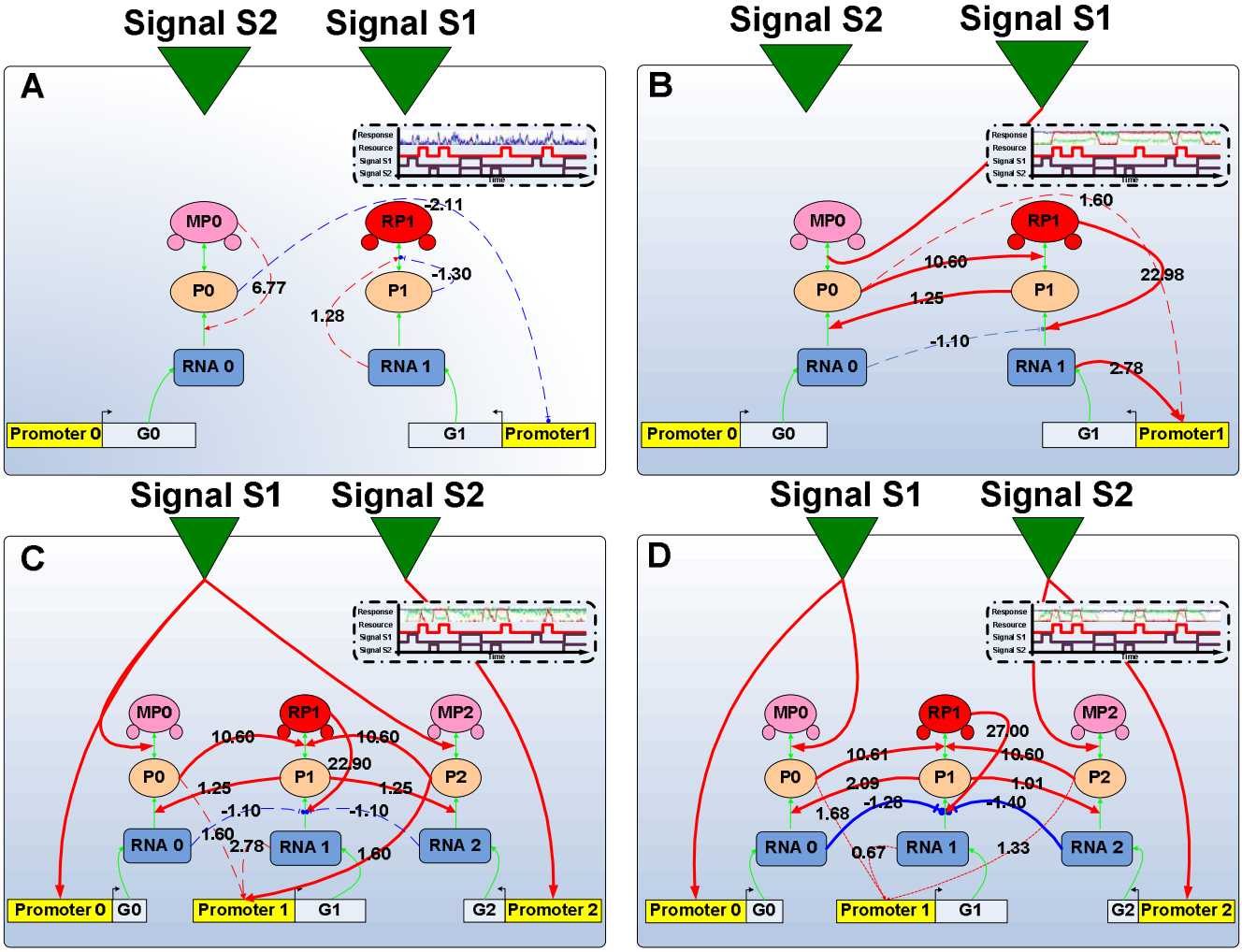


Figure S8 | Network topology of the fittest cell during the evolution of a delayed XOR. The networks portrayed correspond to the fittest cell at epochs **(1-4)** of Fig. S7. Activation and inhibition is represented by red and blue arrows, while solid and dashed lines indicate essential and non-essential links respectively. The phenotypic behavior of these organisms is depicted in the upper right corner. The initially random and unfit network progressively evolves to a network that can accurately predict the resource fluctuation in the environment. **(A)** Initial random network. **(B)** Partial correlation evolved within the first 500 epochs. **(C)** A duplication event of triplet 0 facilitates the evolution of a higher fitness organism. **(D)** Further optimization by less severe mutations and emergence of the final phenotype.

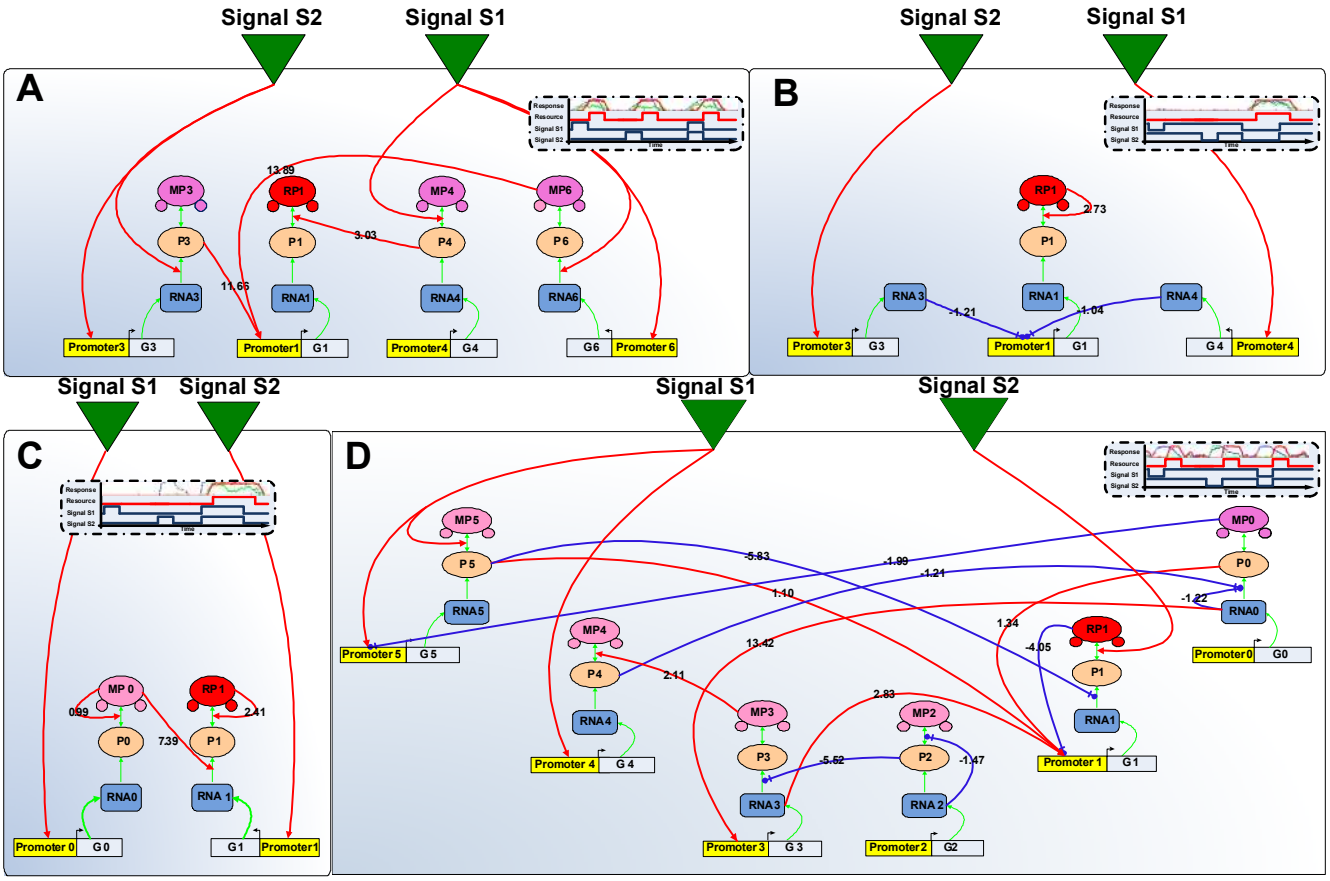


Figure S9 | Minimal networks of delayed dynamic logic gates evolved in environments with low mutation rate. (A) Delayed OR gate where the response pathway is activated through protein P3, modified protein MP6 or both simultaneously. The presence of either signal S1 or S2 is sufficient for the activation of the response pathway. Protein P4 catalyzes the modification of P1 to MP1 and increases the overall fitness of the system (Pearson correlation between MP1 expression and resource presence increases to 0.82 from 0.63) **(B)** Delayed NOR gate where the absence of both signals is necessary for the activation of the response pathway. **(C)** Delayed AND gate where the response protein is expressed only when both signals are present. Signal 2 initiates the transcription of G1, while high MP0 concentration is necessary for the translation of P1. The latter happens only in the presence of signal S1, since it regulates the expression of G0. **(D)** Delayed NAND gate where the response protein is not expressed when both signals are present simultaneously.

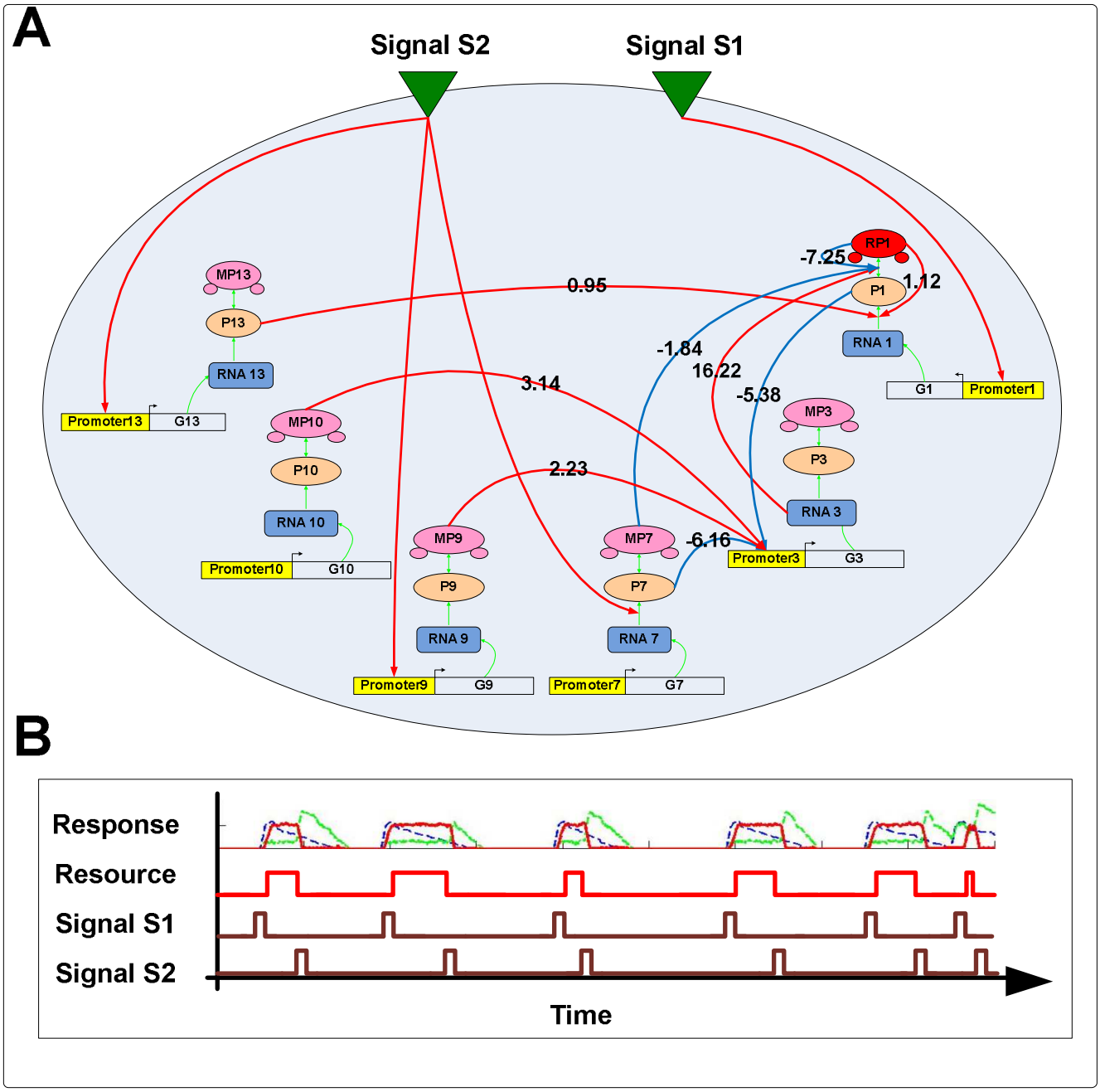


Figure S10 | Minimal network and behavior of an ON/OFF switch. (A) Signal S1 (ON signal) activates the transcription of G1 that leads to the production of RNA1, P1 and RP1. RP1 acts in a positive feedback loop that activates the translation of P1 and locks the creation of RP1 molecules in the absence of Signal S2 (OFF signal). Once a Signal S2 pulse is present, MP7 and P7 directly or indirectly (through inhibition of RNA3, a positive regulator of P1-MP1 modification) inhibit the expression of MP1. **(B)** Signal S1 and Signal S2 pulses indicate the start and the end of resource presence in the environment respectively. Response protein expression (first row, red line) correlates well (>0.9 Pearson Correlation) with resource abundance (second row).

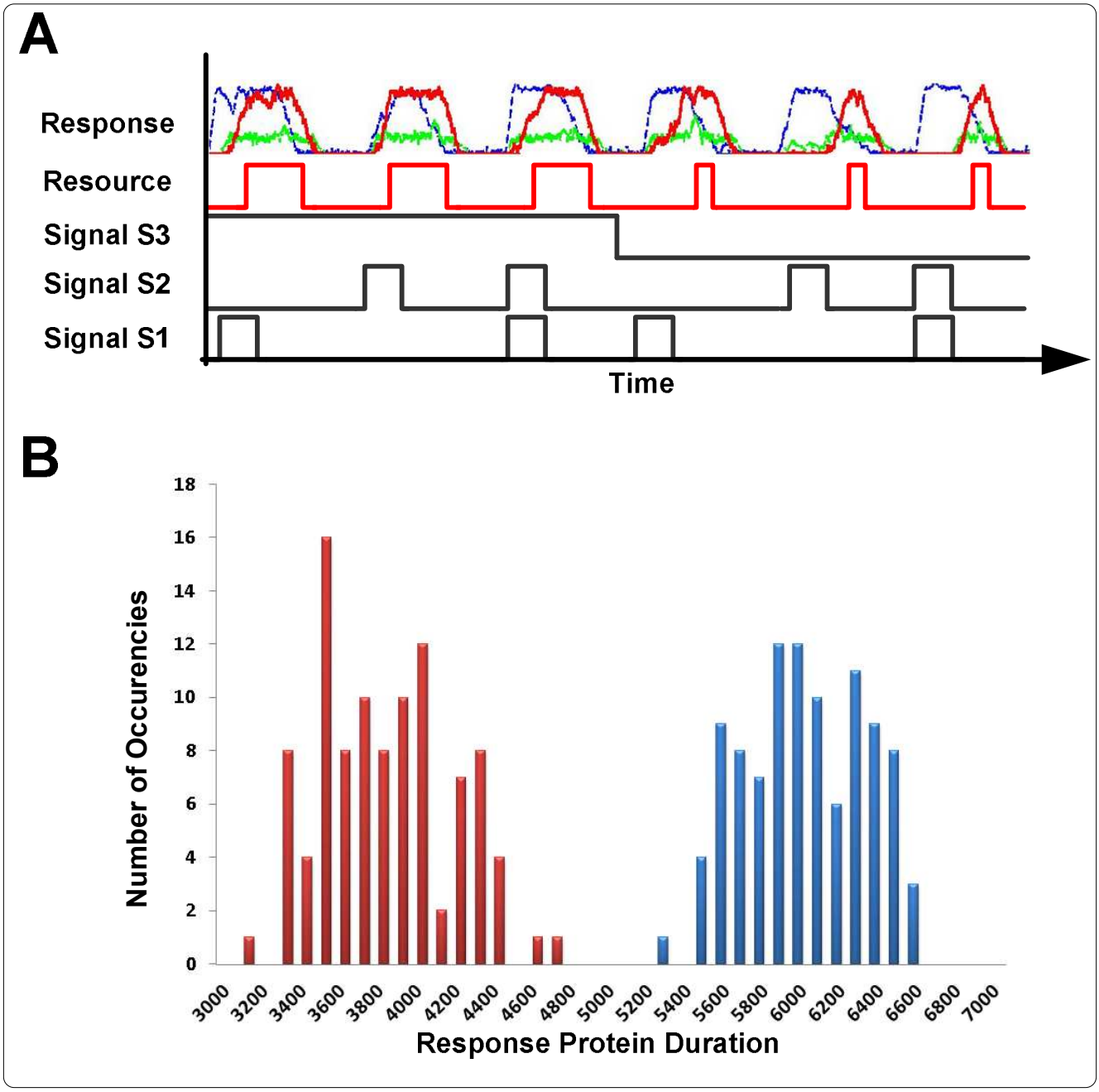


Figure S11 | Evolution of a duration inference mechanism. (A) In an experiment where the duration of the resource presence is variable, organisms have evolved to infer duration variability and highly correlate the expression of response protein with the presence of resource. The duration inference mechanism is based on node coupling to a continuous environmental signal (Signal S3) whose phase is indicative of the environmental variance. **(B)** Distribution over 90 epochs of response protein duration when the duration of resource presence in the environment interchanges between 4000 ± 300 and 6000 ± 300 time units. Response protein expression follows a bimodal distribution with two peaks that correspond to short (red) and long (blue) expression periods. The duration of a response pulse is defined as the time interval between the ascending and descending edge, measured at the 50% of peak expression amplitude.

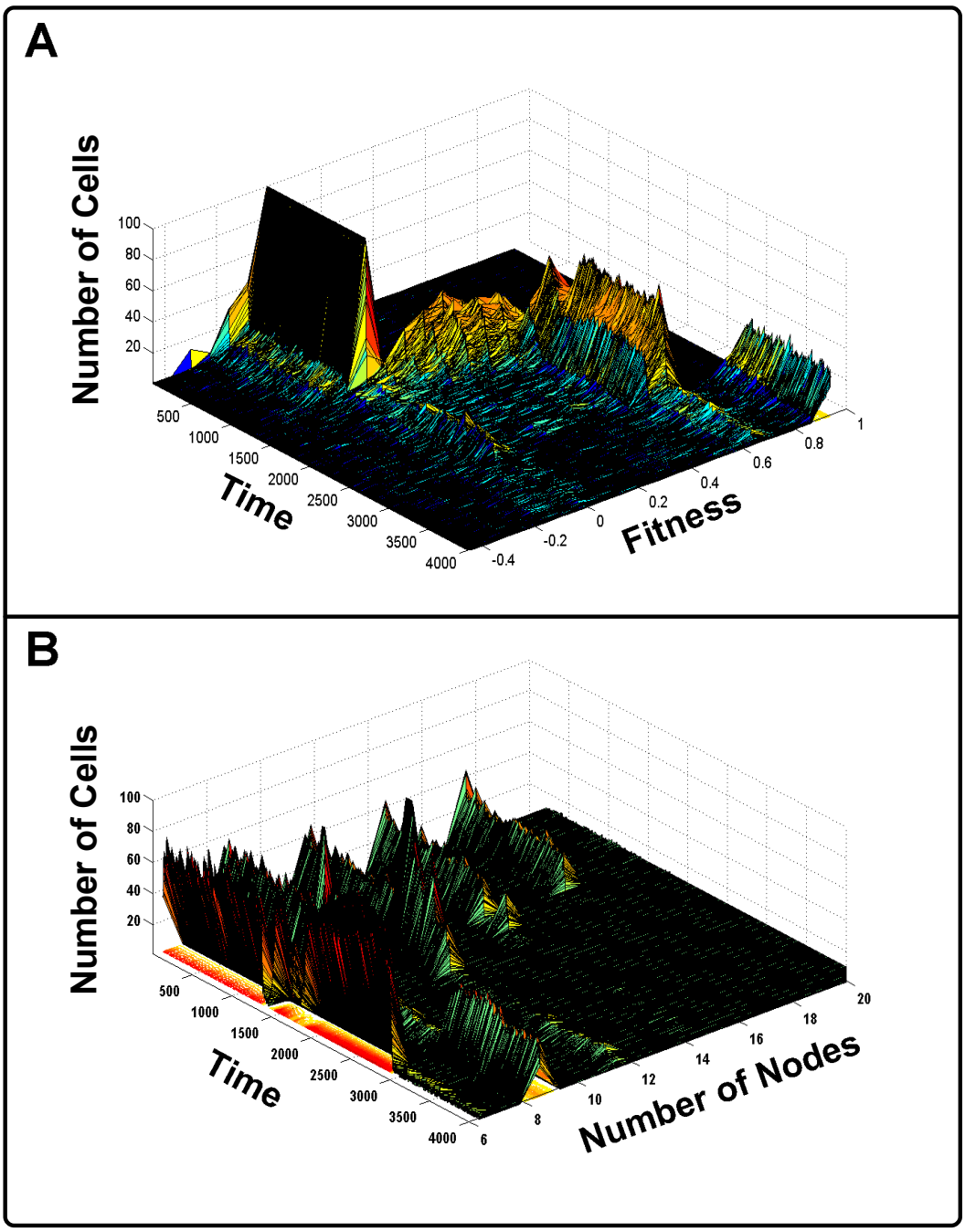


Figure S12 | Fitness and network-size trajectory during the evolution of the delayed XOR network ([Fig. 3]). Each point along the time axis corresponds to an epoch, i.e. 4,500 time units in the experiment. **(A)** A time series depiction of population fitness for all cells (organisms). There are two fitness fixation points at 0.6 and 0.95 that occur around 1200 and 3000 respectively and correspond to the emergence of the one and two signal "mutual exclusive" computational modules, as described in the paper. After the second dominant fixation we observe a decline in the variation of population fitness. **(B)** Node number distribution in the internal non-minimal network of all cells. Nodes are always created and destroyed in triplets (gene, protein and modified protein nodes), hence the discrete nature of the plot. Although the node number variance decreases after the first fixation (where only six nodes comprise the minimal network), there is still triplet duplication events that ultimately lead to the optimal phenotype (which requires a minimum of nine nodes).

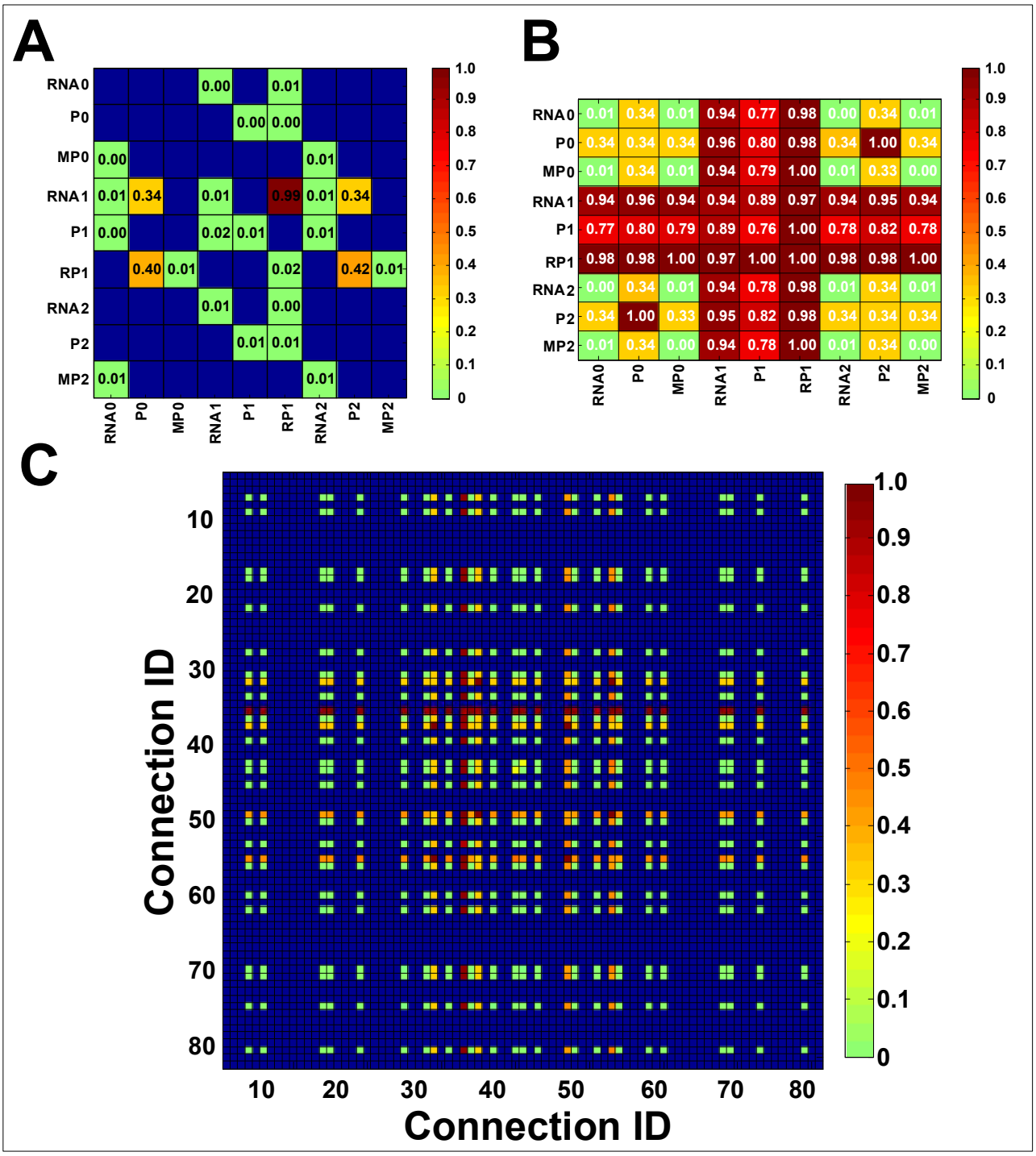


Figure S13 | Network deconstruction in the delayed XOR experiment ([Fig. 3]). (A) Fitness effect of single link knockouts (parallel deconstruction) illustrates link essentiality in the internal network. The fitness impact of links vary from zero (green) to one (dark red) and corresponds to the fitness fraction change that the link severance causes with respect to the fitness of the original network (a 100% fitness change is denoted by 1, 42% by 0.42 and so forth). Matrix elements that depict non-existent links are colored blue. **(B)** Knockouts of node pairs reveal functional similarities between nodes: simultaneous deletion of all links in nodes P0 and P2 reveals strong negative epistasis. **(C)** Epistatic interaction matrix of all 81x81 possible link pair knockouts. Each matrix element shows the fitness effect of a specific link pair knockout. Non-existent link pairs are colored blue.

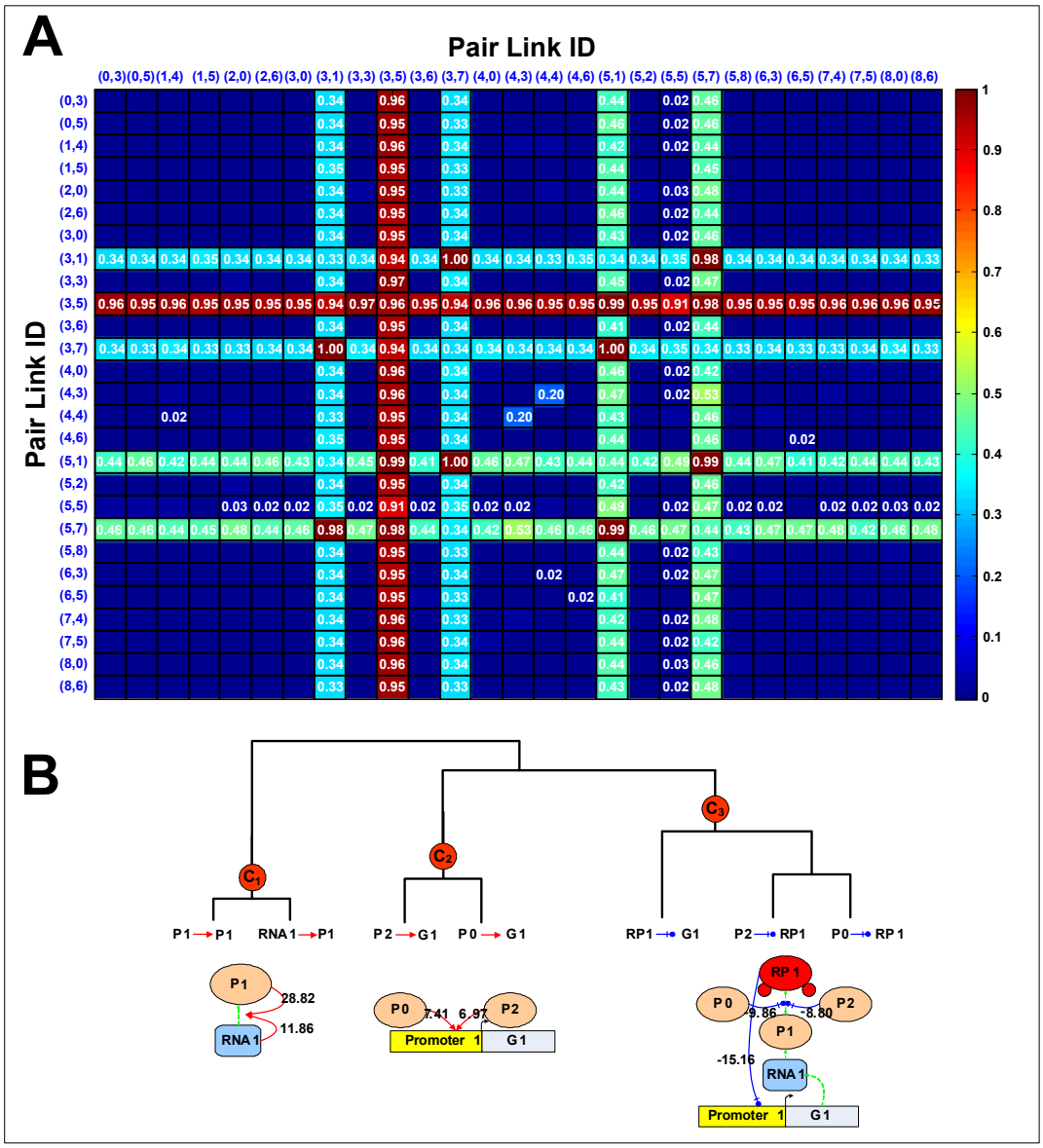


Figure S14 | Epistatic Analysis of a delayed XOR network ([Fig. 4]). (A) Epistatic matrix of existent pair link knockouts derived from the augmented epistatic matrix of Fig. S13c. Only fitness changes more than 1% are shown. Each link pair is represented by the target-regulated and its source-regulator node respectively. Nodes are numbered from 0 to 8 from the lower triplet index to the higher and from RNA to Protein to Modified Protein. For example, matrix row (3,1) denotes the knockout of the direct regulation of G1 by P0. Strong negative epistasis is observed at the severance of both positive regulatory links from P0 and P2 to G1 (link pair (3,1) and (3,7)), both inhibitory interactions between P0, P2 and P1 modification (link pair (5,1) and (5,7)), as well as their crosslink combinations: knockout of P0 (P2) mediated activation of G1 and repression of P1 modification by P2 (P0) – which corresponds to link pairs (3,1)-(5,7) and (3,7)-(5,1) respectively. Positive epistasis occurs during simultaneous deletion of both inhibitory RP1 to G1 and self-activating RP1 to RP1 interactions, during deletion of repression/activation pair P1-RP1 and P1-G1 (link pair (5,1) and (3,1)), and during deletion of P2-RP1 and P2-G1 (link pair (5,7) and (3,7)). **(B)** Monochromatic clustering of epistatic interactions (Segre,D. *et al. Nat. Genet.* **37**, 77-83) reveals the existence of remarkably compact sub-modules that are responsible for keeping P1 expression on after a S1/S2 signal pulse (c_1), response pathway activation (c_2) and delayed behavior of the circuit (c_3).

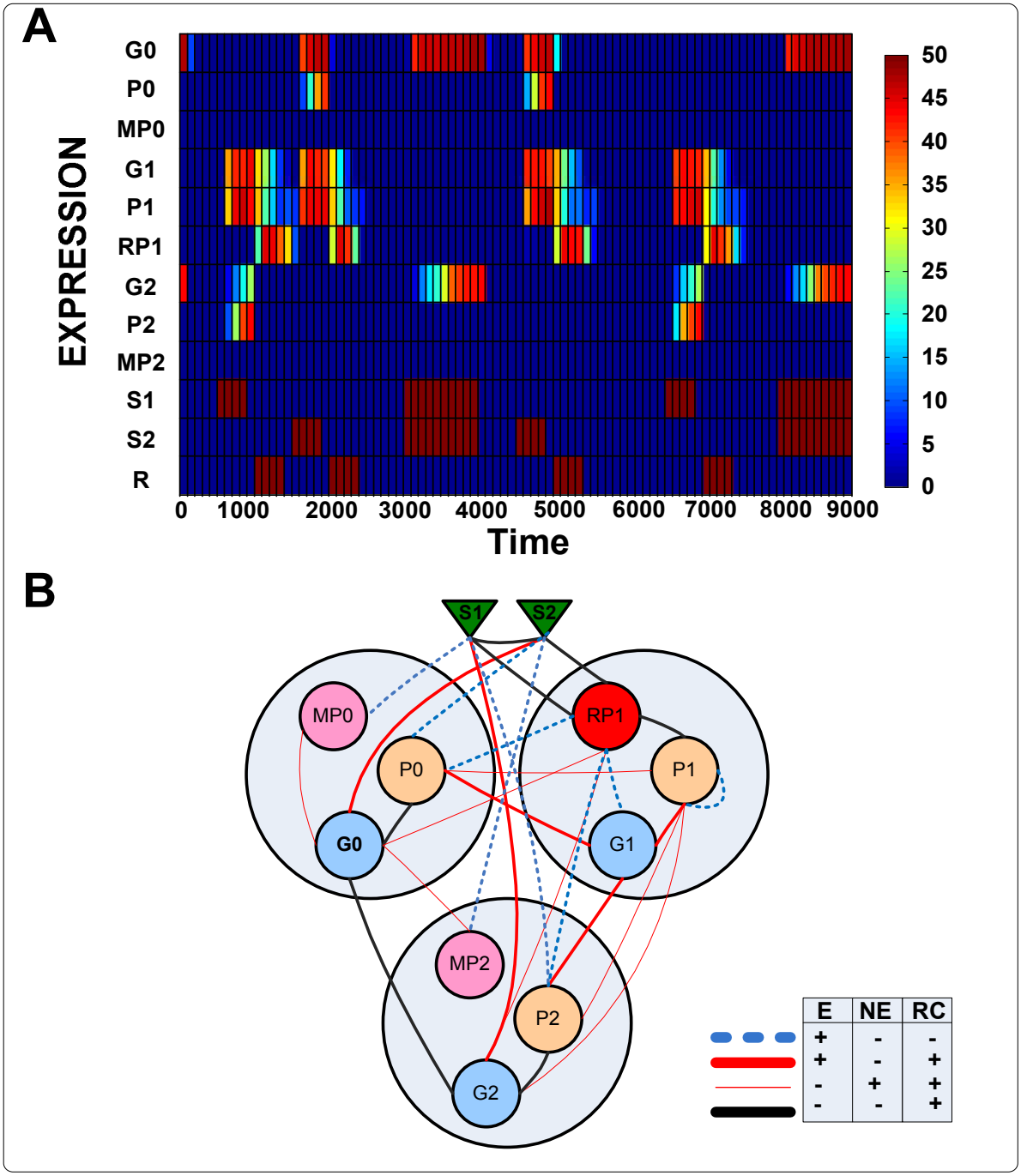


Figure S15 | Microarray analysis and reverse engineering. (A) Depiction of node expression (G0-MP2), signals (S1-S2) and resource (R) abundance of the fittest organism evolved in a delayed XOR environment ([Fig. 4]). The organism successfully predicts resource abundance and expresses RP1 to harvest resources when present. **(B)** Network reconstruction derived from the application of an information-theoretic reconstruction algorithm (ARACNE; Basso,K. *et al. Nat. Genet.* **37**, 382-390) and "microarray" data of network activity for various combinations of exposures to signal S1 and S2. Thick and thin red lines represent essential (E) and non-essential (NE) links that were successfully reconstructed (RC) respectively, while the blue dotted lines map to essential links that were not recovered by the reconstruction algorithm. Black lines indicate links that are predicted in the reconstructed network but which do not exist in the original network.

	Number of Links in Original, Minimal, Reconstructed Networks			Link Percentage i.r.t. Original Network		Reconstructed Statistics i.r.t Original Network				Reconstructed Statistics i.r.t Minimal Network				Reconstructed Accuracy i.r.t. Original Network			Reconstructed Accuracy i.r.t. Minimal Network		
	ORIG.	MIN.	REC.	MIN.	REC.	TP	FP	TN	FN	TP	FP	TN	FN	Precision	Sensitivity	Specificity	Precision	Sensitivity	Specificity
ANDlm	27	5	16	18.50%	59.30%	10	6	88	17	4	12	104	1	62.50%	37.00%	93.60%	25.00%	80.00%	89.70%
ANDmm	31	5	21	16.10%	67.70%	9	12	153	22	4	17	174	1	42.90%	29.00%	92.70%	19.00%	80.00%	91.10%
ANDhm	129	4	60	3.10%	46.50%	18	42	853	111	3	57	963	1	30.00%	14.00%	95.30%	5.00%	75.00%	94.40%
ORlm	24	11	14	45.80%	58.30%	9	5	92	15	6	8	102	5	64.30%	37.50%	94.80%	42.90%	54.50%	92.70%
ORmm	27	11	13	40.70%	48.10%	9	4	90	18	4	9	101	7	69.20%	33.30%	95.70%	30.80%	36.40%	91.80%
ORhm	132	9	50	6.82%	37.90%	24	26	866	108	4	46	969	5	48.00%	18.20%	97.10%	8.00%	44.40%	95.50%
XORlm	77	15	32	19.50%	41.60%	26	6	206	51	6	26	248	9	81.30%	33.80%	97.20%	18.80%	40.00%	90.50%
XORmm	180	19	55	10.60%	30.60%	15	40	1005	165	5	50	1156	14	27.30%	8.33%	96.20%	9.09%	26.30%	95.90%
XORhm	335	14	85	4.18%	25.40%	27	58	1816	308	5	80	2115	9	31.80%	8.06%	96.90%	5.88%	35.70%	96.40%
NORlm	33	12	20	36.40%	60.60%	6	14	149	27	3	17	167	9	30.00%	18.20%	91.40%	15.00%	25.00%	90.80%
NORmm	41	6	20	14.60%	48.80%	9	11	144	32	5	15	175	1	45.00%	22.00%	92.90%	25.00%	83.30%	92.10%
NORhm	117	2	37	1.71%	31.60%	10	27	880	107	2	35	987	0	27.00%	8.55%	97.00%	5.41%	100.00%	96.60%
NANDlm	26	12	23	46.20%	88.50%	7	16	154	19	6	17	167	6	30.40%	26.90%	90.60%	26.10%	50.00%	90.80%
NANDmm	42	8	22	19.00%	52.40%	9	13	141	33	2	20	168	6	40.90%	21.40%	91.60%	9.09%	25.00%	89.40%

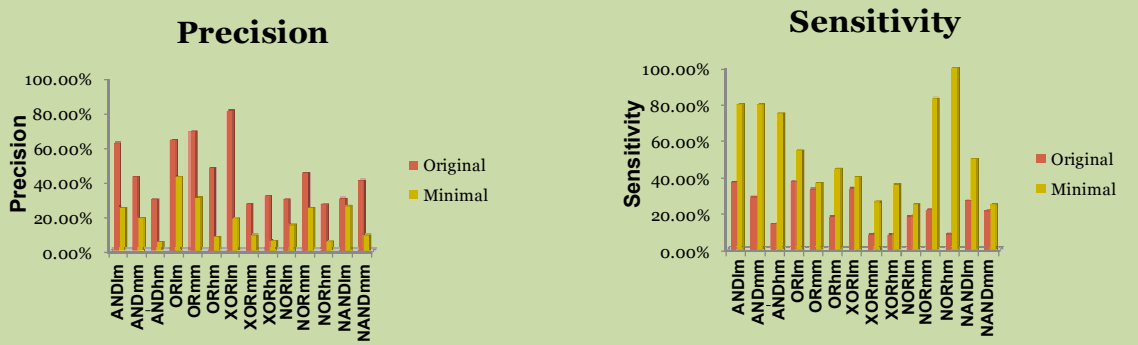


Figure S16 | Statistics of reconstructed networks using a mutual information-based algorithm (ARACNE). We report the total number of links, link percentage, true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), precision ($\frac{TP}{TP+FP}$), sensitivity ($\frac{TP}{TP+FN}$) and specificity ($\frac{TN}{TN+FP}$) of the reconstructed network compared both to the original and minimal network of each organism. The two plots indicate the precision and sensitivity of reconstruction with respect to the original and minimal network. Interestingly, the precision of reconstruction is higher for the original network while the sensitivity of reconstruction is higher for the minimal network. The apparent asymmetry stems from the fact that all links in the minimal network are essential and in that respect they have to be present in the reconstructed network for it to fully exhibit its phenotype.

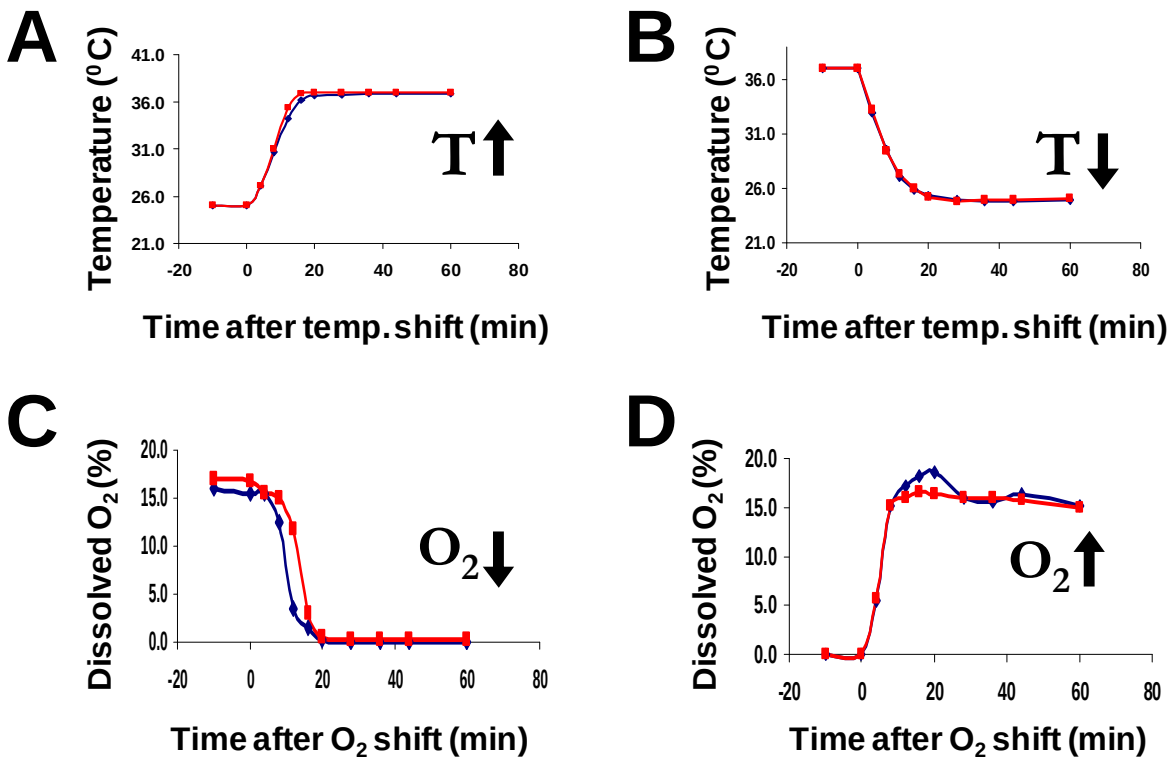


Figure S17. Temperature and oxygen perturbation profiles are highly reproducible. Temperature and oxygen levels were precisely monitored and controlled using a New Brunswick BioFlo 110 bioreactor (New Brunswick Scientific). All experiments were performed in duplicate with high reproducibility (red vs. blue). **(A)** Temperature is shifted from 25^0 C to 37^0 C within a 20 minute interval while the culture is maximally aerated (18% dissolved O_2). **(B)** Temperature is shifted from 37^0 C to 25^0 C within a 20 minute interval while the culture is maximally aerated (18% dissolved O_2). **(C)** Oxygen saturation is down-shifted within a 20 minute interval while temperature is maintained at 25^0 C. (**D)** Oxygen saturation is up-shifted within a 20 minute interval while temperature is maintained at 25^0 C.

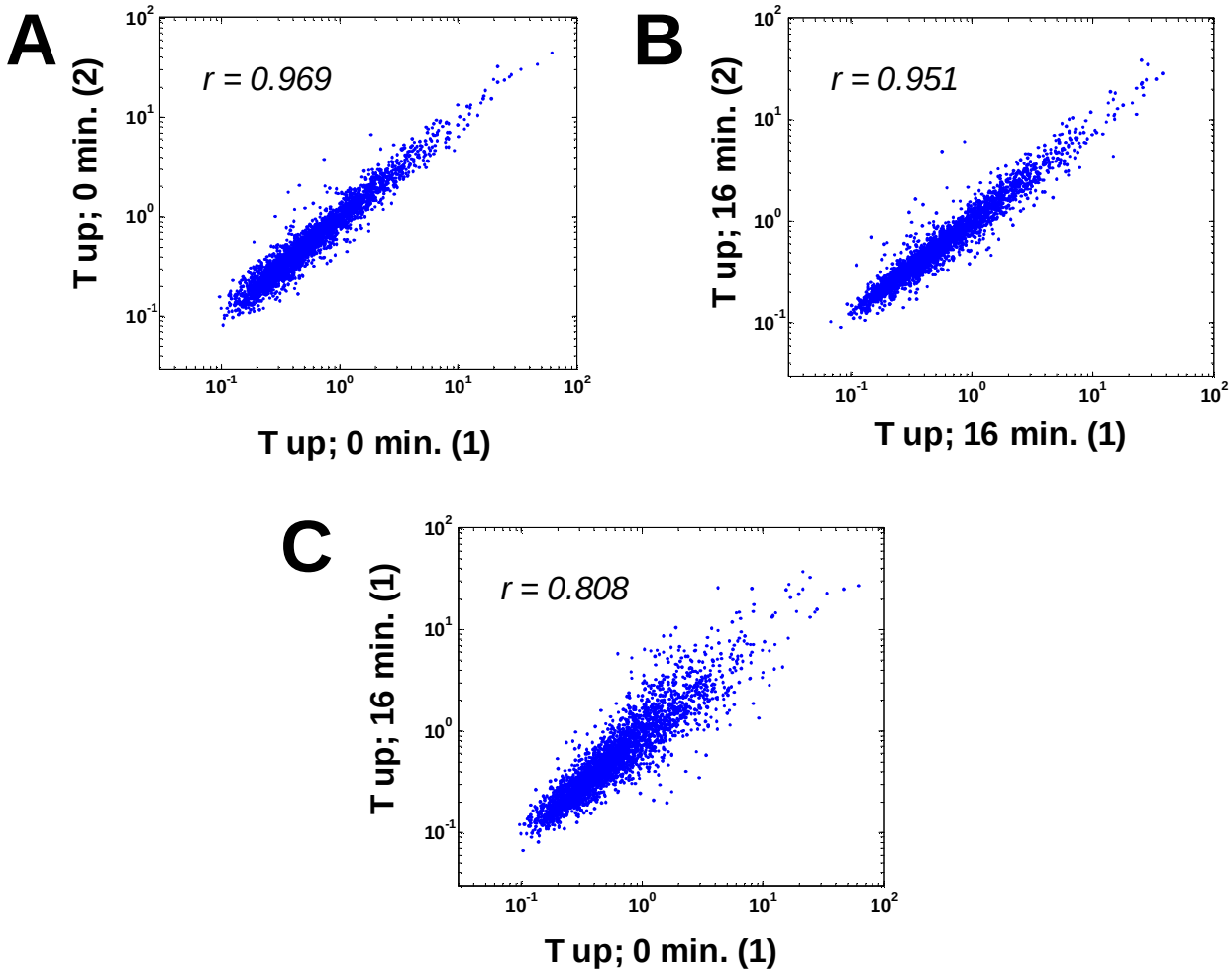


Figure S18. Comparison of microarray expression profiles. High reproducibility of independent experimental replicates: **(A)** temperature up perturbation, 0 minutes, **(B)** temperature up perturbation, 16 minutes. **(C)** Lower correlation reflecting differential expression between 0 and 16 minutes post temperature up transition.

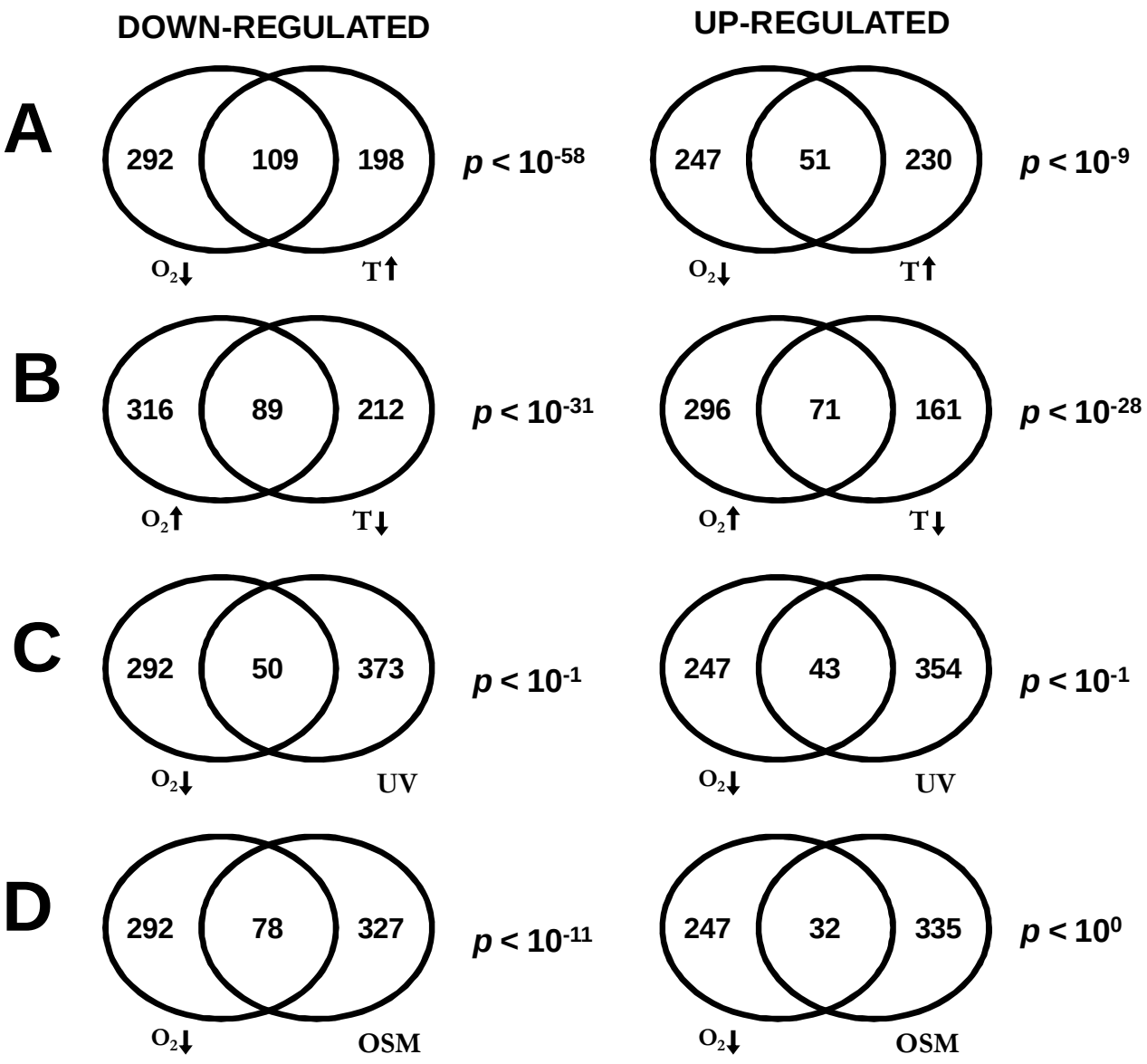


Figure S19. Transcriptional responses to ecologically correlated perturbations in temperature and oxygen show statistically significant overlap. Sets of genes that are differentially expressed by a factor of 1.5 fold show highly significant overlap when temperature and oxygen perturbations correspond to ecologically correlated changes accompanying transitions between the outside environment and the GI-tract (**A, B**). Significantly lower overlap is seen between the oxygen down-shift response and UV response **(C)**, and response to hyper-osmolarity stress **(D).**

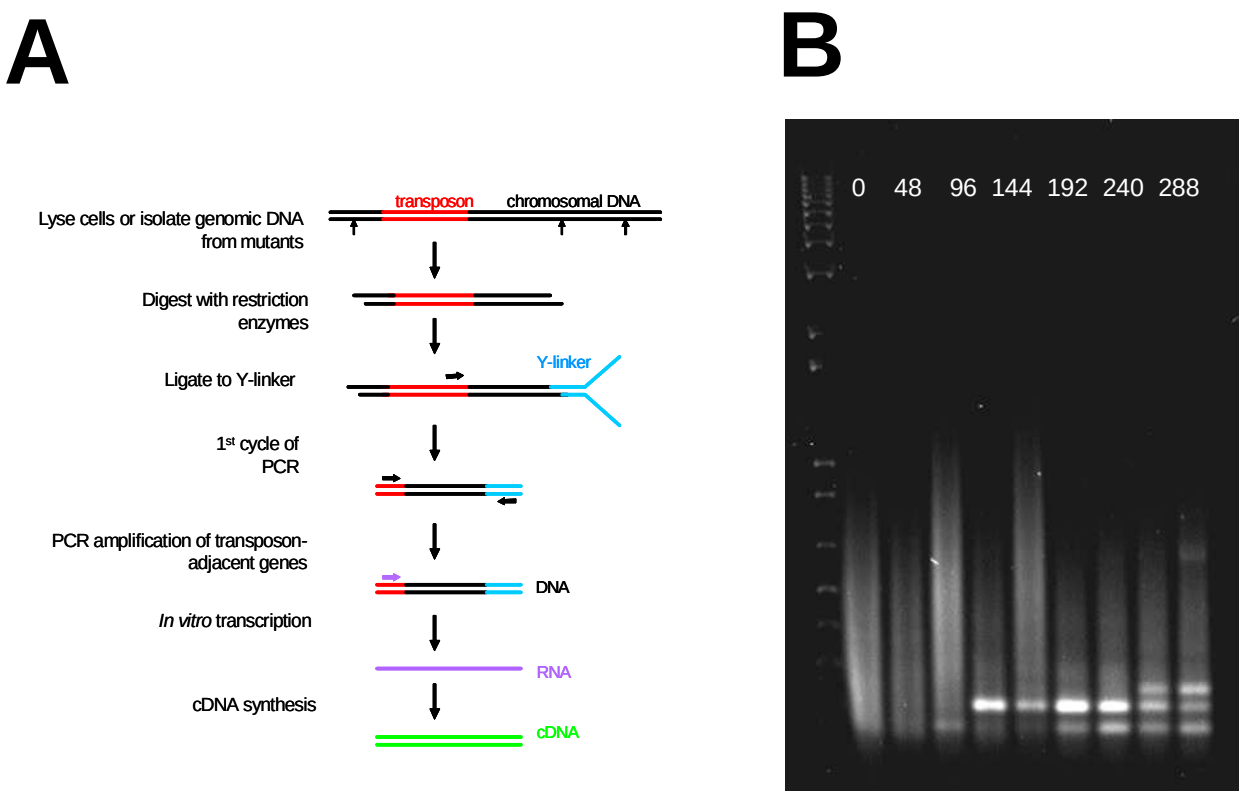


Figure S20. Parallel amplification of transposon-adjacent sequences and monitoring population diversity. (A) Genomic DNA extracted from a population of transposon mutants exposed to a selection is restriction digested, and the resulting fragments are ligated to a partially double-stranded Y-linker (Girgis *et al. PLoS Genetics* 3(9): e154 (2007)). In the initial PCR cycle, primer annealing to a sequence within the transposon and DNA polymerase mediated extension generates a double-stranded DNA template. In subsequent PCR cycles, primers annealing to a region within the transposon and the top strand complement of the Y-linker allows exponential amplification of transposon-adjacent sequences. An *in vitro* transcription reaction using the PCR products generates RNA which, in the presence of reverse transcriptase and modified nucleotides, generates fluorescently labeled cDNA. The fluorescently labeled cDNA can then used in subsequent microarray hybridizations in order to get a genome-wide picture of transposon insertions. **(B)** The DNA products resulting from the PCR amplification of transposon adjacent sequences can be easily visualized on an ethidium-bromide gel, providing a qualitative picture of population diversity throughout the evolution experiment. At the beginning, the population has maximal diversity (smear at time zero). Distinct bands correspond to individual mutants that become increasingly abundant as a function of selection.

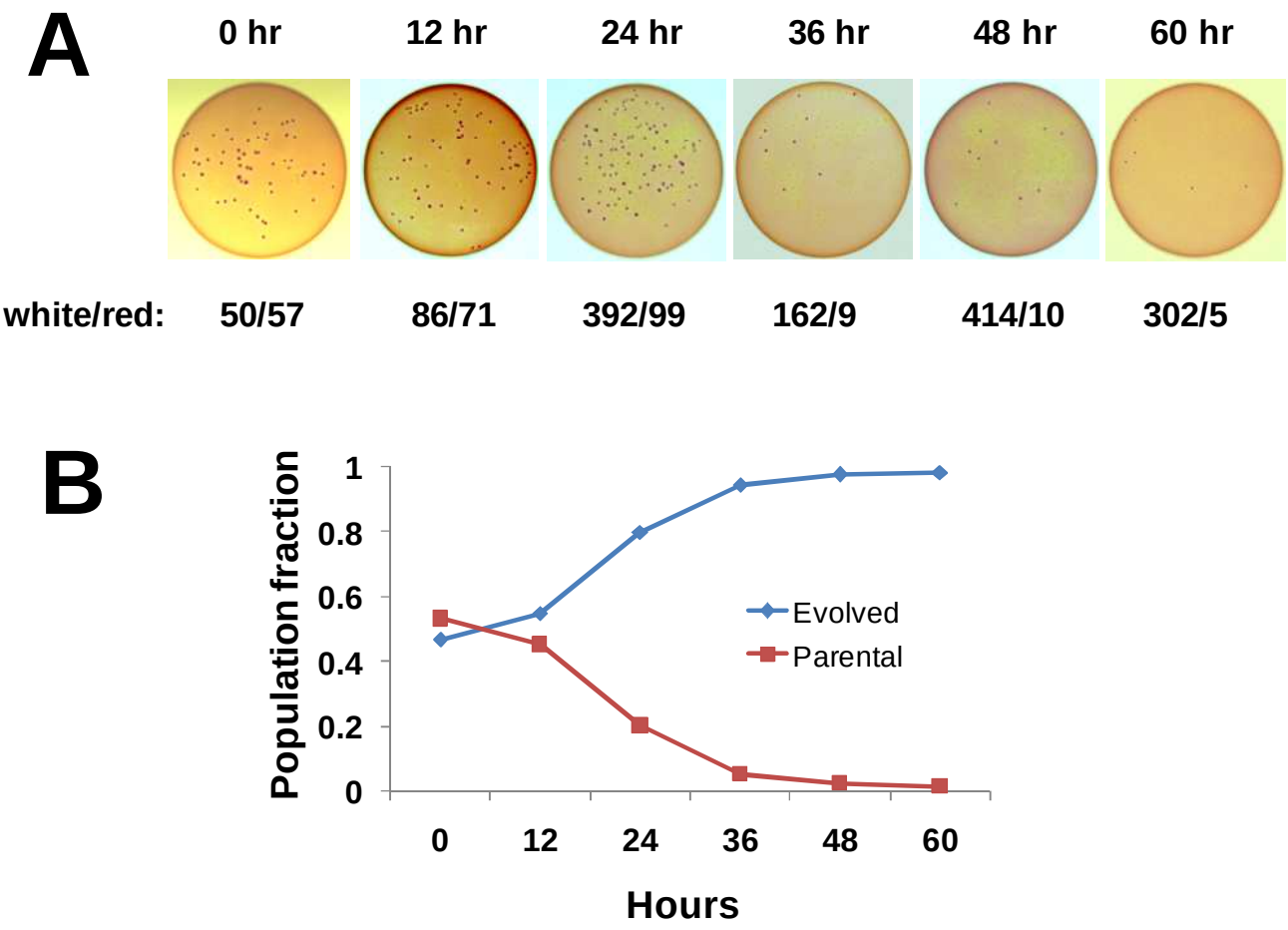


Figure S21. Competition experiments between the parental and the evolved strain. (A) Differential expression of a LacZ marker allows monitoring of relative growth in mixed competition experiments. **(B)** The evolved strain rapidly outcompetes the parental over the course of 60 hours of growth under ecologically incoherent correlations between temperature and oxygen.

	VLM	LM	MM	HM	Total
AND	**14/64**	**17/64**	**40/64**	**22/64**	**93/256 (36.3%)**
OR	**0/64**	**7/64**	**18/64**	**0/64**	**25/256 (9.8%)**
XOR	**0/64**	**4/64**	**13/64**	**0/64**	**17/256 (6.6%)**
NOR	**3/64**	**25/64**	**39/64**	**23/64**	**90/256 (35.2%)**
NAND	**0/64**	**4/64**	**5/64**	**11/64**	**20/256 (7.8%)**
Total	**17/320 (5.3%)**	**57/320 (17.8%)**	**115/320 (35.9%)**	**56/320 (17.5%)**	

Supplementary Table 1 | Success-rate of five delayed dynamic logic gate experiments across four different mutation rates. All experiments were conducted for 4000 epochs ($1.8 \cdot 10^7$ time units). An experiment is considered successful if the input-output characteristics of the fittest organism are representative of the gate we select for. Operationally this translates to having the Pearson Correlation between the resource abundance and the response protein expression higher than 0.7. From left to right the columns represent environments that have very low mutation (VLM, triplet creation probability 10^{-8}), low mutation (LM, triplet creation probability 10^{-7}), medium mutation (MM, triplet creation probability 10^{-6}), high mutation (HM, triplet creation probability 10^{-5}) rate. In each table entry the first value corresponds to the number of successful experiments and the second to the total number of experiments conducted. Other experiments with lower success rate included oscillators with (3/256) and without (1/512) synchronization signal, bistable switches (2/64) and OR-XOR multi-gates (1/128).

References

S1. P. Francois and V. Hakim, Proc.Natl.Acad.Sci.U.S.A 101, 580-585 (2004).

S2. P. Francois, V. Hakim, E. D. Siggia, Mol.Syst.Biol. 3, 154 (2007).

S3. Rodrigo G., Carrera J., A. Jaramillo, CEJB 2, 233-253 (2007).

S4. O. S. Soyer, T. Pfeiffer, S. Bonhoeffer, J.Theor.Biol. 241, 223-232 (2006).

S5. A. Wagner, Proc.Natl.Acad.Sci.U.S.A 102, 11775-11780 (2005).

S6. K. A. Fichthorn and W. H. Weinberg, J.Chem.phys. 95, 1090-1096 (1991).

S7. D. T. Gillespie, J.Phys.Chem. 81, 2340-2361 (1977).

S8. K. Basso et al., Nat.Genet. 37, 382-390 (2005).

S9. H. S. Girgis, Y. Liu, W. S. Ryu, S. Tavazoie, PLoS.Genet. 3, e154 (2007).

S10. J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, P. C. Hanawalt, Genetics 158, 41-64 (2001).

S11. K. J. Cheung, V. Badarinarayana, D. W. Selinger, D. Janse, G. M. Church, Genome Res. 13, 206-215 (2003).

S12. S. F. Elena and R. E. Lenski, Nat.Rev.Genet. 4, 457-469 (2003).