Design principles in two component systems and his-asp phosphorelays

Salvadó Baldiri López

Universitat de Lleida

Design Principles in Two Component Systems and His-Asp Phosphorelays

BALDIRI SALVADÓ i LÓPEZ

Supervisor: Rui Alves
Grup de Bioestadística i Biomatemàtica
Departament de Ciències Mèdiques Bàsiques
Universitat de Lleida & IRB Lleida

Acknowledgements

It is not easy to accept a collaborator as myself: with a lot of enthusiasm, but very little time. I have to thank Rui Alves for believing that someone under my circumstances could take charge of a project and finish it on time. Thank you for your help, for your excellent job as a supervisor, but most of all, thanks for the opportunity to work in this fascinating area of research.

All the members of the research group have been really friendly from my first day. Too bad that I usually arrive at the lab when everybody is leaving it. But, in any case, whenever we meet, Hiren, Anabel, Carles, Verónica, Rui Benfeitas, Montse Martínez, Montse Rué, Ester, Jorge, it is always a pleasure to work with all of you, and to share a casual conversation and/or a visit to the bar during a short break. Everything is easier in pleasant company.

Vull agrair a la Gisela el seu recolzament. No hauria pogut acabar aquesta tesi sense la seva complicitat. De fet, no l'hauria pogut començar. Ha hagut de patir les meves explicacions sobre detalls del treball, m'ha donat opinió i consell, m'ha fet d'assessora artística, … Gràcies per acompanyar-me en aquesta aventura.

Abbreviations

BST: biochemical systems theory

HK: histidine kinase

HPt: histidine phosphotransferase

ODE: ordinary differential equations

PR: histidine-aspartate phosphorelay

RR: response regulator

SK: sensor histidine kinase

TCS: two component system

Summaries

English

The ultimate goal of this thesis is to set the stage for finding general design principles underlying the relationship between network design and network function in two-component (TCS) and His-Asp phosphorelay (PR) signal transduction systems. Design principles are important because i) they can explain the evolution of a particular biological feature, and ii) understanding structure-function relation in molecular systems enables many biotechnological applications.

This thesis starts with a review of the methods for and results from the study of design principles in molecular systems. This review also discusses the importance of studying those design principles.

Next, we home in on TCS and PR, the prevalent signal transduction circuits in bacteria. These circuits are also present in eukaryotes (but not in animals). Their modular structure allowed evolution to generate a great diversity of unique circuit designs for the internal signal transmission within the cascade. Identifying the existing unique designs allows us to set the stage for a systematic comparison of the dynamic responses that are exclusive to each design. This identification is done by surveying the fully sequenced and annotated genomes and proteomes of more than 7000 different organisms. In this survey we identify the operon and protein domain organization of proteins involved in TCS and PR. From this organization we infer the unique topologies of TCS and PR circuits. We find that there is a positive selection for the clustering of HK, RR and HPt domains, both in operons and in multidomain TCS/PR signaling proteins. Regarding the TCS/PR operon composition, we find 530 different combinations of HK, RR and HPt coding genes grouped in operons in the surveyed genomes. As for the TCS/PR gene fusion events, we find 50 unique combinations of

HK, RR and HPt domains that occur in a single polypeptide chain of the surveyed organisms.

Finally, we compare dynamic properties associated with three types of TCS circuit designs through mathematical modeling and analysis of the alternative circuits to be compared, within the framework of Biochemical Systems Theory. The first design is a canonical TCS. The second and third designs are TCS which include an additional protein that can interact either with the sensor kinase preventing its phosphorylation, or with the phosphorylated RR protecting it from dephosphorylation, respectively. We find that the possibility of bistability in the response of the TCS module is decreased by a RR binding third component, but an HK binding third component can either increase or decrease the parameter space of bistability of the network, depending on the monofunctionality or bifunctionality of the HK.

Overall, this thesis represents an example of how bioinformatics and computational biology can be combined to play an important role in molecular systems biology, enabling the systematic characterization of circuit designs and the study of the unique dynamic responses associated to each design.

Català

L’objectiu d’aquesta tesi és trobar principis generals que permetin entendre com l’estructura d’una xarxa molecular de transducció de senyals afecta les seves propietats funcionals. Els principis de disseny són importants perquè i) expliquen l’evolució d’un determinat caràcter biològic, i ii) la comprensió de la relació estructura-funció en sistemes moleculars permet multitud d’aplicacions biotecnològiques.

La tesi s’inicia revisant els mètodes usats per a l’estudi de principis de disseny en sistemes moleculars i alguns dels resultats obtinguts fins ara, i discutint la importància de l’estudi dels principis de disseny.

A continuació ens centrem en els sistemes de transducció de senyals coneguts com *two-component systems* (TCS) i histidina-aspartat *phosphorelays* (PR), predominants en bacteris i també presents en eucariotes no animals. La naturalesa modular d’aquests sistemes moleculars ha facilitat que evolucionin generant un gran nombre de variacions en l’estructura del circuit de transmissió de senyals. La identificació dels diferents dissenys del circuit és el primer pas per a establir les característiques funcionals associades a cada disseny. Per fer aquesta identificació, explorem els proteomes seqüenciats de més de 7000 organismes i fem un inventari dels diferents tipus d’organització en operons i proteïnes dels dominis proteics que intervenen en TCS i PR. A partir d’aquesta informació deduirem alternatives existents en la natura pel que fa al disseny d’aquests circuits moleculars. En aquesta exploració genòmica i proteòmica observem que la selecció natural afavoreix l’agrupament dels dominis proteics implicats (HK, RR i HPt) tant en operons com en proteïnes que contenen diversos dominis. Pel que fa a la composició dels operons, hem trobat 530 tipus diferents de combinacions gèniques en els proteomes explorats. Quant a la fusió

de gens que codifiquen HK, RR o HPt, trobem 50 combinacions diferents d’aquests dominis en les proteïnes dels organismes explorats.

Per acabar, comparem el comportament dinàmic de 3 circuits diferents de TCS mitjançant la modelització matemàtica dels sistemes a comparar, emprant les eines i conceptes aportats per la teoria de sistemes bioquímics. Comparem les respostes d’un TCS canònic amb les d’un TCS on una proteïna addicional interacciona amb la HK evitant la seva fosforilació, o bé amb el RR impedint la seva desfosforilació. Observem que l’espai de valors paramètrics on el sistema presenta biestabilitat es redueix amb la presència d’un tercer component que inhibeix la defosforilació del RR. En canvi, si el tercer component interacciona amb la HK, l’espai de biestabilitat pot ser ampliat o reduït, depenent de si la HK és monofuncional o bifuncional.

Aquesta tesi és, per tant, un exemple de com biologia, informàtica i matemàtiques poden combinar-se en l’àmbit de la biologia de sistemes moleculars per la caracterització de les respostes específicament associades a cada disseny d’un circuit molecular.

Castellano

El objectivo principal de esta tesis es la búsqueda de principios de diseño que relacionen la estructura y la función de redes bioquímicas de transducción de señales, concretamente en *two-component systems* (TCS) y *phosphorelays* (PR). Los principios de diseño nos interesan ya que i) pueden explicar la evolución de un determinado carácter biológico, y ii) el conocimiento de la relación entre estructura y función en sistemas moleculares tiene multitud de aplicaciones biotecnológicas.

La tesis se inicia con una revisión de los métodos usados para el estudio de principios de diseño en sistemas moleculares y algunos de los resultados obtenidos hasta ahora, seguida de una discusión sobre la importancia del estudio de dichos principios de diseño.

A continuación centramos nuestro estudio en TCS y PR, vías de transducción de señal dominantes en bacterias y también presentes en eucariotas no animales. La estructura modular de estas redes moleculares ha permitido que evolucionen dando lugar a una gran variedad de diseños de circuitos transmisores de señales. Identificar esta variedad de diseños nos permite plantear comparaciones entre las propiedades funcionales asociadas a cada diseño. Exploramos los proteomas secuenciados de más de 7000 organismos y hacemos un inventario de los distintos tipos de organización en operones o proteínas de los dominios proteicos implicados en TCS y PR (HK, RR y HPt). A partir de este inventario, trataremos de deducir el repertorio de estructuras existentes en la naturaleza para estos circuitos moleculares. En esta exploración genómica y proteómica observamos que la selección natural provoca la agrupación de los dominios proteicos HK, RR y HPt tanto en operones como en proteínas. En cuanto a la composición de los operones, encontramos 530 combinaciones diferentes de genes

en los operones de TCS/PR de los proteomas explorados. En relación con la fusión de genes que codifican HK, RR o HPt, encontramos 50 combinaciones diferentes de estos dominios en las proteínas de los organismos explorados.

Para terminar, comparamos las propiedades dinámicas de tres circuitos distintos de TCS, mediante modelización matemática en el marco de la teoría de sistemas bioquímicos. Comparamos las respuestas de un TCS canónico con las de un TCS en el cual una proteína adicional se une a la HK inhibiendo su fosforilación, o bien se une al RR inhibiendo su defosforilación. Observamos que el espacio paramétrico de biestabilidad del sistema queda reducido por la presencia de un tercer componente que se une al RR. En cambio, si el tercer componente se une a la HK, el espacio de biestabilidad puede ser ampliado o reducido, dependiendo de si la HK es monofuncional o bifuncional.

Por tanto, esta tesis es un ejemplo de cómo biología, informática y matemáticas pueden combinarse en el campo de la biología de sistemas moleculares para caracterizar las respuestas asociadas a distintos diseños de circuitos moleculares.

Contents

1 Introduction

This thesis aims at setting the stage for a systematic understanding of how specific circuit designs associate with particular dynamic responses in a class of signal transduction biochemical pathways referred to as Two Component Systems (TCS) and Phosphorelays (PR). This will allow the methodical identification of design principles in this class of circuits.

This introduction will present the state of the art and frame the work in the area of molecular systems biology. Sections 1.1 and 1.2 briefly discuss what this discipline encompasses today and how it evolved historically. Then, section 1.3 presents and discusses the mathematical approximations that make the work we do possible. Subsequently, section 1.4 exposes the notion of biological design principles in molecular systems biology and discusses its importance. After this, section 1.5 briefly describes the biological systems in which I focus, TCS and PR. This is followed in section 1.6 by a short presentation of the methods used in the thesis. Section 1.7 describes the organization of the remainder of this thesis. Finally, section 1.8 presents the goals of the thesis.

1.1. Molecular systems biology

A system can be defined as a network of interacting elements that acts as a whole. Biological entities at all scales (ecosystems, organisms, organs, tissues, cells, organelles, biochemical pathways, …) are highly complex systems made up of a set of interconnected subsystems, that is, they are “systems of systems”. For centuries, the predominant approach to the study of such biological entities has been from the philosophical position known as reductionism, grounded on the premise that the properties of a given system can be deduced from the analysis of the properties of their individual components [1, 2]. Therefore, the usual strategy has been the decomposition of a complex system into simpler parts, in order to perform a separate analysis of each of them. This approach has proved to be successful and led to a great progress in cellular and molecular biology, developing a catalog of the building blocks of life (organelles, metabolites, proteins, genes, …) and explaining the physicochemical basis of numerous living processes, but also has its limitations: a system cannot be entirely understood by the sole analysis of the properties of its parts, given that the interactions between different elements, as well as influences from the environment, give rise to the emergence of new system's properties that are not present in the isolated components [3-5]. Moreover, these interactions in biological systems are often nonlinear and make the system's behavior difficult to predict. The focus must be shifted from the system's constituents to their interactions if one is to fully understand the system's dynamics and its response to environmental changes.

Molecular systems biology is a field of biology that approaches the study of biochemical systems (metabolic networks, signaling pathways or gene circuits) from an integrated (holistic) perspective, trying to understand the complexity of the system as a whole, and the rules and principles that govern its function [6].

Since the advent of the "omics" technologies in the 90's, the large amount of data gathered in the databases provides a detailed picture of the molecular state of the cell, if we are able to find a systemic approach in order to integrate this huge amount of information. On the other hand, the current computational power along with the choice of appropriate mathematical methods allows performing simulations of complex nonlinear systems such as biological networks. The integration of experimental data from multiple cellular components combined with computational simulations based on mathematical models that describe a given molecular network allows the comprehension of the functional properties of that network and the prediction of its behavior under different conditions.

Some of the mathematical concepts that allow us to measure important functional properties of biological networks were adapted from other fields of science (analysis of dynamical systems, control theory, electronics, computing). These functional properties include for example stability, sensitivity, signal amplification, response time, the existence of oscillations and stochastic fluctuations.

What are each of these properties? Stability refers to the ability of a system to respond to small perturbations without changing its qualitative behavior. This property comes from the mathematical theory of dynamical systems and was first applied to biological systems by Thom [7]. Coming from the same mathematical field, sensitivity measures the change in a system's variable (a metabolite concentration or flux, for example) with respect to a variation in the value of a parameter (examples of parameters are rate constants, kinetic orders, enzyme levels, …) [8]. Signal amplification (also known as gain), in electronics, is the ratio of the change in the output of a circuit to the change in the input signal. When applied to biochemical systems, it is often computed in its logarithmic form (and then referred to as logarithmic gain). Response time, a magnitude derived from engineering, is the time a system takes to respond to a signal. It can be calculated as the time needed to reach a given percentage of the maximal induction or inhibition. Oscillations can be defined as periodic variations in the steady state values of the output (measured as the concentration of the output molecule, if the system in question is a biochemical system). Stochastic fluctuations of a system cause a random distribution of its response, characteristic observed in all biological systems due to intrinsic and extrinsic sources of noise [9].

The measure of all these functional properties through mathematical tools such as ordinary differential equations, Taylor series, logarithms and power law functions provides a view of an operational biological system as an engineered device whose behavior can be described in a mathematical language. Therefore, to achieve such systemic understanding of a biochemical network is an interdisciplinary task which requires the flux of tools and concepts between many different disciplines, such as biology, mathematics, engineering and computer science.

1.2. Historical origins of systems biology

The rationalist tradition of Western philosophy tends to have a reductionist point of view. In spite of this, problems of integration and organization have always caused great interest in biology. Although systems biology has recently gained popularity (especially from the year 2000 onwards) due to the aforementioned emergence of the omics techniques and the increase in computational power, the importance of this field of study was recognized at least since the nineteenth century.

Claude Bernard was an early precursor of systems biology's theoreticians in the 19th century. This French physician and scientific thinker proposed that mathematics should be used in all fields of science, in order to discover the underlying laws of natural phenomena. However, he also recognized that biology was still too poorly understood to be the subject of a quantitative analysis, and it was necessary to previously gather all the new facts possible: "*The most useful path for physiology and medicine to follow now is to seek to discover new facts instead of trying to reduce to equations the facts which science already possesses …*" but " *… the application of mathematics to natural phenomena is the aim of all science, because the expression of the laws of phenomena should always be mathematical*" [10]. Already in the 20th century, an influential precedent of system's thinking was set by Alexander Bogdanov, who in his book Tectology [11], published in the 1910s, tried to identify universal laws of organization shared by all kind of systems and addressed issues like the emergence of new properties in a complex through the interactions of its components, anticipating many of the ideas that were presented some years later by Ludwig von Bertalanffy in the General Systems Theory [5]. Bertalanffy popularized the assumption of the existence of general principles that can be applied to any system, irrespective of the nature of the entities involved, and the impossibility of understanding the system's behavior by the independent analysis of its components due to the nonlinearity of their interactions. In his General Systems Theory, Bertalanffy presented a conceptual framework for the study of systems in general (biological, physical or social systems), incorporating concepts such as organization and wholeness, until then absent in conventional science.

During the first decades of the 20th century, essentially all of the basic molecular mechanisms and most of the individual enzymes of a living cell such as *E. coli* were defined. Once the biochemical basis for the individual reactions were unraveled, a more synthetic approach was needed in order to attempt the study of the integrated behavior of intact biochemical networks, and an appropriate mathematical method of analysis had to be created which would take into account the nonlinear nature of biological systems. Merging ideas from Bode analysis in electrical circuits and Taylor's theorem, in 1969 Michael Savageau proposed a mathematical formalism based on the power-law representation of all biochemical processes which allows modeling nonlinearities and limits the amount of experimental data required for the description of molecular networks. This mathematical framework is known as

Biochemical Systems Theory, and provides the foundations for the analysis and mathematical modeling of biochemical systems.

1.3. Mathematical modeling of biochemical systems

In order to study the function of complex and highly nonlinear biochemical networks one must use appropriate mathematical and computational tools that are capable of reproducing the dynamics of those networks. Mathematical modeling is an essential means to identify the topology of a system, compare the dynamic behavior of alternative network structures and characterize the underlying rules that govern the systemic behavior of a network.

Biochemical Systems Theory (BST) [12-14] is a mathematical modeling framework for the analysis of molecular systems in biology developed by Michael Savageau and co-workers since the late 1960s. In BST, the dynamic behavior of these networks of biochemical reactions is modeled with systems of ordinary differential equations (ODE), and biochemical processes are described by using the power-law formalism [15-17].

In this formalism, to capture the nonlinear nature of biological systems, the rate at which a given process occurs (production or consumption of a given molecule) is approximated by the first two terms of its Taylor series expansion, in logarithmic space.

$$\log v_i(X_1,\ldots,X_n) = \log v_i(X_{10},\ldots,X_{n0}) + \sum_{j=1}^{n} \frac{\partial[\log \mathrm{vi}(\mathrm{X}_{10},\ldots,\mathrm{X}_{n0})]}{\partial[\log X_j]}(\log X_j - \log X_{j0})$$

(1)

The Xj's are the variables that affect the process and the subscript 0 stands for the value of that variable at a given operating point.

By regrouping terms, Eqn. (1) can be rewritten as

$$\log v_i(X_1,\ldots,X_n) = \log \alpha_i + g_{i1}\log X_1 + \ldots + g_{in}\log X_n \tag{2}$$

where

$$g_{ij} = \frac{\partial(\log v_{i0})}{\partial(\log X_j)} = \frac{X_j}{v_{i0}}\frac{\partial v_{i0}}{\partial X_j} \tag{3}$$

and

$$\alpha_i = v_{i0}\prod_{j=1}^{n} X_{j0}^{-g_{ij}} \tag{4}$$

Eq. 2 can now be transformed back into Cartesian coordinates and expressed as a product of power-law functions:

$$v_i(X_1,\ldots,X_n) = \alpha_i\prod_{j=1}^{n} X_j^{g_{ij}} \tag{5}$$

where αi and gij play the role of the apparent rate constant and the apparent kinetic order with respect to Xj for the net production of Xi, respectively.

By applying this approximation, and considering that the change in the concentration of a variable is given by its aggregate rate of production ($\alpha_i\prod_{j=1}^{n} X_j^{g_{ij}}$) minus its aggregate rate of degradation ($\beta_i\prod_{j=1}^{n} X_j^{h_{ij}}$), the ordinary differential equations describing a system of n chemicals are the following:

$$\frac{\partial X_i}{\partial t} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} \tag{6}$$

These kind of models are called S-systems (because of their ability to capture the synergistic and saturable characteristics of biological complex systems) and they are a special class of generalized mass action (GMA) models.

A GMA model is a system of ODE of the form:

$$\frac{\partial X_i}{\partial t} = \sum_{j=1}^{n} a_{ij} \prod_{k=1}^{d} X_k^{g_{ijk}} \tag{7}$$

In a GMA-system, all the processes of the model that affect the levels of a given species are considered individually, instead of being aggregated into a productive term and a consumption term as in the case of an S-System. Thus, the rate of each individual process that contributes to change the concentration of a given substance Xi is approximated using a power law, derived as described in Eqs 1-5. If the process produces (consumes) Xi, aij will be positive (negative) in Eq 7.

Mathematical modeling allows us to study design principles of molecular networks. Such studies are almost always unfeasible to do experimentally. This is so because mathematical models permit performing an exhaustive comparison between the dynamic behaviors of alternative designs for a given network, in order to identify functional differences related to the topological variations of the systems being compared. Such exhaustive comparisons are very hard, if not impossible, to perform experimentally. To be sure that the differences in the system's behavior are due to differences in the system's architecture, we can use the method known as mathematically controlled comparisons, developed in the early seventies by Michael Savageau [13, 18-22].

In brief, this method requires:

(i) Defining alternative designs for the system under analysis.

(ii) Defining the functional requirements for the biological process the system is involved in.

(iii) Defining internal equivalency conditions: all processes that are identical in both alternative designs that are to be compared must have the same parameter values.

(iv) Defining external equivalency conditions: we fix the alternative parameters so as to impose that a specific functional property is the same in both systems.

Once maximal external equivalency is achieved, the remaining differences in the behavior of the systems must be due to differences in design. Then, functional advantages can be highlighted and related to a specific alternative design. This strategy (modeling and comparison between alternative designs for a given biochemical network) is a useful method in the search of design principles in molecular systems.

1.4. Design principles in molecular circuits

Are there patterns or motifs that are prevalent in a given type of biological system? Can we find rules underlying the structure of specific biochemical systems, as we can find them in engineered designs?

It seems that the answer to those questions is yes. Some quantitative and qualitative features of molecular networks are observed more frequently than expected by chance alone in biological systems performing a given function. The presence of these features can be rationalized as having evolved under a selective pressure to provide a functional advantage to the system. These recurrent motifs that make the system more effective in performing its biological function are known as biological design principles [13, 23], and exist at all levels of organization of life, from the molecular level to the whole organism, populations and ecosystems. Such design principles have been identified in many aspects of molecular circuits [13, 21-27] and allow understanding why a given design was selected for a specific (class of) molecular system(s).

For example, it was established that regulation of a biosynthetic unbranched pathway by overall negative feedback of the end product to the first reaction of the pathway has several physiological advantages with respect to other possible modes of feedback inhibition [28]: a pathway with that control mechanism is more robust to perturbations in parameter values, responds faster to fluctuations in the metabolite concentrations, and the flux through the pathway is more responsive to changes in the demand for the end product.

An example in the context of signal transduction is that Two Component Systems (TCS, see section 1.5) with a bifunctional histidine kinase sensor[1] (SK) are more efficient at amplifying the signal and suppressing crosstalk. Therefore, having a bifunctional SK is more advantageous when crosstalk represents pathological noise, while having a monofunctional SK is more advantageous in situations where the physiological response requires the integration of signals [29].

Another design principle for TCS is that, if the TCS is involved in responses that require hysteresis, it needs to fulfill two conditions: i) the reversible formation of a

[1] A bifunctional SK shows both kinase and phosphatase activities: when phosphorylated, transfers its phosphate group to its cognate response regulator (RR), and when dephosphorylated, it mediates the dephosphorylation of its cognate RR.

dead-end complex between the unphosphorylated forms of SK and response regulator (RR), and ii) RR dephosphorylation mainly done by an alternative phosphatase, independently of unphosphorylated SK [30].

These are three examples of qualitative features of a molecular system's design that give rise to physiological properties that may imply a functional advantage of the system.

The results of this work (see section 4) suggest another principle for the design of TCS: when a third component binds and inhibits the SK (prevents SK autophosphorylation), it increases the possibility of a bistable response, while a third component that binds and stabilizes the active (phosphorylated) form of the RR (prevents RR dephosphorylation) has the opposite effect. Therefore, the first design is advantageous when a switch-like response is required, and the second one is more suitable if the system needs to react in a gradual way.

1.5. Two component systems and His-Asp phosphorelays

In this work we will focus on the analysis of Two Component Systems (TCS). TCS and phosphorelays (PR) are phosphotransfer signaling pathways that enable bacteria to sense and respond to environmental stimuli [31, 32]. In these systems, a sensor histidine kinase (SK), which contains a site of histidine phosphorylation, is the protein that autophosphorylates from ATP in response to an environmental signal. The phosphorylated SK can transfer its phosphate group to an aspartate residue of the response regulator protein (RR) that either mediates the cellular response, mostly through transcription activation (in TCS), or transfers the phosphate to a second SK (containing an HPt - histidine phosphotransferase domain) that will subsequently transfer it to a second RR. This four-step His-Asp phosphotransfer cascade is known as phosphorelay (PR) (See Figure 1). TCS and PR are widely occurring in prokaryotes. However, only PR-like modules have been found in some eukaryotic organisms like protozoa, fungi and plants (although they appear to be absent in animals) [33-37].

In addition to SKs and RRs, some TCS are also known to interact with specific phosphatases that regulate dephosphorylation of the RR [38]. These core components of TCS and phosphorelays are often complemented by auxiliary proteins that play a regulatory role in the activity of the signal transduction module by regulating the transmission of the cognate signal to the SK. For example, the SK CheA is controlled through its interaction with membrane receptors that detect chemical compound in the medium and direct organisms towards higher concentrations of nutrients [39].

Another example is the SK NRII that regulates nitrogen fixation, whose activity is modulated through its interaction with the protein PII [40].

The apparently modular structure of TCS and PR circuits seems to have facilitated the evolution of a variety of designs in different organisms, allowing the module to adapt its performance to the regulation of many different types of biological functions. The prototypical structure of these phosphotransfer signaling cascades permits multiple variations: the SK can be monofunctional or bifunctional; one SK can phosphorylate more than one RR, or one RR can be phosphorylated by more than one SK; there can be one or more than one phosphotransfer step (in TCS and PR respectively); the SK and RR domains can be fused in the same protein; auxiliary proteins (such as an alternative phosphatase or a third component) can be present or not; ...

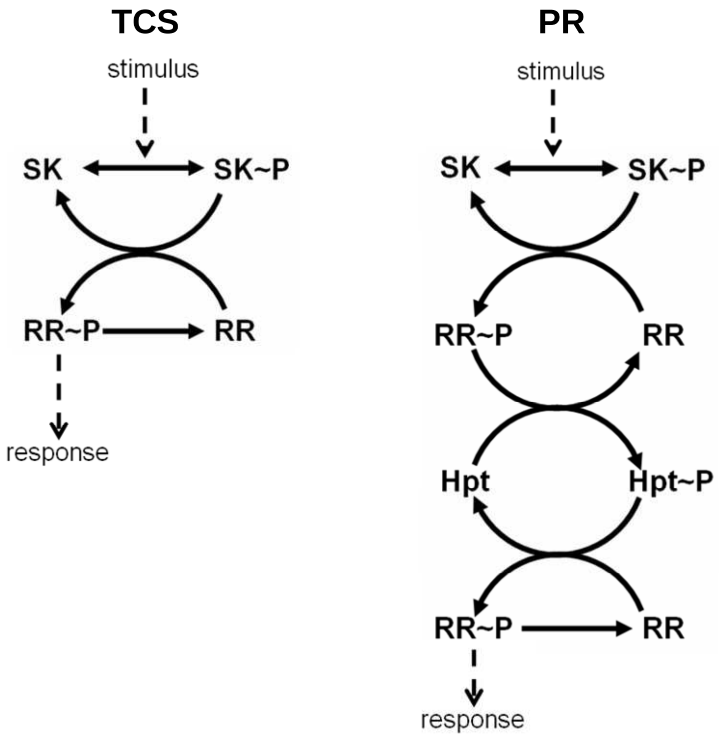


Figure 1. Prototypical Two Component System (TCS), and prototypical phosphorelay (PR). TCS and PR are both His-Asp phosphotransfer signal transduction pathways. TCS, prevalent in prokaryotes, are signaling systems composed of a sensor kinase (SK) and a response regulator (RR). The SK regulates the activity of the pathway changing its phosphorylated state in response to a given stimulus. When phosphorylated, the SK transfers the phosphate group from its histidine residue to an aspartate residue in the cognate RR, and the phosphorylated RR activates a given cellular response. PR are four-step His-Asp phosphotransfer pathways found in prokaryotes and some eukaryotes but not in animals, in which the phosphate group is first transferred from the SK to a receiver domain of a RR, typically attached to the SK, then to a histidine phosphotransfer domain (Hpt) and finally the phosphate is relayed to a second RR, which induces the response.

1.6. Methodology

Since the whole work comprised in this thesis is conceived to be theoretical and computational, the methodology that will be used is based upon the following computational tools:

Materials

The genome database at NCBI provides more than 10000 fully sequenced genomes, which will be surveyed in order to analyze the clustering of TCS/PR genes in each organism and the combination of TCS/PR protein domains in their corresponding gene products, and try to deduce from these data the structure of the TCS/PR circuits found in nature.

Identification of circuit design

Bioinformatics methods will be used to identify the relevant proteins and domains in the surveyed proteomes and compare the HK, RR, and HPt ortholog sequences found at PROSITE (http://prosite.expasy.org/) to the sequences of proteins in the fully sequenced and annotated genomes. Blast [41] and hmmer [42] will be used to survey a proteome or genome database, searching for homolog sequences, while Mega 6 [43] will be used to perform alignments of homolog sequences. Wolfram Mathematica [44] will be used for the statistical analysis of the results.

Model Building

Once the various types of design for TCS and PR circuits have been computationally identified from the genome exploration, mathematical models will be built for each of them. The mathematical models will be built using the formalisms of BST [12-14]. This will ensure that general models can be built and analyzed. In addition, it will guarantee that the results will be general and independent of parameter values. The analysis will combine Mathematically Controlled Comparisons and other methods identified and reviewed in the section 2 of this thesis.

Wolfram Mathematica [44] and Copasi [45] will be used to build the mathematical models of molecular systems and perform the simulations.

PERL will be used to make scripts, to automate database searches, handle large amounts of data, perform repeated actions, ...

1.7. Organization of this thesis

We start our search of design principles in TCS and PR circuits by reviewing in section 2 the methods used for the study of design principles in molecular systems and the results achieved thus far. This chapter led to the publication of a paper in Mathematical Biosciences [46].

Next, in section 3 we perform an extensive survey of the fully sequenced and annotated genomes and proteomes of 7609 organisms belonging to all domains of live with the purpose of identifying genes coding for TCS/PR proteins, analyzing their clustering in the genome and the composition of protein domains in the individual TCS/PR proteins. From the results of that genomic survey, we hope to be able to trace some alternative architectures of the TCS/PR circuits found in nature. This chapter led to the publication of a paper in PeerJ [47].

In section 4 we analyze one of those alternative TCS circuit designs: a TCS in which an additional third protein interacts either to the SK or to the RR. Through mathematical modelling we try to identify the differences in the dynamic behavior of the TCS caused by the presence of that auxiliary protein. This chapter led to the publication of a paper in PLoS One [48].

In section 5, the overall results obtained from the whole work are discussed, and finally, in section 6 the final conclusions derived from those results are concisely exposed.

1.8. Goals

The general goal of this thesis is to contribute for the systematic identification of design principles in TCS and PR circuits. In order to achieve that general goal, more specific goals were posed. These more specific goals are:

1. Review the conceptual methods developed in the context of BST for the study of design principles in molecular networks.
2. Review the results of the application of the methods mentioned above in studying design principles in gene circuits, cellular rhythms, molecular metabolic pathways and signal transduction networks.
3. Perform an extensive survey of the presence of TCS/PR protein domains (HK, RR and Hpt) in the proteomes of species from all taxonomic groups with fully sequenced and annotated genomes.
4. Analyze how TCS/PR protein domains organize in the individual signaling proteins. These protein domains can either be found in independent proteins

or can be fused in a single multi-domain protein. We want to carry out a phylogenetic study of the domain composition of TCS/PR proteins in each phylum.

5. Measure the tendency of genes encoding TCS/PR domain containing proteins to be clustered in the genome in neighboring positions forming operons, in each phylum.
6. Find all unique types of operons of TCS/PR coding genes that occur in the organisms with fully sequenced genomes.
7. Try to deduce, from the phylogenetic study of domain organization in TCS/PR proteins and genomic location of TCS/PR protein coding genes, alternative TCS and PR circuit topologies.
8. Start the systematic analysis of alternative TCS/PR circuit topologies by studying the physiological effect on the dynamics of a canonical TCS of an additional protein which can either interact with the SK (preventing SK activation through phosphorylation) or with the RR (stabilizing the activated form of the RR).

1.9. References

1. Descartes, *Discourses, Part V,* 1637.
2. Nagel, E., *The structure of science; problems in the logic of scientific explanation,*1961, New York,: Harcourt. 618 p.
3. Aristotle, *Metaphysics, Book H*.
4. Huxley, T.H. and J. Huxley, *Evolution and ethics, 1893-1943,* 1947, London,: Pilot Press. vii, 235 p.
5. Bertalanffy, L.V., *General Theory of Systems: Foundations, Development, Applications,* . New York Penguin1968.
6. Mesarovic, M.D. and Case Institute of Technology. Systems Research Center., *Systems theory and biology. Proceedings of the 3rd Systems Symposium at Case Institute of Technology*1968, Berlin, New York etc.: Springer. xii, 403 p. with illus.
7. Thom, R., ed. *Structural Stability and Morphogenesis; An Outline of a General Theory of Models"*. ed. R. Benjamin/Addison-Wesley, Massachusetts1975.
8. Cruz, J.B., *System sensitivity analysis*1973: Dowden, Hutchinson & Ross.
9. Swain, P.S., M.B. Elowitz, and E.D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12795-800.
10. Bernard, C., ed. *Introduction à l'étude de la médecine expérimentale*. 1865: Paris.
11. Bogdanov, A., ed. *Tektology: Universal Organization Science*. 1913-1922.
12. Savageau, M.A., *The behaviour of intact biochemical control systems.* Curr.Tops.Cell.Reg, 1972. **6**: p. 63-130.
13. Savageau, M.A., *Biochemical Systems Analysis: A study of function and design in molecular biology.* Reading, Mass.: Addison-Wesley, 1976.
14. Savageau, M.A., *Biochemical systems theory: operational differences among variant representations and their significance.* J Theor Biol, 1991. **151**(4): p. 509-30.
15. Savageau, M.A., *Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions.* J.Theor.Biol., 1969. **25**: p. 365-369.
16. Savageau, M.A., *Biochemical systems analysis II. Steady state solutions for an n- poll system using a power-law approximation.* J.Theor.Biol., 1969. **25**: p. 370-379.
17. Savageau, M.A., *Biochemical systems analysis. III. Dynamic solutions using a power-law approximation.* J Theor Biol, 1970. **26**(2): p. 215-26.
18. Savageau, M.A., *Genetic regulatory mechanisms and the ecological niche of Escherichia coli.* Proc Natl Acad Sci U S A, 1974. **71**(6): p. 2453-5.
19. Savageau, M.A., *Optimal design of feedback control by inhibition.* J Mol Evol, 1974. **4**(2): p. 139-56.
20. Savageau, M.A., *Comparison of classical and autogenous systems of regulation in inducible operons.* Nature, 1974. **252**(5484): p. 546-9.
21. Savageau, M.A., *Optimal design of feedback control by inhibition: dynamic considerations.* J Mol Evol, 1975. **5**(3): p. 199-222.
22. Alves, R. and M.A. Savageau, *Extending the method of mathematically controlled comparison to include numerical comparisons.* Bioinformatics, 2000. **16**(9): p. 786-98.
23. Alon, U., *Biological networks: the tinkerer as an engineer.* Science, 2003. **301**(5641): p. 1866-7.
24. Hlavacek, W.S. and M.A. Savageau, *Completely uncoupled and perfectly coupled gene expression in repressible systems.* J Mol Biol, 1997. **266**(3): p. 538-58.
25. Savageau, M.A., *Significance of autogenously regulated and constitutive synthesis of regulatory proteins in repressible biosynthetic systems.* Nature, 1975. **258**(5532): p. 208-14.

26. Savageau, M.A., *Demand theory of gene regulation. I. Quantitative development of the theory.* Genetics, 1998. **149**(4): p. 1665-76.
27. Alves, R. and M.A. Savageau, *Irreversibility in unbranched pathways: preferred positions based on regulatory considerations.* Biophys J, 2001. **80**(3): p. 1174-85.
28. Alves, R. and M.A. Savageau, *Effect of overall feedback inhibition in unbranched biosynthetic pathways.* Biophys J, 2000. **79**(5): p. 2290-304.
29. Alves, R. and M.A. Savageau, *Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function.* Mol Microbiol, 2003. **48**(1): p. 25-51.
30. Igoshin, O.A., R. Alves, and M.A. Savageau, *Hysteretic and graded responses in bacterial two-component signal transduction.* Mol Microbiol, 2008. **68**(5): p. 1196-215.
31. Hoch, J.A., *Two-component and phosphorelay signal transduction.* Curr Opin Microbiol, 2000. **3**(2): p. 165-70.
32. Parkinson, J.S., *Signal transduction schemes of bacteria.* Cell, 1993. **73**(5): p. 857-71.
33. Appleby, J.L., J.S. Parkinson, and R.B. Bourret, *Signal transduction via the multi-step phosphorelay: not necessarily a road less traveled.* Cell, 1996. **86**(6): p. 845-8.
34. Loomis, W.F., A. Kuspa, and G. Shaulsky, *Two-component signal transduction systems in eukaryotic microorganisms.* Curr Opin Microbiol, 1998. **1**(6): p. 643-648.
35. D'Agostino, I.B. and J.J. Kieber, *Phosphorelay signal transduction: the emerging family of plant response regulators.* Trends Biochem Sci, 1999. **24**(11): p. 452-6.
36. Sakakibara, H., M. Taniguchi, and T. Sugiyama, *His-Asp phosphorelay signaling: a communication avenue between plants and their environment.* Plant Mol Biol, 2000. **42**(2): p. 273-8.
37. Caffrey, D.R., O'Neil, L.A., et al., *The evolution of MAP kinase pathways: coduplication of interacting proteins leads to new signaling cascades.* J. Mol. Evol., 1999. **49**(5): p. 567-582.
38. Silversmith, R.E., *Auxiliary phosphatases in two-component signal transduction.* Curr Opin Microbiol, 2010. **13**(2): p. 177-83.
39. Hazelbauer, G.L. and W.C. Lai, *Bacterial chemoreceptors: providing enhanced features to two-component signaling.* Curr Opin Microbiol, 2010. **13**(2): p. 124-32.
40. Atkinson, M.R. and A.J. Ninfa, *Role of the GlnK signal transduction protein in the regulation of nitrogen assimilation in Escherichia coli.* Mol Microbiol, 1998. **29**(2): p. 431-47.
41. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
42. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
43. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
44. Wolfram Research, I., *Mathematica, Version 8.0.* Champaign, IL, 2010.
45. Hoops, S., et al., *COPASI--a COmplex PAthway SImulator.* Bioinformatics, 2006. **22**(24): p. 3067-74.
46. Salvado, B., et al., *Methods for and results from the study of design principles in molecular systems.* Math Biosci, 2011. **231**(1): p. 3-18.
47. Salvado B, V.E., Sorribas A, Alves R, *A survey of HK, HPt, and RR domains and their relative organization in two-component systems and phosphorelay proteins of organisms with fully sequenced genomes.* PeerJ, 2015.
48. Salvado, B., et al., *Two component systems: physiological effect of a third component.* PLoS ONE, 2012. **7**: p. e31095.

2 Methods for and results from the study of design principles in molecular systems

2.1. Abstract

Most aspects of molecular biology can be understood in terms of biological design principles. These principles can be loosely defined as qualitative and quantitative features that emerge in evolution and recur more frequently than one would expect by chance alone in biological systems that perform a given type of process or function. Furthermore, such recurrence can be rationalized in terms of the functional advantage that the design provides to the system when compared with possible alternatives. This chapter focuses on those design features that can be related to improved functional effectiveness of molecular and regulatory networks. We begin by reviewing assumptions and methods that underlie the study of such principles in molecular networks. We follow by discussing many of the design principles that have been found in genetic, metabolic, and signal transduction circuits. We concentrate mainly on results in the context of Biochemical Systems Theory, although we also briefly discuss other work. We conclude by discussing the importance of these principles for both, understanding the natural evolution of complex networks at the molecular level and for creating artificial biological systems with specific features.

2.2. Introduction

One of the most important goals in biology is the understanding of how the molecular features of biological systems have emerged and become fixed. Emergence of these features during evolution is random, due to different mechanisms such as mutation and recombination. Fixation of the different alternative features can be accidental, due to chance. Another possibility is that they become fixed because they generate molecular variants that make fitter organisms that survive and reproduce better. In this process, the predominance of a given molecular feature is a consequence of natural selection acting as a process that increases the frequency of designs with a better functional performance. Differentiating between the two possibilities allows researchers to identify the biological design principles of the molecular systems of interest [1, 2]. In this context, biological design principles can be defined as repeated qualitative and quantitative features of biological components and their interactions that are observed in molecular systems at high frequencies and improve the functional performance of a system that executes a specific process. Such principles have been found in many aspects of molecular biology [3].

For example, sequence biases that facilitate the control of gene expression under different conditions, with appropriate timing, are recurrent and can be rationalized as having evolved under a selective pressure to minimize the metabolic cost associated with the process of synthesis [4-6]. As another example, certain types of protein domains that are more abundant in proteomes and are associated with specific functional requirements for protein stability suggest a recurrent evolutionary design associated with that specific function [7, 8]. As a final example, an interesting structural design principle is found in glycogen. This molecule has evolved to provide a reservoir of glucose and to make glucose available quickly and in large amounts when required. Melendez-Hevia and co-workers showed that the branching in the structure of glycogen is an optimal solution to the problem of optimizing storage space and fast glucose mobilization [9, 10].

If design principles emerge from the evolution of complex biological systems, one may expect to identify such principles at all organizational levels, from metabolic and gene networks, organs, and physiology, to organisms and their interactions [2].

Here we center our attention on the evolution of structure and regulation of molecular networks in cells. These include gene networks, metabolic pathways, signal transduction cascades, cell cycle, immune response, and other molecular networks. The study of design principles in the context of regulatory molecular networks started as early as the seventies (see, for example, [11-20]). These early studies were performed in the context of BST (*Biochemical Systems Theory*) [20, 21], a body of work providing a set of tools that facilitate the creation and analysis of mathematical models for biological systems [21]. The current surge in interest towards network motifs and the modular structure of molecular networks is partially a consequence of those early studies. However, it is also a consequence of the amount and complexity of the biological information that continuously accumulates and becomes available to us. This creates a situation where learning how biology works hinges on the possibility of understanding general organizational principles in biological systems, rather than through memorizing massive ‘‘grocery lists’’ of biological facts.

There have been several methods developed within the BST framework specifically to studying design principles in molecular networks. One of these methods is that of the *Mathematically Controlled Comparisons* [11, 12, 22, 23]. This technique facilitates the analysis and comparison of the differences in systemic behavior between alternative designs for the same network. The compared behaviors range from steady-state characteristics to dynamic behavior and parameter robustness. The comparisons are mathematically constrained in a way that ensures that any differences in behavior are a consequence of the differences in design and not of other spurious changes between systems. Recently, design space representations that provide a simplified way to analyze the different phenotypic regions of systemic behavior were developed within the same theoretical framework [24-26].

In this chapter, we review some of these methods, and results of their application to the study of molecular networks. We focus mainly on studies within the context of BST, although we also briefly discuss other relevant work. We conclude by discussing the importance of biological design principles and how they can be organized in the future.

2.3. Molecular circuits vs the molecular network of the cell

This special issue of Mathematical Biosciences focuses mostly on design principles that can be inferred for the structure and regulation of molecular circuits that are responsible for specific biological functions, and on how such principles correlate to those functions.

These biological design principles are a consequence of evolution selecting for particular features that make some circuits more effective in performing their biological function. Given that evolution acts on organisms and populations [27], it is fair to ask two questions about the previous sentence. The first question is how appropriate is it to identify functional effectiveness of a specific circuit or module rather than that of the entire network of circuits. The second question is how can we be sure that a particular design is a consequence of selection because it is functionally more effective, rather than an accident of evolution.

To answer the first question one must consider two aspects. To begin, one must admit that at the molecular level living organisms seem to evolve in a modular way. Several examples point to this. Proteins evolve mostly through domain recombination, and specific functions are associated with each type of protein domains [28]. The expression of genes coding for proteins of many pathways is coordinated in operons and regulons [29, 30]. In addition, recent work suggests that, given the parallel and multiple demands that biological systems have to cope with during evolution, it is likely that their functionality has evolved in a modular fashion [31-35]. Considering that such modularity appears to be extended in biology, one must also consider that most mutations in a circuit are likely to cause malfunctioning of that circuit. The malfunction of the circuit contributes to decrease the fitness of the organism (see, for example, [36]). These two considerations suggest that it is indeed appropriate to consider functional effectiveness of circuits, when isolated from the entire molecular network of the cell.

To answer the second question one must consider that alternative designs come about randomly for any given molecular circuit, through the natural forces and events that generate diversity in biology (mutation, cross-over, ...). If various alternatives are selected during evolution under different conditions, this implies that not all network designs are functionally equivalent and that each design would provide for a better functionality under the conditions in which it was selected[2].

2.4. Functional effectiveness of molecular networks

A fundamental aspect in the study of biological design principles is how to define functional effectiveness criteria for a given circuit and how to analyze the effect of changes in the design of the circuit on those criteria. In essence, one should

[2] In this discussion we disregard the effect of random drift and population size. We assume that the population is always sufficiently large so that the effect of Muller's ratchet in fixing deleterious mutations is small.

understand the biology that a given type of circuit is involved in, find out what the specific role of that circuit is, and identify physiologically relevant aspects of that role that can be associated with improved or decreased functionality.

It is hard to propose an algorithm to define functional criteria that are applicable to every type of circuit one could be interested in. For example, in signal transduction circuits, one should consider specific criteria related to signal interpretation, such as amplification, delay, frequency response, noise propagation, correlation between input and output [37, 38]. In contrast, in some moiety conservation cycles, one would apply considerations that are similar to those engineers apply when designing batteries [24, 39]. However, there are some general criteria that are applicable, in a broad sense, to different modules. For example, the ability to maintain performance under small perturbations in parameters values (robustness) seems to be a desirable characteristic for many different systems [40, 41].

In essence, several physiologically relevant criteria are simultaneously important for the appropriate function of a circuit. Analyzing the molecular circuits that perform a given biological process provides important insights into what interactions determine that the circuit functions as it should under different conditions [42]. Furthermore, it helps understanding if different design characteristics of those circuits are linked, making it so that if one characteristic is selected for or against, others are automatically implemented or excluded (see, for example, [43, 44]). In addition, contradictory functional demands may be placed upon a molecular biological network of interest, constraining its evolution. These considerations imply that it is difficult to intuitively understand how a design may have been selected for or against. Such an understanding requires the use of appropriate analytical tools to evaluate how changes in the design can simultaneously affect all relevant functional aspects of the circuit.

2.5. Methods to analyze design principles

An accurate analysis of the effect of alternative designs of a molecular circuit on the circuit's function requires that one understands what that function is. This knowledge is essential to identify the relevant performance criteria that need to be analyzed in order to understand the selection of alternative designs for the circuit. As stated above, some of these criteria will be quite general (robustness, stability, ...), while others will be system specific.

When characterizing the effect of a circuit's design on the performance of that circuit, one is typically interested in understanding either (a) the functional performance limits of a given design or (b) why analogous systems have alternative designs under different conditions or in diverse organisms. The performance limits of the circuit can be analyzed from qualitative (for example, can a given network structure generate oscillatory behavior or multistationarity?) or quantitative (for example, by how much should gene expression change during some adaptive response in order to ensure organism survival?) perspectives. Whatever the perspective, the analysis is typically done by building mathematical models that represent the circuit and analyzing these models using one or two of an array of different methods.

Methods to determine the functional performance limits of a given circuit include, for example, approaches such as *Reaction Network Theory* (RNT) [45-57]. RNT permit identifying necessary conditions in the structure of mass actions circuits that lead to robustness, oscillations and multistationarity, independent of the parameter values. This is done using a combination of graph theory and differential equations theory in order to analyze the stoichiometric matrix of the circuit [54, 55, 58-66]. RNT calculates (a) the rank of the stoichiometric matrix of the network, (b) the number of different sets of reactants and/or products of individual reactions in the network, and (c) the number of isolated subnetworks in which the circuit can be decomposed. With these three numbers, a deficiency is calculated for the network and, based on this deficiency, the necessary conditions for different types of dynamic behavior are determined (Fig. 1).

Other qualitative methods, such as the pentose phosphate pathway (PPP) game [58, 59], have been used to understand what is being optimized during the evolution of a particular solution for the structure of a circuit or network. The PPP game considers all possible reaction paths that a set of biological enzymes can generate between different metabolites. Then, it compares these alternative paths to the ones that naturally evolved in organisms (Fig. 2). These comparisons have led to the inference that minimization of the number of steps is a significant driving force in the evolution of metabolic circuits [59, 60].

Limits of functional performance of circuits can also be characterized through the use of numerical methods. For example, the physiological constraints that may shape the evolution of changes in gene expression during heat shock response of the yeast *Saccharomyces cerevisiae* have been systematically studied [61-63]. An initial approach to the problem led to the creation of a mathematical model representing the main metabolic pathways involved in this response. Then, the numerical criteria that represented minimal requirements for survival were identified. Finally, a large scale Monte Carlo (MC) sampling of the parameters of the system was performed, eliminating all parameter combinations that generate systems that did not meet the minimal criteria. Once this was done, an analysis of the parameter sets generating systems that were feasible led to the identification of numerical design principles for

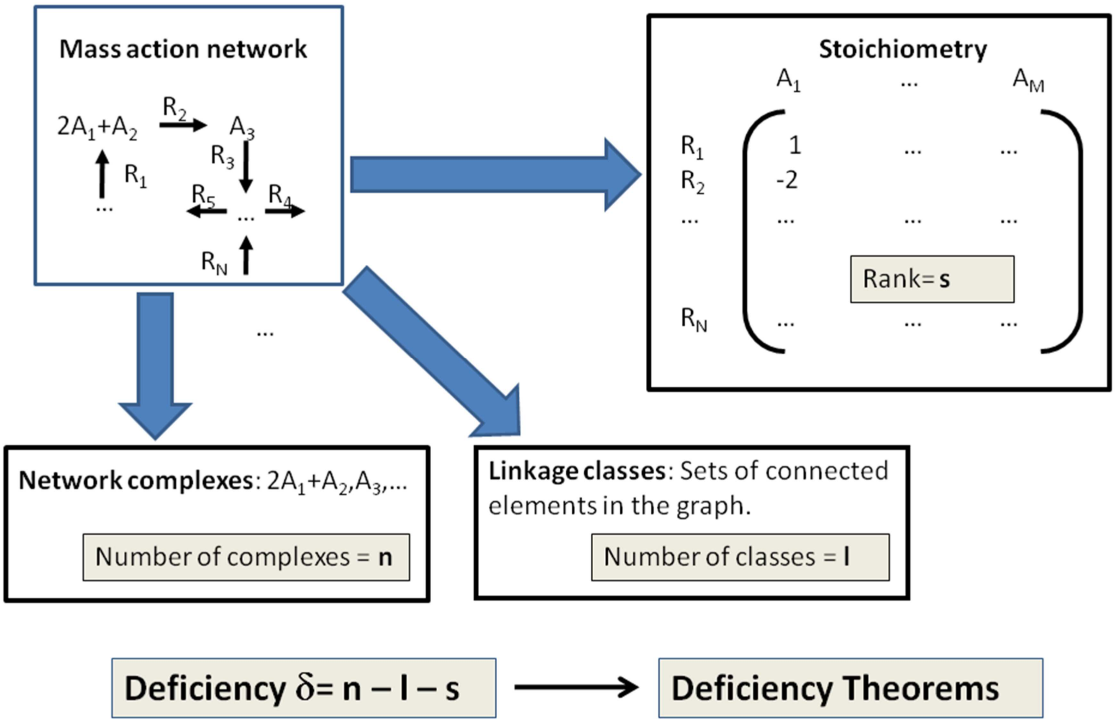


Figure 1. Reaction Network Theory (RNT). By analyzing the structure of (usually mass action) reaction networks, RNT derives deficiency-related theorems that, depending on that deficiency δ, certify existence of single or multiple steady state and/or limit cycles. The theory, in its basic forms, requires knowing the rank of the stoichiometic matrix of the network, as well as the number of complexes in the reactions and the number of linkage classes.

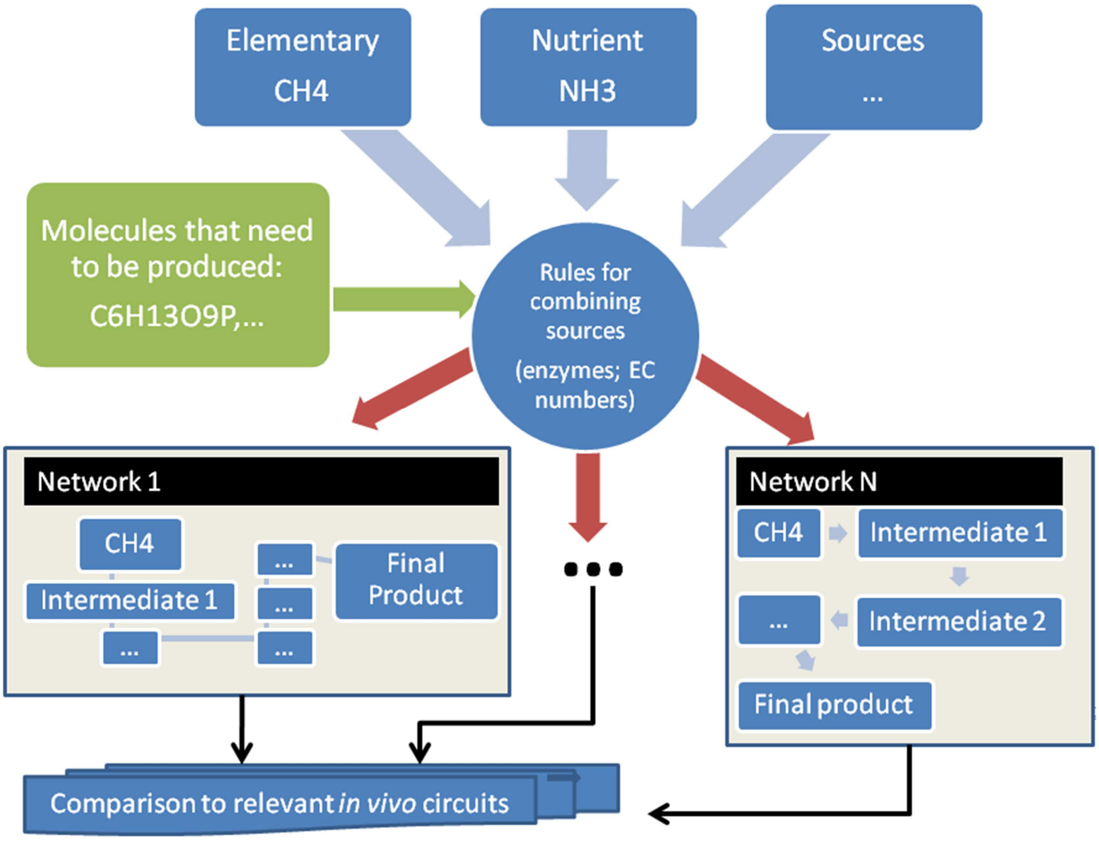


Figure 2. Pentose Phosphate Pathway (PPP) like games. By starting from elementary nutritional sources and using all enzymatic activities that are known, these games determine all possible reaction pathways that lead from the elementary carbon sources to the biological molecules. A comparison of these pathways to those occuring in living beings supports the notion that nature selects for the shortest paths between elementary nutritional sources and biological building blocks (see text for details).

these parameters that can be justified by the functionality of the system. This approach can be applied to similar problems, although it can become computationally demanding for systems with increasing dimensions. Recently, a more efficient approach to the problem was developed and applied. Instead of using large scale MC sampling, global optimization methods are used to map the parameter space in such a way that all regions of this space that meet minimal functionality criteria can be identified [61-64].

The concept of *design spaces* has been recently systematized and proposed by Michael Savageau and co-workers as an alternative to fully characterize the different phenotypical regimes of a molecular circuit [24-26]. These different regimes are identified with regions in the parameter space in which different elementary processes dominate the dynamic change in the level of each variable of the circuit. In short, one creates a model for the circuit of interest and then performs dimensional reduction on the model in such a way that the number of parameters is minimized. Then, the parameter space of the reduced model is divided into regions where different dominant elementary processes regulate the production and consumption of each of the variables in the system. The borders between regions identify approximate boundaries for the different phenotypes of the model in the parameter space (Fig. 3).

There are also methods that are specifically tailored to address questions about why alternative designs exist for analogous circuits performing the same function. The first method specifically developed to address these questions was published in the early 1970s by Michael Savageau [14-16, 18-20]. In this pioneering work, he developed the first version of what is now known as *Mathematically Controlled Comparisons*. Later, this method was further developed and applied to various biological problems [11, 12, 22, 23]. These comparisons can be done in fully

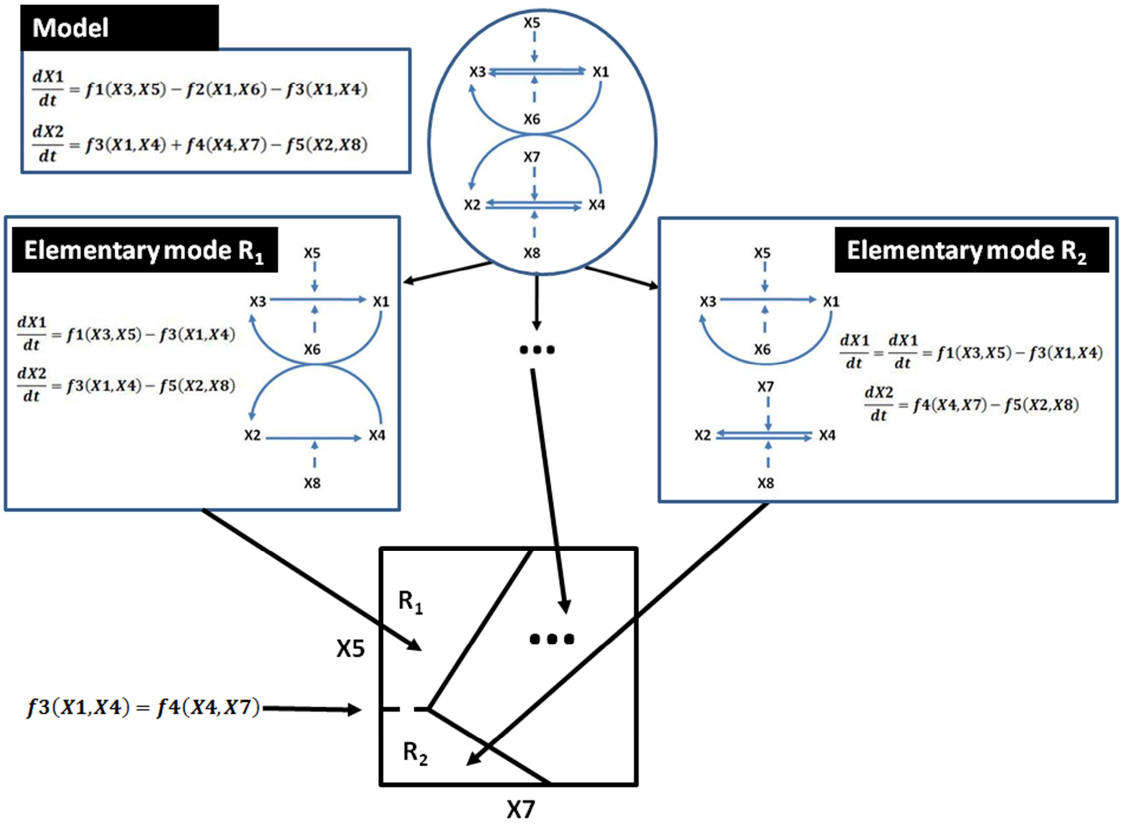

Model
dX1/dt = f1(X3,X5) − f2(X1,X6) − f3(X1,X4)
dX2/dt = f3(X1,X4) + f4(X4,X7) − f5(X2,X8)
X5
X3
X1
X6
X7
X2
X4
X8
Elementary mode R1
dX1/dt = f1(X3,X5) − f3(X1,X4)
dX2/dt = f3(X1,X4) − f5(X2,X8)
Elementary mode R2
dX1/dt = dX1/dt = f1(X3,X5) − f3(X1,X4)
dX2/dt = f4(X4,X7) − f5(X2,X8)
•••
R1
R2
•••
X5
X7
f3(X1,X4) = f4(X4,X7)

Figure 3. Design spaces. This method decomposes the mathematical model of the system into the elementary modes of consumption and production of each of the variables. Each combination of these elementary modes is a region Ri in the design space. By analyzing the simpler models in the space of variables and/or parameters one is interested in, one can identify ‘‘pure’’ possible phenotypes for the circuits. The borders between the regions correspond to zones where two production (or consumption) terms for a given variable have the same numerical value (see text for details).

analytical form or numerically [22, 23, 65], depending on the models being compared and on the questions one is asking [46, 49, 67]. Mathematical details can be found in the literature [11, 12, 66, 68-70]. Briefly, the use of this method requires (see also Fig. 4):

(i) Defining the functional requirements for the biological process or network under analysis.

(ii) Defining alternative designs for the system.

(iii) Defining basic criteria of *internal equivalency* between alternative designs. In general, all processes that are identical in the alternative networks are considered to have exactly the same parameter values in the two systems. This is equivalent to making control experiments in a wet lab.

(iv) For each pair of comparisons, the system in which the process with alternative designs has the largest number of parameters is usually taken as the reference, while the other system is taken as the alternative. Then, one defines *external equivalency* conditions. The reasoning underlying such conditions is as follows. There are certain behaviors of the system that are important for its function. If the reference process had mutated in such a way that it became the alternative process, then, in the best of all possible worlds nature could mutate the parameters of this alternative such that it would make both systems equivalent with respect to at least some of those behaviors. Therefore, if one takes each of the parameters of the alternative system and imposes that a specific behavioral trait is the same in the alternative and in the reference, one can fix the value for each of the alternative parameters. This comparison process assumes that evolution has an infinite amount of tries and time to make alternative designs as equivalent as possible when a specific functionality is required. Although this may not be the case in biological evolution, the results obtained by using the method, so far, indicate that these assumptions are reasonable for successfully identifying design principles in many cases.

(v) When maximal external equivalency is achieved, any remaining differences in the behavior of the systems are exclusively attributable to the differences in design. Then, advantages in the functional performance of the system can be highlighted and related to the emergence of a particular design under specific conditions.

Other approaches to identify and study biological design principles are also available. For example, one can study a catalog of network designs to identify functional alternatives that have been implemented by nature during evolution [71]. For example, this approach was used to identify design principles for biochemical oscillators [72]. The analysis and classification of network motifs according to their dynamical behavior also follows this strategy [73, 74].

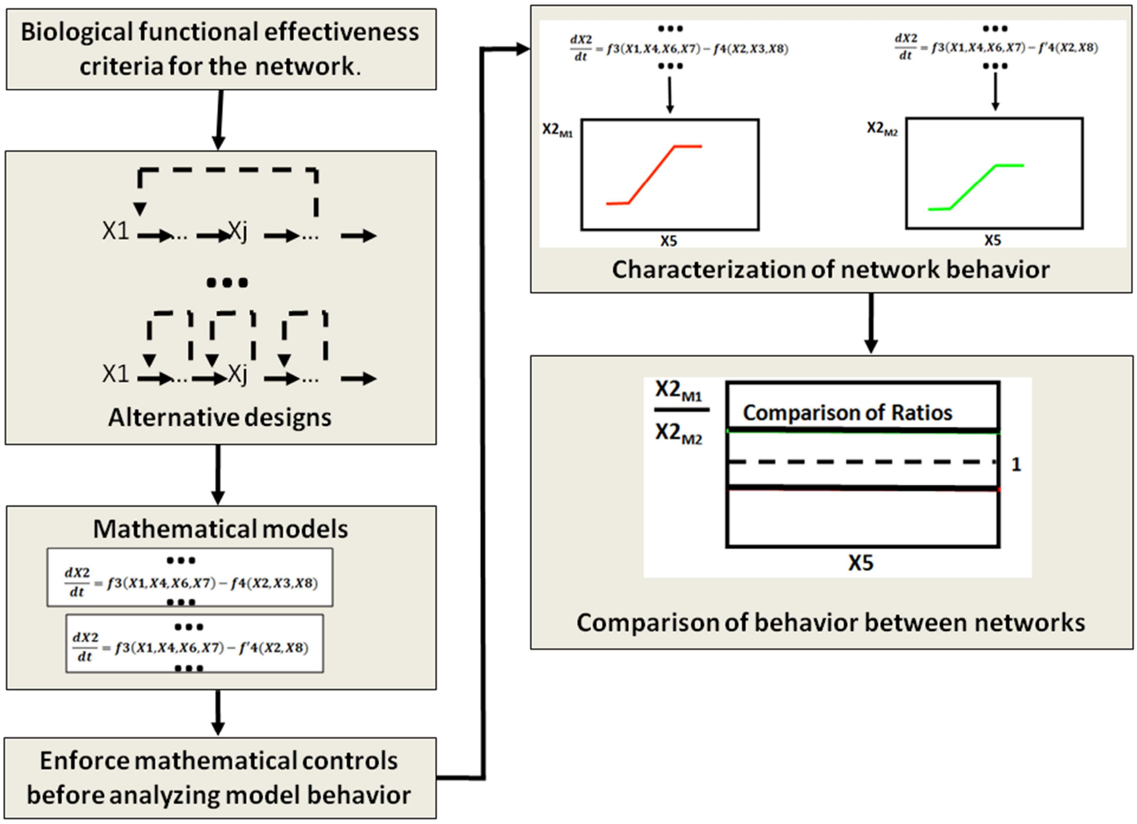

Biological functional effectiveness criteria for the network.
X1
Xj
Alternative designs
Mathematical models
$\frac{dX2}{dt} = f3(X1,X4,X6,X7) - f4(X2,X3,X8)$
$\frac{dX2}{dt} = f3(X1,X4,X6,X7) - f'4(X2,X8)$
Enforce mathematical controls before analyzing model behavior
$X2_{M1}$
$X2_{M2}$
X5
Characterization of network behavior
$\frac{X2_{M1}}{X2_{M2}}$
Comparison of Ratios
1
X5
Comparison of behavior between networks

Figure 4. Mathematically controlled comparisons. This method permits comparing the functional effectiveness of alternative circuits for biological networks that perform the same function. This is done through the creation of mathematical models for the alternative designs. Then, one implements a set of controls to ensure that any differences between the behavior of the two models is only due to the differences in network structure. Typically, the comparison is done by taking the ratio of the property of interest in the reference system [M1] to the corresponding property in the alternative system (s) [M2]. If the property is always larger in the reference system, the ratio will always be larger than one [upper line in the last panel of the figure]. If the property is always smaller in the reference system, the ratio will always be smaller than one [lower line in the last panel of the figure] (see text for details).

2.6. Design principles in molecular systems

2.6.1. Design principles in gene circuits

Gene regulation networks show a number of recurrent motifs that could represent a fundamental topology of regulatory circuits that is independent of the specific genes involved in the circuit. One of the most prevalent motifs is a feed forward loop in which a transcription factor X regulates another transcription factor Y and both regulate a given gene Z. This motif can generate eight different basic designs. In four of these designs the direct effect of X on the gene expression of Z is similar to the indirect of X on the gene expression of Y compounded with the effect of Y on the gene expression of Z. These are called coherent designs. Four other designs are incoherent (Fig. 5). An initial theoretical analysis of the different designs shows that coherent loops are advantageous for delaying response to a signal, while incoherent loops work more effectively as accelerators of response to a signal [67]. This lead to the suggestion that coherent feed-forward loops should be selected in environments where the distribution of the input pulse duration is sufficiently broad [75]. More recent work shows that both types of loop can accelerate or delay response to a signal, depending on parameter values [76]. Incoherent loops have also been proposed as a functionally more effective mechanism for detecting fold-change in gene regulation [14].

One of the earliest case studies where design principles have been identified in molecular circuits regards the relationship between mode of regulation for gene expression and the demand for the gene product, leading to the proposal of the *demand theory for gene expression* [14, 69, 70, 77]. The theoretical results correctly predict that positive regulation is preferentially selected for genes whose product is required over a large fraction of the life cycle of the individual (high demand genes), while negative regulation is preferentially selected for genes whose product is required for a small fraction of that life cycle (low demand genes) [66, 68, 78-80]. The biological explanation for the prediction boils down to a ‘‘use it or lose it’’ principle. The effect of losing the binding site for the regulation is proportional to the fraction of time that it is under use. For example, if a positively regulated gene is under low demand, there is a much smaller fraction of the life cycle of the individual when losing this regulation will affect the individual. Conversely, if a negatively regulated gene is under high demand, there is a much smaller fraction of the life cycle of the individual when losing this regulation will also affect the individual. In other words, this theory proposes that rate at which Muller’s ratchet will turn for deleterious mutations in the binding sites is proportional to the fraction of time that those sites are inactive (Fig. 6).

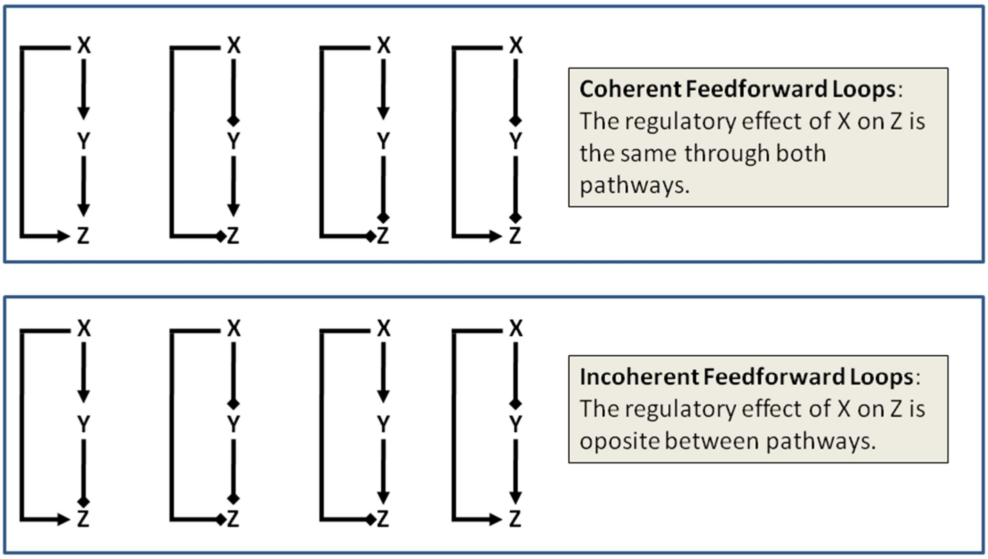

X
Y
Z
Coherent Feedforward Loops: The regulatory effect of X on Z is the same through both pathways.
Incoherent Feedforward Loops: The regulatory effect of X on Z is oposite between pathways.

Figure 5. All possible types of feedforward loops in three-species genetic circuits. Arrows with triangular heads indicate activation, while heads with square heads indicate inhibition.

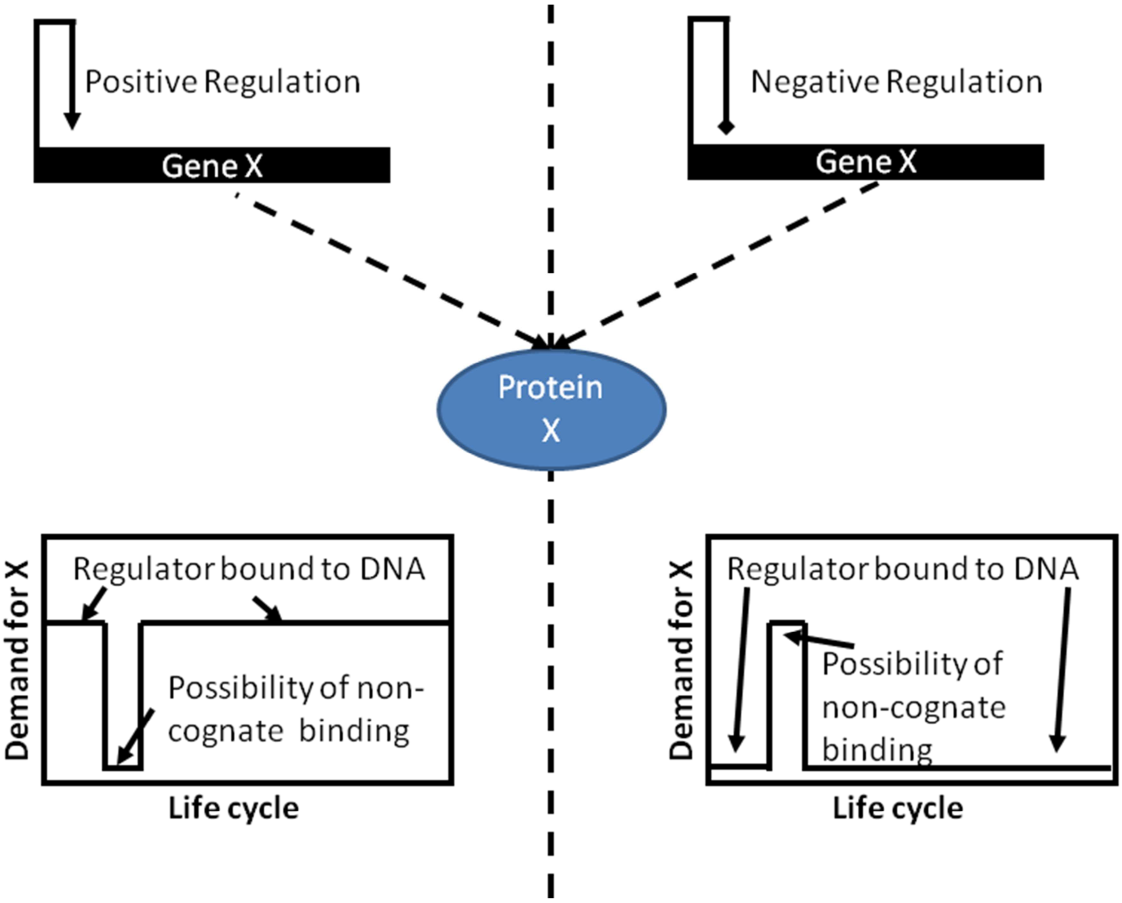


Figure 6. Demand Theory: Arrows with triangular heads indicate activation, while arrows with square heads indicate inhibition. Originally, demand theory predicted that negative regulation of gene expression is observed under low demand, while positive regulation is observed under high demand. This was explained by higher probability of losing negative regulation sites under high demand and positive regulation sites under low demand. More recently, it was suggested that the correlation between mode of regulation and demand is a consequence of the fraction of the life cycle in which binding of non-cognate regulators could lead to noise in gene expression (see text for details).

Recently, however, some doubts have been presented with respect to this interpretation, and similar predictions were shown to arise if one considers how the different modes of regulation minimize errors during transcription. Systems in which free sites are more error-prone (exposed to binding by non-specific factors) than sites bound to their cognate partner, will tend to evolve mechanisms that keep the sites bound most of the time, thus minimizing errors [81]. Noise filtering was also put forward as a possible explanation for the different modes of gene regulation [37, 82, 83]. Approaching the problem from an alternative perspective showed that gene circuits with negative regulation are better at filtering noise out of signals with high intensity, while positively regulated circuits are more efficient in filtering noise out of low intensity signals.

These explanations for selection between alternative modes of gene regulation may not be mutually exclusive. Classical demand theory [69] predicts that loss of binding sites has a smaller effect on fitness if those binding sites are rarely used. Therefore, to keep regulation, it should be implemented using the type of binding site that is used most often for the gene in question in the organism of interest. The noise-related variation of the theory states that fitness is affected mostly because of inappropriate binding in the absence of the cognate regulator, resulting in deleterious gene expression [81]. However, these two aspects are complementary. Under low demand, with a positive regulator, the binding site would be available for binding. If this binding leads to expression of the gene when it is not needed, there would be a deleterious effect that would select for sites where such binding would not occur. This could cause loss of the positive regulatory effect through selection, while classical demand theory argues that such loss could come about even by drift. A similar argument can be made for negative regulation in a high demand environment. It is conceivable that both evolutionary effects could contribute for the observed regulatory pattern under different conditions. There are studies that hint at such complementarity. Effective population size and the typical time scale of environmental variations appear to be key parameters in determining the fitness advantage of the different modes of regulation [84]. The ‘‘use-it-or lose-it’’ principle that underlies classical demand theory is valid for small populations with long time scales of environmental variations. Conversely, a complementary principle will be valid for populations with large effective sizes in rapidly changing environments [84]. Under these conditions, one would expect that both, positive and negative regulation, be stable.

Design principles have also been identified for other aspects of how gene circuits function and for the interplay between genotype and phenotype. One example of this are the design principles described for the organization of the gene networks that are responsible for regulating the development of sea urchin embryos, suggesting a number of strategies that may play similar roles in different organisms [85, 86]. Another example is the aptitude of polyphasic positive feedback loops [3] to work as count-up cellular timers used to defer the response to stimuli, counteracting protein dilution during cell growth and proliferation [87].

The quantitative design aspects of the regulation of gene expression have also been analyzed. One example of this is the study that shows that the minimal requirement for network dosage compensation to exist in genetic circuits is that the circuit is regulated by both, a positive and a negative regulator [88]. Another example has to do with regulation of changes in gene expression during stress response. Such changes enable organisms to regulate pathway fluxes and metabolite concentrations in ways that permit an appropriate adaptive response to changing environmental conditions. Adaptive responses are fundamental for survival and can be achieved following different strategies that change gene expression from a given reference initial state to the adapted state. Analyzing these strategies reveals that, in *Escherichia coli* amino acid biosynthetic pathways, genes from the same transcript are translated into proteins in such a way that each subsequent enzyme in a pathway becomes

[3] A polyphasic feedback loop is an architecture that temporally divides a feedback loop into different phases, through modification of the feedback signal strength.

available when enough of its substrate is produced by the previous enzyme of the pathway (Fig. 7) [89].

Operative changes in gene expression that are required to attain a given adaptive response while maintaining a set of basic physiological requirements have been investigated by Sorribas and co-workers [61, 62, 64, 80]. Based on previous work by Voit and Radivoyevitch [63], they have identified the physiological requirements that constrain the quantitative changes in gene expression during the adaptive response of yeast to heat shock, using a Monte-Carlo based approach [62]. More recently a global optimization method that exactly maps the operating regions of gene expression space that meet the physiological requirements for cell survival has been developed [61, 64]. The results of applying this method to the analysis of changes in gene expression during yeast stress response are consistent with those from the Monte-Carlo approach (Fig. 8).

This new technique allows for identifying feasibility regions in the enzyme activities so that a number of physiological constraints required for cell survival are met. These feasibility regions contain many admissible expression values for the genes that are compatible with a given set of physiological requirements. As such, one expects that evolution selects gene expression patterns that fall within these regions. The available experimental data is consistent with the computational predictions, suggesting that the physiological constraints that were used to identify the feasibility regions are close to those that are active *in vivo*.

The technique described above maps the gene expression space, identifying regions in this space which give rise to phenotypes that fulfill physiological constraints. The opposite approach would be to use single-cell techniques to measure the position of individual cells in the gene expression space, and analyze the geometry of the region containing the actual single-cell gene expression data of a given cell-type. Using this approach led to the observation that cell populations performing multiple tasks fall within the high dimensional gene expression space in simple low dimensional polyhedrons with a number of vertices equal to the number of tasks cells must confront. The vertices of these polytopes are optimal gene expression profiles for each of the tasks [90]. This geometry of gene expression profiles is related to the concept of Pareto optimality: no gene expression profile can be optimal for all tasks faced by the cell, and the trade-off between these tasks shapes the distribution of cells in the gene expression space. This technique can be used to infer the biological tasks represented by the vertices of the polytopes [91].

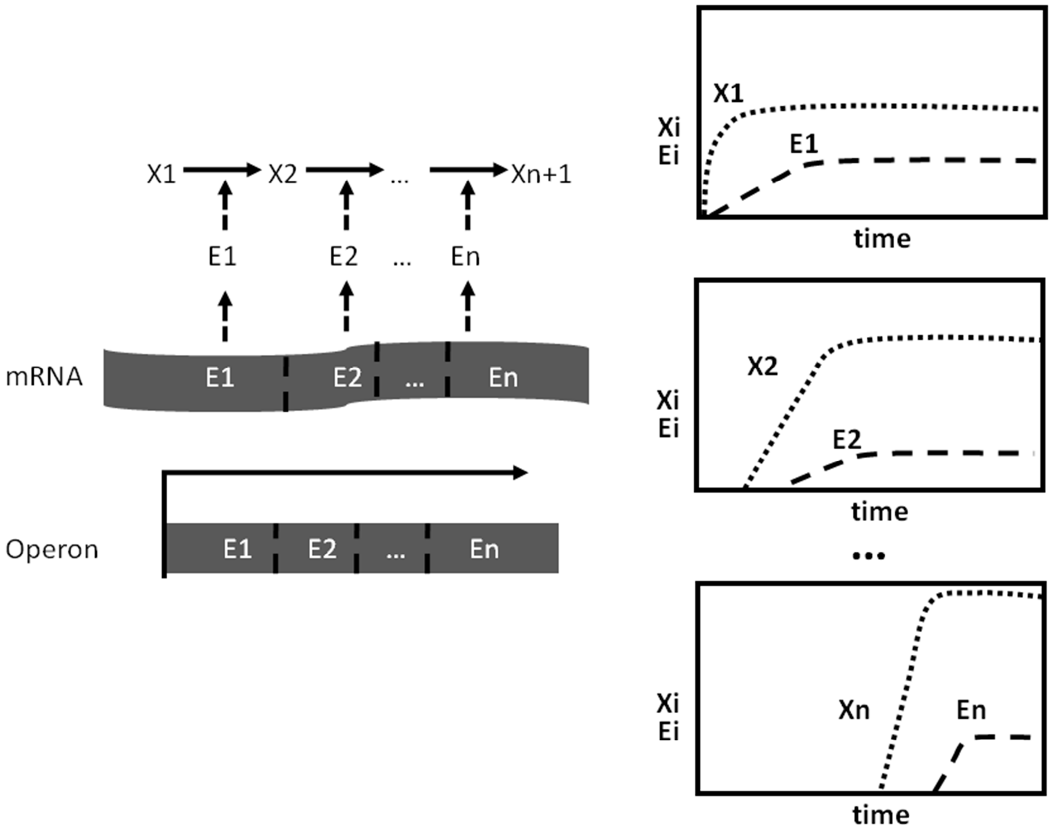


Figure 7. Design principles in translation of multicystronic mRNAs. In biosynthetic pathways, it appears that the accumulation of enzymes after translation lags behind the accumulation of the substrate for that enzyme. This makes biological sense, as the cell would not spend resources building enzymes before it needs them at sufficiently high concentrations (see text for details).

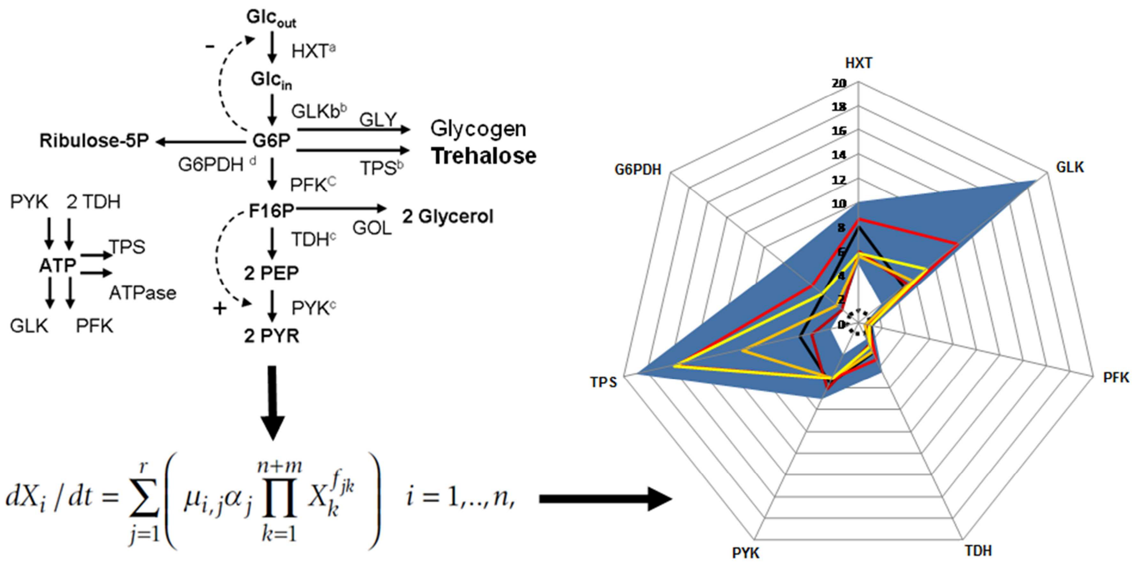


Figure 8. Design principles for changes in gene expression during stress response. A minimal model of metabolism that still accounts for important changes was built. Subsequently, this model was cast into non-linear form. Finally, global optimization methods were used to determine the ranges of changes in gene expression with respect to the basal level that would allow the cell to survive. These ranges are represented in blue in the spider plot on the right of the figure. Each axis of the graph represents one of the different genes in the model. Full lines indicate experimentally measured microarray profiles.

2.6.2. Design principles in RNA circuits

In the 1970s, Michael Savageau and co-workers found evidence for parallel processing as a design principle in RNA splicing. Such processing decreases the losses of immature intermediates, has shorter processing times, and is more amenable to evolutionary refinements [92]. The current surge of interest in RNA circuits has led to the identification of additional design principles in new types of RNA circuits [93-96]. For example, consider the following three regulatory mechanisms for riboswitch action: transcriptional termination, translational repression and mRNA destabilization. The ratio between reversible and irreversible rate constants is shown to have a critical impact on the performance of the circuit, establishing three operating regimes with distinct tuning properties.

Regulation of gene expression by small RNAs has also been analyzed [97, 98]. It was found that such regulation has features that are distinct from protein-mediated gene regulation. The strength of repression is set by the ratio between transcription rates of sRNA and the target gene: at target's high expression, sRNA may have no effect. The threshold value is tunable through controlling the rate of sRNA transcription. The model predicts reduced variance in protein level for sRNA-mediated regulation (attenuation of noise), and high sensitivity to changes in sRNA near the threshold. Different mRNA species are expected to compete for binding with the same pool of sRNA in a hierarchical crosstalk where targets of a given binding strength affect (but are not affected by) targets of lower binding strength. This form of regulation also provides a very fast temporal responsiveness, making sRNA mediated repression a good system when levels of mRNA need to shift reversibly and quickly in response to signals [99].

MicroRNAs (miRNA) are a class of short non-coding RNAs that post-transcriptionally control mRNA expression through degradation or translational repression. A distinctive feature of these molecules is that individual miRNAs can regulate a large number of mRNA targets, and each target gene can be regulated by multiple miRNAs, forming complex regulatory networks with target hubs. Pairs of miRNAs with very close binding sites show cooperative or synergistic behavior, while single miRNAs typically induce only mild repression to their targets. Mathematical modelling suggests that such collective miRNA repression induce fine-tuning and noise buffering in the regulation of gene expression, and is a means to overcome the low specificity inherent to regulation by each individual miRNA [100, 101]. Recent work suggests that miRNA cooperativity is a frequent mechanism for enhanced and efficient gene silencing by pairs of miRNAs in the human genome [102].

2.6.3. Design principles in metabolic networks

One of the first problems to be analyzed by means of Mathematical Controlled Comparisons was the regulation of a biosynthetic pathway by overall negative feedback of the end product to the first reaction of the pathway (Fig. 9). By comparing this design to other possible modes of feedback inhibition, it is seen that the overall negative feedback from the final product of an unbranched pathway to the first reaction of the pathway had several physiological advantages [15, 19, 103, 104]. These advantages include a production of the pathways' end product that is better regulated by cellular demand and less sensitive to spurious interactions with the environment. Later on, it was shown that overall feedback was the most functionally advantageous regulatory loop by inhibitory feedback that such pathways can have [105].

It was also found that a feedforward inhibition of the Amino-acyl-tRNA synthase by an intermediate of the amino acid biosynthesis pathway stabilizes that biosynthesis [103, 104]. Additionally, it was found that when reversible reactions are at the beginning of these pathways, regulation by demand is more effective, as is speed of adaptation to cellular demand signal [106].

Recently, it was found that the robustness of the activity of one of the enzyme isocitrate dehydrogenase in the glyoxylate bypass regulation relies, in addition to other known features of the system, on the existence of a ternary protein complex where the kinase activity is higher than the phosphatase activity. This model is quite general: it may apply to other systems with a bifunctional enzyme that catalyzes antagonistic reactions [107].

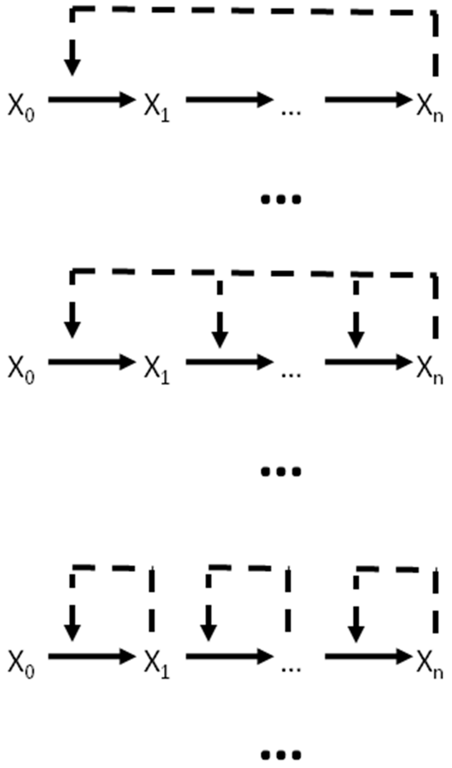


Figure 9. Design principles for negative feedback in unbranched biosynthetic pathways. All possible alternatives were considered. Even in the presence of additional feedback, overall feedback (top-most reaction scheme) increases the functional effectiveness of the circuit (see text for detail).

Other metabolic modules that have been analyzed in search for design principles are moiety conservation cycles. An analysis of the glucose 6-phosphate dehydrogenase (G6PD)–glutathione reductase (GSR) pathway, which catalyzes the reversible redox cycle of NADPH/NADP, found that each enzyme is designed with different functional demands. The activity of the NADP-reductive G6PD far exceeds the capacity of human erythrocytes for a steady NADPH supply, which is limited upstream of G6PD. The analysis indicates that maintaining such a surplus of G6PD activity ensures sufficient robustness of the NADPH concentration and responsiveness of the NADPH supply. These results suggest that large excess capacities found in some biochemical and physiological systems, rather than representing large safety factors, may reflect a close match of system design to unscrutinized performance requirements [44]. These results are complemented by the analysis of the kinetic activity of the GSR enzyme. The normal activity of GSR is under selective pressure by virtue of its ability to minimize the accumulation of oxidized glutathione. Contrary to the assumption of a single functional requirement, natural selection for the normal activities of the distinct enzymes in the pathway is mediated by different requirements. Much, if not most, of the enzymes may thus be fulfilling functional demands other than flux [43].

It was also found that even though negative feedback is often used in biochemical networks to achieve homeostasis, under certain conditions this feedback can cause the steady state to lose stability and be replaced by spontaneous oscillations of metabolites. The conditions for oscillation are: sufficient ‘‘memory’’ (or time delay) in the negative feedback loop, sufficient nonlinearity in the reaction kinetics, and proper balancing of the timescales of components in the loop [72]. Another interesting and well known result is that the coupling of positive feedback loops and the decrease of negative feedback loops in a network increase the stability of its steady state. A recent report that support this design principle analyzes both random networks and models of specific biological networks to conclude that concatenate negative feedback loops decrease the stability of steady states while concatenated positive feedback loops increase that stability [108].

Melendez-Hevia and co-workers used the pentose phosphate pathway game (Fig. 2; see above) to understand how some of the more central metabolic pathways have evolved. These researchers developed and used the pentose phosphate pathway game (Fig. 2) to build alternative pathways to get from one metabolite to another in a metabolic network. By combining constraints about the minimal number of carbon atoms that could be exchanged between metabolites with optimality principles favoring a minimal number of pathway steps between metabolites, they concluded that the principle of the minimal number of steps is consistent with pathway evolution in general [50, 51, 58-60]. Later, this method was combined with thermodynamic constraints and used to argue that glycolysis is quantitatively designed in an optimal way with respect to flux optimization, ATP production and ATP usage [109-111]. However, it should be stressed that different *a priori* thermodynamic constraints could change the results of this analysis. Sometime later Mittenthal and co-workers developed a more complex version of the game [112-114]. They generated alternative networks relaxing the number of carbons that could be exchanged between metabolites, included a larger fraction of irreversible reactions in the networks and considered additional types of reactions and inputs. Pathway evolution was shown to be consistent with the rules of the modified game, because the predicted pathway was the same as those observed in real organisms. Recently, an evolution of this method was applied to study if central metabolism in *E. coli* follows a similar optimality principle. The new rules consider that exchange of chemical groups between metabolites is limited by the functionality of enzymes described in the EC classification. With these rules, it was found that central metabolism is structured in a way that uses the minimal number of steps to connect the key precursor metabolites essential for biomass and energy production. Paths between consecutive precursors cannot be made shorter. The non-precursor compounds in the network form the shortest possible bridges between the precursors. Thus, central metabolism appears to be a minimal walk in chemical space between precursors [115]. This minimization of the number of steps between precursors could be driven by constraints imposed to the growth of *E. coli* by protein synthesis [116]. This biosynthetic process is often growth limiting, which would imply that cells with shorter pathways may have a competitive advantage due to their economy in proteins. Furthermore, short pathways have fewer intermediate and generate higher flux than long pathways of equally effective enzymes [117, 118]. This optimality principle allows making predictions: in organisms where a precursor is no longer essential, a shortcut would evolve that bypasses that precursor compound; and if a longer-than-minimal path is found between two compounds, an essential metabolite lies on that path. One question is that most pairs of precursors separated by more than one step could have been connected by several other alternative paths of the same length (but not shorter). Why the particular minimal path that occurs in the cell was selected out of these alternatives? Possibilities to explore include effects that can differentiate between paths of equal length, such as energy and reduction potential, toxicity effects of intermediate compounds and differential enzyme efficiency in each possible path.

Design principles at the molecular level have also started to be linked to macroscopic organism fitness. For example, ammonia was used to analyze a fitness tradeoff between resource abundant and resource limited environments for *S. cerevisiae*. This was done by analyzing the level of noise in Gdh1p expression and correlating it to the relative balance between resistance to toxic levels of ammonia and fitness in lower levels. It was found that as the noise in Gdh1p expression increased, this conferred enhanced resistance to ammonia toxicity. On the other hand, lower variation (noise) in Gdh1p levels exhibits greater fitness in physiological concentrations of ammonia [36].

Global metabolic responses have also been analyzed in search for design principles. For example, analyzing yeast data, it was found that the metabolic pathway map and the protein–protein interaction network (PIN) have significant positive correlation between the shortest paths across both network types. The sub-systems of the entire PIN appear to follow specific organizing principles: while physical interactions between proteins are generally dissortative (proteins of high degree interact with proteins of low degree), interactions between metabolic enzymes were observed to be assortative (enzymes frequently interact with other enzymes of similar degree or number of links associated with a node)[119].

Simple and robust growth laws connect growth rate with cell composition [120]. Growth rate of cells is maximized by interlocking two regulatory loops. They coordinate the amino acid flux between supply (amino acid synthesis and transport) and consumption (protein synthesis). One of the loops is the negative feedback by end-production inhibition of amino acids biosynthesis [19], discussed above. The second loop is the aminoacid supply-driven feedforward activation of ribosomal protein synthesis, restoring flux balance [121].

An analysis of *E. coli* cells cultured under different growth-limiting conditions shows that the regulation of cellular proteome can be understood in terms of the general function of proteins. If the proteins are partitioned into several types of function, activity of the proteins in each partition is regulated in a coordinated fashion to respond to the specific metabolic challenge limiting the growth rate. Despite the complexity in the molecular details of the adaptive response, this coarse-grained approach suggests a principle for resource allocation in proteome economy [122]. Such top-down view, as in the characterization of the state of a gas through macroscopic measures such as temperature and pressure, captures the collective behavior of the metabolic network and gives a simple quantitative picture of the global regulation of the metabolic response that can be profitably used in future omics studies.

2.6.4. Design principles in cellular rhythms

The presence of a negative feedback loop in a network is a necessary condition for that network to be able to generate oscillations. Thus, different topological circuits can be associated with this dynamical behavior [72]. Oscillatory phenomena are the basis of cellular rhythms and may be found in different contexts, from metabolism [123] to development [124] and circadian rhythms [125]. Understanding the fundamental biological design principles underlying the networks generating such cyclic behaviors is an important question.

One of the most well studied cellular oscillators is cell cycle. Basic design principles have been identified for the networks regulating the cellular process.

Models created by using molecular information suggest that the molecular mechanism regulating the eukaryotic cell cycle is composed of two bistable switches (governing G1-S and G2-M transitions) and an oscillator (controlling mitotic exit) [126]. The bistable switches are controlled by a molecular antagonism between CDKs and their antagonists. This switch has two alternative states: G1 (low CDK activity) and S-G2-M (high CDK activity) [127-129]. ''Starter Kinases'' (SKs) and ''Exit Proteins'' (EPs) flip the switch back and forth. Transitions between these states are controlled by two negative-feedback loops. The Start transition (G1-S) is triggered by a class of SKs that are down-regulated by the very species they are aiding. The Exit transition (M to G1) is promoted by a class of EPs that kill the very species they depend on. This topology creates a dynamic of irreversible transitions. Start and Exit checkpoints block progression through the cycle if any serious problems are encountered (DNA damage blocks Start, incorrect chromosome alignment block Exit). A size checkpoint at the Start transition ensures balanced growth and division. This control system of cell cycle regulation has four fundamental properties: alternation of S and M; check-points; irreversibility; balanced growth and division. Variations of this model also account for alternative modes of cell division, such as oogenesis (cell growth without division), fertilized egg division (rapid mitotic cycle without growth), endoreplication (repeated rounds of DNA synthesis without mitosis) and meiosis. Recent work by the groups of Nurse and Cross suggests that the different cell cycles have evolved from duplication and divergence from a primordial cell cycle with a single cyclin. The accumulation of this cyclin throughout the cell cycle allowed for the progression of the cycle. Cell division led to an abrupt decrease in that concentration, restarting the cycle [128, 130-133].

Other important biological oscillators are the networks responsible for regulating the circadian rhythm of organisms. These biological processes appear to have evolved independently for different groups of organisms [134]. For example, the proteins that regulate the circadian clocks of cyanobacteria and those of multicellular organisms evolved from different ancestors and generated networks that have diverse regulatory loops. On top of a stable oscillation, the networks of genes and proteins responsible for the circadian clock need also appropriate mechanisms for input signals that are required to reset and entrain the clock when conditions change. Inputs that are known to entrain the clock include light, temperature, and food. All known circadian clock networks use a multi-loop structure to obtain circadian oscillations that can, in principle be obtained with a single negative feedback loop. The presence of these multiple feedback loops appear to provide the clocks with higher flexibility that allows these clocks to be entrained and have their phase more easily reset by the input signals, while remaining fairly insensitive to noise and having a robust period [135, 136]. This makes evolutionary and biological sense, because organisms on earth have a constant circadian period that often requires phase resets either due to changes in the day–night cycle or to moves between different time zones. A linear analysis of a non-mechanistic model for the circadian clock of *Arabidopsis* further suggests that the circadian clock of this plant requires a mechanism for rapid light inputs if the clock is to adjust to photoperiod-dependent changes [137]. More complex instances of circadian clocks have also been analyzed. For example, in mammals, several thousand neurons of the suprachiasmic nucleus generate rhythms of approximately 24 h [138]. A mathematical model of the system suggests that the neurotransmitter feedback loop plays an important role in the appropriate synchronization of the ligth/dark cycles, allowing the network to resynchronizing the clock after a perturbation that simulates a 'jet-lag' of several hours. Other design principles have been proposed for networks to achieve phase-splitting behavior [139].

Another important issue about the networks that regulate biological rhythms is to understand in which situations one can expect the networks that regulate each autonomous rhythm to interact. Furthermore, how does that interaction benefit the fitness of the organism? Finally, are there specific modes of interaction (design principles) that improve the functional effectiveness of the interactions under different conditions?

The answer to the first question is positive [140]. Cell cycle is also regulated by the circadian clock in *Synechococcus elongatus* [141] and in mice [140]. This regulation is consistent with a model where cell cycle rate decreases during the night [142]. The structure of the network that integrates both oscillators is still unclear. Thus, the answers to the second and third question are still missing. Nevertheless, in *S. elongatus*, a phosphorylation cascade of circadian clock proteins that signal to the putative transcription factor RpaA is involved in linking the two processes. It is tempting to speculate that in a photosynthetic organism such as *S. elongatus* it would make physiological sense to decrease the rate of cell cycle during the night, as the main source of energy for the cell is shut-off. If availability of resources is an important selective pressure in the coupling of the circadian and cell cycle oscillators, one might expect that cells from diurnal animals will go through cell cycle faster during the day, while cells from nocturnal animals will have a faster cell cycle during the night. An analysis of available data for nocturnal rodents is consistent with this prediction (see figures in [143, 144]).

Another strategy that can be used to generate oscillations provides a simple mechanism for coupling these oscillations to cell cycle. This strategy is based on the transient gene dosage imbalance caused by the location of two genes in opposite sides of the bacterial chromosome. This simple mechanism has been observed in the phosphorelay controlling sporulation in *Bacillus subtilis* and, along with a delayed negative feedback-loop between the proteins of the phosphorelay, is responsible for cell-cycle coordinated pulses of the sporulation master regulator Spo0A following DNA replication [145].

2.6.5. Design principles in signal transduction networks

Signal transduction is another area where design principles have been studied, both in prokaryotes and eukaryotes. The identification of many types of design principles for these networks has been reported. Here, we will discuss only a few of these reports, focusing mostly on phosphorylation cascades, both in prokaryotes and in eukaryotes.

In prokaryotes, signal transduction through phosphorylation events is mediated by Two Component Systems (TCS) or Phosphorelays (PR). In these systems, a sensor protein modifies its own phosphorylation state in response to some signal from the environment. The phosphate is then transferred to a response regulator protein that either modulates physiological response (in TCS) or transfers it again to a second histidine kinase that will subsequently transfer the same phosphate to a second response regulator (in PR; see Fig. 10). TCS are ubiquitous in bacteria, and homologous pathways have been identified in several eukaryotic organisms as well, including *S. cerevisiae*, *Arabidopsis thaliana, Neurospora crassa and Dictyostelium discoideum*.

The modular aspect of TCS and PR circuits has facilitated the evolution of a variety of signal transduction modules. One circuitry motif that exemplifies this versatility is the four-step His-Asp-His-Asp PR. Different PR show the same alternating pattern of histidine and aspartate phosphorylation sites, but can utilize a different pattern of covalent linkage between individual protein domains: the four phosphorylation sites of the Kin-Spo0 pathway (in *Bacillus subtilis*) are found in independent proteins, whereas one protein can join the first two or three members of the PR (Sln1p-Ypd1p-Ssk1p and BvgS-BvgA pathways, in *S. cerevisiae* and *Bordetella pertussis*, respectively). The discovery that the yeast Sln1 pathway employs a PR mechanism with the same His-Asp-His-Asp configuration reported for the Kin-SpoO and BvgS-BvgA systems suggests that this signaling strategy may be widely utilized by eukaryotes as well as prokaryotes. However, it appears to be absent in mammals [146-149].

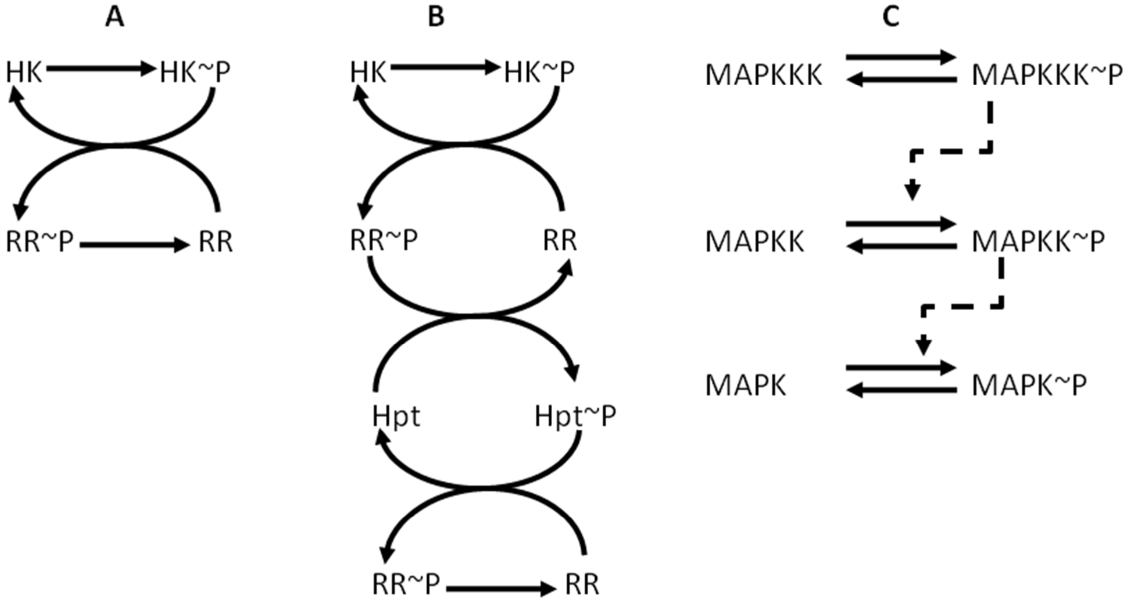


Figure 10. Prototypical Two Component Systems (A), Phosphorelays (B), and MAP kinase cascades (C). HK – histine kinases; RR – response regulators; HPt – intermediate phosphotransfer protein, accepting phosphate on a histidine. MAPK – MAP kinase; MAPKK – MAPK kinase; MAPKKK – MAPKK kinase; ~P – phosphorylated form of the proteins. See text for mechanistic details.

Several aspects of the physiological regulation by TCS have been analyzed. One of these is the apparent insensitivity of the input–output relationship of TCS modules to changes in the concentrations of the system's components [54]. It was found that this insensitivity can justify a design of the TCS that require three biochemical features: (i) ATP dependence of dephosphorylation; (ii) sensor kinase bifunctionality (the sensor catalyzes the phosphorylation of the response-regulator but also the dephosphorylation of the phosphorylated RR); and finally, (iii) the two-step nature of the sensor kinase (autophosphorylation and phosphotransfer) [150]. In contrast, it was found that TCS mediating responses that require hysteresis should have a channel for response regulator (RR) dephosphorylation that is independent from the sensor protein. In addition it is also required that the dephosphorylated forms of sensor and RR form a reversible dead-end complex [38, 151]. It has also been shown that TCS modules where the sensor kinase is bifunctional should be preferentially selected in physiological responses that need to be buffered against crosstalk, while TCS with monofunctional sensors should be selected in situations where the physiological response requires the integration of signals [151]. However, the use of signaling pathways with multiple inputs and a single output entails a loss of information about input signals. How cells integrate information from multiple inputs to modulate their gene expression states is poorly understood. Information theory can be adapted to study a biological circuit performing information processing and signal integration. The analysis of quorum sensing in *Vibrio harveyi* revealed that information transmission is primarily limited by interference from other signals, not by noise. Cells must tune the kinase activity of each signaling branch of the quorum sensing circuit to simultaneously learn about individual inputs. Cells can increase how much they learn about individual signals by manipulating the different autoinducer production rates. Bacteria can learn preferentially about a particular input in a particular environment by using simple feedback loops to control receptor numbers. This analysis suggests that the need to minimize interference between signals probably imposes strong constraints on the design of signal-integration networks [152].

Some TCSs are positively autoregulated: the regulon controlled by active RR often includes the TCS operon, leading to a feedback loop. Positive autoregulation does not necessarily give rise to overall positive feedback. Mathematical model analysis shows that the effective sign of this feedback is determined by the values of the kinetic parameters of the system, making TCSs capable of tuning feedback sign, switching between positive and negative feedback to achieve appropriate outputs in different circumstances. Attainment of negative feedback depends on sensor bifunctionality (so that the sensor protein of the TCS can both increase and decrease the fraction of active RR) and RR activation independent of its cognate sensor. The feedback sign is physiologically relevant, since negative feedback reduces noise and gives rise to fast overshooting responses and positive loops lead to bistability, phenotypic heterogeneity and a stronger learning effect [153].

How does feedback lead to bistability [154]? The effect of the interplay of two positive feedbacks on the network bistability has been studied theoretically and experimentally. One example is the mycobacterial stress-response network which consists of the MprA/MprB TCS along with the σ^E-RseA sigma/anti-sigma factor system, involved in persistence in mycobacteria. This network contains two positive feedback loops. Positive autoregulation of the *mpr*AB operon by MprA-P gives rise to a positive feedback. A second positive feedback arises from the transcriptional activation of σ^E by MprA-P and subsequent upregulation of *mpr*AB operon by σ^E. The analysis of reduced versions of the network, to understand the role of each component, shows that the second feedback involving σ^E makes the network bistable, but only due to the post-translational regulation of σ^E by its anti-sigma factor RseA, which increases effective cooperativity and leads to bistability [155]. Bifunctionality of the sensor kinase avoids bistability in the positively regulated TCS.

Recently, the effect of the number of steps in the signaling of PR cascades was analyzed [156]. Under simplifying mechanistic assumptions, models for cascades with less than four steps are not capable of ultrasensitivity responses to signals. Thus, the authors suggest that 4-step PR cascades are the simplest evolutionary solution to the problem of high signal amplification in bacterial signal transduction.

Despite the simplicity of regulatory loops in TCS and PR signaling pathways, these are capable of exhibiting complex temporal dynamics both on short and long timescales [157]. For example, positive autoregulation in the PhoB/PhoR TCS provides a regulatory mechanism that allows cells to adapt to changing environments by expressing different optimal levels of PhoB and PhoR proteins [158]. Another example of a sophisticated dynamics is found when, upon nutrient-limited conditions, the PR responsible for sporulation initiation in *B. subtilis* shows a pulsatile level of its output molecule, the phosphorylated master regulator Spo0A [159]. These series of pulses are successively larger and span over the course of several cell cycles, until a threshold is reached and cells commit to sporulation. As mentioned in the previous subsection, this pulsatile behavior is the consequence of a negative feedback-loop in the PR between Spo0F and KinA (substrate inhibition of the histidine kinase), which makes the system very sensitive to the ratio of KinA and Spo0F concentrations. As a result, any perturbation of this ratio can force the system to produce a pulsed response [145]. On the evolutionary timescale, these pathways have expanded in many species to respond to a wide range of stimuli. This expansion has been driven by gene duplication and the subsequent diversification of specificity residues in the HK and RR, coevolving to retain their interaction while becoming insulated from their counterparts [160].

The eukaryotic equivalent of TCS and PR are MAP cascades (Fig. 10). These cascades are composed of three proteins. The first step in the cascade is the MAPKKK protein. It becomes phosphorylated in response to some signal and it in turn phosphorylates the second proteins of the cascade, the MAPKK. MAPKKs in turn phosphorylate MAPK, which then regulate the physiological response. Unlike TCS and phosphorelays, ATP is consumed in each phosphorylation event in MAPKs. It was shown that this type of signal transmission could account for high signal amplification [161, 162], and that the most energy efficient way to regulate this signal transduction is by signaling both the phosphorylating and dephosphorylating enzymes that control the cascade [163]. Such amplification depended on the existence of a highly cooperative mechanism in the phosphorylation of the proteins in the cascade and on an increase in the concentration of protein in each subsequent step of the cascade. Nevertheless, several questions about the design of these cascades remain unanswered.

For example, why do MAPK cascades use three kinases instead of one? (other membrane-to-nucleus signaling pathways, such as the cAMP/protein kinase A and the Jak/Stat pathways, employ a single kinase). A numerical analysis of a MAPK cascade model shows that, with typical parameter values, the three step cascade behaves like a highly cooperative enzyme, even if none of the individual enzymes is regulated cooperatively. The degree of ultrasensitivity increases as the cascade is descended and depends critically on the assumption that the dual phosphorylation of MAPKK and MAPK occurs through a two-collision mechanism [164]. Thus, MAP cascades can convert graded inputs into switch-like outputs, filter out noise and flip from off to on over a narrow range of input stimuli. This sort of behavior would be appropriate for a signaling system that mediates processes where cells switch rapidly between discrete states without assuming intermediate positions, like in mitogenesis, cell-fate induction, and oocyte maturation.

Other questions that regard the design of MAP cascades concern the relationship between the concentrations of the enzymes in the three steps of the cascade [165-167]. Computational analysis provides rationale for why the MAPK and MAPKK concentrations are similar. The response time of the cascade is critically dependent on specific combinations of ranges of cellular MAPK and MAPKK concentrations. Concentrations of these signaling components fall within a region where the cascade seems to achieve optimal efficiency and rapid activation. When the MAPKK concentration becomes very different from the concentration of MAPK an undesirable delay is predicted in the response. Both increases and decreases in the MAPK and MAPKK concentrations result in a reduction in the efficiency of this initial response [166]. The way that MAPK cascades interact has also been analyzed. Interacting MAPK cascades are capable of implementing useful logic and amplitude-dependent signal processing functions (‘‘exclusive-or’’ function and an in-band detector or two-sided threshold) and their implementation requires only limited crosstalk. This behavior cannot be achieved with a single cascade or with non-interacting cascades. A significant challenge still remaining is to determine if this potential is actually realized in the cell and if the computationally evolved solution resembles the solution chosen in the evolution of life. We also have yet to consider the cascade in a larger context, embedded in feedback loops, engaged in crosstalk with other signaling networks or protected from crosstalk by scaffolds [167].

As mentioned above, signal transduction networks regulate their response using (typically negative) feedback loops. Such down-regulation of the response to signals can increase the correlation between the input and the output of the network [168, 169]. Recent work suggests that evolution of feedback as a mechanism to regulate the response in signal transduction networks must optimize opposing goals. On one hand this mechanism should increase the correlation between signal and output. On the other hand, it should be able to decrease the transmission of noise through the network. A network that maximizes the correlation signal-output also increases the effect of noise on that output [170]. This is easy to understand because by perfectly correlating input and output, a network will also perfectly correlate noise in the input to noise in the output. Thus, depending on the particular system one might expect feedback loops that preferably buffer the response of the network against noise, while in other systems the feedback loops will preferably maximize the correlation between input and output.

2.7. Final remarks

To be able to write this chapter we struggled with the question of what is a biological design principle. The definition we gravitated towards is by no means the only one available. However, once we accepted it as a working definition, we could review some of the work that has improved our understanding of such principles in molecular circuits. The importance of that work is justified because it improves our understanding of how biology works. The appropriateness of considering functional effectiveness of molecular circuits rather than fitness of the whole organism in the analysis is also discussed in this review. After establishing a framework for thinking about design principles, we discuss the different theoretical and mathematical methods that are usually applied to study them. We finish by presenting examples of those principles in different types of molecular circuits. We restricted the discussion mostly to intracellular networks, with some exceptions [86]. This means that most of the work that deals with design principles in molecular networks that regulate development is not included (for example, see [22, 74, 171-173]). Nevertheless, the examples given here present a general view of the research in this field.

Considering the work reviewed and presented here, one could feel that many of the design principles are somewhat *ad hoc* and too system specific. This view raises the important question of whether, over time, something like a ‘‘periodic table’’ of universal design principles that are valid for all types of biological circuits can be built. In other words, can we identify network elements, either qualitative or quantitative, that are almost always associated with specific types of behavior?

There appear to be cases where the answer is positive. For example, it is well known that the existence of a positive feedback loop is a necessary condition for multistability in molecular networks [154]. Also, a sort of ‘‘uncertainty principle’’ was proposed for feedback in biological systems [170]. This principle roughly states that feedback can be used to maximize correlation between input and output of a biological system at the cost of increasing noise amplification or used to decrease noise amplification at the cost of decreasing correlation between input and output. This imposes fundamental limits to how much evolution can optimize responses to noise in molecular systems through the evolution of feedback interactions. Results of Reaction Network Theory that relate the structure of the network with the possibility of different types of dynamical behaviors may also fit into this category of basic design principles [40, 48, 51, 54, 55]. The common link between all these principles is the fact that they are independent of the specific function of the circuit being analyzed and represent hard constraints to dynamical behavior imposed by network structure.

As opposed to these ‘‘elementary’’ design principles, most of the principles discussed in this review hinge heavily on understanding the function of the circuit under analysis. Showing that a given feature improves the function of the circuit is crucial to explain why that feature is fixed during evolution. Such features are specific elements in the network (for example, bifunctionality in bacterial two component systems [151]), particular ranges of parameter values that enable a given dynamic response (for example, survival during heat shock adaptation in yeast [62]), or both (for example, only specific network designs with a given range of parameter values permit creating a developmental system with one stripe [74]). Take the analogy of a ‘‘periodic table of design principles’’ a bit further, many of the principles discussed in this special issue may be more like ‘‘molecules’’, for which no periodic table exists, rather than like ‘‘atoms’’, for which it does.

This does not in any way demeans the usefulness of these principles for understanding the way biological systems work and how they came to be as they are. If fact, an engineer might argue that proof of understanding of a system comes from building instances of the system that work under different regimes and demand specifications. From this perspective, creating more restricted catalogues that associate a specific functional behavior in a given type of system to a specific design element for that system may be more useful that a general periodic table. Such catalogues could become extremely useful for Synthetic Biology, enabling the construction of artificial biological circuits of a certain type with specific properties and behavior.

Synthetic Biology is the major body of work that is absent from this review. This choice was made because many good and extensive reviews on the subject have been published recently. We refer the readers to some of those reviews for more details [174-190]. Researchers are using decades of accumulated molecular knowledge to engineer new circuits within organisms that either implement new functionality or test some of the predictions made in the past through the analysis of design principles (see, for example, [16, 191]). Synthetic biologists design and implement non-naturally occurring biological networks that perform a given function. Identification of design principles, on the other hand, focuses on understanding the emergence of these designs from evolution. Both activities are complementary and design principles can greatly assist and guide the development of Synthetic Biology applications (see, for example, [192] for a more detailed discussion on this subject). The merging of Design Principle analysis to Synthetic Biology creates a field of opportunities that may immensely potentiate our understanding of how organisms work at the molecular level and why they came to work like they do [193].

Biomedical research is another area that may in the future benefit from the study of biological design principles. If principles that guide shifts between pathogenic and healthy states can be identified, these can be used to devise strategies for better treatments. Furthermore, host-pathogen interactions might also have evolved in such a way that these interactions and their regulation can be classified into a small set of principles that can be used to facilitate host survival.

In summary, it seems to us that there may come a time when a hierarchy of design principles will need to be established and accepted for molecular networks. It is hard to imagine what such a hierarchy will look like. One possibility is that it becomes organized along the lines discussed above. It could be that there will be a set of design principles that are universal and constrained by network structure. Then, on top of these, and specific to the networks that regulate the biological processes of interest, one will identify principles that explain if and why such networks have been selected to perform the process. If this is the case, then we believe that the work reviewed here constitutes a very encouraging head start towards the goal of such a classification.

2.8. References

1. Alon, U., *Biological networks: the tinkerer as an engineer.* Science, 2003. **301**(5641): p. 1866-7.
2. *The tinkerer's accomplice: how design emerges from life itself.* Choice: Current Reviews for Academic Libraries, 2007. **44**(8): p. 1364-1364.
3. Alon, U., *An Introduction to Systems Biology: Design Principles of Biological Circuits*2006: Chapman and Hall/CRC.
4. Banerjee, R. and D. Roy, *Codon usage and gene expression pattern of Stenotrophomonas maltophilia R551-3 for pathogenic mode of living.* Biochemical and Biophysical Research Communications, 2009. **390**(2): p. 177-181.
5. Sharp, P.M., L.R. Emery, and K. Zeng, *Forces that influence the evolution of codon bias.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2010. **365**(1544): p. 1203-1212.
6. Vilaprinyo, E., R. Alves, and A. Sorribas, *Minimization of Biosynthetic Costs in Adaptive Gene Expression Responses of Yeast to Environmental Changes.* PLoS Comput Biol, 2010. **6**(2).
7. Szilagyi, A., D. Gyorffy, and P. Zavodszky, *The twilight zone between protein order and disorder.* Biophysical Journal, 2008. **95**(4): p. 1612-1626.
8. Minary, P. and M. Levitt, *Probing protein fold space with a simplified model.* J Mol Biol, 2008. **375**(4): p. 920-933.
9. Melendez, R., et al., *Physical constraints in the synthesis of glycogen that influence its structural homogeneity: A two-dimensional approach.* Biophysical Journal, 1998. **75**(1): p. 106-114.
10. Melendez, R., E. MelendezHevia, and M. Cascante, *How did glycogen structure evolve to satisfy the requirement for rapid mobilization of glucose? A problem of physical constraints in structure building.* Journal of Molecular Evolution, 1997. **45**(4): p. 446-455.
11. Irvine, D.H. and M.A. Savageau, *Network Regulation of the Immune-Response - Modulation of Suppressor Lymphocytes by Alternative Signals Including Contrasuppression.* Journal of Immunology, 1985. **134**(4): p. 2117-2130.
12. Irvine, D.H. and M.A. Savageau, *Network Regulation of the Immune-Response - Alternative Control Points for Suppressor Modulation of Effector Lymphocytes.* Journal of Immunology, 1985. **134**(4): p. 2100-2116.
13. Savageau, M.A., *Concepts Relating Behavior of Biochemical Systems to Their Underlyin Molecular Properties.* Archives of Biochemistry and Biophysics, 1971. **145**(2): p. 612-&.
14. Savageau, M.A., *Genetic Regulatory Mechanisms and Ecological Niche of Escherichia-Coli.* Proc Natl Acad Sci U S A, 1974. **71**(6): p. 2453-2455.
15. Savageau, M.A., *Optimal design of feedback control by inhibition.* Journal of Molecular Evolution, 1974. **4**(2): p. 139-56.
16. Savageau, M.A., *Comparison of classical and autogenous systems of regulation in inducible operons.* Nature, 1974. **252**(5484): p. 546-9.
17. Savageau, M.A., *Kinetic Organization of Biosynthetic Regulatory Systems in Bacteria.* Abstracts of Papers of the American Chemical Society, 1974: p. 26-26.
18. Savageau, M.A., *Selection of Positive and Negative Mechanisms of Genetic-Control in Enteric Bacteria.* Federation Proceedings, 1974. **33**(5): p. 1464-1464.
19. Savageau, M.A., *Optimal design of feedback control by inhibition: dynamic considerations.* Journal of Molecular Evolution, 1975. **5**(3): p. 199-222.

20. Savageau, M.A., *Biochemical systems analysis: a study of function and design in molecular biology*1976: Addison-Wesley.

21. Voit, E.O., *Canonical nonlinear modelling*1991: Springer.

22. Alves, R. and M.A. Savageau, *Extending the method of mathematically controlled comparison to include numerical comparisons.* Bioinformatics, 2000. **16**(9): p. 786-98.

23. Schwacke, J.H. and E.O. Voit, *Improved methods for the mathematically controlled comparison of biochemical systems.* Theor Biol Med Model, 2004. **1**: p. 1.

24. Coelho, P.M., A. Salvador, and M.A. Savageau, *Quantifying global tolerance of biochemical systems: design implications for moiety-transfer cycles.* PLoS Comput Biol, 2009. **5**(3): p. e1000319.

25. Savageau, M.A., et al., *Phenotypes and tolerances in the design space of biochemical systems.* Proc Natl Acad Sci U S A, 2009. **106**(16): p. 6435-40.

26. Savageau, M.A. and R.A. Fasani, *Qualitatively distinct phenotypes in the design space of biochemical systems.* FEBS Lett, 2009. **583**(24): p. 3914-22.

27. Clutton-Brock, T. and B.C. Sheldon, *Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology.* Trends Ecol Evol, 2010. **25**(10): p. 562-73.

28. Vogel, C., S.A. Teichmann, and J. Pereira-Leal, *The relationship between domain duplication and recombination.* J Mol Biol, 2005. **346**(1): p. 355-365.

29. Price, M.N., P.S. Dehal, and A.P. Arkin, *Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli.* Genome Biol, 2008. **9**(1): p. R4.

30. Price, M.N., A.P. Arkin, and E.J. Alm, *The life-cycle of operons.* PLoS Genet, 2006. **2**(6): p. e96.

31. Kashtan, N., et al., *Extinctions in heterogeneous environments and the evolution of modularity.* Evolution, 2009. **63**(8): p. 1964-75.

32. Kashtan, N., et al., *An analytically solvable model for rapid evolution of modular structure.* PLoS Comput Biol, 2009. **5**(4): p. e1000355.

33. Parter, M., N. Kashtan, and U. Alon, *Facilitated Variation: How Evolution Learns from Past Environments To Generalize to New Environments.* PLoS Comput Biol, 2008. **4**(11).

34. Parter, M., N. Kashtan, and U. Alon, *Environmental variability and modularity of bacterial metabolic networks.* Bmc Evolutionary Biology, 2007. **7**.

35. Kashtan, N. and U. Alon, *Spontaneous evolution of modularity and network motifs.* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13773-13778.

36. Bayer, T.S., et al., *Synthetic control of a fitness tradeoff in yeast nitrogen metabolism.* J Biol Eng, 2009. **3**: p. 1.

37. Cagatay, T., et al., *Architecture-dependent noise discriminates functionally analogous differentiation circuits.* Cell, 2009. **139**(3): p. 512-22.

38. Igoshin, O.A., R. Alves, and M.A. Savageau, *Hysteretic and graded responses in bacterial two-component signal transduction.* Mol Microbiol, 2008. **68**(5): p. 1196-215.

39. Coelho, P.M., A. Salvador, and M.A. Savageau, *Relating mutant genotype to phenotype via quantitative behavior of the NADPH redox cycle in human erythrocytes.* PLoS One, 2010. **5**(9).

40. Savageau, M.A., *Parameter Sensitivity as a Criterion for Evaluating and Comparing Performance of Biochemical Systems.* Nature, 1971. **229**(5286): p. 542-&.

41. Kitano, H., *Towards a theory of biological robustness.* Mol Syst Biol, 2007. **3**: p. 137.

42. Rao, C.V. and A.P. Arkin, *Control motifs for intracellular regulatory networks.* Annu Rev Biomed Eng, 2001. **3**: p. 391-419.

43. Salvador, A. and M.A. Savageau, *Evolution of enzymes in a series is driven by dissimilar functional demands.* Proc Natl Acad Sci U S A, 2006. **103**(7): p. 2226-31.

44. Salvador, A. and M.A. Savageau, *Quantitative evolutionary design of glucose 6-phosphate dehydrogenase expression in human erythrocytes.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14463-8.
45. Craciun, G. and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: I. The injectivity property.* Siam Journal on Applied Mathematics, 2005. **65**(5): p. 1526-1546.
46. Craciun, G. and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models.* Syst Biol (Stevenage), 2006. **153**(4): p. 179-86.
47. Craciun, G. and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: II. The species-reaction graph.* Siam Journal on Applied Mathematics, 2006. **66**(4): p. 1321-1338.
48. Craciun, G. and M. Feinberg, *Multiple Equilibria in Complex Chemical Reaction Networks: Semiopen Mass Action Systems.* Siam Journal on Applied Mathematics, 2010. **70**(6): p. 1859-1877.
49. Craciun, G., Y.Z. Tang, and M. Feinberg, *Understanding bistability in complex enzyme-driven reaction networks.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8697-8702.
50. Feinberg, M., *Reaction Network Structure and Multiple Steady-States in Complex Isothermal Reactors.* Abstracts of Papers of the American Chemical Society, 1985. **189**(Apr-): p. 35-Inde.
51. Feinberg, M., *The existence and uniqueness of steady states for a class of chemical reaction networks.* Archive for Rational Mechanics and Analysis, 1995. **132**(4): p. 311-370.
52. Feinberg, M., *Multiple steady states for chemical reaction networks of deficiency one.* Archive for Rational Mechanics and Analysis, 1995. **132**(4): p. 371-406.
53. Schlosser, P.M. and M. Feinberg, *A Theory of Multiple Steady-States in Isothermal Homogeneous Cfstrs with Many Reactions.* Chemical Engineering Science, 1994. **49**(11): p. 1749-1767.
54. Shinar, G., U. Alon, and M. Feinberg, *Sensitivity and Robustness in Chemical Reaction Networks.* Siam Journal on Applied Mathematics, 2009. **69**(4): p. 977-998.
55. Shinar, G. and M. Feinberg, *Structural Sources of Robustness in Biochemical Reaction Networks.* Science, 2010. **327**(5971): p. 1389-1391.
56. Knight, D., G. Shinar, and M. Feinberg, *Sharper graph-theoretical conditions for the stabilization of complex reaction networks.* Math Biosci, 2015. **262**: p. 10-27.
57. Shinar, G. and M. Feinberg, *Concordant chemical reaction networks and the Species-Reaction Graph.* Math Biosci, 2013. **241**(1): p. 1-23.
58. Melendez-Hevia, E., *The game of the pentose phosphate cycle: a mathematical approach to study the optimization in design of metabolic pathways during evolution.* Biomed Biochim Acta, 1990. **49**(8-9): p. 903-16.
59. Melendez-Hevia, E. and A. Isidoro, *The game of the pentose phosphate cycle.* J Theor Biol, 1985. **117**(2): p. 251-63.
60. Melendez-Hevia, E. and N.V. Torres, *Economy of design in metabolic pathways: further remarks on the game of the pentose phosphate cycle.* J Theor Biol, 1988. **132**(1): p. 97-111.
61. Sorribas, A., et al., *Optimization and evolution in metabolic pathways: global optimization techniques in Generalized Mass Action models.* J Biotechnol, 2010. **149**(3): p. 141-53.
62. Vilaprinyo, E., R. Alves, and A. Sorribas, *Use of physiological constraints to identify quantitative design principles for gene expression in yeast adaptation to heat shock.* Bmc Bioinformatics, 2006. **7**.

63. Voit, E.O. and T. Radivoyevitch, *Biochemical systems analysis of genome-wide expression data.* Bioinformatics, 2000. **16**(11): p. 1023-1037.

64. Guillen-Gosalbez, G. and A. Sorribas, *Identifying quantitative operation principles in metabolic pathways: a systematic method for searching feasible enzyme activity patterns leading to cellular adaptive responses.* Bmc Bioinformatics, 2009. **10**.

65. Igoshin, O.A., C.W. Price, and M.A. Savageau, *Signalling network with a bistable hysteretic switch controls developmental activation of the sigma(F) transcription factor in Bacillus subtilis.* Mol Microbiol, 2006. **61**(1): p. 165-184.

66. Hlavacek, W.S. and M.A. Savageau, *Rules for coupled expression of regulator and effector genes in inducible circuits.* J Mol Biol, 1996. **255**(1): p. 121-139.

67. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-11985.

68. Hlavacek, W.S. and M.A. Savageau, *Completely uncoupled and perfectly coupled gene expression in repressible systems.* J Mol Biol, 1997. **266**(3): p. 538-58.

69. Savageau, M.A., *Demand theory of gene regulation. I. Quantitative development of the theory.* Genetics, 1998. **149**(4): p. 1665-76.

70. Savageau, M.A., *Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of Escherichia coli.* Genetics, 1998. **149**(4): p. 1677-91.

71. Bose, I., B. Ghosh, and R. Karmakar, *Motifs in gene transcription regulatory networks.* Physica a-Statistical Mechanics and Its Applications, 2005. **346**(1-2): p. 49-57.

72. Novak, B. and J.J. Tyson, *Design principles of biochemical oscillators.* Nature Reviews Molecular Cell Biology, 2008. **9**(12): p. 981-991.

73. Alon, U., *Network motifs: theory and experimental approaches.* Nature Reviews Genetics, 2007. **8**(6): p. 450-461.

74. Cotterell, J. and J. Sharpe, *An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients.* Mol Syst Biol, 2010. **6**.

75. Dekel, E., S. Mangan, and U. Alon, *Environmental selection of the feed-forward loop circuit in gene-regulation networks.* Physical Biology, 2005. **2**(2): p. 81-88.

76. Wall, M.E., M.J. Dunlop, and W.S. Hlavacek, *Multiple functions of a feed-forward-loop gene circuit.* J Mol Biol, 2005. **349**(3): p. 501-14.

77. Savageau, M.A., *Regulation of differentiated cell-specific functions.* Proc Natl Acad Sci U S A, 1983. **80**(5): p. 1411-5.

78. Hlavacek, W.S. and M.A. Savageau, *Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits.* J Mol Biol, 1995. **248**(4): p. 739-55.

79. Wall, M.E., W.S. Hlavacek, and M.A. Savageau, *Design principles for regulator gene expression in a repressible gene circuit.* J Mol Biol, 2003. **332**(4): p. 861-76.

80. Wall, M.E., W.S. Hlavacek, and M.A. Savageau, *Design of gene circuits: lessons from bacteria.* Nature Reviews Genetics, 2004. **5**(1): p. 34-42.

81. Shinar, G., et al., *Rules for biological regulation based on error minimization.* Proc Natl Acad Sci U S A, 2006. **103**(11): p. 3999-4004.

82. Libby, E., T.J. Perkins, and P.S. Swain, *Noisy information processing through transcriptional regulation.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7151-6.

83. Sprinzak, D., et al., *Cis-interactions between Notch and Delta generate mutually exclusive signalling states.* Nature, 2010. **465**(7294): p. 86-U95.

84. Gerland, U. and T. Hwa, *Evolutionary selection between alternative modes of gene regulation.* Proc Natl Acad Sci U S A, 2009. **106**(22): p. 8841-6.

85. Peter, I.S. and E.H. Davidson, *Modularity and design principles in the sea urchin embryo gene regulatory network.* FEBS Lett, 2009. **583**(24): p. 3948-58.

86. Davidson, E.H., *Network design principles from the sea urchin embryo.* Curr Opin Genet Dev, 2009. **19**(6): p. 535-40.

87. Levine, J.H. and M.B. Elowitz, *Polyphasic feedback enables tunable cellular timers.* Current Biology, 2014. **24**(20): p. R994-5.

88. Acar, M., et al., *A general mechanism for network-dosage compensation in gene circuits.* Science, 2010. **329**(5999): p. 1656-60.

89. Zaslaver, A., et al., *Just-in-time transcription program in metabolic pathways.* Nature Genetics, 2004. **36**(5): p. 486-491.

90. Korem, Y., et al., *Geometry of the Gene Expression Space of Individual Cells.* PLoS Comput Biol, 2015. **11**(7): p. e1004224.

91. Hart, Y., et al., *Inferring biological tasks using Pareto analysis of high-dimensional data.* Nature Methods, 2015. **12**(3): p. 233-+.

92. Vonheijne, G. and M.A. Savageau, *Rna Splicing - Advantages of Parallel Processing.* Journal of Theoretical Biology, 1982. **98**(4): p. 563-574.

93. Beisel, C.L. and C.D. Smolke, *Design Principles for Riboswitch Function.* PLoS Comput Biol, 2009. **5**(4).

94. Win, M.N., J.C. Liang, and C.D. Smolke, *Frameworks for Programming Biological Function through RNA Parts and Devices.* Chemistry & Biology, 2009. **16**(3): p. 298-310.

95. Win, M.N. and C.D. Smolke, *Higher-order cellular information processing with synthetic RNA devices.* Science, 2008. **322**(5900): p. 456-460.

96. Pfleger, B.F., et al., *Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes.* Nature Biotechnology, 2006. **24**(8): p. 1027-1032.

97. Levine, E. and T. Hwa, *Small RNAs establish gene expression thresholds.* Curr Opin Microbiol, 2008. **11**(6): p. 574-579.

98. Levine, E., et al., *Quantitative characteristics of gene regulation by small RNA.* Plos Biology, 2007. **5**(9): p. 1998-2010.

99. Flynt, A.S. and E.C. Lai, *Biological principles of microRNA-mediated regulation: shared themes amid diversity.* Nature Reviews Genetics, 2008. **9**(11): p. 831-842.

100. Lai, X., et al., *Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs.* Nucleic Acids Res, 2012. **40**(18): p. 8818-34.

101. Friedman, Y., O. Balaga, and M. Linial, *Working together: combinatorial regulation by microRNAs.* Cellular Oscillatory Mechanisms, 2013. **774**: p. 317-37.

102. Schmitz, U., et al., *Cooperative gene regulation by microRNA pairs and their identification using a computational workflow.* Nucleic Acids Res, 2014. **42**(12): p. 7539-52.

103. Savageau, M.A., *Feedforward Inhibition in Biosynthetic Pathways - Inhibition of the Aminoacyl Transfer Rna-Synthetase by the Penultimate Product.* J Theor Biol, 1979. **77**(4): p. 385-404.

104. Savageau, M.A. and G. Jacknow, *Feedforward Inhibition in Biosynthetic Pathways - Inhibition of the Aminoacyl Transfer Rna-Synthetase by Intermediates of the Pathway.* J Theor Biol, 1979. **77**(4): p. 405-425.

105. Alves, R. and M.A. Savageau, *Effect of overall feedback inhibition in unbranched biosynthetic pathways.* Biophysical Journal, 2000. **79**(5): p. 2290-2304.

106. Alves, R. and M.A. Savageau, *Irreversibility in unbranched pathways: Preferred positions based on regulatory considerations.* Biophysical Journal, 2001. **80**(3): p. 1174-1185.

107. Shinar, G., J.D. Rabinowitz, and U. Alon, *Robustness in Glyoxylate Bypass Regulation.* PLoS Comput Biol, 2009. **5**(3).

108. Kwon, Y.K. and K.H. Cho, *Coherent coupling of feedback loops: a design principle of cell signaling networks.* Bioinformatics, 2008. **24**(17): p. 1926-1932.

109. Heinrich, R., et al., *The structural design of glycolysis: an evolutionary approach.* Biochemical Society Transactions, 1999. **27**(2): p. 294-298.

110. MelendezHevia, E., et al., *Theoretical approaches to the evolutionary optimization of glycolysis - Chemical analysis.* European Journal of Biochemistry, 1997. **244**(2): p. 527-543.

111. Heinrich, R., et al., *Theoretical approaches to the evolutionary optimization of glycolysis - Thermodynamic and kinetic constraints.* European Journal of Biochemistry, 1997. **243**(1-2): p. 191-201.

112. Mittenthal, J.E., et al., *A new method for assembling metabolic networks, with application to the Krebs citric acid cycle.* J Theor Biol, 2001. **208**(3): p. 361-382.

113. Mittenthal, J.E., *An algorithm to assemble pathways from processes.* Pacific Symposium on Biocomputing '97, 1996: p. 292-303.

114. Mittenthal, J.E., et al., *Designing metabolism: Alternative connectivities for the pentose phosphate pathway.* Bulletin of Mathematical Biology, 1998. **60**(5): p. 815-856.

115. Noor, E., et al., *Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy.* Molecular Cell, 2010. **39**(5): p. 809-820.

116. Kurland, C.G. and H.J. Dong, *Bacterial growth inhibition by overproduction of protein.* Mol Microbiol, 1996. **21**(1): p. 1-4.

117. Cascante, M., et al., *The metabolic productivity of the cell factory.* J Theor Biol, 1996. **182**(3): p. 317-325.

118. Cascante, M., et al., *Control Analysis of Transit-Time for Free and Enzyme-Bound Metabolites - Physiological and Evolutionary Significance of Metabolic Response-Times.* Biochemical Journal, 1995. **308**: p. 895-899.

119. Durek, P. and D. Walther, *The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles.* Bmc Systems Biology, 2008. **2**.

120. Scott, M. and T. Hwa, *Bacterial growth laws and their applications.* Current Opinion in Biotechnology, 2011. **22**(4): p. 559-565.

121. Scott, M., et al., *Emergence of robust growth laws from optimal regulation of ribosome synthesis.* Mol Syst Biol, 2014. **10**(8).

122. Hui, S., et al., *Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria.* Mol Syst Biol, 2015. **11**(1): p. 784.

123. Ghosh, A. and B. Chance, *Oscillations of Glycolytic Intermediates in Yeast Cells.* Biochemical and Biophysical Research Communications, 1964. **16**(2): p. 174-&.

124. Palmeirim, I., et al., *Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis.* Cell, 1997. **91**(5): p. 639-648.

125. Leloup, J.C. and A. Goldbeter, *Chaos and birhythmicity in a model for circadian oscillations of the PER and TIM proteins in Drosophila.* J Theor Biol, 1999. **198**(3): p. 445-459.

126. Csikasz-Nagy, A., B. Novak, and J.J. Tyson, *Reverse engineering models of cell cycle regulation.* Cellular Oscillatory Mechanisms, 2008. **641**: p. 88-97.

127. Csikasz-Nagy, A., et al., *Cell cycle regulation by feed-forward loops coupling transcription and phosphorylation.* Mol Syst Biol, 2009. **5**.

128. Drapkin, B.J., et al., *Analysis of the mitotic exit control system using locked levels of stable mitotic cyclin.* Mol Syst Biol, 2009. **5**.

129. Tyson, J.J. and B. Novak, *Temporal organization of the cell cycle.* Current Biology, 2008. **18**(17): p. R759-R768.

130. Kiang, L., et al., *Cyclin-Dependent Kinase Inhibits Reinitiation of a Normal S-Phase Program during G(2) in Fission Yeast.* Molecular and Cellular Biology, 2009. **29**(15): p. 4025-4032.

131. Moseley, J.B., et al., *A spatial gradient coordinates cell size and mitotic entry in fission yeast.* Nature, 2009. **459**(7248): p. 857-U8.

132. Lu, Y. and F.R. Cross, *Periodic Cyclin-Cdk Activity Entrains an Autonomous Cdc14 Release Oscillator.* Cell, 2010. **141**(2): p. 268-279.

133. Charvin, G., et al., *Origin of Irreversibility of Cell Cycle Start in Budding Yeast.* Plos Biology, 2010. **8**(1).

134. Young, M.W. and S.A. Kay, *Time zones: A comparative genetics of circadian clocks.* Nature Reviews Genetics, 2001. **2**(9): p. 702-715.

135. Rand, D.A., et al., *Design principles underlying circadian clocks.* Journal of the Royal Society Interface, 2004. **1**(1): p. 119-130.

136. Rand, D.A., et al., *Uncovering the design principles analysis of flexibility of circadian clocks: Mathematical and evolutionary goals.* J Theor Biol, 2006. **238**(3): p. 616-635.

137. Dalchau, N., et al., *Correct biological timing in Arabidopsis requires multiple light-signaling pathways.* Proc Natl Acad Sci U S A, 2010. **107**(29): p. 13171-13176.

138. Locke, J.C.W., et al., *Global parameter search reveals design principles of the mammalian circadian clock.* Bmc Systems Biology, 2008. **2**.

139. Indic, P., W.J. Schwartz, and D. Paydarfar, *Design principles for phase-splitting behaviour of coupled cellular oscillators: clues from hamsters with 'split' circadian rhythms.* Journal of the Royal Society Interface, 2008. **5**(25): p. 873-883.

140. Pando, B.F. and A. van Oudenaarden, *Coupling cellular oscillators-circadian and cell division cycles in cyanobacteria.* Curr Opin Genet Dev, 2010. **20**(6): p. 613-618.

141. Mori, T., B. Binder, and C.H. Johnson, *Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 hours.* Proc Natl Acad Sci U S A, 1996. **93**(19): p. 10183-10188.

142. Yang, Q., et al., *Circadian Gating of the Cell Cycle Revealed in Single Cyanobacterial Cells.* Science, 2010. **327**(5972): p. 1522-1526.

143. Matsuo, T., et al., *Control mechanism of the circadian clock for timing of cell division in vivo.* Science, 2003. **302**(5643): p. 255-259.

144. Nagoshi, E., et al., *Circadian gene expression in individual fibroblasts: Cell-autonomous and self-sustained oscillators pass time to daughter cells.* Cell, 2004. **119**(5): p. 693-705.

145. Narula, J., et al., *Chromosomal Arrangement of Phosphorelay Genes Couples Sporulation and DNA Replication.* Cell, 2015. **162**(2): p. 328-37.

146. Appleby, J.L., J.S. Parkinson, and R.B. Bourret, *Signal transduction via the multi-step phosphorelay: not necessarily a road less traveled.* Cell, 1996. **86**(6): p. 845-8.

147. Hoch, J.A., *Two-component and phosphorelay signal transduction.* Curr Opin Microbiol, 2000. **3**(2): p. 165-70.

148. D'Agostino, I.B. and J.J. Kieber, *Phosphorelay signal transduction: the emerging family of plant response regulators.* Trends in Biochemical Sciences, 1999. **24**(11): p. 452-456.

149. Loomis, W.F., A. Kuspa, and G. Shaulsky, *Two-component signal transduction systems in eukaryotic microorganisms.* Curr Opin Microbiol, 1998. **1**(6): p. 643-648.

150. Shinar, G., et al., *Input-output robustness in simple bacterial signaling systems.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 19931-19935.

151. Alves, R. and M.A. Savageau, *Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function.* Mol Microbiol, 2003. **48**(1): p. 25-51.

152. Mehta, P., et al., *Information processing and signal integration in bacterial quorum sensing.* Mol Syst Biol, 2009. **5**.

153. Ray, J.C.J. and O.A. Igoshin, *Adaptable Functionality of Transcriptional Feedback in Bacterial Two-Component Systems.* PLoS Comput Biol, 2010. **6**(2).

154. Plahte, E., T. Mestl, and S.W. Omholt, *A methodological basis for description and analysis of systems with complex switch-like interactions.* Journal of Mathematical Biology, 1998. **36**(4): p. 321-348.
155. Tiwari, A., et al., *The interplay of multiple feedback loops with post-translational kinetics results in bistability of mycobacterial stress response.* Physical Biology, 2010. **7**(3).
156. Csikasz-Nagy, A., L. Cardelli, and O.S. Soyer, *Response dynamics of phosphorelays suggest their potential utility in cell signalling.* Journal of the Royal Society Interface, 2011. **8**(57): p. 480-488.
157. Salazar, M.E. and M.T. Laub, *Temporal and evolutionary dynamics of two-component signaling pathways.* Curr Opin Microbiol, 2015. **24**: p. 7-14.
158. Gao, R. and A.M. Stock, *Evolutionary tuning of protein expression levels of a positively autoregulated two-component system.* PLoS Genet, 2013. **9**(10): p. e1003927.
159. Levine, J.H., et al., *Pulsed feedback defers cellular differentiation.* Plos Biology, 2012. **10**(1): p. e1001252.
160. Capra, E.J., et al., *Adaptive Mutations that Prevent Crosstalk Enable the Expansion of Paralogous Signaling Protein Families.* Cell, 2012. **150**(1): p. 222-232.
161. Goldbeter, A. and D.E. Koshland, *An Amplified Sensitivity Arising from Covalent Modification in Biological-Systems.* Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 1981. **78**(11): p. 6840-6844.
162. Goldbeter, A. and D.E. Koshland, *Ultrasensitivity in Biochemical Systems Controlled by Covalent Modification - Interplay between Zero-Order and Multistep Effects.* Journal of Biological Chemistry, 1984. **259**(23): p. 14441-14447.
163. Goldbeter, A. and D.E. Koshland, *Energy-Expenditure in the Control of Biochemical Systems by Covalent Modification.* Journal of Biological Chemistry, 1987. **262**(10): p. 4460-4471.
164. Huang, C.Y.F. and J.E. Ferrell, *Ultrasensitivity in the mitogen-activated protein kinase cascade.* Proc Natl Acad Sci U S A, 1996. **93**(19): p. 10078-10083.
165. Schwacke, J.H. and E.O. Voit, *Computation and analysis of time-dependent sensitivities in generalized mass action systems.* J Theor Biol, 2005. **236**(1): p. 21-38.
166. Schwacke, J.H. and E.O. Voit, *Concentration-dependent effects on the rapid and efficient activation of the MAP kinase signaling cascade.* Proteomics, 2007. **7**(6): p. 890-899.
167. Schwacke, J.H. and E.O. Volt, *The potential for signal integration and processing in interacting MAP kinase cascades.* J Theor Biol, 2007. **246**(4): p. 604-620.
168. Shankaran, H., H. Resat, and H.S. Wiley, *Cell surface receptors for signal transduction and ligand transport: A design principles study.* PLoS Comput Biol, 2007. **3**(6): p. 986-999.
169. Shankaran, H., H.S. Wiley, and H. Resat, *Receptor downregulation and desensitization enhance the information processing ability of signalling receptors.* Bmc Systems Biology, 2007. **1**.
170. Lestas, I., G. Vinnicombe, and J. Paulsson, *Fundamental limits on the suppression of molecular fluctuations.* Nature, 2010. **467**(7312): p. 174-178.
171. Ben-Zvi, D. and N. Barkai, *Scaling of morphogen gradients by an expansion-repression integral feedback control.* Proc Natl Acad Sci U S A, 2010. **107**(15): p. 6924-6929.
172. Ben-Zvi, D., et al., *Scaling of the Bmp morphogen gradient in Xenopus embryos.* Developmental Biology, 2009. **331**(2): p. 425-425.
173. Barkai, N. and B.Z. Shilo, *Robust Generation and Decoding of Morphogen Gradients.* Cold Spring Harbor Perspectives in Biology, 2009. **1**(5).

174. Kampf, M.M. and W. Weber, *Synthetic biology in the analysis and engineering of signaling processes.* Integrative Biology, 2010. **2**(1): p. 12-24.

175. Khalil, A.S. and J.J. Collins, *Synthetic biology: applications come of age.* Nature Reviews Genetics, 2010. **11**(5): p. 367-379.

176. Rothschild, L.J., *A powerful toolkit for synthetic biology: Over 3.8 billion years of evolution.* Bioessays, 2010. **32**(4): p. 304-313.

177. Aubel, D. and M. Fussenegger, *Mammalian synthetic biology - from tools to therapies.* Bioessays, 2010. **32**(4): p. 332-345.

178. Bashor, C.J., et al., *Rewiring Cells: Synthetic Biology as a Tool to Interrogate the Organizational Principles of Living Systems.* Annual Review of Biophysics, Vol 39, 2010. **39**: p. 515-537.

179. Young, E. and H. Alper, *Synthetic Biology: Tools to Design, Build, and Optimize Cellular Processes.* Journal of Biomedicine and Biotechnology, 2010.

180. Alterovitz, G., T. Muso, and M.F. Ramoni, *The challenges of informatics in synthetic biology: from biomolecular networks to artificial organisms.* Briefings in Bioinformatics, 2010. **11**(1): p. 80-95.

181. Mukherji, S. and A. van Oudenaarden, *Synthetic biology: understanding biological design from synthetic circuits.* Nature Reviews Genetics, 2009. **10**(12): p. 859-871.

182. Marner, W.D., 2nd, *Practical application of synthetic biology principles.* Biotechnol J, 2009. **4**(10): p. 1406-19.

183. Agapakis, C.M. and P.A. Silver, *Synthetic biology: exploring and exploiting genetic modularity through the design of novel biological networks.* Mol Biosyst, 2009. **5**(7): p. 704-13.

184. Purnick, P.E.M. and R. Weiss, *The second wave of synthetic biology: from modules to systems.* Nature Reviews Molecular Cell Biology, 2009. **10**(6): p. 410-422.

185. Endler, L., et al., *Designing and encoding models for synthetic biology.* Journal of the Royal Society Interface, 2009. **6**.

186. Martin, C.H., et al., *Synthetic Metabolism: Engineering Biology at the Protein and Pathway Scales.* Chemistry & Biology, 2009. **16**(3): p. 277-286.

187. Lee, S.K., et al., *Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels.* Current Opinion in Biotechnology, 2008. **19**(6): p. 556-563.

188. Saito, H. and T. Inoue, *Synthetic biology with RNA motifs.* International Journal of Biochemistry & Cell Biology, 2009. **41**(2): p. 398-404.

189. Brenner, K., L.C. You, and F.H. Arnold, *Engineering microbial consortia: a new frontier in synthetic biology.* Trends in Biotechnology, 2008. **26**(9): p. 483-489.

190. Channon, K., E.H.C. Bromley, and D.N. Woolfson, *Synthetic biology through biomolecular design and engineering.* Current Opinion in Structural Biology, 2008. **18**(4): p. 491-498.

191. Becskei, A. and L. Serrano, *Engineering stability in gene networks by autoregulation.* Nature, 2000. **405**(6786): p. 590-593.

192. Skerker, J.M., J.B. Lucks, and A.P. Arkin, *Evolution, ecology and the engineered organism: lessons for synthetic biology.* Genome Biology, 2009. **10**(11).

193. Atkinson, M.R., et al., *Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli.* Cell, 2003. **113**(5): p. 597-607.

3 A survey of HK, HPt and RR domains and their relative organization in two-component systems and phosphorelays of organisms with fully sequenced genomes

3.1. Abstract

Two Component Systems and Phosphorelays (TCS/PR) are environmental signal transduction cascades in prokaryotes and, less frequently, in eukaryotes. The internal domain organization of proteins and the topology of TCS/PR cascades play an important role in shaping the responses of the circuits. It is thus important to maintain updated censuses of TCS/PR proteins in order to identify the various topologies used by nature and enable a systematic study of the dynamics associated with those topologies.

To create such a census, we analyzed the proteomes of 7609 organisms from all domains of life with fully sequenced and annotated genomes. To begin, we survey each proteome searching for proteins containing domains that are associated with internal signal transmission within TCS/PR: Histidine Kinase (HK), Response Regulator (RR) and Histidine Phosphotranfer (HPt) domains, and analyze how these domains are arranged in the individual proteins. Then, we find all types of operon organization and calculate how much more likely are proteins that contain TCS/PR domains to be coded by neighboring genes than one would expect from the genome background of each organism. Finally, we analyze if the fusion of domains into single TCS/PR proteins is more frequently observed than one might expect from the background of each proteome.

We find 50 alternative ways in which the HK, HPt, and RR domains are observed to organize into single proteins. In prokaryotes, TCS/PR coding genes tend to be clustered in operons. 90% of all proteins identified in this study contain just one of the three domains, while 8% of the remaining proteins combine one copy of an HK, a RR, and/or an HPt domain. In eukaryotes, 25% of all TCS/PR proteins have more than

one domain. These results might have implications for how signals are internally transmitted within TCS/PR cascades. These implications could explain the selection of the various designs in alternative circumstances.

3.2. Introduction

Historically, Two Component Systems and Phosphorelays (TCS/PR) have been considered as primary environmental signal transduction cascades in prokaryotes [1, 2]. In TCS/PR, environmental signals regulate the autophosphorylation state of a sensor histidine kinase. In TCS this sensor transfers its phosphate to a response regulator, which will in turn directly regulate the relevant cellular responses to the signal. The sensor and response regulator may be two independent proteins. They may also be the same protein, containing independent domains that are responsible for each of the two functions. In PR, additional phosphotransfer steps may happen before the phosphate reaches the response regulator protein(s) that directly controls cellular responses (Figure 1). PR are considered to be a main form of signal transduction in bacteria [3, 4] . They are less frequently present in eukaryotes and absent in animals [5-8] .

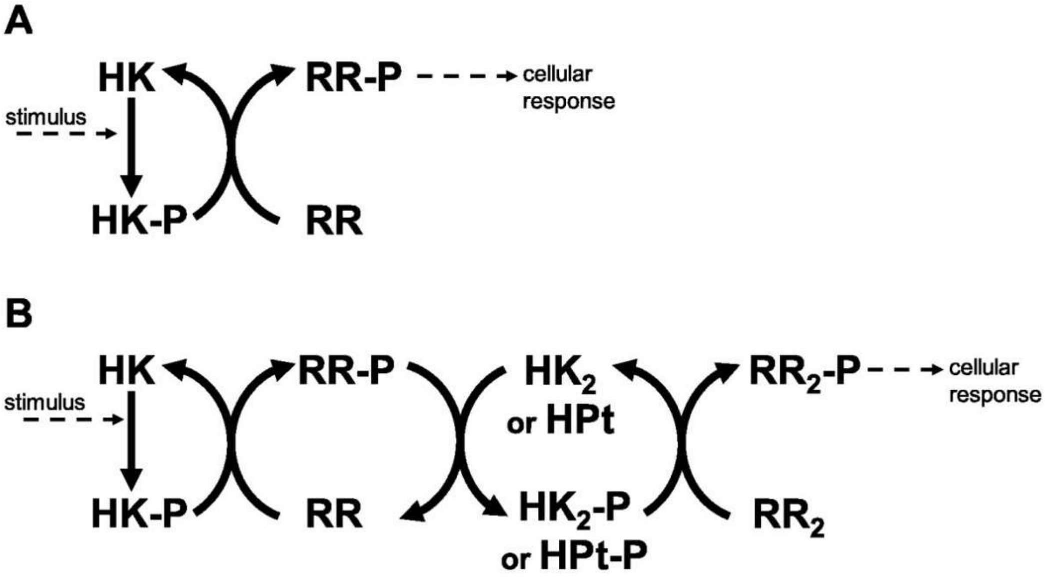


Figure 1.Two component systems. A– Prototypical two component system with one phosphotransfer step between HK and RR. B – 3-step phosphorelay, with four protein domains involved in the signal transduction process and 3 phosphotransfer steps.

The mechanism of signal sensing in the various types of TCS/PR have been studied with great detail and is reviewed elsewhere [9-11]. Extensive and insightful reviews have also been published about the topology (pattern of molecular interactions between the proteins in the cascade), crosstalk and signal transmission in TCS/PR [10, 12-30], as well as about the domain structure and evolution of the proteins involved in the cascades [1, 2, 11, 27, 31-43].

There are several protein types and domains that nature uses in TCS/PR cascades. For example, CHEW adapter proteins permit transmitting information about nutrient gradients to the TCS that regulates bacterial response to those gradients [44]. In another example, the PII protein regulates the activity of the TCS that responds to nitrogen depletion in the environment [45]. There are other cases where external proteins bind proteins from a TCS/PR cascade and modulate their stability [46]. These protein types are used in TCS/PR with specific biological functions and are not common to all TCS/PR cascades.

Nevertheless, there are four types of protein domains that are common to all TCS/PR cascades. Sensor domains, with wide sequence variability, are responsible for capturing the environmental changes and adjusting the activity of the cascade [2, 22]. Irrespective of protein domain organization, signal transmission within a TCS/PR circuit is done using histidine kinase (HK) domains, response regulator (RR) domains, and/or histidine phosphotransfer (HPt) domains. These last three domains are responsible for internal signal transmission (IST) within the cascade and represent the focus of the current work. Because they are common to all TCS/PR cascades, the results from our study are generally applicable and do not depend on the specific environmental signal or biological response mediated by the cascades.

Several examples demonstrate that the dynamic range and signal-response curve that a given cascade might exhibit is closely related to the interactions between the various proteins and to the organization of IST domains within each cascade protein [46-52]. For example, circuits where each IST domain is in an independent protein are more likely to participate in crosstalk and branching is more likely to occur in the signal transduction process [27, 37]. In addition, noise propagates differently in a cascade of independent IST domain proteins than in a cascade where IST domains are found within the same protein [53] (Figure 2). Also, TCS where phosphatases are involved in dephosphorylating the response regulator protein may show hysteretic behavior. In contrast, TCS where the sensor protein works both as phosphodonor and phosphatase for the response regulator may only exhibit graded responses to changes in the signal [49].

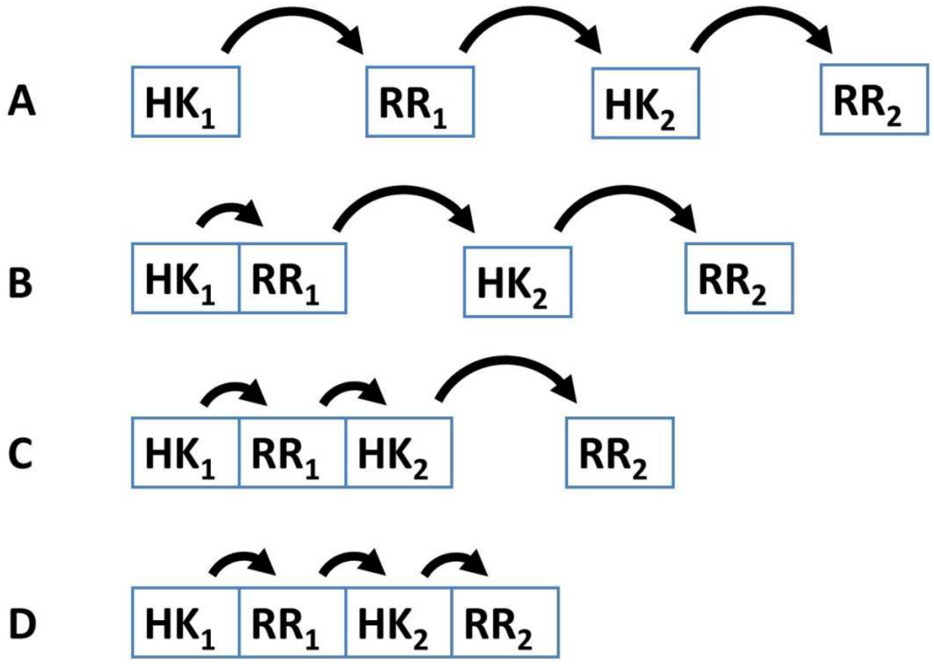


Figure 2.Four different patterns of covalent linkage between the protein domains involved in phosphorelays. A – A four protein phosphorelay. B – A phosphorelay with and hybrid kinase at the beginning of the cascade. C – A two protein phosphorelay where the first two phosphotransfer steps between domains occur in a single protein. D – A one protein phosphorelay, where all phosphotransfer steps take place between domains of a single protein.

These and other examples [52, 54-56] show that connectivity of the TCS/PR circuits and domain organization of the proteins play an important role in shaping the responses of the cascades to their cognate signals. It is thus important to maintain censuses of TCS/PR proteins in order to identify the various network topologies used by nature and enable a systematic study of the internal signal transduction dynamics associated with those topologies. Information about such topologies can be retrieved for a detailed analysis from several databases [57, 58].

While MIST2 [57] contains information about less than 3000 genomes, Pfam [59] contains a few hundred sequences divided among the HK, RR, and HPt domain families involved in TCS/PR cascades. Currently, at the NIH there are over 10000 fully sequenced and annotated genomes that are freely accessible to the public. Because of this, obtaining a more up to date census of the TCS/PR in these genomes is an important task that we set out to do. We analyzed the TCS/PR proteins of 7609 organisms from all domains of life with fully sequenced and annotated genomes. We focus on the IST domain families HK, RR, and HPt of TCS/PR cascades. First, we survey the number of TCS/PR domains in each organism and how these domains are arranged into individual proteins. Then, we find all different type of operon organizations and analyze how much more likely are proteins that contain TCS/PR domains to be coded by neighboring genes than one would expect from the genome background. Finally, we analyze how the percentage IST domain fusion within TCS/PR proteins changes among all analyzed genomes.

Our census finds that there are 50 alternative ways in which the HK, HPt, and RR domains are observed to organize into single proteins. 90% of all proteins identified in this study contain just one RR or HK domain, while 8% of the remaining proteins combine one copy of a HK, a RR, and/or a HPt domain. We also find that more than

25% of all TCS/PR eukaryotic proteins have more than one domain. Our results are consistent with previous works and identify TCS/PR proteins in all non-animal phyla. Overall, our results set the stage for a systematic study to compare the internal dynamic behavior of signal transduction associated with each circuit topology in TCS/PR cascades.

3.3. Material and Methods

3.3.1. Identification of proteins containing TCS/PR domains

The fully annotated proteomes of 9961 organisms were downloaded from NCBI's genome database (January 2014 version). 2352 of these proteomes were eliminated because they belonged to phages, virus, satellite DNA sequences, or organisms whose taxonomic classification was still not fully resolved. The remaining 7609 proteomes belonging to 35 phyla from Bacteria, 6 phyla from Archaea and 11 eukaryotic phyla (Supplementary Table 1) were further analyzed in search for proteins containing domains of types HK, RR and HPt. These domains are associated with IST in all TCS/PR cascades. Other protein domains (such as the CHEW adaptor domain or the P2 protein from NRI/NRII, among many others) were not included in the analysis because they are specific of certain TCS/PR cascades. The sensor domain of TCS/PR cascade proteins was also not included due to its sequence variability. Thus, the results from our study are general for all TCS/PR cascades.

We used PROSITE (http://prosite.expasy.org) to obtain a set of well curated sequences that can be used as a seed to identify TCS/PR proteins in the relevant proteomes. We downloaded a multiple alignment of all relevant ortholog sequences for each protein domain (HK – PS50109 PROSITE Domain, RR – PS50110 PROSITE Domain and HPt – PS50894 PROSITE Domain) from PROSITE. We then used these three multiple alignments as a set of query sequences for two independent searches. One was done using HMMER [60]. For each multiple alignment downloaded from PROSITE, we built a profile HMM using hmmbuild, and performed the search of the profile HMM against all proteomes selected from the NCBI database using jackhmmer. The second search was done in parallel using PSI-BLAST [61] and the three multiple alignments downloaded from PROSITE as a query. HMMER finds homologues that are more distantly related than those found by BLAST.

We simultaneously use BLAST and HMMER because they have different sensitivities in detecting sequence similarities. BLAST generates a higher number of false negatives, while HMMER generates a higher number of false positives. By using both and filtering the results, we hope to obtain a more precise picture of the conserved domains. In each search we queried the 7609 proteomes in order to identify proteins with domains that are homologous to those used as queries.

In addition, the consensus sequence was calculated for each domain (HK, RR and HPt) independently. Using an in-house PERL script, the most common residue in each position was identified for each of the three multiple alignments. This residue was taken as the consensus value for that position in the corresponding protein domain. Subsequently the three consensus sequences were used to search each proteome using PSI-BLAST [61]. In all three searches, the hits selected were the ones with an e-value lower than 10^{-6} and with a domain coverage of at least 80%.

After performing these three searches, a PERL script was also used to perform a fourth text-mining search and identify the proteins that were annotated in each proteome as being histidine kinases, sensory kinases, hybrid kinases, response regulators or histidine phosphotransferases.

The results of the four searches were merged into a non-redundant set. A total amount of 469421 proteins containing HK, HPt and/or RR domains were identified. This set was curated in the following way:

1 – First we manually looked at the annotation of the proteins to identify functions that are not involved in TCS/PR cascades (e.g. serine kinase).

2 – Then, we build a PERL script that automatically eliminates proteins annotated with those functions from the list.

3 – We finish by automatically comparing the number of proteins in the list and the number of proteins containing terms related to TCS/PR cascades.

4 – We repeat steps 1-3 until the number of proteins in the list and the number of proteins containing only terms related to TCS/PR cascades are the same.

In this way, we semi manually identified 36169 proteins that were annotated as being something other than a TCS/PR protein. These proteins were eliminated. Frequent protein types found in the discarded set of proteins are serine/threonine kinases and several types of regulatory transcription factors.

The remaining 433255 proteins were then reanalyzed and an additional set of 17727 proteins were found to be annotated as being hypothetical or partial proteins. For each of the three domains, the set of 17727 hypothetical and partial proteins were aligned using Clustal X in order to identify the conserved histidine motif in the HK and HPt domains, and the conserved aspartate residue in RR domains. Those sequences without a conserved histidine or aspartate residue were eliminated from the data, leaving a grand total of 415525 annotated proteins and 17724 partial/hypothetical proteins containing HK, RR and/or HPt domains.

A PERL script was developed to filter the curated data sets and determine both, the domain composition of each protein and, when they belonged to the same organism, the relative position of their corresponding genes with respect to each other in the genome.

Once we had identified all proteins containing HK, RR or HPt domains, and the relative genomic position of their corresponding genes, we looked for all type of operons of TCS/PR coding genes that occur in the organisms with fully sequenced genomes. For this purpose, we performed a search of all genes coding HK, RR or HPt protein domains that are located in consecutive positions on prokaryotic genomes. We assumed that they constitute a transcription unit, although this may introduce a small error, as consecutive operons coding for independent TCS/PR exceptionally exist. In our search, we allow the presence of a gap in the operon, that is, a gene which does not encode any HK, RR or HPt domain, because this could be a gene with regulatory functions in the operon.

The statistical treatment of data was carried out independently with and without taking into account the hypothetical and partial proteins found. Both results are qualitatively the same. In the Results section of this chapter we give the results from the analysis of the set of proteins without the hypothetical and partial proteins. The sequence files for all domains are also available at http://web.udl.es/usuaris/pg193845/Salvadoretal.html.

3.3.2. Numerical and Statistical Data analysis

To estimate how the clustering of the various TCS and PR proteins in a genome differed from what one would expect by chance in the context of that genome, we took the following approach. First, we calculated how frequently one would expect proteins containing TCS/PR domains to be coded by neighboring genes in a genome if the order of genes was fully random, given the total number of proteins in that genome, and the number of proteins involved in TCS/PR cascades. The expected neighboring frequencies under this assumption can be computed by Eqs. 1-6. In these equations F(P1↔P2) represents the expected frequency of the neighboring events in a genome for genes coding proteins of types P1 and P2, n_{RR} represents the number of proteins containing one RR domain in the proteome, n_{HK} represents the number of proteins containing one HK domain in the proteome, and P represents the total number of proteins annotated to the proteome.

$$\mathrm{F(HK} \leftrightarrow \mathrm{RR)} = \frac{\mathrm{n}_{RR}}{\mathrm{P\ -1}} + \frac{\mathrm{n}_{RR}}{\mathrm{P\ -2}} \qquad \text{Eq. 1}$$

Eq. 1 represents the probability that a gene localized in position j of the genome is located next to a gene coding for a protein that contains an RR domain, either in positions j-1 or j+1, if gene order is random in a genome. The first term of the sum represents the probability of the presence of an RR gene in one of the two possible locations irrespective of its presence also in the other genome location, and the second term is the probability of the presence of the RR gene in one of the two genome locations when it is not found in the other one. We note that we are not calculating the probability of having a consecutive gene pair containing HK and RR domains. Rather, for any genomic position j, we ask what the probability of its neighboring a gene containing an RR domain is. Eq. 1 gives a good estimation of this random probability, given that the total number of protein coding genes is tens to hundreds of times larger than the number of IST domain coding genes, and assuming that position j represents neither the first nor the last genomic position. This expected RR neighboring frequency will be compared with the actual fraction of HK genes that are found next to RR genes in order to study their genomic distribution.

$$\mathrm{F(HK \leftrightarrow RR \leftrightarrow HK_2)} = 6 \times \frac{\mathrm{n}_{HK}}{\mathrm{P\ -1}} \times \frac{\mathrm{n}_{RR}}{\mathrm{P\ -2}} \qquad \text{Eq. 2}$$

Eq. 2 computes the probability of finding an RR gene and a second HK gene in the genomic neighborhood of a given HK gene. Because these three consecutive genes can be sorted in 6 different ways, we must multiply by 6 the probability of an individual neighboring event. Again, note that we assume having an HK domain containing gene, and ask what the probability of its neighboring genes containing additional HK and RR domains is.

$$\mathrm{F(HK \leftrightarrow RR \leftrightarrow HK_2 \leftrightarrow RR_2)} = 12 \times \frac{\mathrm{n}_{RR}}{\mathrm{P}-1} \times \frac{\mathrm{n}_{HK}-1}{\mathrm{P}-2} \times \frac{\mathrm{n}_{RR}-1}{\mathrm{P}-3} \qquad \text{Eq. 3}$$

Similarly, in Eq. 3 we compute the probability that, considering that we have found a gene containing an HK domain in a given place in the genome, we also find in consecutive genomic positions around that HK gene location another HK gene and two RR genes, if gene organization is random. These four genes can be sorted in 24 different ways, but we don't differentiate between the two RR genes and therefore there are only 12 possible spatial arrangements of these series of four genes.

$$\mathrm{F(HKRR \leftrightarrow HK \leftrightarrow RR)} = 6 \times \frac{\mathrm{n}_{HK}}{\mathrm{P\ -1}} \times \frac{\mathrm{n}_{RR}}{\mathrm{P\ -2}} \qquad \text{Eq. 4}$$

In Eq. 4, the probability of the event is computed in exactly the same way as in Eq. 2.

$$F(HKRRHPt \leftrightarrow RR) = \frac{n_{RR}}{P-1} + \frac{n_{RR}}{P-2}$$ Eq. 5

$$F(HKRRHK \quad \leftrightarrow RR) = \frac{n_{RR}}{P-1} + \frac{n_{RR}}{P-2}$$ Eq. 6

Eq. 5 and Eq. 6 compute the probability of finding and RR gene placed in the genome next to an HKRRHPt or an HKRRHK gene respectively, exactly in the same way as described above for Eq. 1.

Once these expected frequencies were computed using Eqs. 1-6, we calculated the odds ratios of the observed neighboring events with respect to the expected neighboring event. All numerical and statistical calculations were done using Mathematica [62].

3.3.3. Statistical Models

To analyze the relationship between the number of TCS/PR gene fusion events and the proteome size, we built a linear model that would better fit our data for % of fused HK (RR, HPt) domains vs. total number of HK (respectively, RR, HPt). We also built linear models of total number of IST domains in an organism vs. total number of proteins in the proteome and phylogeny (prokaryote, eukaryote). In other words, we fit the data to Eq. 7:

$$Number\ of\ IST\ domains = \alpha_1 \left(Total\ number\ of\ proteins\ in\ proteome\right) + \alpha_2\ Phy\log eny + \varepsilon$$

Eq. 7.

In Eq. 7, the variable phylogeny can assume two values. If the organism is a prokaryote, the variable has value 1; otherwise it has value 2. An ANOVA analysis was used to determine whether the coefficients for each control variable of the linear model are significantly different from zero. If so, this implies that the variable is relevant in explaining the variation observed in the dependent variable.

When fitting the data to the linear models we also calculated the R^2 and adjusted R^2 of the models. R^2 shows how well terms (data points) fit a curve or line; adjusted R^2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model.

3.4. Results

3.4.1. Survey of proteomes containing proteins with domains involved in internal signal transduction (IST) in TCS/PR cascades

Bacteria

Table 1 summarizes the full set of results for bacteria. Proteins with HK and RR domains are present in the proteome of 100% of the species analyzed from the following bacterial phyla: Aquificae, Chlorobi, Verrucomicrobia, Chloroflexi, Cyanobacteria, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Acidobacteria, Nitrospirae, Planctomycetes, Epsilonproteobacteria, Spirochaetes, Thermodesulfobacteria, and Thermotogae. In contrast, proteins containing HK and/or

RR domains were not identified in a small percentage of species in the following phyla: Actinobacteria – 0.63% (4 out of 635 species surveyed), Bacteroidetes – 9.36% (22 out of 235 species), Firmicutes – 0.68% (14 out of 2066), Fusobacteria – 5.26% (2 out of 38), Alphaproteobacteria – 3.55% (16 out of 451), Betaproteobacteria – 1.64% (6 out of 366), Deltaproteobacteria – 1.22% (1 out of 82), Epsilonproteobacteria – 0.24% (1 out of 410), Gammaproteobacteria – 1.83% (41 out of 2246), Synergistetes – 9.09% (1 out of 11). Interestingly, no proteins containing HK or RR domains were identified in most Tenericutes species. Only 18 out of the 111 surveyed Tenericutes species have proteins with HK and RR domains.

The percentage of species in each phylum with proteins containing HPt domains is lower than the percentage of species with HKs and RRs, and ranges from less than 10% (Chlamydiae, Tenericutes) to more than 90% (Deferribacteres, Acidobacteria, Nitrospirae, Planctomycetes, Deltaproteobacteria, Epsilonproteobacteria, Gammaproteobacteria, Spirochaetes, Thermodesulfobacteria, and Thermotogae). It should be noted that HPt domains are also used by proteins that import PTS sugars [63], which means that not all HPt domains we found are involved in PR or TCS signal transduction.

Archaea

Proteins with HK and RR domains were identified in the proteome of 154 out of 179 Euryarchaeota species, 9 of the 11 Taumarchaeota species and only 2 out of 51 Crenarchaeota species surveyed. Proteins with HPt domains were identified in the proteome of 115 Euryarchaeota species and in 7 of the 11 Taumarchaeota species surveyed. No proteins containing HK, RR, or HPt domains were identified in Nanoarchaeota, Nanohaloarcheota, and Korarchaeota (Table 1).

Table 1. Percentage of species in each phylum with TCS/PR proteins.

Domain	Phylum	Abbreviaton	nº of species surveyed	% of species with HK and RR domains	% of species with HPt domains
Bacteria	Actinobacteria	At	635	99.37	14.49
Bacteria	Aquificae	Aq	13	100.00	76.92
Bacteria	Armatimonadetes	Ar	1	100.00	100.00
Bacteria	Bacteroidetes	Ba	235	89.79	49.79
Bacteria	Chlorobi	Cb	14	100.00	71.43
Bacteria	Caldiserica	Cd	1	100.00	0.00
Bacteria	Chlamydiae	Cm	108	98.15	1.85
Bacteria	Lentisphaerae	L	1	100.00	100.00
Bacteria	Verrucomicrobia	V	10	100.00	80.00
Bacteria	Chloroflexi	Cf	23	100.00	65.21
Bacteria	Chrysiogenetes	Cr	1	100.00	100.00
Bacteria	Cyanobacteria	Cy	118	100.00	75.42
Bacteria	Deferribacteres	Df	4	100.00	100.00
Bacteria	Deinococcus-Thermus	Dt	20	100.00	35.00
Bacteria	Dictyoglomi	Dc	2	100.00	0.00
Bacteria	Elusimicrobia	El	1	100.00	0.00
Bacteria	Acidobacteria	Ac	9	100.00	100.00
Bacteria	Fibrobacteres	Fb	1	100.00	100.00
Bacteria	Firmicutes	Fi	2066	99.42	37.80
Bacteria	Fusobacteria	Fu	38	94.74	28.95
Bacteria	Gemmatimonadetes	Ge	1	100.00	100.00
Bacteria	Nitrospinae	Ni	1	100.00	100.00
Bacteria	Nitrospirae	Nt	4	100.00	100.00
Bacteria	Planctomycetes	Pl	20	100.00	100.00
Bacteria	Alphaproteobacteria	A	451	96.67	58.31
Bacteria	Betaproteobacteria	B	366	98.36	59.56
Bacteria	Deltaproteobacteria	D	82	98.78	98.78
Bacteria	Epsilonproteobacteria	E	410	100.00	98.54
Bacteria	Gammaproteobacteria	G	2246	98.31	95.46
Bacteria	Zetaproteobacteria	Z	1	100.00	100.00
Bacteria	Spirochaetes	S	274	100.00	99.64
Bacteria	Synergistetes	Sy	11	90.91	63.64
Bacteria	Tenericutes	T	111	15.32	7.21
Bacteria	Thermodesulfobacteria	Th	2	100.00	100.00
Bacteria	Thermotogae	Tt	17	100.00	100.00
Archaea	Crenarchaeota	C	51	3.92	3.92
Archaea	Euryarchaeota	Eu	179	86.03	64.25
Archaea	Korarchaeota	K	1	0.00	0.00
Archaea	Thaumarchaeota	Ta	11	81.82	63.64
Archaea	Nanoarchaeota	N	1	0.00	0.00
Archaea	Nanohaloarchaeota	Nh	1	0.00	0.00
Eukarya	Alveolates	Av	5	0.00	20.00
Eukarya	Amoeboflagellates	Am	1	100.00	100.00
Eukarya	Euglenozoa	Eg	5	40.00	0.00
Eukarya	Microsporidians	Mi	2	50.00	0.00
Eukarya	Ascomycetes	As	31	54.84	96.77
Eukarya	Basidiomycetes	Bs	2	100.00	100.00
Eukarya	Eudicots	Ed	2	100.00	100.00
Eukarya	Monocots	M	1	0.00	100.00
Eukarya	Nematodes	-	1	0.00	0.00
Eukarya	Arthropods	-	7	0.00	0.00
Eukarya	Chordates	-	10	0.00	0.00

Eukaryotes

HK and RR domains were identified in the proteomes of 20 in 35 fungi species. 32 fungi species contain proteins where the HPt-domain was identified. HK, HPt, and RR domains were identified in the proteomes of the 2 eudicot species surveyed, but only HK and HPt, and not RR domains, were identified in *Oryza sativa*.

There are only two surveyed protist phyla that contain proteins with IST domains. These phyla are Euglenozoa and Amoeboflagellates. We analyzed five Euglenozoa species. Out of these, only *Leishmania donovani* and *Leishmania major* contain proteins with HK and RR IST domains. These domains are always found in separate proteins. Interestingly, only one RR domain containing protein was identified in each of the two species. Surprisingly, only HK domains were identified in proteins from *Leishmania infantum* and *Trypanosoma brucei*. No IST domains were identified in *Leishmania braziliensis*. In *Dictyostelium discoideum* (Amoeboflagellates), the HK domain was only identified in hybrid HKRR, HKRR1RR2, or HKRR1HK2RR2 proteins. In contrast, RR domains also appear in proteins where no other IST domains are identified.

No HK, RR, or HPt domains were found in animal proteomes in the context of TCS/PR cascades

.

3.4.2. Percentage of proteins with HK, RR or HPt domains in the surveyed proteomes

For simplicity, hereafter we shall refer to proteins containing IST domains typical from TCS/PR cascades as TCS/PR proteins. On average, between 1 and 2% of a prokaryotic proteome is composed of TCS/PR proteins (mean = 1.37%). In contrast, when an eukaryotic proteome contains TCS/PR proteins, they account for between 0.05% and 0.2% of the entire proteome (mean = 0.11%). In bacteria, Deltaproteobacteria is the group with the highest average percentage of TCS/PR proteins (Figure 3). In contrast Tenericutes and Chlamydiae almost tie with the lowest average percentage of TCS/PR proteins (Figure 3).

It has been observed in previous analyses that the number of proteins containing IST domains associated with TCS/PR cascades increases almost quadratically with the number of total proteins in a proteome [64, 65]. We further wanted to assess if this dependency is significantly different between eukaryotes and prokaryotes. To do so we fit the data to the linear model described by Eq. 7. An ANOVA analysis shows that phylogeny is important in explaining the variation in total number of IST domains found in a proteome ($p<10^{-25}$). Because of this we divided the dataset in prokaryotes and eukaryotes, and fit each dataset to the linear model shown in Figure 4. We find that the fraction of variability in number of IST domains explained by proteome size in eukaryotes doubles that of prokaryotes. This suggests that the number of IST domains could evolve differently in prokaryotic and eukaryotic organisms.

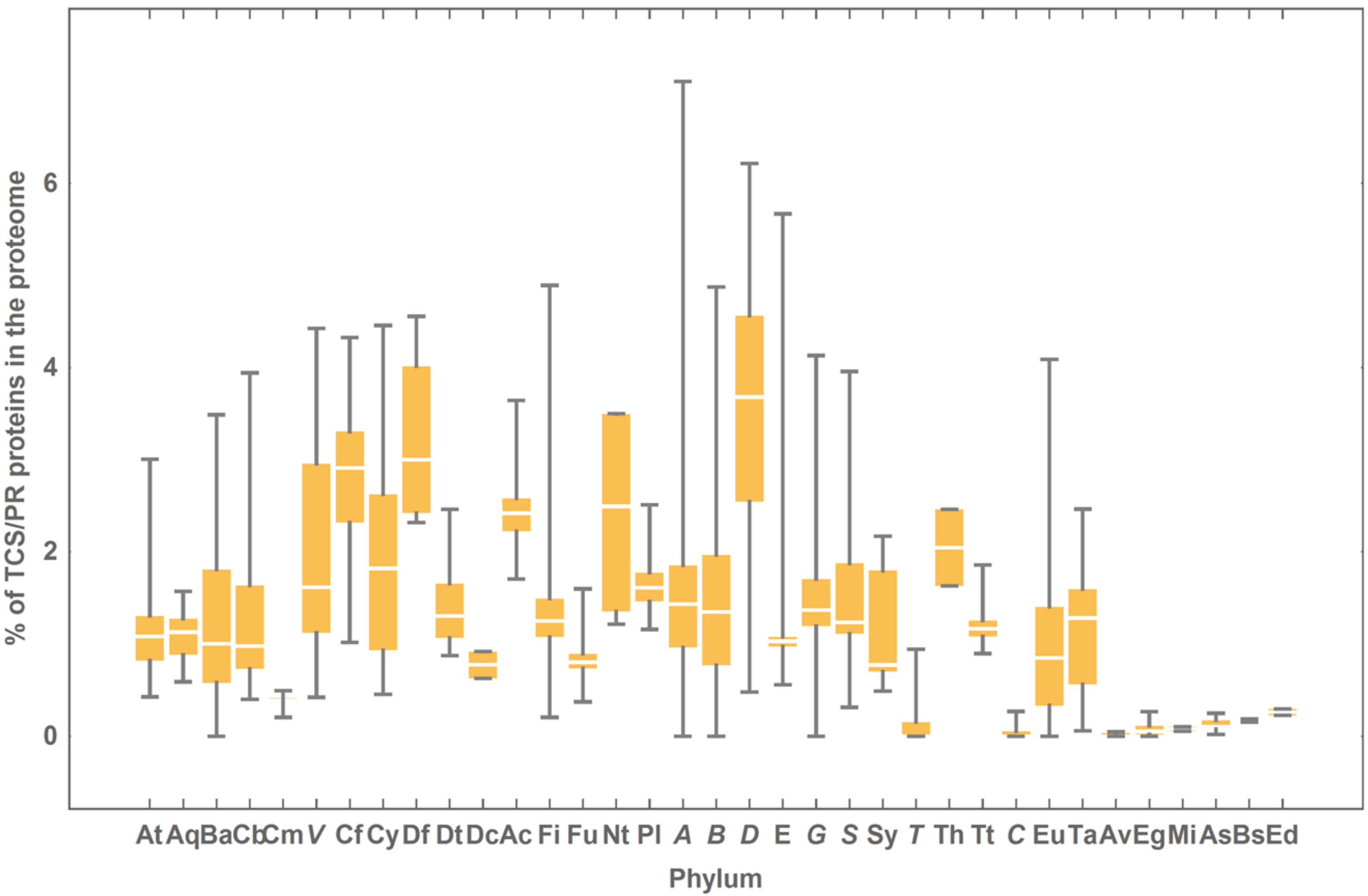


Figure 3.Percentage of TCS/PR proteins in the proteome per phylum. The colored box represents the range of percentage values comprised between the 25% and the 75% quantiles, and the edges of the vertical bar denote the upper and lower percentage values for each phylum. Phylum abbreviations are given in Table 1. Phyla with only one species surveyed are not represented in the figure. Their percentage of TCS/PR proteins per phylum are: Ar (0.93), Cd (0.89), L (0.68), Cr (3.73), El (0.78), Fb (0.81), Ge (3.15), Z (2.47), Ni (1.97), K (0), N (0), Nh (0), Am (0.17) and M (0.04).We have found only 2 TCS/PR proteins in Av (5 sp): 1 HPt in *T. annulata* and 1 HK in *T. parva*.

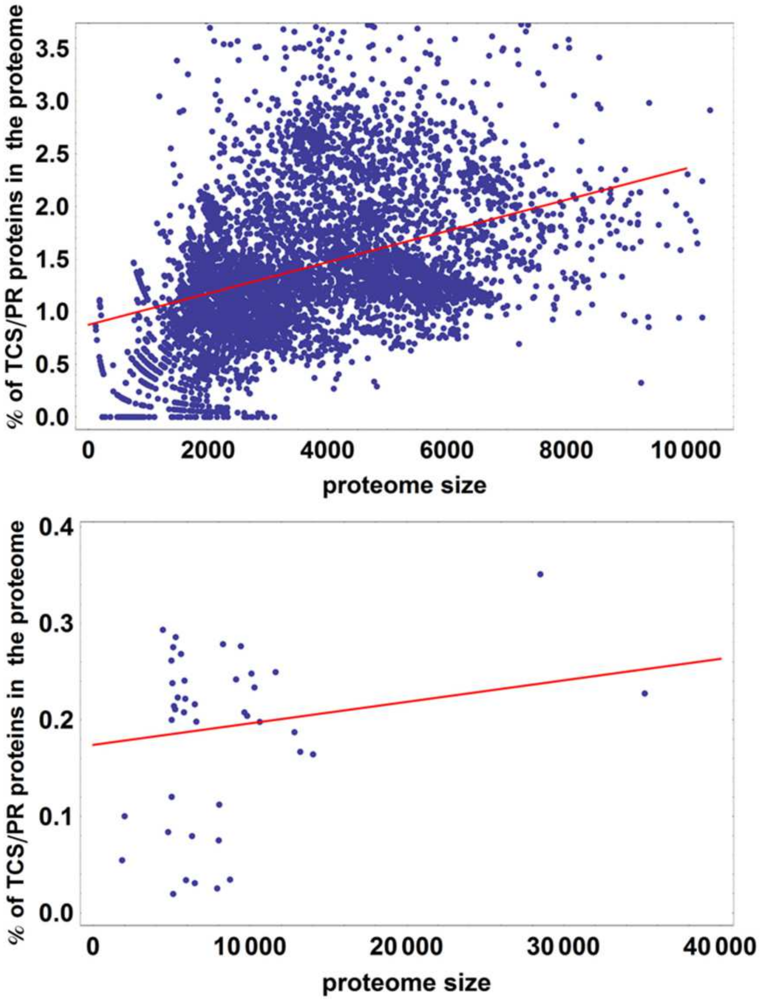


Figure 4.Percentage of TCS/PR proteins in the proteome versus total number of proteins in the proteome. The upper graph depicts the data from prokaryotes, and the lower graph depicts the data from eukaryotes. R^2 is 0.21 for prokaryotes and 0.49 for eukaryotes. This means that proteome size explains 21% of the variation in the percentage of TCS/PR in prokaryotes and 49% in eukaryotes.

3.4.3. Survey of TCS/PR protein types

We find fifty unique types of TCS/PR proteins, when it comes to IST domain organization within a single polypeptide chain. These unique types of TCS/PR proteins are shown in Table 2, sorted by abundance. In that table, the protein identifier describes the type of IST domain (HK, HPt, or RR) and the number describes how many domains of a given IST type are found in each protein. Hereafter we shall refer to proteins containing only one HK IST domain as HK protein type, proteins containing one HK domain and one RR domain as HKRR protein type, and so on and so forth.

Overall, all phyla where IST domains associated with TCS/PR cascades were identified have RR and HK protein types, with the exception of Monocots, which lack RR domains. HKRR protein type (also known as hybrid HK) is present in all phyla where TCS/PR proteins were identified, except in Aquificae, Tenericutes, Chlamydiae, and Crenarchaeota (Supplementary Table 2). Together, HK, RR, and HKRR proteins represent 94% of all TCS/PR proteins that were identified.

In prokaryotes, RR or HK protein types are the most abundant. Together, they represent more than 90% of all TCS/PR proteins found in the genomes of many organisms (Supplementary Table 2). HKRR represent the third most abundant type of TCS/PR protein, oscillating between less than 1% (Firmicutes) and more than 10% (Cyanobacteria) of all TCS/PR proteins (Supplementary Table 2). The remaining protein types (HPt, HKRRHPt, HK_1RRHK_2, $HKRR_1HPtRR_2$, $HK_1RR_1HK_2RR_2$, ...) range from less than 1% to 5 % of all TCS/PR proteins identified in a phylum. Of these less abundant protein types, the three-domain HKRRHPt protein is more abundant than HK_1RRHK_2. The HPt domain is more frequently found in combination with other IST TCS/PR protein domains than alone in a protein, with the exception of Firmicutes, Tenericutes,

Table 2. Types of TCS/PR proteins found in the 7609 surveyed species. The protein identifier describes the type (HK, HPt, or RR) and number of TCS/PR domains fused in each protein.

Protein type	Total number of proteins found	Percentage of proteomes with this type of protein	Number of species with this type of protein	Average number of proteins/organism
RR	219436	97,07	7386	29,71
HK	151849	95,98	7303	20,79
HKRR	18383	48,57	3696	4,97
HKRRHPt	9097	40,85	3108	2,93
HKHPt	5506	41,99	3195	1,72
HPt	3534	28,05	2134	1,66
RR1RR2	2034	17,60	1339	1,52
HKRR1RR2	2017	13,59	1034	1,95
HKRR1HPtRR2	982	8,12	618	1,59
HK1RR1RR2RR3	580	6,58	501	1,16
HK1HK2	450	4,07	310	1,45
HK1RRHK2	392	3,30	251	1,56
RRHPt	312	3,47	264	1,18
HKRRHPt1HPt2HPt3	141	1,85	141	1,00
RR1RR2HPt	130	1,45	110	1,18
HKRRHPt1HPt2HPt3HPt4	108	1,42	108	1,00
HK1RR1HK2RR2	90	0,79	60	1,50
RR1RR2RR3HPt	72	0,51	39	1,85
HKRRHPt1HPt2HPt3HPt4HPt5	61	0,80	61	1,00
HKRRHPt1HPt2	58	0,72	55	1,05
HK1HK2RRHPt	39	0,50	38	1,03
HK1HK2HPt	39	0,50	38	1,03
HKHPt1HPt2	36	0,46	35	1,03
RR1RR2RR3	34	0,32	24	1,42
HKRR1RR2RR3HPt	33	0,37	28	1,18
HPt1HPt2	21	0,20	15	1,40
HKHPt1HPt2HPt3	16	0,20	15	1,07
HK1HK2RR1RR2RR3	9	0,12	9	1,00
HK1HK2HK3	9	0,04	3	3,00
HKRR1RR2RR3RR4RR5HPt	7	0,09	7	1,00
HKRRHPt1HPt2HPt3HPt4HPt5HPt6HPt7	7	0,09	7	1,00
HKRR1RR2RR3RR4	6	0,08	6	1,00
HK1HK2HK3HK4RR1RR2	6	0,08	6	1,00
HK1HK2RRHPt1HPt2	5	0,07	5	1,00
HKRR1RR2RR3RR4HPt	5	0,07	5	1,00
RR1RR2RR3RR4	2	0,03	2	1,00
HK1HK2RR1RR2HPt1HPt2	2	0,03	2	1,00
HK1HK2RR1RR2RR3RR4	2	0,03	2	1,00
HK1HK2HK3HK4	2	0,03	2	1,00
HK1HK2HPt1HPt2	2	0,03	2	1,00
HKRR1RR2HPt1HPt2	2	0,03	2	1,00
HK1HK2HK3RR	1	0,01	1	1,00
HPt1HPt2HPt3	1	0,01	1	1,00
HK1HK2RRHPt1HPt2HPt3	1	0,01	1	1,00
HKRR1RR2RR3HPt1HPt2HPt3	1	0,01	1	1,00
HPt1HPt2HPt3HPt4	1	0,01	1	1,00
HKRR1RR2HPt1HPt2HPt3	1	0,01	1	1,00
HK1HK2RR1RR2HPt	1	0,01	1	1,00
HK1HK2RR1RR2RR3RR4RR5RR6HPt	1	0,01	1	1,00
HKRRHPt1HPt2HPt3HPt4HPt5HPt6	1	0,01	1	1,00

Actinobacteria, Bacteroidetes and Spirochaetes. We also observe that $HKRR_1HPtRR_2$ is more abundant than $HK_1RR_1HK_2RR_2$ (Supplementary Table 2).

The relative abundances of proteins containing IST domains associated with TCS/PR cascades in eukaryotes are different from those of prokaryotes. In broad terms, HK and RR protein types tend to make for a smaller fraction of TCS/PR proteins in eukaryotes than in prokaryotes, while the opposite is observed for HKRR proteins. Another clear distinction between prokaryotes and eukaryotes refers to HPt-containing proteins: HPt protein type represents more than 10% of all TCS/PR proteins in eukaryotes. In prokaryotes, except in Tenericutes, HPt proteins typically account for less than 1% of TCS/PR proteins. Moreover, no HKRRHPt or $HKRR_1HPtRR_2$ protein types were found in eukaryotes (Supplementary Table 2).

Among protists, Euglenozoa proteomes contain mostly HK protein type, although HKRR type is the most abundant in *D. discoideum* (Amoeboflagellate). There are cases of inactive HK domains that have lost their histidine. When identified, these proteins were eliminated from the analysis as described in Methods. However, there is always the possibility that some such proteins have passed our filters. To control for that possibility we created a multiple alignment of the Euglenozoa HK proteins. We found that the HK domains contained the conserved histidine motif that is needed for HK signal transduction. Hence, these proteins could be active HK proteins. Furthermore, if we lower our e-value for cut-off to 10^{-4}, many of these proteins will also be flagged as containing RR domains with conserved aspartate residues, suggesting that such proteins could be HKRR types with a high degree of sequence divergence from other HKRR proteins we identified. Thus, the HK proteins in this clade might either be hybrid HKs or be active in a context that does not involve a TCS/PR cascade. TCS/PR proteins are almost absent in Alveolates. In the fungi phyla (Table 1), HKRR is the most abundant protein type in Basidiomycetes, making up for almost 50% of total TCS/PR proteins. In contrast, RR, HK and HPt protein types are relatively more abundant than HKRR protein type in Ascomycetes. A remarkable result in fungi is the relative abundance of HK_1RRHK_2 and $HK_1RR_1HK_2RR_2$, which are much more frequent in eukaryotes (above 10%) than in prokaryotes. In plants, RR is the most abundant protein type in Eudicots, making for about 60% of all TCS/PR proteins.

3.4.4. Distribution of genes coding for TCS/PR protein types in the genomes

Previous surveys found that many of the TCS/PR proteins are mostly organized in operons and/or regulons in prokaryotes [33, 64, 66, 67]. Consistent with this, we find that between 60% and 90% of genes containing HK domains are neighbors to genes containing RR domains. Exact percentages depend on the phylum, but below 20% of the total prokaryotic HK coding genes are orphan, that is, they are not neighboring any other gene coding for a protein that contains at least one IST domain. We also have found some clusters of genes coding HK, RR or HPt domains in eukaryotes, but all of them are a succession of genes with identical domain composition. Although the existence of operons has been reported in the eukaryote *C. elegans* [68], the gene clusters identified in our search have independent promoters.

Altogether, we found 530 different types of gene clusters coding for TCS/PR proteins. We now briefly describe these results, shown in Supplementary Table 9.

Neighborhood analysis for HK and RR protein types

In most prokaryotes neighboring genes coding for HK and RR protein types are between 50 and 100 times more frequent than one might expect by chance alone. In some species, this frequency is even higher (Supplementary Figure 1, Supplementary Table 3). Several phyla have a small percentage of species containing only orphan HK and RR protein types in their genomes (20 out of 2066 species in Firmicutes, 2 out of 635 in Actinobacteria, 6 out of 235 in Bacteroidetes, 11 out of 2246 in Gammaproteobacteria, 48 out of 451 in Alphaproteobacteria, 7 out of 366 in Betaproteobacteria, 4 out of 108 in Chlamydiae, 3 out of 118 in Cyanobacteria and 9 out of 179 in Euryarchaeota). Most of these species have a number of TCS/PR proteins below the average of their phylum.

Neighborhood analysis for HK-RR-HK2

Approximately 20% of all prokaryotic species have HK-RR-HK2 consecutive genes in their genomes at least 10 (and sometimes 50) times more frequently than one might expect by chance alone. Conversely, the frequency of this gene neighborhood organization is what one would expect by chance alone in the remaining 80% prokaryotic species (Supplementary Table 4).

Neighborhood analysis for HK-RR-HK2-RR2

In most prokaryotic phyla, between 10% and 60% of species have *HK-RR-HK2-RR2* genes at least 100 times more frequently than one would expect by chance alone (Supplementary Table 5).

Neighborhood analysis for HKRR-HK2-RR2

In the majority of prokaryotic species, genes coding for proteins of type HKRR have no neighboring genes coding for proteins of types HK or RR. Nevertheless, in more than 20% of the species of some prokaryotic phyla, such as Proteobacteria or Spirochaetes, genes coding for HKRR-protein type are neighbors to genes coding for HK or RR protein type with a frequency more than 100 times higher than expected by chance alone (Supplementary Table 6).

Neighborhood analysis for HKRRHPt next to RR2

In most of the prokaryotic species where HKRRHPt protein types are present, the observed frequency of HKRRHPt-RR genetic neighborhoods is between 10 and 50 times more frequent than one would expect by chance alone (Supplementary Table 7).

Neighborhood analysis for HK1RRHK2 next to RR2

In prokaryotes, HK_1RRHK_2 is a scarce protein, present only in a few species (Table 2). If present, it is located in the genome next to a RR protein type on average 31% of the times (Table 3). In Gammaproteobacteria, HK_1RRHK_2 is present only in 28 out of 2246 species surveyed, and in 9 of these 28 species, the observed frequency of HK_1RRHK_2 genes placed in the chromosome next to RR genes is more than 100 times higher than the random expected frequency (Supplementary Table 8).

Table 3. Total number of HKRRHPt and HKRRHK proteins found in prokaryotic phyla. Phyla in bold are from the bacterial domain. Italicized phyla are from the archaeal domain.

Phylum	Number of HKRRHPt/HK$_1$RRHK$_2$ proteins found	Number of HKRRHPt/HK$_1$RRHK$_2$ genes with a neighboring RR gene	% of HKRRHPt / HK$_1$RRHK$_2$ genes with a neighboring RR gene
Actinobacteria	12/4	9/1	75.00/25.00
Aquificae	0/0	0/0	-/-
Armatimonadetes	0/0	0/0	-/-
Bacteroidetes	107/9	62/4	57.94/44.44
Chlorobi	4/0	0/0	0.00/-
Caldiserica	0/0	0/0	-/-
Chlamydiae	2/0	1/0	50.00/-
Lentisphaerae	1/0	0/0	0.00/-
Verrucomicrobia	12/2	9/1	75.00/50.00
Chloroflexi	16/0	8/0	50.00/-
Chrysiogenetes	1/0	0/0	0.00/-
Cyanobacteria	193/28	41/9	21.24/32.14
Deferribacteres	9/0	7/0	77.78/-
Deinococcus-Thermus	0/4	0/1	-/25.00
Dictyoglomi	0/0	0/0	-/-
Elusimicrobia	0/0	0/0	-/-
Acidobacteria	1/5	1/2	100.00/40.00
Fibrobacteres	0/0	0/0	-/-
Firmicutes	65/97	44/69	67.69/71.13
Fusobacteria	2/0	2/0	100.00/-
Gemmatimonadetes	3/0	3/0	100.00/-
Nitrospinae	0/0	0/0	-/-
Nitrospirae	4/0	3/0	75.00/-
Planctomycetes	40/0	18/0	45.00/-
Alphaproteobacteria	337/10	233/5	69.14/50.00
Betaproteobacteria	364/9	274/4	75.27/44.44
Deltaproteobacteria	208/29	131/1	62.98/3.45
Epsilonproteobacteria	399/0	389/0	97.49/-
Gammaproteobacteria	7239/28	3336/15	46.08/53.57
Zetaproteobacteria	2/0	1/0	50.00/-
Spirochaetes	53/147	16/3	30.19/2.04
Synergistetes	6/0	6/0	100.00/-
Tenericutes	0/0	0/0	-/-
Thermodesulfobacteria	2/0	1/0	50.00/-
Thermotogae	6/0	5/0	83.33/-
Crenarchaeota	0/0	0/0	-/-
Euryarchaeota	9/1	3/0	33.33/0.00
Thaumarchaeota	0/0	0/0	-/-
Total	**9097/373**	**4603/115**	**50.60/30.83**

3.4.5. Gene fusion of TCS/PR proteins

Gene fusion events

The number of gene fusion events observed in a genome is expected to be proportional to genome size, in a model for neutral evolution of protein domain fusion [36, 69]. Thus, if gene fusion events in the case of HK and RR are random one would expect that the linear model that would best fit the data for % of fused HK (RR, HPt) domains vs. total number of HK (respectively, RR, HPt) domains has slope zero. In contrast, if these events are favored, the slope of that model should be positive, and if the events are disfavored, that slope should be negative.

We analyze fusion events of IST domains associated with TCS/PR cascades in the individual phyla by creating a linear model of percentage of fused HK (or RR) domains as a function of the total number of HK (or RR) domains in the genome and calculate the likelihood that the slope is different from zero. The results are shown in Table 4. We find that the percentage of fused HK (or RR) domains increases with the number of HK (or RR) domains in the genomes. This is consistent with a positive selection for fused HKRR proteins.

Table 4. Percentage of RR and HK domains in hybrid proteins as a function of the total number of HK and RR proteins in the genome. Phyla in bold are from the bacterial domain. Italicized phyla are from the archaeal domain. Other phyla are from the eukaryotic domain.

Phylum	RR	SK
Gammaproteobacteria	$6.97 + 0.2\,x^{**}$	$14 + 0.31\,x^{**}$
Betaproteobacteria	$1.90 + 0.22x^{**}$	$4.6 + 0.37x^{**}$
Epsilonproteobacteria	$15.3 - 0.06x^{+}$	$0.17 + 0.36x^{*}$
Deltaproteobacteria	$17.3 + 0.1x^{*}$	$31.9 + 0.06x^{+}$
Alphaproteobacteria	$4.1 + 0.29x^{**}$	$3.8 + 0.4x^{**}$
Firmicutes	$-0.6 + 0.1x^{**}$	$1.7 + 0.09x^{*}$
Tenericutes	---	---
Actinobacteria	$-4.5 + 0.32x^{**}$	$-0.38 + 0.2x^{*}$
Chlamydiae	---	---
Spirochaetes	$5 + 0.47x^{*}$	$27.2 + 0.07x^{+}$
Acidobacteria	$-5.7 + 0.26x$	$-16 + 0.53x^{*}$
Bacteroidetes	$30.7 + 0.05x^{+}$	$32 + 0.09x^{+}$
Fusobacteria	$-5.8 + x$	$-7.1 + 1.5x$
Verrumicrobia	$6.7 + 0.3x^{+}$	$6.6 + 0.4x^{+}$
Planctomycetes	$32.8 - 0.1x^{+}$	$49.8 - 0.21x^{+}$
Synergistetes	---	---
Cyanobacteria	$2.2 + 0.3x^{**}$	$5.8 + 0.4x^{**}$
Green sulfur bacteria	$31.6 + 0.7x^{+}$	$33.5 + 0.5x^{+}$
Green non-sulfur bacteria	$5.2 + 0.2x$	$8.9 + 0.2x$
Deinococcus-Thermus	$-1.2 + 0.2x$	$-1.6 + 0.2x$
Euryarchaeota	$6.9 + 0.6x^{*}$	$15.4 + 0.1x^{+}$
Crenarchaeota	---	---
Nanoarchaeota	---	---
Korarchaeota	---	---
Oomycetes	---	---
Diatoms	---	---
Parabasilids	---	---
Diplomonads	---	---
Euglenozoa	---	---
Alveolates	---	$9.5 + 0.4x^{+}$
Amoeboflagellates	---	---
Choanoflagellates	---	---
Microsporideans	---	---
Basidiomycetes	$25.7 + 3.7x$	$88 + 1.1x^{+}$
Ascomycetes	$37.4 + 2.5x^{**}$	$92.4 - 0.07x^{+}$
Red algae	---	---
Green algae	$29.2 + x^{+}$	$114.7 - 9x^{+}$
Mosses	---	---
Monocots	$12.3 + 0.2x^{+}$	$32.9 + 1.9x^{+}$
Eudicots	$17.4 + 0.1x^{+}$	$71 - 0.5x^{+}$

**p-value<10^{-8}; *p-value<10^{-3};$^{+}$ non-significant (p-value>0.1)

3.5. Discussion

3.5.1. Scope, caveats, and limitations of our analysis

In this work we analyze the distribution and prevalence of different types of TCS/PR proteins in 7609 organisms belonging to 52 phyla. These proteins are responsible for sensing and adequately regulating the cellular responses to environmental cues. To date, this is the largest survey of TCS/PR proteins we are aware of. We confirm that these proteins are predominantly prokaryotic, although they are also present in many eukaryotic phyla. However, functional TCS/PR cascades appear to be absent in animals. This is also consistent with previous findings [39] .

An important feature in this study is that we include all organisms with fully sequenced and annotated genomes in our analysis. For example, on the order of one thousand *Escherichia coli* strains are included in our analysis. This would clearly bias any deletion/duplication or horizontal gene transfer study of TCS/PR proteins that one might make in the full dataset. However, considering all strains and subspecies in our analysis is fundamental for identifying extremely low-frequency unique IST domain and operon organization types.

3.5.2. Identifying unique types of IST domain organization in TCS/PR cascades

The main goal of this analysis is to identify the unique types of organization for IST domains in proteins of TCS/PR cascades. In addition we also perform a less thorough identification of operon organization for TCS/PR proteins. This study was independently made in two ways: first, we eliminate all proteins annotated as hypothetical or partial. Subsequently we include such proteins in the analysis. Results are qualitatively similar in both cases, and the raw sequences in FASTA format can be downloaded from http://web.udl.es/usuaris/pg193845/Salvadoretal.html.

Our analysis identifies 50 unique types of TCS/PR proteins, when it comes to intra protein IST domain organization. The most frequent types of proteins with fused IST domains are the hybrid histidine kinases, a design with one HK and one RR protein domains fused in a single protein. This organization has been observed in most of the eukaryotic PRs that have been well characterized genetically and biochemically (for example the Sln1p-Ypd1p-Ssk1p pathway in *S. cerevisiae* [6] or the ETR1 system in *A. thaliana* [5]). It is also present in some prokaryotic systems (for example, the RcsC/YojN/RcsB pathway, involved in the regulation of capsular polysaccharide synthesis in *E. coli* [70] , and the Lux pathway regulating bioluminescence in *V. Harveyi* [71]). Another relatively frequent type of IST domain organization is when one HK, one HPt, and one RR domain are found within a single protein. Such proteins are called unorthodox histidine kinase or tripartite HK. Some examples of systems with this design are: BvgS-BvgA [72], EvgS/EvgA [73] , ArcB/ArcA [74], TorS/TorR [75], BarA/UvrY [76], TodS/TodT [77] and GacS/GacA [76].

We also identify 530 unique types of possible operons in prokaryotes and some eukaryotes, such as ascomycetes and eudicots (Supplementary Table 9). This variety will be used in subsequent works to infer naturally occurring variations in the pattern of regulatory interactions between the proteins involved in TCS/PR networks. For example, if we find a gene cluster formed by one HK and two RR coding genes, we can infer that the signaling pathway has a branching point in which the HK phosphorylates both RR. This alternative circuitry is important because it has been proved that network architecture affects network dynamics and can define the operational limits of the system in a way that is independent of the specific biological processes being regulated [46, 49, 52, 78, 79].

We have no way of identifying TCS/PR cascades at the regulon level using only sequence data. Many examples for this type of organization exist, such as the Kin-SpoO pathway [80].

Why do we focus only on the IST domains of TCS/PR cascades, rather than also including also other protein domain that are involved in TCS/PR signal transduction? By focusing on these domains and their organization, our results set the stage for an analysis of general dynamics organization principles in the internal transmission of signals within TCS/PR cascades. The organization of IST domains, either within a protein or within an operon, plays an important role in determining the dynamics of the signal transmission in a cascade [49, 51, 52, 81]. Hence, that organization is likely to be the subject of natural selection. Had we included other types of domains, we would be also analyzing aspects of the input and output of the cascades that are case specific and not general to all cascades of a given type.

3.5.3. Some physiological, phylogenetic and evolutionary considerations

In prokaryotes, approximately 90% of all PR proteins have only one HK domain or one RR domain (Table 2 and Supplementary Table 2), and most of the genes encoding these proteins are located in the chromosome next to other PR/TCS genes, forming operons. In contrast to this, in eukaryotes proteins of types HK and RR are less common, and genes encoding these proteins are never located next to other TCS/PR genes in the species surveyed. On the other hand, in eukaryotes there is a higher fraction of TCS/PR proteins containing a combination of the HK and RR domains (the HPt domain was not found in these eukaryotic multi domain TCS/PR proteins), such as HKRR, HK_1RRHK_2 and $HK_1RR_1HK_2RR_2$. This implies that TCS/PR signal transduction in eukaryotes is in principle less prone to crosstalk and noise, as the signal is internally transmitted within the same peptide chain [79, 82].

Our analysis confirms the *ad hoc* observation that coordinated expression of IST domains and/or TCS/PR proteins involved in the same cascade is frequent. We also quantify how much more frequent this coordinated expression is with respect to what one would expect by chance alone. Although this is not unexpected [66, 82-84], to our knowledge, such quantification had not been done before on such a large dataset.

This suggests that alternative IST regimes might be favored by evolution in prokaryotic or eukaryotic TCS/PR cascades. This can be inferred from the fact that the three types of gene expression coordination (regulon, operon, or gene fusion) imply different characteristics when it comes to internal signal transmission within the cascade. In general, fused genes will have a lower level of noise in signal transduction, followed by genes coded in the same operon, and with genes coded in the same regulon permitting the highest level of noise to enter the signal transduction process [83].

Why is this so? TCS/PR proteins whose expression is coordinated either at the regulon or operon levels are potentially translated in different amounts. RR proteins typically are orders of magnitude more abundant than HK proteins [49]. This leads to a type of signal transduction where amplification of the signal can be high, as many RR molecules can be modified by a single HK protein. In contrast, in hybrid kinases where the HK and RR domains are fused in the same protein, the ratio of HK/RR domains is one to one. This means that each HK domain will likely only phosphorylate one RR domain. Moreover, independent HK protein types might also be leakier, phosphorylating non-cognate RRs. Similarly, independent RR protein types can be more prone to phosphorylation by non-cognate sources. Such non-cognate phosphorylation events are physically harder to achieve in HKRR protein types. Thus, proteins with fused TCS/PR domains represent a design that will on average transduce signals with smaller amplification, but higher fidelity than TCS/PR cascades composed only of proteins with individual TCS/PR domains.

Taking these considerations into account, one might think that maximization of internal signal amplification is likely to be an important selective pressure for the evolution of TCS/PR cascades in prokaryotes, while fidelity of internal signal transmission appears to be a more important selective pressure for the evolution of TCS/PR cascades in eukaryotes. These two functional requirements for IST in TCS/PR cascades are generic and independent of more specific pressures, such as the type of signal they transduce, whether the organism is uni- or multi-cellular, or other similar considerations [15, 16, 34, 64, 66, 67]. If and why amplification and fidelity of internal signal transmission are indeed shaping the general organization of TCS/PR cascades is a matter to be investigated further in the future. This will be done in a forthcoming study by creating mathematical models for the TCS/PR cascade architectures identified in this study and comparing the dynamic behavior of each of the alternatives.

3.6. References

1. Wolanin, P.M., P.A. Thomason, and J.B. Stock, *Histidine protein kinases: key signal transducers outside the animal kingdom.* Genome Biology, 2002. **3**(10).
2. Cheung, J. and W.A. Hendrickson, *Sensor domains of two-component regulatory systems.* Current Opinion in Microbiology, 2010. **13**(2): p. 116-123.
3. Hoch, J.A., Silhavy, T.J., *Two-Component Signal Transduction.* American Society for Microbiology, Washington, D.C., 1995.
4. Parkinson, J.S., *Signal transduction schemes of bacteria.* Cell, 1993. **73**(5): p. 857-71.
5. Chang, C., et al., *Arabidopsis ethylene-response gene ETR1: similarity of product to two-component regulators.* Science, 1993. **262**(5133): p. 539-44.
6. Maeda, T., S.M. Wurgler-Murphy, and H. Saito, *A two-component system that regulates an osmosensing MAP kinase cascade in yeast.* Nature, 1994. **369**(6477): p. 242-5.
7. Appleby, J.L., J.S. Parkinson, and R.B. Bourret, *Signal transduction via the multi-step phosphorelay: not necessarily a road less traveled.* Cell, 1996. **86**(6): p. 845-8.
8. Thomason, P. and R. Kay, *Eukaryotic signal transduction via histidine-aspartate phosphorelay.* J Cell Sci, 2000. **113 (Pt 18)**: p. 3141-50.
9. Simon, M.I., B.R. Crane, and A. Crane, *Two-component signaling systems.* 2007.
10. Gross, R. and D. Beier, *Two-component systems in bacteria.* 2012: p. 426.
11. Inouye, M. and R. Dutta, *Histidine kinases in signal transduction.* 2003: p. 520.
12. Oka, A., H. Sakai, and S. Iwakoshi, *His-Asp phosphorelay signal transduction in higher plants: receptors and response regulators for cytokinin signaling in Arabidopsis thaliana.* Genes & genetic systems, 2002. **77**: p. 383-91.
13. Majdalani, N. and S. Gottesman, *The Rcs phosphorelay: a complex signal transduction system.* Annual review of microbiology, 2005. **59**: p. 379-405.
14. Bekker, M., M.J. Teixeira de Mattos, and K.J. Hellingwerf, *The role of two-component regulation systems in the physiology of the bacterial cell.* Science progress, 2006. **89**: p. 213-42.
15. Podgornaia, A.I. and M.T. Laub, *Determinants of specificity in two-component signal transduction.* Current Opinion in Microbiology, 2013. **16**: p. 156-162.
16. Laub, M.T. and M. Goulian, *Specificity in two-component signal transduction pathways.* Annual review of genetics, 2007. **41**: p. 121-45.
17. Szurmant, H., R.A. White, and J.A. Hoch, *Sensor complexes regulating two-component signal transduction.* Curr Opin Struct Biol, 2007. **17**(6): p. 706-15.
18. Ortiz de Orué Lucana, D. and M.R. Groves, *The three-component signalling system HbpS-SenS-SenR as an example of a redox sensing pathway in bacteria.* Amino acids, 2009. **37**: p. 479-86.
19. Casino, P., V. Rubio, and A. Marina, *The mechanism of signal transduction by two-component systems.* Current opinion in structural biology, 2010. **20**: p. 763-71.
20. Krell, T., et al., *Bacterial sensor kinases: diversity in the recognition of environmental signals.* Annual review of microbiology, 2010. **64**: p. 539-59.
21. Buelow, D.R. and T.L. Raivio, *Three (and more) component regulatory systems - auxiliary regulators of bacterial histidine kinases.* Molecular microbiology, 2010. **75**: p. 547-66.
22. Hazelbauer, G.L. and W.-C. Lai, *Bacterial chemoreceptors: providing enhanced features to two-component signaling.* Current opinion in microbiology, 2010. **13**: p. 124-32.
23. Szurmant, H. and J.A. Hoch, *Interaction fidelity in two-component signaling.* Current opinion in microbiology, 2010. **13**: p. 190-7.

24. Kobir, A., et al., *Protein phosphorylation in bacterial signal transduction.* Biochimica et biophysica acta, 2011. **1810**: p. 989-94.
25. Porter, S.L., G.H. Wadhams, and J.P. Armitage, *Signal processing in complex chemotaxis pathways.* Nature reviews. Microbiology, 2011. **9**: p. 153-65.
26. Schaller, G.E., S.-H. Shiu, and J.P. Armitage, *Two-component systems and their co-option for eukaryotic signal transduction.* Current biology : CB, 2011. **21**: p. R320-30.
27. Seshasayee, A.S.N. and N.M. Luscombe, *Comparative genomics suggests differential deployment of linear and branched signaling across bacteria.* Molecular bioSystems, 2011. **7**: p. 3042-9.
28. Jung, K., et al., *Histidine kinases and response regulators in networks.* Current opinion in microbiology, 2012. **15**: p. 118-24.
29. Fassler, J.S. and A.H. West, *Histidine phosphotransfer proteins in fungal two-component signal transduction pathways.* Eukaryotic cell, 2013. **12**: p. 1052-60.
30. Mascher, T., *Bacterial (intramembrane-sensing) histidine kinases: signal transfer rather than stimulus perception.* Trends in microbiology, 2014.
31. Jenal, U. and M.Y. Galperin, *Single domain response regulators: molecular switches with emerging roles in cell organization and dynamics.* Current opinion in microbiology, 2009. **12**: p. 152-60.
32. Galperin, M.Y. and A.N. Nikolskaya, *Identification of sensory and signal-transducing domains in two-component signaling systems.* Methods in enzymology, 2007. **422**: p. 47-74.
33. Galperin, M.Y., *Diversity of structure and function of response regulator output domains.* Current opinion in microbiology, 2010. **13**: p. 150-9.
34. Capra, E.J. and M.T. Laub, *Evolution of two-component signal transduction systems.* Annual review of microbiology, 2012. **66**: p. 325-47.
35. Perry, J., K. Koteva, and G. Wright, *Receptor domains of two-component signal transduction systems.* Molecular bioSystems, 2011. **7**: p. 1388-98.
36. Whitworth, D.E. and P.J.A. Cock, *Evolution of prokaryotic two-component systems: insights from comparative genomics.* Amino acids, 2009. **37**: p. 459-66.
37. Catlett, N.L., O.C. Yoder, and B.G. Turgeon, *Whole-genome analysis of two-component signal transduction genes in fungal pathogens.* Eukaryotic cell, 2003. **2**: p. 1151-61.
38. Cock, P.J.A. and D.E. Whitworth, *Evolution of prokaryotic two-component system signaling pathways: gene fusions and fissions.* Molecular biology and evolution, 2007. **24**: p. 2355-7.
39. Attwood, P.V., *Histidine kinases from bacteria to humans.* Biochemical Society transactions, 2013. **41**: p. 1023-8.
40. Kim, S., et al., *Identification and classification of a two-component system based on domain structures in bacteria and differences in domain structure between Gram-positive and Gram-negative bacteria.* Bioscience, biotechnology, and biochemistry, 2010. **74**: p. 716-20.
41. Wuichet, K., B.J. Cantwell, and I.B. Zhulin, *Evolution and phyletic distribution of two-component signal transduction systems.* Current opinion in microbiology, 2010. **13**: p. 219-25.
42. Sheng, X., et al., *Evolutionary characteristics of bacterial two-component systems.* Advances in experimental medicine and biology, 2012. **751**: p. 121-37.
43. Ortet, P., et al., *P2CS: updates of the prokaryotic two-component systems database.* Nucleic acids research, 2015. **43**: p. D536-41.
44. Jones, C.W. and J.P. Armitage, *Positioning of bacterial chemoreceptors.* Trends in microbiology, 2015. **23**: p. 247-256.

45. Huergo, L.F., et al., *PII signal transduction proteins: pivotal players in post-translational control of nitrogenase activity.* Microbiology (Reading, England), 2012. **158**: p. 176-90.

46. Salvado, B., et al., *Two component systems: physiological effect of a third component.* PLoS ONE, 2012. **7**: p. e31095.

47. Igoshin, O.A., et al., *A biochemical oscillator explains several aspects of Myxococcus xanthus behavior during development.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 15760-5.

48. Ray, J.C.J., J.J. Tabor, and O.A. Igoshin, *Non-transcriptional regulatory processes shape transcriptional network dynamics.* Nature reviews. Microbiology, 2011. **9**: p. 817-28.

49. Igoshin, O.A., R. Alves, and M.A. Savageau, *Hysteretic and graded responses in bacterial two-component signal transduction.* Mol Microbiol, 2008. **68**(5): p. 1196-215.

50. Eswaramoorthy, P., et al., *Single-cell measurement of the levels and distributions of the phosphorelay components in a population of sporulating Bacillus subtilis cells.* Microbiology (Reading, England), 2010. **156**: p. 2294-304.

51. Narula, J., et al., *Ultrasensitivity of the Bacillus subtilis sporulation decision.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**: p. E3513-22.

52. Alves, R. and M.A. Savageau, *Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function.* Mol Microbiol, 2003. **48**(1): p. 25-51.

53. Swain, P.S., *Efficient attenuation of stochasticity in gene expression through post-transcriptional control.* J Mol Biol, 2004. **344**(4): p. 965-76.

54. Kuchina, A., et al., *Reversible and noisy progression towards a commitment point enables adaptable and reliable cellular decision-making.* PLoS computational biology, 2011. **7**: p. e1002273.

55. Kuchina, A., et al., *Temporal competition between differentiation programs determines cell fate choice.* Molecular systems biology, 2011. **7**: p. 557.

56. Süel, G.M., et al., *An excitable gene regulatory circuit induces transient cellular differentiation.* Nature, 2006. **440**: p. 545-50.

57. Ulrich, L.E. and I.B. Zhulin, *The MiST2 database: a comprehensive genomics resource on microbial signal transduction.* Nucleic acids research, 2010. **38**: p. D401-7.

58. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic acids research, 2014. **42**: p. D222-30.

59. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.

60. Johnson, L.S., S.R. Eddy, and E. Portugaly, *Hidden Markov model speed heuristic and iterative HMM search procedure.* BMC bioinformatics, 2010. **11**: p. 431.

61. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**: p. 403-10.

62. Wolfram, S., *The MATHEMATICA ® Book, Version 4.* 1999: p. 1496.

63. Clore, G.M. and V. Venditti, *Structure, dynamics and biophysics of the cytoplasmic protein-protein complexes of the bacterial phosphoenolpyruvate: sugar phosphotransferase system.* Trends in biochemical sciences, 2013. **38**: p. 515-30.

64. Galperin, M.Y., R. Higdon, and E. Kolker, *Interplay of heritage and habitat in the distribution of bacterial signal transduction systems.* Molecular bioSystems, 2010. **6**: p. 721-8.

65. Ulrich, L.E., E.V. Koonin, and I.B. Zhulin, *One-component systems dominate signal transduction in prokaryotes.* Trends Microbiol, 2005. **13**(2): p. 52-6.

66. Alm, E., K. Huang, and A. Arkin, *The evolution of two-component systems in bacteria reveals different strategies for niche adaptation.* PLoS computational biology, 2006. **2**: p. e143.
67. Williams, R.H.N. and D.E. Whitworth, *The genetic organisation of prokaryotic two-component system signalling pathways.* BMC genomics, 2010. **11**: p. 720.
68. Blumenthal, T., P. Davis, and A. Garrido-Lecca, *Operon and non-operon gene clusters in the C. elegans genome.* WormBook : the online review of C. elegans biology, 2015: p. 1-20.
69. Durrens, P., M. Nikolski, and D. Sherman, *Fusion and fission of genes define a metric between fungal genomes.* PLoS computational biology, 2008. **4**: p. e1000200.
70. Takeda, S., et al., *A novel feature of the multistep phosphorelay in Escherichia coli: a revised model of the RcsC --> YojN --> RcsB signalling pathway implicated in capsular synthesis and swarming behaviour.* Mol Microbiol, 2001. **40**(2): p. 440-50.
71. Freeman, J.A. and B.L. Bassler, *Sequence and function of LuxU: a two-component phosphorelay protein that regulates quorum sensing in Vibrio harveyi.* J Bacteriol, 1999. **181**(3): p. 899-906.
72. Uhl, M.A. and J.F. Miller, *Central role of the BvgS receiver as a phosphorylated intermediate in a complex two-component phosphorelay.* J Biol Chem, 1996. **271**(52): p. 33176-80.
73. Bock, A. and R. Gross, *The unorthodox histidine kinases BvgS and EvgS are responsive to the oxidation status of a quinone electron carrier.* Eur J Biochem, 2002. **269**(14): p. 3479-84.
74. Georgellis, D., A.S. Lynch, and E.C. Lin, *In vitro phosphorylation study of the arc two-component signal transduction system of Escherichia coli.* J Bacteriol, 1997. **179**(17): p. 5429-35.
75. Bordi, C., et al., *Anticipating an alkaline stress through the Tor phosphorelay system in Escherichia coli.* Mol Microbiol, 2003. **48**(1): p. 211-23.
76. Sahu, S.N., et al., *The bacterial adaptive response gene, barA, encodes a novel conserved histidine kinase regulatory switch for adaptation and modulation of metabolism in Escherichia coli.* Mol Cell Biochem, 2003. **253**(1-2): p. 167-77.
77. Silva-Jimenez, H., J.L. Ramos, and T. Krell, *Construction of a prototype two-component system from the phosphorelay system TodS/TodT.* Protein Eng Des Sel, 2012. **25**(4): p. 159-69.
78. Cagatay, T., et al., *Architecture-dependent noise discriminates functionally analogous differentiation circuits.* Cell, 2009. **139**(3): p. 512-22.
79. Tiwari, A., et al., *The interplay of multiple feedback loops with post-translational kinetics results in bistability of mycobacterial stress response.* Physical Biology, 2010. **7**(3).
80. Burbulys, D., K.A. Trach, and J.A. Hoch, *Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay.* Cell, 1991. **64**(3): p. 545-52.
81. Ray, J.C.J. and O.A. Igoshin, *Adaptable Functionality of Transcriptional Feedback in Bacterial Two-Component Systems.* PLoS Comput Biol, 2010. **6**(2).
82. Tiwari, A. and O.A. Igoshin, *Coupling between feedback loops in autoregulatory networks affects bistability range, open-loop gain and switching times.* Physical biology, 2012. **9**: p. 055003.
83. Ray, J.C. and O.A. Igoshin, *Interplay of gene expression noise and ultrasensitive dynamics affects bacterial operon organization.* PLoS Comput Biol, 2012. **8**(8): p. e1002672.
84. Price, M.N., A.P. Arkin, and E.J. Alm, *The life-cycle of operons.* PLoS Genet, 2006. **2**(6): p. e96.

3.7. Supplementary materials

Supplementary Table 1: list of the 7609 species, classified per phylum. Due to space limitations, this table is not printed here. See Table S1 in the digital version of this thesis.

Supplementary Table 2: Percentage of each TCS/PR protein type per phylum. Phylum abbreviations are given in Table 1. Only phyla with proteins containing HK, RR or HPt domains are represented. Korarchaeota, Nanoarchaeota, Nanohaloarchaeota and phyla from the animal kingdom do not appear in the table because we have not found any protein containing HK, RR or HPt domains in the surveyed species classified in these phyla.

Phylum	RR	HK	HKRR	HPt	HKRRHPt	HK_1RRHK_2	$HKRR_1HPtRR_2$	$HK_1RR_1HK_2RR_2$
At	55.04	43.23	1.00	0.30	0.04	0.01	0.05	0.00
Aq	55.80	38.41	0.00	1.81	0.00	0.00	0.00	0.00
Ar	61.54	30.77	3.85	0.00	0.00	0.00	0.00	0.00
Ba	51.36	36.04	9.16	1.04	0.91	0.08	0.15	0.00
Cb	38.37	33.06	18.37	0.82	0.82	0.00	2.24	0.00
Cd	42.86	50.00	7.14	0.00	0.00	0.00	0.00	0.00
Cm	33.19	66.18	0.00	0.21	0.42	0.00	0.00	0.00
L	42.50	42.50	7.50	5.00	2.50	0.00	0.00	0.00
V	46.21	29.66	15.07	0.99	1.18	0.20	0.39	0.20
Cf	50.14	39.02	7.00	0.32	0.72	0.00	0.05	0.00
Cr	57.29	32.29	4.17	0.00	1.04	0.00	1.04	0.00
Cy	44.97	31.94	13.53	1.03	1.80	0.26	0.88	0.20
Df	51.25	34.69	6.88	0.31	2.81	0.00	0.63	0.00
Dt	54.14	41.77	1.28	0.23	0.00	0.47	0.00	0.00
Dc	50.00	46.43	3.57	0.00	0.00	0.00	0.00	0.00
El	50.00	41.67	8.33	0.00	0.00	0.00	0.00	0.00
Ac	53.43	33.04	8.79	0.87	0.10	0.48	0.39	0.29
Fb	48.00	16.00	16.00	4.00	0.00	0.00	4.00	0.00
Fi	54.90	42.69	0.44	0.60	0.08	0.12	0.04	0.00
Fu	59.74	37.65	1.09	0.87	0.22	0.00	0.00	0.00
Ge	47.58	32.26	14.52	0.81	2.42	0.00	0.00	0.00
Ni	53.42	24.66	15.07	2.74	0.00	0.00	1.37	0.00
Nt	56.99	29.72	3.50	1.40	1.40	0.00	1.40	0.00
Pl	53.67	25.84	10.84	1.47	1.89	0.00	1.14	0.00
A	53.36	31.75	7.66	1.28	1.15	0.03	0.24	0.07
B	54.56	34.34	5.83	0.54	1.43	0.04	0.51	0.05
D	49.17	30.06	11.86	1.39	1.63	0.23	0.68	0.10
E	59.14	33.36	0.83	0.65	4.07	0.00	0.07	0.00
G	52.58	34.48	3.99	0.91	4.36	0.02	0.27	0.00
Z	46.07	22.47	26.97	0.00	2.25	0.00	1.12	0.00
S	49.08	32.84	9.93	1.40	0.33	0.91	0.09	0.00
Sy	57.14	34.43	1.10	1.83	2.20	0.00	0.73	0.00
T	54.31	40.10	0.00	5.58	0.00	0.00	0.00	0.00
Th	50.00	31.25	11.25	1.25	2.50	0.00	0.00	0.00
Tt	53.73	39.55	0.25	0.00	1.49	0.00	0.75	0.00
C	63.64	30.30	0.00	6.06	0.00	0.00	0.00	0.00
Eu	38.20	47.45	11.49	0.08	0.15	0.02	0.05	0.00
Ta	59.12	36.82	0.34	2.36	0.00	0.00	0.00	0.00
Av	0.00	50.00	0.00	50.00	0.00	0.00	0.00	0.00
Am	22.73	0.00	40.91	9.09	0.00	0.00	0.00	4.55
Eg	6.06	93.94	0.00	0.00	0.00	0.00	0.00	0.00
Mi	66.67	33.33	0.00	0.00	0.00	0.00	0.00	0.00
As	28.10	18.60	15.29	19.83	0.00	6.61	0.00	2.48
Bs	18.18	4.55	45.45	9.09	0.00	9.09	0.00	13.64
Ed	60.12	7.36	13.50	16.56	0.00	0.61	0.00	0.00
M	0.00	63.64	0.00	36.36	0.00	0.00	0.00	0.00

Supplementary Table 3. Percentage of HK genes and RR genes that are neighbors in the genome to other TCS/PR genes. Phylum abbreviations are given in Table 1. Only phyla with HK and RR genes are represented. Alveolates are omitted because we found only 1 HK protein and 1 HPt protein in 5 species surveyed.

Phylum	Orphan HK	HK next to RR	HK next to RR and HK$_2$	HK next to RR, HK$_2$ and RR$_2$	Orphan RR	RR next to HK	RR next to HPt	RR next to HKRRHPt	RR next to HKRRHK
At	20,64	75,82	0,33	0,40	34,72	59,55	0,08	0,01	0,00
Aq	37,74	58,49	0,94	0,00	46,10	40,26	0,00	0,00	0,00
Ar	50,00	50,00	0,00	0,00	62,50	25,00	0,00	0,00	0,00
Ba	31,33	54,98	1,53	0,61	47,78	38,57	0,10	0,94	0,05
Cb	37,04	40,74	3,09	0,00	31,38	35,11	0,53	0,00	0,00
Cd	28,57	71,43	0,00	0,00	0,00	83,33	0,00	0,00	0,00
Cm	65,62	34,38	0,00	0,00	27,04	68,55	0,00	0,63	0,00
L	41,18	52,94	0,00	0,00	35,29	52,94	0,00	0,00	0,00
V	24,92	59,14	0,66	0,00	36,46	37,95	0,43	0,64	0,00
Cf	32,41	51,50	1,04	1,16	36,04	40,09	0,09	0,27	0,00
Cr	48,39	16,13	0,00	3,23	70,91	9,09	0,00	0,00	0,00
Cy	56,50	25,99	0,47	0,06	50,66	18,46	0,10	0,46	0,06
Df	36,04	53,15	1,80	0,90	39,63	35,98	0,00	3,66	0,00
Dt	31,01	60,89	1,12	0,84	37,93	46,98	0,00	0,00	0,22
Dc	30,77	61,54	0,00	0,00	42,86	57,14	0,00	0,00	0,00
El	60,00	20,00	0,00	0,00	66,67	16,67	0,00	0,00	0,00
Ac	28,65	58,48	0,00	1,46	41,41	36,17	0,18	0,18	0,18
Fb	0,00	75,00	0,00	0,00	33,33	25,00	0,00	0,00	0,00
Fi	12,15	84,82	0,34	0,37	29,32	65,96	0,08	0,06	0,14
Fu	10,40	86,99	0,29	0,00	40,80	54,83	0,55	0,18	0,00
Ge	20,00	65,00	0,00	2,50	28,81	44,07	0,00	3,39	0,00
Ni	33,33	61,11	0,00	0,00	46,15	28,21	0,00	0,00	0,00
Nt	11,76	61,18	2,35	1,18	30,06	31,90	0,00	1,23	0,00
Pl	37,91	48,90	0,37	0,18	53,53	23,54	0,88	1,41	0,00
A	32,43	51,81	0,84	2,84	42,09	30,83	0,72	0,76	0,01
B	13,04	77,21	0,46	0,94	27,85	48,59	0,20	1,31	0,02
D	35,58	39,04	1,12	0,83	40,22	23,87	0,46	1,02	0,00
E	24,20	72,50	0,37	0,31	48,85	40,90	0,02	6,68	0,00
G	11,86	81,24	0,41	0,25	26,41	53,27	0,46	3,28	0,02
Z	10,00	90,00	0,00	0,00	21,95	43,90	0,00	0,00	0,00
S	35,93	45,26	0,09	0,02	50,51	30,28	0,04	0,13	0,00
Sy	21,28	75,53	0,00	1,06	39,74	45,51	0,64	1,92	0,00
T	1,27	98,73	0,00	0,00	27,10	72,90	0,00	0,00	0,00
Th	36,00	60,00	0,00	0,00	37,50	37,50	0,00	0,00	0,00
Tt	38,36	58,49	0,00	0,63	42,59	43,06	0,00	2,31	0,00
C	50,00	50,00	0,00	0,00	76,19	23,81	0,00	0,00	0,00
Eu	69,37	18,61	0,43	0,11	51,56	23,12	0,00	0,04	0,00
Ta	35,78	40,37	0,92	0,00	50,86	25,14	0,00	0,00	0,00
Eg	45,16	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
Mi	100,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
As	100,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
Bs	100,00	0,00	0,00	0,00	100,00	0,00	0,00	0,00	0,00
Ed	100,00	0,00	0,00	0,00	59,18	0,00	0,00	0,00	0,00

Supplementary Table 4. Odds ratios (ratio between the observed and the randomly expected frequency) of HK genes located next to RR and HK_2 genes in the genome. Only species with HK and RR genes are taken into account in the percentages. Alveolates and Monocots do not appear in the table because we have not found RR proteins in the surveyed species belonging to these phyla. Amoeboflagellates do not appear because we have not found HK proteins in the surveyed species classified in this phylum. Phylum abbreviations are given in Table 1.

Phylum	% of species with 2<odds ratio<10	% of species with 10<odds ratio<50	% of species with 50<odds ratio<100	% of species with odds ratio>100
At	0.80	6.38	2.23	5.42
Aq	0.00	0.00	0.00	7.69
Ar	0.00	0.00	0.00	0.00
Ba	1.93	15.94	5.80	6.76
Cb	0.00	7.14	0.00	21.43
Cd	0.00	0.00	0.00	0.00
Cm	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00
V	0.00	20.00	10.00	20.00
Cf	13.04	39.13	0.00	0.00
Cr	0.00	100.00	0.00	0.00
Cy	7.63	13.56	2.54	2.54
Df	0.00	25.00	0.00	0.00
Dt	0.00	15.00	0.00	0.00
Dc	0.00	0.00	0.00	0.00
El	0.00	0.00	0.00	0.00
Ac	0.00	55.56	0.00	0.00
Fb	0.00	0.00	0.00	0.00
Fi	1.72	4.62	1.48	2.71
Fu	0.00	0.00	0.00	5.56
Ge	0.00	100.00	0.00	0.00
Ni	0.00	0.00	0.00	0.00
Nt	0.00	50.00	0.00	0.00
Pl	0.00	25.00	10.00	5.00
A	1.43	15. 71	9.05	12.38
B	3.64	11.20	1.40	1.12
D	25.64	32.05	0.00	0.00
E	1.22	1.47	0.00	16.87
G	1.00	13.80	20.00	8.02
Z	0.00	0.00	0.00	0.00
S	0.40	1.62	2.02	0.81
Sy	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00
Th	0.00	0.00	0.00	0.00
Tt	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00
Eu	0.00	8.03	7.30	8.03
Ta	0.00	25.00	12.50	12.50
Eg	0.00	0.00	0.00	0.00
Mi	0.00	0.00	0.00	0.00
As	0.00	0.00	0.00	0.00
Bs	0.00	0.00	0.00	0.00
Ed	0.00	0.00	0.00	0.00

Supplementary Table 5. Odds ratios (ratio between the observed and the randomly expected frequency) of HK genes located in the genome next to RR, HK_2 and RR_2 genes. Only species with HK and RR genes are taken into account in the percentages. Alveolates and Monocots do not appear in the table because we have not found RR proteins in the surveyed species belonging to these phyla. Amoeboflagellates do not appear because we have not found HK proteins in the surveyed species classified in this phylum. Phylum abbreviations are given in Table 1.

Phylum	% of species with 2<odds ratio<10	% of species with 10<odds ratio<50	% of species with 50<odds ratio<100	% of species with odds ratio>100
At	0.00	0.00	0.00	11.96
Aq	0.00	0.00	0.00	0.00
Ar	0.00	0.00	0.00	0.00
Ba	0.00	0.48	0.00	20.77
Cb	0.00	0.00	0.00	14.29
Cd	0.00	0.00	0.00	0.00
Cm	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	20.00
Cf	0.00	0.00	8.70	34.78
Cr	0.00	0.00	0.00	100.00
Cy	0.00	0.00	0.00	10.17
Df	0.00	0.00	0.00	25.00
Dt	0.00	0.00	0.00	30.00
Dc	0.00	0.00	0.00	0.00
El	0.00	0.00	0.00	0.00
Ac	0.00	0.00	0.00	66.67
Fb	0.00	0.00	0.00	0.00
Fi	0.00	0.10	0.20	7.53
Fu	0.00	0.00	0.00	0.00
Ge	0.00	0.00	0.00	100.00
Ni	0.00	0.00	0.00	0.00
Nt	0.00	0.00	0.00	50.00
Pl	0.00	0.00	0.00	25.00
A	0.00	0.00	0.24	61.22
B	0.00	0.28	0.28	26.99
D	0.00	2.56	7.69	62.82
E	0.00	0.24	0.00	18.09
G	0.00	0.00	0.00	16.60
Z	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	2.02
Sy	0.00	0.00	0.00	20.00
T	0.00	0.00	0.00	0.00
Th	0.00	0.00	0.00	0.00
Tt	0.00	0.00	0.00	5.88
C	0.00	0.00	0.00	0.00
Eu	0.00	0.00	0.00	6.57
Ta	0.00	0.00	0.00	0.00
Eg	0.00	0.00	0.00	0.00
Mi	0.00	0.00	0.00	0.00
As	0.00	0.00	0.00	0.00
Bs	0.00	0.00	0.00	0.00
Ed	0.00	0.00	0.00	0.00

Supplementary Table 6. Odds ratios (ratio between the observed and the randomly expected frequency) of HKRR genes located in the genome next to HK_2 and RR_2 genes. Only species with HKRR genes are taken into account in the percentages. Phyla without this type of protein (Aquificae, Chlamydiae , Tenericutes, Crenarchaeota, Alveolates, Euglenozoa, Microsporidians and Monocots) do not appear in the table. Amoeboflagellates do not appear because we have not found HK proteins in the surveyed species classified in this phylum. Phylum abbreviations are given in Table 1.

Phylum	% of species with 2<odds ratio<10	% of species with 10<odds ratio<50	% of species with 50<odds ratio<100	% of species with odds ratio>100
At	0.00	0.00	0.72	9.42
Ar	0.00	0.00	0.00	0.00
Ba	0.00	8.18	4.40	8.18
Cb	0.00	14.29	7.14	0.00
Cd	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00
V	14.29	14.29	0.00	0.00
Cf	0.00	25.00	5.00	10.00
Cr	0.00	100.00	0.00	0.00
Cy	1.10	34.07	8.79	4.40
Df	0.00	25.00	0.00	0.00
Dt	0.00	0.00	0.00	0.00
Dc	0.00	0.00	0.00	0.00
El	0.00	0.00	0.00	100.00
Ac	0.00	33.33	0.00	44.44
Fb	0.00	0.00	0.00	0.00
Fi	0.00	1.17	2.34	4.68
Fu	0.00	0.00	0.00	0.00
Ge	0.00	0.00	0.00	0.00
Ni	0.00	0.00	100.00	0.00
Nt	0.00	50.00	0.00	0.00
Pl	0.00	5.26	26.32	26.32
A	0.00	8.94	8.94	12.57
B	0.00	9.31	11.27	25.00
D	16.88	42.86	2.60	2.60
E	0.00	9.09	9.09	9.09
G	0.00	2.98	2.09	17.65
Z	0.00	0.00	0.00	0.00
S	0.00	1.27	6.36	31.36
Sy	0.00	0.00	0.00	0.00
Th	0.00	0.00	0.00	0.00
Tt	0.00	0.00	0.00	0.00
Eu	0.00	3.77	8.49	14.15
Ta	0.00	0.00	0.00	0.00
As	0.00	0.00	0.00	0.00
Bs	0.00	0.00	0.00	0.00
Ed	0.00	0.00	0.00	0.00

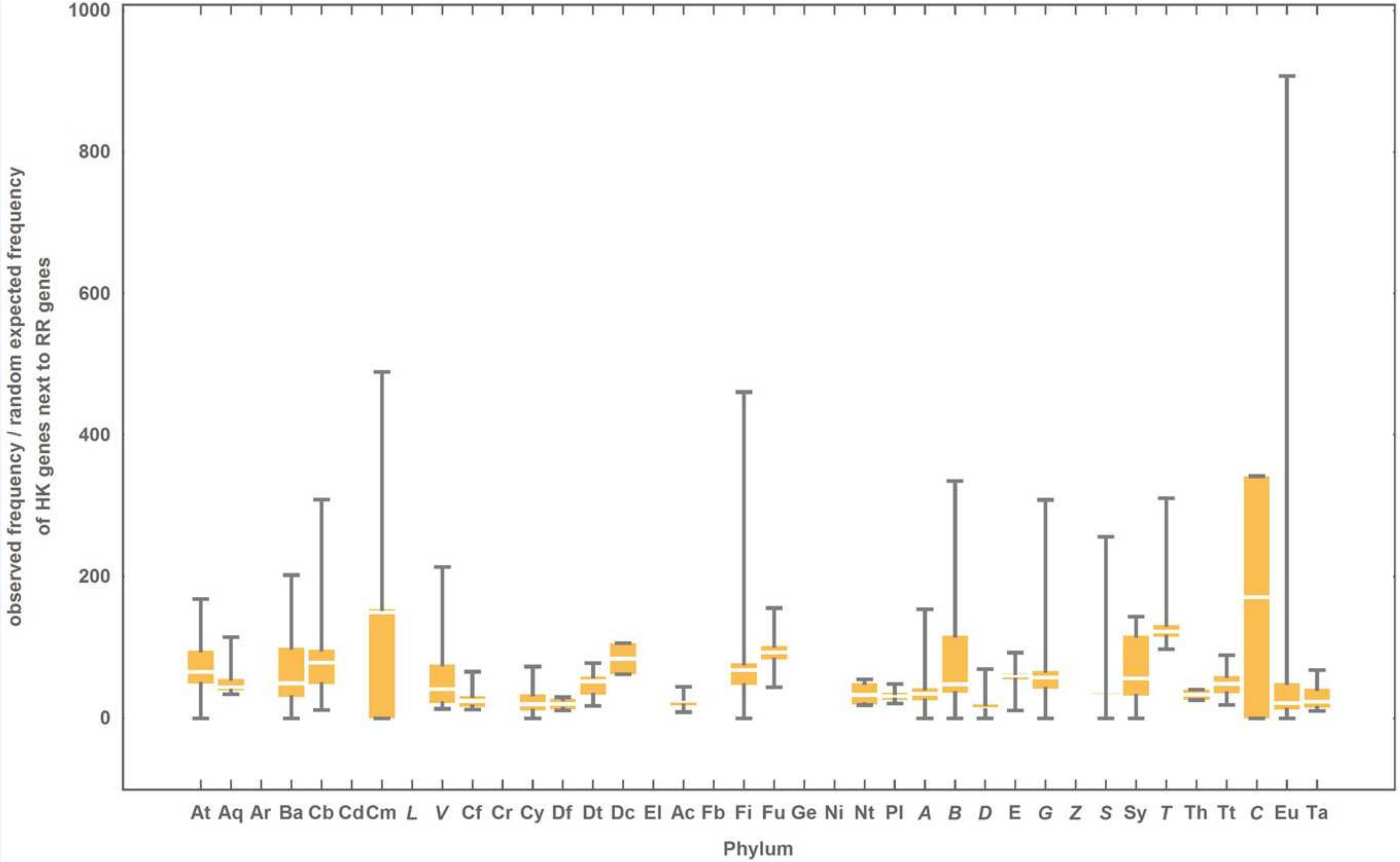
Supplementary Table 7. Odds ratios (ratio between the observed and the randomly expected frequency) of HKRRHPt genes located in the genome next to RR_2 genes. Only species with HKRRHPt proteins are taken into account in the percentages. Prokaryotic phyla without this type of protein (Aquificae, Armatimonadetes, Caldiserica, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Fibrobacteres, Nitrospinae, Tenericutes, Crenarchaeota and Thaumarchaeota) do not appear in the table. Eukaryotes are not included in the table since we have not found any HKRRHPt protein in this domain. Phylum abbreviations are given in Table 1.

Phylum	% of species with 2<odds ratio<10	% of species with 10<odds ratio<50	% of species with 50<odds ratio<100	% of species with odds ratio>100
At	0.00	85.71	14.29	0.00
Ba	1.59	33.33	25.40	4.76
Cb	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00
V	0.00	80.00	20.00	0.00
Cf	0.00	44.44	0.00	0.00
Cr	0.00	0.00	0.00	0.00
Cy	15.38	32.31	3.08	0.00
Df	25.00	25.00	0.00	0.00
Ac	0.00	100.00	0.00	0.00
Fi	0.00	51.02	22.45	2.04
Fu	0.00	0.00	0.00	100.00
Ge	0.00	100.00	0.00	0.00
Nt	0.00	25.00	50.00	0.00
Pl	0.00	62.50	0.00	0.00
A	3.42	48.63	21.92	0.00
B	1.86	60.25	26.09	0.00
D	16.92	64.62	0.00	0.00
E	0.00	1.82	95.32	1.30
G	5.37	68.49	1.87	0.05
Z	0.00	100.00	0.00	0.00
S	14.29	35.71	14.29	0.00
Sy	0.00	100.00	0.00	0.00
Th	0.00	50.00	0.00	0.00
Tt	0.00	0.00	0.00	0.00
Eu	0.00	12.50	12.50	12.50

Supplementary Table 8. Odds ratios (ratio between the observed and the randomly expected frequency) of HKRRHK$_2$ genes located in the genome next to RR$_2$ genes. Only species with HKRRHK$_2$ proteins are taken into account in the percentages. Phyla without this type of protein (Aquificae, Armatimonadetes, Chlorobi, Caldiserica, Chlamydiae, Lentisphaerae, Chloroflexi, Chrysiogenetes, Deferribacteres, Dictyoglomi, Elusimicrobia, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Nitrospinae, Nitrospirae, Planctomycetes, Epsilonproteobacteria, Zetaproteobacteria, Synergistetes, Tenericutes, Thermodesulfobacteria, Thermotogae, Crenarchaeota, Thaumarchaeota, Alveolates, Amoeboflagellate, Euglenozoa, Microsporidians and Monocots) do not appear in the table. Eukaryotic phyla with HKRRHK$_2$ genes are not included in this statistics because none of those HKRRHK$_2$ genes have been found neighboring an RR gene. Phylum abbreviations are given in Table 1.

Phylum	% of species with 2<odds ratio<10	% of species with 10<odds ratio<50	% of species with 50<odds ratio<100	% of species with odds ratio>100
At	0.00	0.00	0.00	25.00
Ba	0.00	25.00	0.00	0.00
V	0.00	50.00	0.00	0.00
Cy	0.00	38.89	0.00	0.00
Dt	0.00	0.00	33.33	0.00
Ac	0.00	33.33	0.00	0.00
Fi	1.64	37.70	44.26	0.00
A	0.00	55.56	0.00	0.00
B	0.00	22.22	22.22	0.00
D	0.00	8.33	0.00	0.00
G	0.00	17.86	3.57	32.14
S	0.00	3.49	0.00	0.00
Eu	0.00	0.00	0.00	0.00

Supplementary Table 9: List of all clusters of genes containing IST domains found in our search. Due to space limitations, this table is not printed here. See Table S9 in the digital version of this thesis.



Supplementary Figure 1.Ratio between the observed and the randomly expected frequency of HK genes located next to RR genes in the genome. Phylum abbreviations are explained in Table 1. The colored box represents the range of percentage values comprised between the 25% and the 75% quantiles, and the edges of the vertical bar denote the upper and lower percentage values for each phylum.

4 Two Component Systems: Physiological effect of a third component

4.1. Abstract

Signal transduction systems mediate the response and adaptation of organisms to environmental changes. In prokaryotes, this signal transduction is often done through Two Component Systems (TCS). These TCS are phosphotransfer protein cascades, and in their prototypical form they are composed by a kinase that senses the environmental signals (SK) and by a response regulator (RR) that regulates the cellular response. This basic motif can be modified by the addition of a third protein that interacts either with the SK or the RR in a way that could change the dynamic response of the TCS module.

In this work we aim at understanding the effect of such an additional protein (which we call "third component") on the functional properties of a prototypical TCS. To do so we build mathematical models of TCS with alternative designs for their interaction with that third component. These mathematical models are analyzed in order to identify the differences in dynamic behavior inherent to each design, with respect to functionally relevant properties such as sensitivity to changes in either the parameter values or the molecular concentrations, temporal responsiveness, possibility of multiple steady states, or stochastic fluctuations in the system. The differences are then correlated to the physiological requirements that impinge on the functioning of the TCS. This analysis sheds light on both, the dynamic behavior of synthetically designed TCS, and the conditions under which natural selection might favor each of the designs.

We find that a third component that modulates SK activity increases the parameter space where a bistable response of the TCS module to signals is possible, if SK is monofunctional, but decreases it when the SK is bifunctional. The presence of a

third component that modulates RR activity decreases the parameter space where a bistable response of the TCS module to signals is possible.

4.2. Introduction

Two component systems (TCS) are biochemical signaling modules that are ubiquitous in bacteria and are also present in some eukaryotes. Prototypical TCS are composed of two proteins: a sensor kinase (SK) and a response regulator (RR). The SK phosphorylates a histidine residue and subsequently transfers the phosphate to an aspartate residue in the RR. There are many variations around this prototype, ranging from phosphorelays that can concatenate up to three phosphotransfers (His→Asp→His→Asp) between different proteins to hybrid kinases in which the SK and the RR domains are fused in the same protein [1,2]. In prototypical TCS, the SK can be ***bifunctional*** if, when unphosphorylated, it increases the dephosphorylation rate of the RR. Otherwise, the SK is ***monofunctional***. The majority of well characterized SKs are bifunctional, with a few, such as the chemotaxis regulating CheA, being monofunctional.

In addition to SKs and RRs, some TCS are also known to interact with specific phosphatases that regulate dephosphorylation of the RR [3]. These core components of TCS and phosphorelays are also complemented by auxiliary proteins that play a regulatory role in the activity of some TCS, transmitting the cognate signal to the SK. For example, the SK CheA is regulated through its interaction with membrane receptors that detect chemical compounds in the medium and direct organisms towards higher concentrations of nutrients [4] and the activity of the SK NRII that regulates nitrogen fixation is modulated through its interaction with the protein PII [5].

In recent years, interactions between the TCS and auxiliary proteins were identified as a strategy to integrate non-cognate signals in the regulation of TCS [6]. For example, the orphan SK RetS interacts with the GacS SK, preventing the response of the latter to its cognate signal [7,8,9,10,11] and the peptide PmrD binds to and protects the phosphorylated form of the RR PmrA from the phosphatase activity of its cognate SK, PmrB [12]. The GacS/GacA TCS regulates virulence in *Pseudomonas aeruginosa* [13,14], while the PmrB/PmrA TCS is required for resistance of *Salmonella* to acidic and antibiotic stresses, among others [12,15]. These systems raise the question of understanding the effect of such interactions with the core TCS module in the operating regime of the module and what consequences these effects may have on the influence of the module on the cellular physiology of the organism [16,17,18,19,20,21,22]. Previous studies suggested that a third component that binds to and protects the phosphorylated form of the RR causes delays in the response of autogenous TCS systems that regulate their own expression [12,17,22]. However, to our knowledge, no studies were made about the effect that binding of a third component to the SK has on the potential dynamic behavior of the TCS module. In addition, the effect of both types of third component proteins was not studied in TCS that do not regulate their own expression.

In previous work we have used mathematical models to characterize the effect of diverse architectures on the signaling response of prototypical TCS. The analysis of such models enables understanding if particular physiological responses are more effectively achieved by one of several alternative designs of the network that executes the biological process of interest [23]. Such studies are difficult, if not impossible to do without the assistance of those mathematical models. In the case of the TCS, we showed that TCS with bifunctional SKs are more effective in buffering the TCS against crosstalk, while monofunctional SK are more effective in integrating different signals [24,25]. We have also identified necessary conditions for the existence of post-translational bistable responses in prototypical TCS [25]. If a system is capable of bistable responses, this means that its output variable can assume one of two possible values as a consequence of the same input. The specific value that the variable assumes depends on the value that the variable had before the stimulus. Post-translational bistability is only possible in TCS in which the affinity between the phosphorylated SK and unphosphorylated RR is similar to that between the unphosphorylated forms of the proteins. In addition, a large fraction of the dephosphorylation flux of the RR must be independent of any phosphatase activity of the SK [25].

Given these considerations, in this work our goal is to understand the physiological effect of a third protein, such as RetS or PmrD, on the function of canonical TCS in the absence of auto-regulation of gene expression. To achieve this, we built and analyzed mathematical models for the alternative designs of TCS with and without such a third component, and compared the dynamic behavior of the different systems. This analysis identifies specific physiological behaviors that are more effectively executed by each alternative design for the TCS.

Our study reveals that a RR-binding third component (TC_{RR}) decreases the region in parameter space where a bistable response is possible, while a SK-binding third component (TC_{SK}) increases the parametric region where a bistable response is possible when the SK is monofunctional and decreases it when the SK is bifunctional.

4.3. Methods

In order to understand the physiological effect of a third component (TC) on the function of a prototypical TCS, we built models of TCS with and without that TC and compared the dynamical behavior of those models. Figure 1 shows a schematic representation of the three models used in our analysis. These models are mathematically described by using a mass action system of ordinary differential equations (ODE) [26]. The resulting ODE systems for each of the three alternative models can be analyzed and compared numerically by running appropriate simulations on a computer.

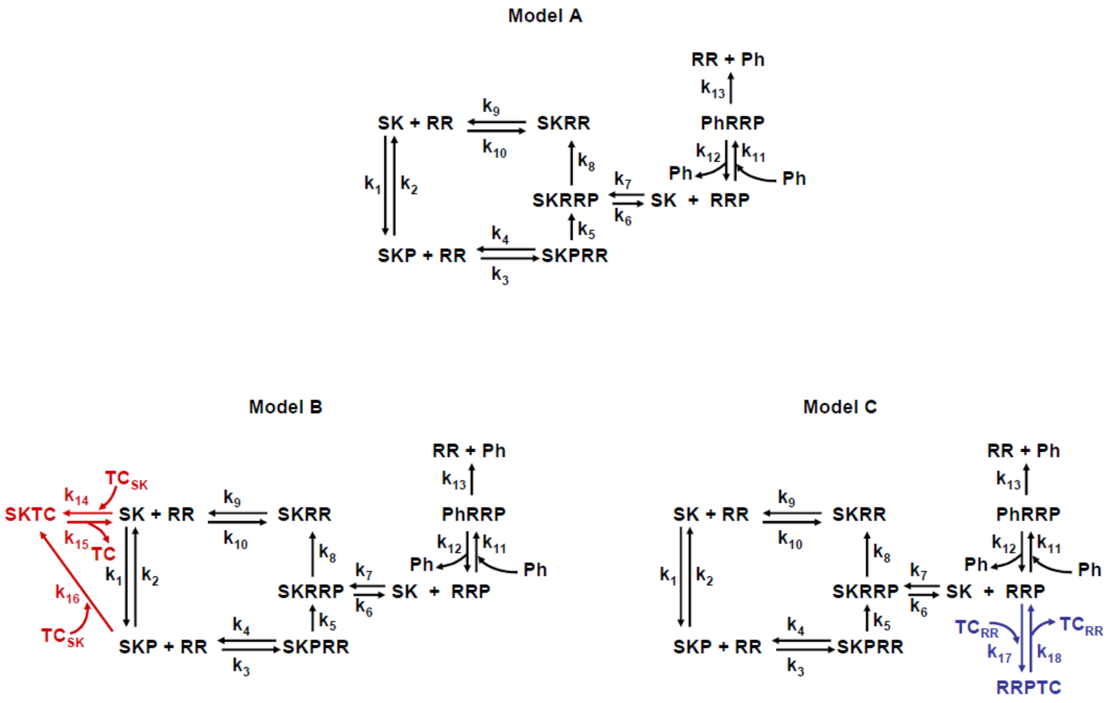


Figure 1. Analyzed Two Component Systems modules. Model A represents a prototypical TCS. Model B represents a TCS with a SK-binding third component (TC_{SK}). Model C represents a TCS with a RR-binding third component (TC_{RR}). SK: sensor kinase; RR: response regulator; SKP: phosphorylated SK; RRP: phosphorylated RR; Ph: alternative phosphatase that dephosphorylates RRP; SKRR: dead-end complex, resulting from the binding of SK and RR; SKPRR: protein complex formed by the binding of SKP and RR; SKRRP: protein complex formed by the binding of SK and RRP; PhRRP: protein complex formed by the binding of Ph and RRP; SKTC and RRPTC: protein complexes formed by the binding of the third component to SK and RRP, respectively; (k_1, …, k_{18}): kinetic constants of the individual reactions. For simplicity, ATP and the release of inorganic phosphate are omitted. To analyze TCS modules with monofunctional sensors, k_8 is set to 0. To analyze TCS modules with bifunctional sensors, k_8 is set to be different from 0.

4.3.1. Models and comparisons

The network model that we use to describe the prototypical TCS in our analysis is that defined in Igoshin et al. [25], which is based on earlier work [27]. In Model A, shown in Figure 1, the SK can autophosphorylate and/or autodephosphorylate in response to an external signal. Both phosphorylated and unphosphorylated forms of SK are allowed to bind RR with similar affinities, as reported in [28,29,30]. Binding of unphosphorylated SK and RR is reversible and forms a dead-end complex (SKRR). Phosphorylated SK (SKP) can transfer its phosphate to the RR. The phosphorylated RR (RRP) will modulate the biological levels and activity of relevant proteins.

This network for the prototypical TCS was modified to study the effect of a TC binding to either the SK or the RR. The changes in the network are also shown in Figure 1. Model B represents a TCS where a third component binds to the SK (TCSK), inactivating it. Model C represents a TCS where a third component binds to the phosphorylated RR (TCRR) and stabilizes this phosphorylated form. In prototypical TCS modules with bifunctional sensors, the unphosphorylated SK can destabilize the phosphorylated form of the RR and it increases the dephosphorylation rate of RRP (k8>0 in Figure 1). In prototypical TCS modules with monofunctional sensors, the unphosphorylated SK has no effect upon the dephosphorylation rate of RRP (k8=0 in Figure 1). The model includes a phosphatase that dephosphorylates RRP independently of the SK. This is done for generality. In the cases where no such phosphatase exists, this set of reactions can be replaced by a single reaction where the unstable RRP phosphate bond hydrolyzes over time. An appropriate choice of parameter values will make the results of the analysis similar to those described for the full model.

In this study we analyze the potential effect of a TC in the physiological behavior of TCS modules with bifunctional and monofunctional sensors independently. If the TC has no effect on the physiological behavior of the TCS, then the presence of TC in particular instances of TCS should be understood as an evolutionary accident. If the TC has an effect on the physiological behavior of the TCS, this could provide a rationale for the selection of a TCS design that includes a TC. To perform the analysis, we compare the dynamical behavior of Model A to that of Models B or C, independently. This comparison is done in two ways.

First, Models A and B (or C) are compared ensuring that the parameter values of all processes that are common are the same in the two models. This guarantees that whatever differences are found are only due to the addition of the TC. This comparison is equivalent to comparing an organism where the TCS interacts with a TC to another where the TC has been deleted from the genome. This situation could occur, for example during the creation of a new biological circuit by genetic manipulation in a biotechnological context. Thus, this type of comparison is relevant for understanding the differences in behavior of biological circuits created using synthetic biology techniques.

Second, we also perform a mathematically controlled comparison between Models A and B (or C). This is a well established method for evaluating the irreducible effect of a change in the design of a biological circuit on the dynamic behavior of the network [31]. In this comparison, in addition to ensuring that Models A and B (or C) have the same values for corresponding parameters of all processes that are common, we use the differences between the designs as degrees of freedom that evolution can use as a substrate to minimize differences between the dynamic behavior of the two systems. If the alternative designs can be made equivalent by using these degrees of freedom, then one may argue that they cannot be distinguished by natural selection. If, after making the systems as equivalent as possible, there are still irreducible differences in the physiological behavior between designs, then one may expect one of them to be preferably selected when its functionality provides better adaptive advantage. In the models under comparison, the difference is the deletion of a protein from the module between Model B (or C) and Model A. In this situation, the protein burden caused by Model A is lower than that caused by its alternative designs. Hence, we allow that the system changes the total concentrations of the remaining proteins (SK and/or RR). The details for this comparison are given in subsection 4.3.3. This comparison is thus relevant for understanding the differences in the dynamic behavior that are intrinsic to the differences in design between Models A and B (or C), and to those alone, in evolutionary terms.

4.3.2. Equations

In order to compare the physiological behavior of the three systems in Figure 1, we must create a mathematical representation for each of the networks. The positive and negative terms of each ODE correspond to individual reactions that give rise to the synthesis and degradation of the reactant, respectively. Each reaction is considered to be mass action.

Because the turnover times for protein synthesis and degradation are much higher than those for the phosphorylation-dephosphorylation reactions, we consider the total amount of each participating protein to be approximately constant. Thus,

SKt = SK+SKP+SKPRR+SKRRP+SKRR

RRt = RR+RRP+SKPRR+SKRRP+SKRR+PhRRP

Pht = Ph +PhRRP

$TC_{SK}t = TC_{SK} + SKTC$

$TC_{RR}t = TC_{RR} + RRPTC$

where SKt, RRt, Pht, $TC_{SK}t$ and $TC_{RR}t$ are constant and denote the total amount of SK, RR, Ph, TC_{SK} and TC_{RR} respectively.

Applying all simplifications, the differential equations for Model A become:

$$\frac{dSKP}{dt} = (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_1 - SKP\ k_2 - SKP\ (RRt\text{-}RRP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR\text{-}PhRRP)\ k_3 + SKPRR\ k_4$$

$$\frac{dRRP}{dt} = SKRRPk_6 - RRP(SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)k_7 - (Pht\text{-}PhRRP)RRP\ k_{11} + PhRRP\ k_{12}$$

$$\frac{dSKPRR}{dt} = SKP\ (RRt\text{-}RRP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR\text{-}PhRRP)\ k_3 - SKPRR\ (k_4 + k_5) \qquad (11)$$

$$\frac{dSKRRP}{dt} = SKPRR\ k_5 - SKRRP\ (k_6 + k_8) + RRP\ (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_7$$

$$\frac{dSKRR}{dt} = SKRRP\ k_8 - SKRR\ k_9 + (RRt\text{-}RRP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR\text{-}PhRRP)\ (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_{10}$$

$$\frac{dPhRRP}{dt} = (Pht\text{-}PhRRP)\ RRP\ k_{11} - PhRRP\ (k_{12} + k_{13})$$

Applying all simplifications, the differential equations for Model B become:

$$\frac{dSKP}{dt} = (SKt-SKP-SKPRR-SKRRP-SKRR-SKTC)\ k_1 - SKP\ k_2 - SKP\ (RRt-RRP-SKPRR-SKRRP-SKRR-PhRRP)\ k_3 + SKPRR\ k_4 - (TC_{SK\ total} -SKi)\ SKP\ k_{16}$$

$$\frac{dRRP}{dt} = SKRRP\ k_6 - RRP\ (SKt-SKP-SKPRR-SKRRP-SKRR-SKTC)\ k_7 - (Pht-PhRRP)\ RRP\ k_{11} + PhRRP\ k_{12}$$

$$\frac{dSKTC}{dt} = (TC_{SK\ total}-SKTC)\ (SKt-SKP-SKPRR-SKRRP-SKRR-SKTC)\ k_{14} - SKTCk_{15} + (TC_{SK\ total} - SKTC)\ SKP\ k_{16}$$

$$\frac{dSKPRR}{dt} = SKP\ (RRt-RRP-SKPRR-SKRRP-SKRR-PhRRP)\ k_3 - SKPRR\ (k_4 + k_5) \tag{12}$$

$$\frac{dSKRRP}{dt} = SKPRR\ k_5 - SKRRP\ (k_6 + k_8) + RRP\ (SKt-SKP-SKPRR-SKRRP-SKRR-SKTC)\ k_7$$

$$\frac{dSKRR}{dt} = SKRRP\ k_8 - SKRR\ k_9 + (RRt-RRP-SKPRR-SKRRP-SKRR-PhRRP)\ (SKt-SKP-SKPRR-SKRRP-SKRR-SKTC)\ k_{10}$$

$$\frac{dPhRRP}{dt} = (Pht-PhRRP)\ RRP\ k_{11} - PhRRP\ (k_{12}+k_{13})$$

Applying all simplifications, the differential equations for Model C become:

$$\frac{dSKP}{dt} = (SKt-SKP-SKPRR-SKRRP-SKRR)\ k_1 - SKP\ k_2 - SKP\ (RRt-RRP-SKPRR-SKRRP-SKRR-PhRRP-RRPTC)\ k_3 + SKPRR\ k_4$$

$$\frac{dRRP}{dt} = SKRRP\ k_6 - RRP\ (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_7 - (Pht\text{-}PhRRP)\ RRP\ k_{11} + PhRRP\ k_{12} - RRP\ (TC_{RR\ total}\text{-}RRPTC) k_{17} + RRPTC\ k_{18}$$

$$\frac{dSKPRR}{dt} = SKP\ (RRt\text{-}RRP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR\text{-}PhRRP\text{-}RRPTC)\ k_3 - SKPRR\ (k_4 + k_5)$$

$$\frac{dSKRRP}{dt} = SKPRR\ k_5 - SKRRP\ (k_6 + k_8) + RRP\ (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_7 \qquad (13)$$

$$\frac{dSKRR}{dt} = SKRRP\ k_8 - SKRR\ k_9 + (RRt\text{-}RRP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR\text{-}PhRRP\text{-}RRPTC)\ (SKt\text{-}SKP\text{-}SKPRR\text{-}SKRRP\text{-}SKRR)\ k_{10}$$

$$\frac{dRRPTC}{dt} = RRP\ (\ TC_{RR\ total}\ \text{-}RRPTC)\ k_{17} - RRPTC\ k_{18}$$

$$\frac{dPhRRP}{dt} = (Pht\text{-}PhRRP)\ RRP\ k_{11} - PhRRP\ (k_{12}+k_{13})$$

The parameters for the models are given in Table 3. All these parameters have an experimental basis, clearly presented in Igoshin *et al.* [25].

4.3.3. Mathematically controlled comparisons

We aim at comparing the physiological behavior of the three models in order to understand if the presence of a TC in a TCS module causes intrinsic differences to the potential physiological responses that the modules can have. To make sure that the differences observed in the behavior of the systems that are being compared are due to the presence of the TC, the comparisons must be made in a controlled way. For this we use the method of mathematically controlled comparisons [31]. This method requires that all components and processes that are common to the alternative models that are to be compared are made numerically equal, making the models internally equivalent. In contrast, the components and processes that are different between the alternative models are degrees of freedom that nature could potentially use to compensate the changes in the physiological responses caused by the differences between systems. In this case, the systems with a TC invest additional resources to synthesize a new protein that binds either the SK or the RR and modulates their phosphorylation state. All new processes of Models B and C with respect to Model A are due to the presence of this TC. In order to control the comparison between TCS with TC and the prototypical TCS, the prototypical system (Model A) should also be allowed to invest additional resources in adjusting the total amount of the SK or the RR. These adjustments will allow the prototypical system to have a physiological response that is as similar as possible to that of the model with a SK-binding or a RR-binding TC (Models B and C, respectively). This control condition ensures maximal external equivalency between the models. Once the maximum equivalency is achieved between the compared models, the remaining behavioral differences can be related to the presence of the TC.

To determine the changes in the total amount of SK or RR that make the physiological responses between Model A and Models B or C as similar as possible, we have used a minimum square differences method. We have calculated the steady state responses of the system in Models B and C to changes in the input phosphorylation or dephosphorylation rate of the modules, by calculating the steady state concentration of RRP in Models B and C, at input signal strengths between 10^{-6} and 10. These curves were then used individually to fit Model A and calculate the concentration of SK and/or RR that would minimize the differences in the steady state RRP concentration between Model A and Models B or C, independently. All calculations were done using Mathematica. The best fits are achieved by allowing the total amount of SK to change in Model A. The values for the total amount of SK in Model A that minimize the differences between the responses of this model and Model B or Model C are shown in Table 1.

Table 1. Values of SK_{total} in Model A used in the mathematically controlled comparisons.

[SK_{total}] in Model A (µM)

	Monofunctional		**Bifunctional**	
	k_1	**k_2**	**k_1**	**k_2**
Model A|B	0.13	0.13	0.14	0.14
Model A|C	0.50	0.50	0.90	0.90

These values are chosen to make the signal-response curves of the prototypical TCS (Model A) and the system with a third component (Models B or C) as similar as possible, for responses to an environmental stimulus that modulates either k_1 (SK autophosphrylation kinetic constant) or k_2 (SKP autodephosphrylation kinetic constant). A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

4.3.4. Calculations

All simulations were performed in Mathematica [45] and COPASI [46]. Analyses of regions of bistability were done in Mathematica, using in-house scripts.

4.4. Results

4.4.1. Effect of a third component on TCS signal amplification and bistability

Signal amplification is an important physiological property of TCS. TCS with appropriate signal amplification can provide evolutionary advantages to organisms harboring them. Thus, understanding how signal amplification is affected by adding a TC to a TCS would help in predicting under which conditions to expect such a design to be selected. Figure 2 shows that all models can achieve the same signal amplification, whether the environmental signal modulates the autophosphorylation (k_1) or the autodephosphorylation (k_2) of the SK. This can be seen because the difference between the amount of RRP (phosphorylated RR) when k_1 is low (k_2 is high) and when k_1 is high (k_2 is low) can be similar for all models. Nevertheless, Model B responds at higher signal intensities and Model C responds at lower signal intensities than Model A, when the stimulus modulates the SK autophosphorylation reaction rate (compare the curves for k_1 response of Model A to those of Models B and C in Figure 2). When the signal modulates the SK autodephosphorylation reaction rate, Model B responds at lower signal intensity and Model C at higher signal intensity than Model A (compare the curves for k_2 response of Model A to those of Models B and C in Figure 2). However, mostly, the differences in signal intensity at which the systems are turned ON or OFF are small.

In addition, the prototypical TCS shown in Model A can show bistable behavior [25], making it possible that a signal can lead to one of two alternative responses, depending on the history of the system. Such a response may have some evolutionary advantages, for example in situations like sporulation where an irreversible developmental decision is made by cells. Bistable regions in the curves of Figure 2 have three values of RRP for a single value of signal intensity. The two extreme values are the alternative stable steady states, while the middle value is a biologically irrelevant unstable steady state that is mathematically required to exist if two stable steady states are present. In the figure one can see that the signaling ranges where bistability exists are different if the environmental signal modulates the autophosphorylation (k_1) or the autodephosphorylation (k_2) of the SK.

Necessary, although not sufficient, conditions for the existence of such bistable behavior in the prototypical TCS are i) the formation of a dead-end complex between the dephosphorylated forms of SK and RR and ii) that a sufficiently high fraction of the flux for the dephosphorylation of RRP is independent of SK. To understand how the presence of a TC affects the possibility of a bistable response in the prototypical TCS, we analyzed Models B and C in search of the existence of multiple steady states, followed by a comparison of the physiological behavior between Models A and B, and between Models A and C.

Given that signals can in principle modulate either the autophosphorylation (k_1) or the autodephosphorylation (k_2) rate of SK, we performed parallel computational experiments independently modulating their intensity. These experiments were done independently for models with monofunctional and bifunctional SK (Figure 2).

Our results show that, in an uncontrolled comparison, the range of bistability for the bifunctional prototypical TCS is larger than if a TC binds any of the proteins of the module (compare panel B to panels D and F of Figure 2). Bistability for Model B in panel D is only observed for k_1 signaling, while no bistability is observed for Model C in panel F. On the other hand, the range of bistability for the monofunctional prototypical TCS is larger than if a TC binds the RR of the module (compare panel A to panel E of Figure 2), but smaller than if the TC binds the SK (compare panel A to panels C of Figure 2). Differences among the three systems are more pronounced when the signal induces dephosphorylation of the SK (k_2), rather than inducing SK autophosphorylation (k_1).

An additional definition is needed before presenting and discussing additional results. Hereafter the system is said to be in an ON state if most of its RR is in the phosphorylated RRP form. If most of the RR is in its dephosphorylated form, the system is said to be in its OFF state. With this in mind, and as one might expect, systems with a TC_{SK} are in an ON state for a smaller signaling range (panels C and D) and systems with a TC_{RR} are in an ON state for a larger signaling range (panels E and F), in comparison with the uncontrolled Model A (panels A and B).

When the comparisons are controlled we see that the response of Model A can become similar to that of Model B or C by adjusting the total amount of available SK. If the response of Model B is to be mimicked, the total amount of SK in Model A is decreased (Figure 2, panels C and D, see methods for the exact values of the total amount of SK), while mimicking the response of Model C leads to an increase in the concentration of SK (Figure 2, panels E and F, see methods for the exact values of the total amount of SK).

The k_2-response curves in Figure 2 panels B and C show that the switch from ON to OFF (from high to low levels of RRP) in these models could be irreversible or very difficult to reverse. In other words, modulation of the autodephosphorylation rate of SK by an external signal could generate nearly irreversible biological switches.

Our simulations also show that the necessary conditions for bistability in prototypical TCS remain necessary in the TCS with a TC. If either no independent phosphatase is present in the system (Ph=0) or no dead-end complex is formed (k_{10}=0) all TCS modules analyzed here are monostable (see section “Effect of changes in SK-independent RRP dephosphorylation and SKRR affinity on bistability” below).

In summary, a TC_{RR} causes a reduction in the TCS parameter space of bistability and an increase in the signaling range in which the system is in the ON state (responds at lower k_1-signal intensity and at higher k_2-signal intensity), whether the SK is monofunctional or bifunctional. This can be more effectively compensated by prototypical TCS through a change (an increase) in the concentration of the SK. In contrast, TC_{SK} increases the signaling range in which the TCS can show a bistable response if and only if the SK is monofunctional and the environment modulates k_2 (SK dephosphorylation rate). The behavior of TCS with a TC_{SK} can be mimicked by prototypical TCS through a change (a decrease) in the concentration of the SK.

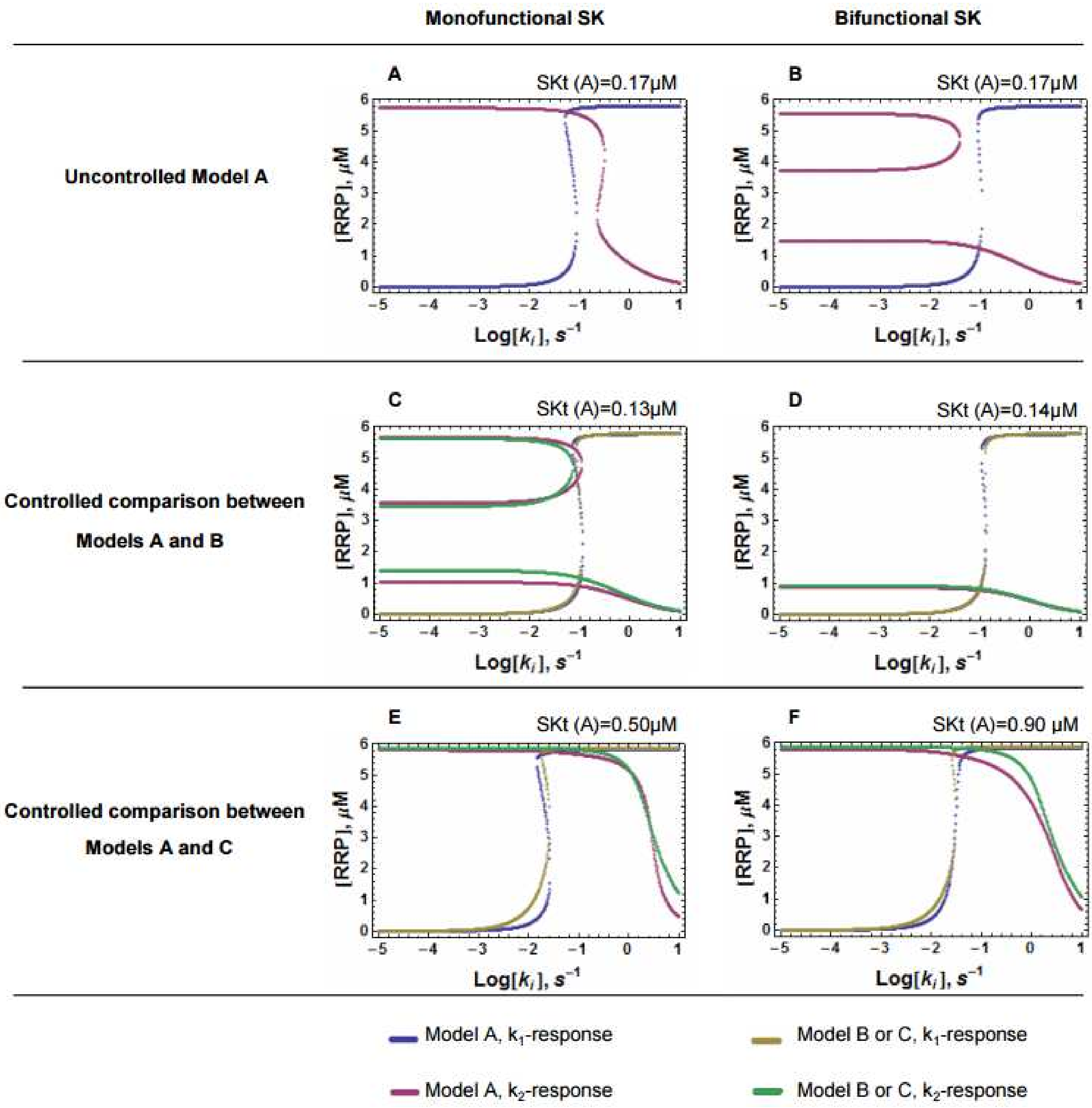


Figure 2. Steady state signal-response curves for the various TCS modules. Each plot shows the steady state levels of the phosphorylated RR in the y axis at different values of the signal k_1 (SK autophosphorylation rate constant) or k_2 (SKP dephosphorylation rate constant) in the x axis. When the signal modulates SK dephosphorylation (changes in k_2), the system behaves symmetrically to when SK phosphorylation (changes in k_1) is modulated. In the first case, increases in signal intensity cause the fraction of RRP to decrease, while in the latter, increases in signal intensity cause the fraction of RRP to increase. A, C, E: Response curves of TCS modules with monofunctional sensor. B, D, F: Response curves of TCS modules with bifunctional sensor. A, B, Response curves of Model A. C, D: Mathematically controlled comparison between the response curves of Model B and those of Model A. E, F: Mathematically controlled comparison between the response curves of Model C and those of Model A. Mathematical controls are implemented to make sure that the differences in response between the alternative modules are due to the presence of third component and not to other spurious differences.

4.4.2. Effect of a third component on TCS response time

In addition to signal amplification, the response time to signals is an important physiological property of TCS. In evolutionary terms, a change in response time may have important consequences to the fitness of the system. Therefore, we analyzed the effect of a TC on the response times of the TCS. To do this we performed four independent sets of experiments for each of the models, and independently considering systems with a monofunctional SK and with a bifunctional SK. In experiments 1 and 2 we instantaneously change the signal k_1 and measure how long the system takes to come within 90% of its new steady state. This measures the response time of the system if the physiological signal modulates SK phosphorylation. In experiments 3 and 4, we instantaneously change the signal k_2 and measure how long the system takes to come within 90% of its new steady state. This measures the response time of the system if the physiological signal modulates SK dephosphorylation. The details about how the experiments were run are as follows:

1 - We set each system to its OFF state, with k_1=10^{-5} s^{-1}. Then, we increased the value of k_1 to a value $k_{1\ higher}$ and measured how long the system took to get to within 90% of its new steady state value. $k_{1\ higher}$ was systematically changed between 10^{-5} and 10 s^{-1}.

2 - We set each system to its ON state, with k_1=10 s^{-1}. Then, we decreased the value of k_1 to a value $k_{1\ lower}$ and measured how long the system took to get to within 90% of its new steady state value. $k_{1\ lower}$ was systematically changed between 10^{-5} and 10 s^{-1}.

3 - We set each system to its OFF state, with k_2=10 s^{-1}. Then, we decreased the value of k_2 to a value $k_{2\ lower}$ and measured how long the system took to get to within

90% of its new steady state value. $k_{2\ lower}$ was systematically changed between 10^{-5} and 10 s^{-1}.

4 - We set each system to its ON state, with $k_2=10^{-5}$ s^{-1}. Then, we increased the value of k_2 to a value $k_{2\ higher}$ and measured how long the system took to get to within 90% of its new steady state value. $k_{2\ higher}$ was systematically changed between 10^{-5} and 10 s^{-1}.

Results are shown in Figure 3. We see that the response times increase by more than two orders of magnitude when the new parameter value $k_{\cdot lower}$ or $k_{\cdot higher}$ approaches the threshold value for exiting the bistability region of a system. The peaks of slower response in the curves in Figure 3 are in the region of signal intensity that lies immediately beyond the border of the bistability ranges shown in Figure 2. Given that the peaks of slower response are located at the exit of the bistable region, there is no peak in the signal-response time curve when the response is monostable or when there is an irreversible turning OFF of the system. Model B and Model A|B (A controlled for B) don't have a peak in their OFF to ON k_2-response times (Panel C of Figure 3) because these models irreversibly turn OFF after an increase in k_2 (as depicted in Figure 2 panel C). Model C also has no peak in the response time (Panels C and D of Figure 3) because this model has a monostable response to changes in k_2 (see Figure 2 panel E). In panels G and H of Figure 3, neither of the three systems shows a peak in their signal-response time curve because of the lack of bistability in their signal-response steady state curve (see Figure 2 panels D and F). When Model A is compared to Model B in an uncontrolled manner, the time response peaks of Model A appear at signal intensities that are always lower than those where the peak appears in the response of Model B. When Model A is compared to Model C in an uncontrolled manner, the time response peaks of Model A appear at signal intensities that are always higher than those where the peak appears in the response of Model C (see Figure S1).

In order to have a proxy of the integral temporal responsiveness of each system, we calculated the area under each of the signal- response time curves shown in Figure 3. This area is the sum of all the transient response times for each signaling response. The values of these areas are given in Table 2 and show that overall response times are similar between Models A and B. In contrast, Model A has a faster response than Model C. When the comparison is not controlled, differences between integrated response times of the three models are small, when the signal modulates autophosphorylation of SK. However, if SK dephosphorylation is modulated, Model B has the fastest integrated response, followed by Model A. Model C is, again, the slowest responder (Table S1).

In summary, Model B is a faster overall responder than the prototypical TCS when the system is turned ON by modulating the phosphorylation rate of the SK, and it is a slower responder in any other case. In contrast, Model C is always slower to turn ON or turn OFF than the prototypical TCS, under controlled comparison conditions.

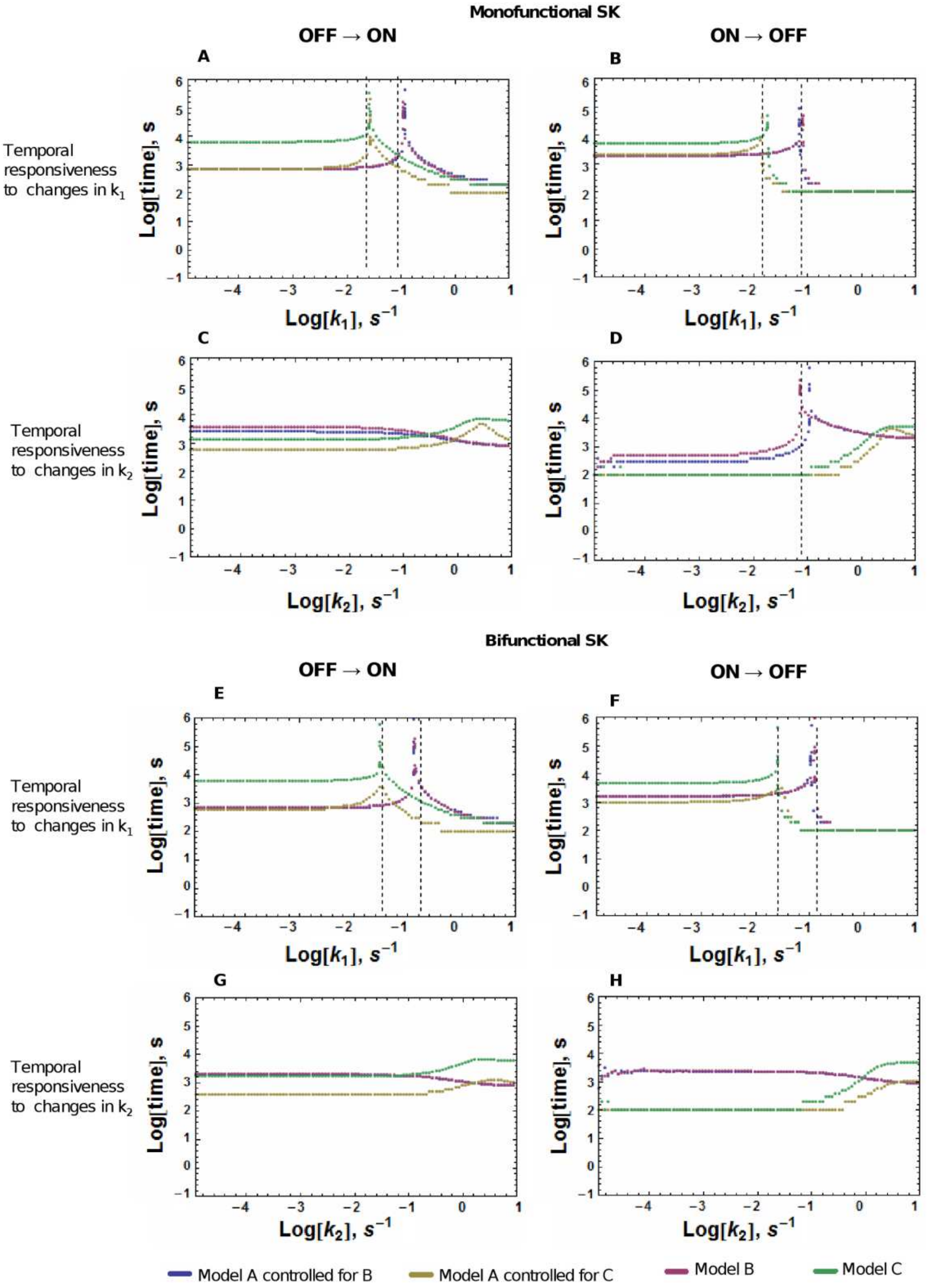


Figure 3. Temporal responsiveness curves of Models A, B, and C. The systems are at an initial steady state and, at time zero, the signal, represented in the x axis, changes instantaneously and the time it takes for the system to get to within 90% of the new steady state is measured and plotted in the y axis. A-D: Response times of TCS with monofunctional SK. E-H: Response times of TCS with bifunctional SK. The OFF to ON plots start with the systems at an OFF steady state (low levels of RRP) corresponding to a low value of k_1 (A, C, E, G) or a high value of k_2 (B, D, F, H). The signal is then changed to increase the steady state level of RRP. The ON to OFF plots start with the systems at an ON steady state (high levels of RRP) corresponding to a high value of k_1 or a low value of k_2. The signal is then changed to decrease the steady state level of RRP. Peaks that indicate slower response times are located immediately outside the range of bistability. The lack of a peak in a curve can be due to monostability or irreversibility. The dashed lines indicate the signal value at which Models B and C exit its bistable range. Absence of a dashed line indicates irreversible turning ON or OFF of the system (Model B in panel C) or absence of bistability (see the signal-response curves of Figure 2).

Table 2. Controlled comparison of the overall response times between Models A and B, and between Models A and C[a].

	Modulation of SK autophosphorylation (k_1)		Modulation of SKP dephosphorylation (k_2)	
	OFF → ON	**ON → OFF**	**OFF → ON**	**ON → OFF**
Monofunctional				
Model A|B	3 646.18	1 244.27	9 129.47	24 524.50
Model B	3 406.48	1 337.95	9 467.02	24 801.00
Bifunctional				
Model A|B	3 917.63	1 501.14	8 656.10	10 565.20
Model B	3 672.27	1 739.08	8 695.38	10 672.20
Monofunctional				
Model A|C	1 351.02	1 003.90	21 984.30	26 656.70
Model C	3 125.05	1 091.73	57 574.80	43 048.20
Bifunctional				
Model A|C	1 152.38	1 029.89	10 647.20	8 972.97
Model C	3 358.06	1 195.35	57 212.80	40 114.40

[a] The reported values represent the area below each curve in Figure 3, that is, the sum of the transient times for each response. A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

4.4.3. Stochastic effects of a third component

Fluctuations in the amount of proteins that participate in biological reactions can lead to stochastic effects in the system's behavior, when the total number of proteins participating in reactions is small. We performed stochastic simulations to understand the role of stochasticity on the effect of the TC on the physiological response of the TCS networks. These simulations take into account that the number of TCS proteins present in the cell are typically in the 10-1000 molecules range.

The simulation experiments performed were similar to those described in experiments 1-4 of the previous section, although with a smaller number of data points. Figures 4 and 5 show the results of these simulations.

The OFF → ON plots start with the system at the OFF steady-state (low concentration of active RR) corresponding to a low value of k_1 (k_1=10^{-5} s^{-1}) or a high value of k_2 (k_2=10 s^{-1}), and depict the temporal trajectory of the RRP concentration after an instantaneous increase in k_1 or decrease in k_2, for three different values of k_1 and k_2.

The ON → OFF plots start with the system at the ON steady-state (high concentration of active response regulator) corresponding to a high value of the signal k_1 (k_1=10 s^{-1}) or a low value of k_2 (k_2= $5\cdot10^{-6}$ s^{-1}), and depict the temporal trajectory of the RRP concentration after an instantaneous decrease in k_1 or increase in k_2, for three different values of k_1 and k_2.

The simulation results for three different signal intensities are plotted in Figures 4 and 5. Three independent simulations are shown for each signal intensity. The values of k_1 and k_2 in each trajectory are chosen to be below, next to and above the threshold value at which the system switches from OFF to ON, or from ON to OFF (in the cases in which this threshold exists). Because each system has a different threshold value, the parameter scan is different for each plot.

The results from the analysis of the continuous model are consistent with the stochastic simulations: as discussed in the previous section (Figure 3), in systems with a signal range of bistability the response times increase when the signal intensity is near the threshold value at which the system exits the bistability region. One can see in Figures 4 and 5 that, in many cases, the curves that correspond to a signal that is just outside of the bistability range do not reach steady state during the simulation time. These curves correspond with the peaks in Figure 3.

Furthermore, our simulations predict that the systemic response becomes noisier as the signal intensity approaches the threshold value for bistability. Just above and just below this value there is an increase in the stochastic fluctuations of the system. This can be seen because the triplicate curves corresponding to these values in Figures 4 and 5 are much more different among themselves than the triplicate curves for the signals away from this threshold.

The response in the systems A, B and C is noisier when k_1 is modulated than when k_2 is modulated. The OFF to ON trajectories of Model B after an instantaneous decrease in k_2 confirm that the turn OFF of this system due to an increase in k_2 is irreversible and the system can’t return to the ON state (see Figure 2 panel C). The system C does not have a bistability region in its k_2-response curve (see Figure 2 panels E and F). Therefore, we don’t find a range of k_2 values for which the systemic response becomes slower and noisier.

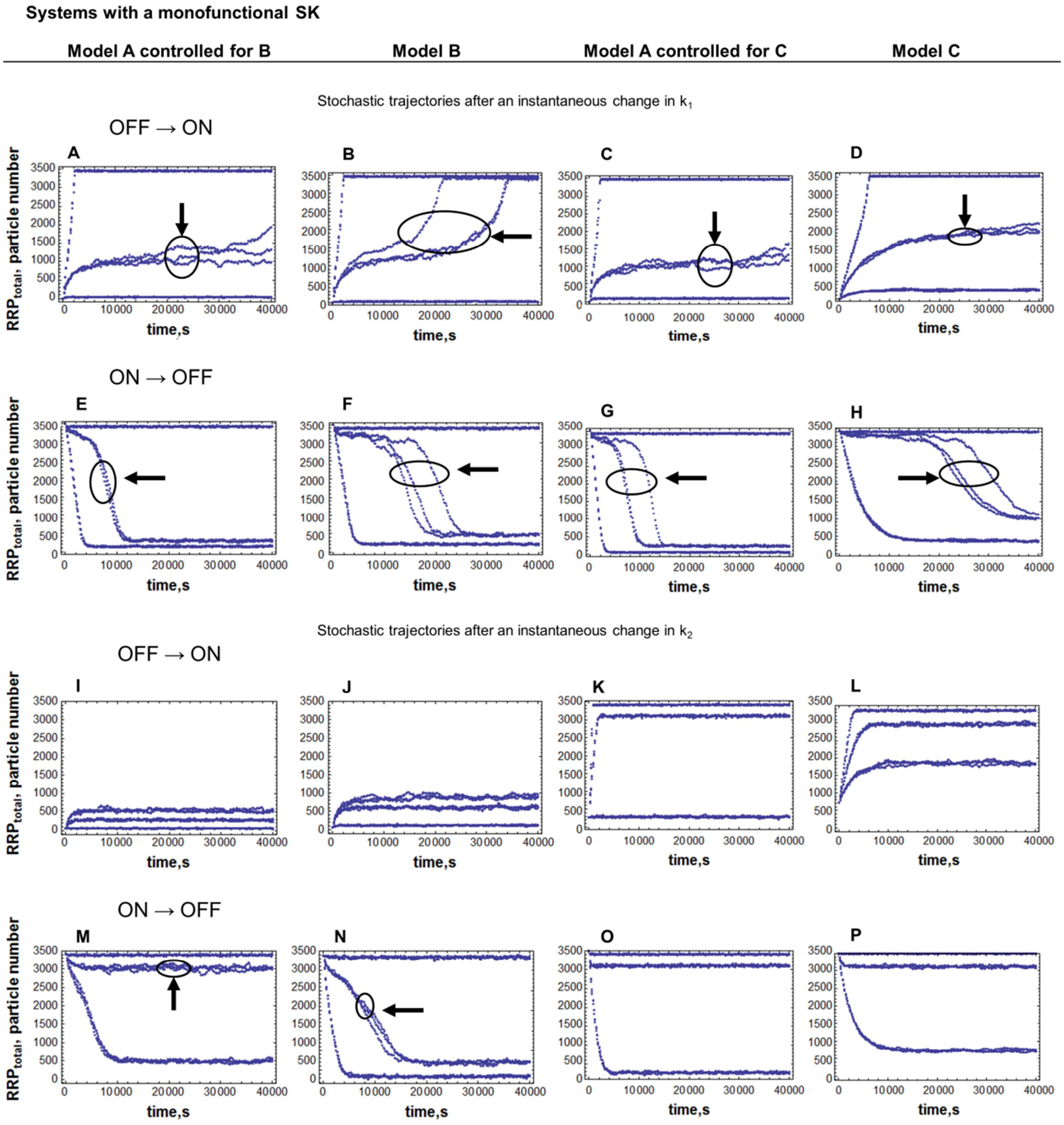


Figure 4. Stochastic time trajectories after an instantaneous change in the signal, for the three systems modeled with a monofunctional SK. A mathematically controlled comparison between Models A and B, and between Models A and C was performed as described in methods. The results for three individual runs for each value of k_1 or k_2 are plotted in each panel. Panels in the first column correspond to Model A controlled to be as similar as possible to Model B. Panels in the second column correspond to Model B. Panels in the third column correspond to Model A controlled to be as similar as possible to Model C. Panels in the fourth column correspond to Model C. The circles indicate lines that are replicates of the same simulation. Simulations marked with an arrow correspond to a signal intensity close to the bistability threshold and show slower and noisier responses. The OFF to ON plots start with the systems at an OFF steady state (low levels of RRP) corresponding to a low value of k_1 or a high value of k_2. At time zero, there is an instantaneous increase in k_1 or decrease in k_2. The ON to OFF plots start with the systems at an ON steady state (high levels of RRP) corresponding to a high value of k_1 or a low value of k_2. At time zero, there is an instantaneous decrease in k_1 or increase in k_2. The values for k_1 or k_2 are chosen to be below, next to and above the threshold value at which the system switches from OFF to ON, or from ON to OFF. See text for further details.

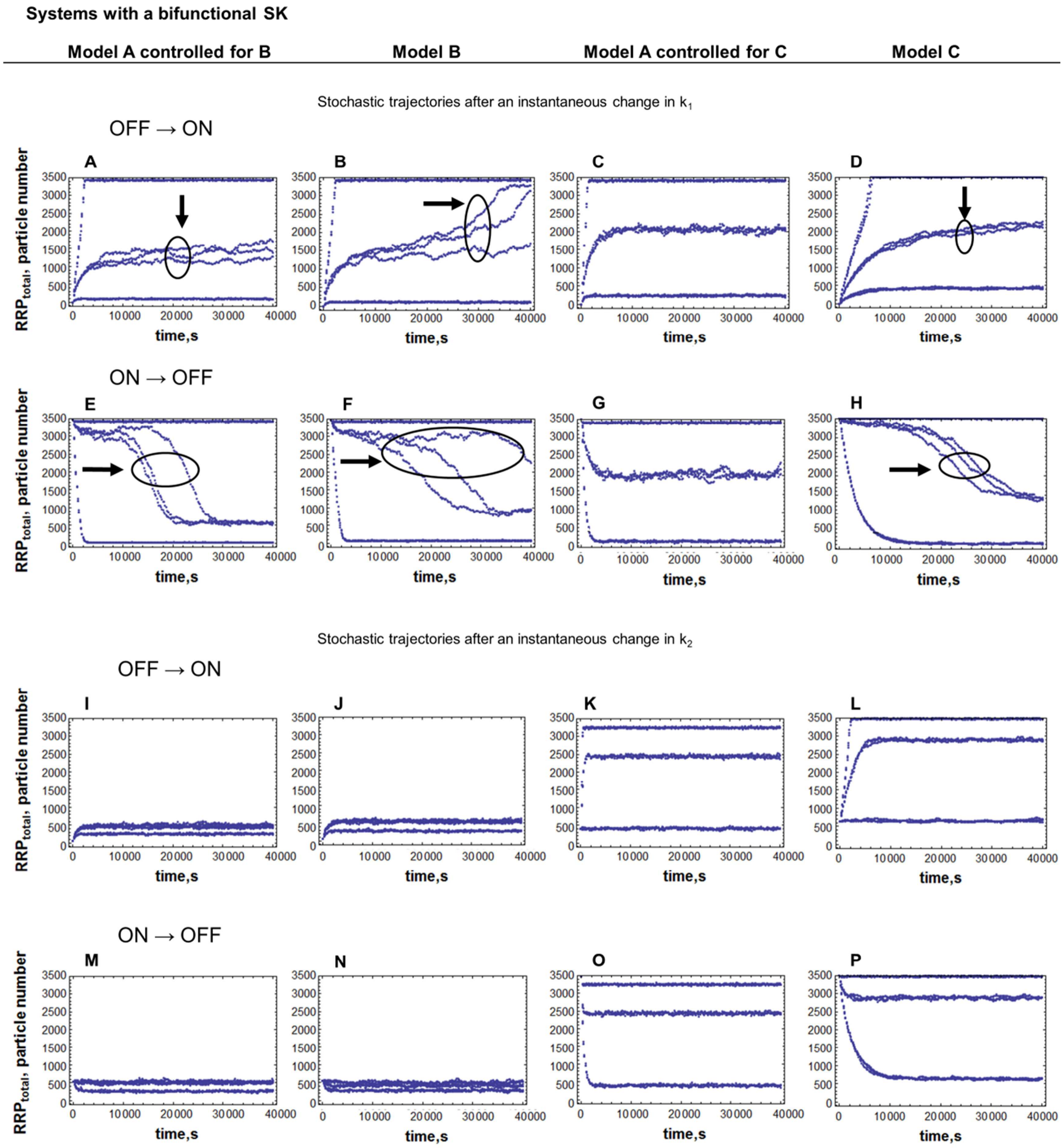


Figure 5. Stochastic time trajectories after an instantaneous change in the signal, for the three systems modeled with a bifunctional SK. A mathematically controlled comparison between Models A and B, and between Models A and C was performed as described in methods. The results for three individual runs for each value of k_1 or k_2 are plotted in each panel. Panels in the first column correspond to Model A controlled to be as similar as possible to Model B. Panels in the second column correspond to Model B. Panels in the third column correspond to Model A controlled to be as similar as possible to Model C. Panels in the fourth column correspond to Model C. The circles indicate lines that are replicates of the same simulation. Simulations marked with an arrow correspond to a signal intensity close to the bistability threshold and show slower and noisier responses. The OFF to ON plots start with the systems at an OFF steady state (low levels of RRP) corresponding to a low value of k_1 or a high value of k_2. At time zero, there is an instantaneous increase in k_1 or decrease in k_2. The ON to OFF plots start with the systems at an ON steady state (high levels of RRP) corresponding to a high value of k_1 or a low value of k_2. At time zero, there is an instantaneous decrease in k_1 or increase in k_2. The values for k_1 or k_2 are chosen to be below, next to and above the threshold value at which the system switches from OFF to ON, or from ON to OFF. See text for further details.

4.4.4. Robustness of the analysis

The analysis thus far was done using the specific set of parameter values reported in Table 3. In order to study the generality of the results we performed sensitivity analyses of the bistability to changes in the different parameter values and concentrations of the systems. The results of the controlled and uncontrolled comparison between Model A and Model B or C with respect to the effect of changing parameter values on a possible bistable response of the TCS are summarized in Table 4. The detailed results are shown in Figure S2, where we show a set of two-dimensional sections of the multidimensional parameter space in which bistability is observed.

Overall, a system with a TC_{SK} appears to have a wider parameter range of bistability if the SK is monofunctional, and a lower parameter range of bistability if the SK is bifunctional, while a system with a TC_{RR} appears to have a lower parameter range of bistability, for systems with either a monofunctional or a bifunctional SK, when either system is compared to a prototypical TCS. However, if the comparison between Model A and Model B or C is controlled, then we see that the robustness of the parameter range of bistability is larger in the prototypical TCS (Model A) with only one exception: in systems with a bifunctional SK, Model C has a more robust parameter range of bistability.

Table 3. Basal values for the parameters and concentrations of the models in Figure 1.

Kinetic constant	**Value**
k_1	[e] 0.1 s^{-1}
k_2	0.0005 s^{-1}
k_3	0.5 $\mu M^{-1}s^{-1}$
k_4	0.5 s^{-1}
k_5	1.5 s^{-1}
k_6	0.5 s^{-1}
k_7	0.05 $\mu M^{-1}s^{-1}$
k_8	0 s^{-1} (monofunctional SK)/ [f] 0.05 s^{-1} (bifunctional SK)
k_9	[f]0.5 s^{-1}
k_{10}	[g]0.5 $\mu M^{-1}s^{-1}$
k_{11}	0.5 $\mu M^{-1}s^{-1}$
k_{12}	0.5 s^{-1}
k_{13}	0.025 s^{-1}
[a]k_{14}	0.5 $\mu M^{-1}s^{-1}$
k_{15}	0.5 s^{-1}
[b]k_{16}	0.005 $\mu M^{-1}s^{-1}$
[a]k_{17}	0.5 $\mu M^{-1}s^{-1}$
k_{18}	0.5 s^{-1}
Proteins	**Total Concentrations**
RR	6μM
SK	0.17μM
Ph	0.17μM
[c]TC_{SK}	1.17 μM
[d]TC_{RR}	10μM

[a] These values were chosen in such a way that the affinity of the TCS proteins with the third component would be similar to the affinity between the SK and the RR.

[b] The value for this parameter was chosen to be one order of magnitude larger than that representing SK autodephosphorylation, because the TC_{SK} enhances SK autodephosphorylation.

[c] $TC_{SK\ total}$ is the total amount of the third component in Model B. This third component protein binds the SK of the TCS module. The amount for this protein was chosen taking into account that basal mRNA levels for RetS in GEO micro profiles of *Pseudomonas aeruginosa* are between 2 and 10 times higher than those of GacS. GacS is an SK and RetS is its cognate TC_{SK} [47].

[d] $TC_{RR\ total}$ is the total amount of the third component in Model C. This third component protein binds the phosphorylated RR of the TCS module. The amount for this protein was chosen to be in the same order of magnitude as that of the RR, as is done in reference [43].

[e] This is the average value for the autophosphorylation catalytic constant between *Salmonella typhimurium* and *Escherichia coli* [16].

[f] It should be noted that, for Model C, this value for the phosphatase rate constant could be as high as 0.14 in *Escherichia coli* [16].

[g] Although some measurements have suggested that the affinity between non-phosphorylated forms of the SK and RR is much lower than the affinity between phosphorylated forms of the proteins [48], more recent measurements suggest the opposite [10].

Table 4. Percentage of parameter space where bistable responses are possible[a].

	Model A	Model A|B	Model B	Model A|C	Model C
Monofunctional					
Input signal: change in k_1	8	7.56	6.04	8.98	6.74
Input signal: change in k_2	11.36	21.87	17.52	9.11	4.01
Bifunctional					
Input signal: change in k_1	4.85	4.89	3.81	2.24	4.98
Input signal: change in k_2	11.44	7.77	4.11	1.84	4.31

[a] Some bidimensional sections of the multidimensional parameter space of bistability are shown in Figure S2. The results show that in TCS with a bifunctional SK, both a TC_{SK} and a TC_{RR} cause a decrease in the size of the parametric region of bistability, with one exception: Model C has a larger parametric region of bistability when the signaling target is SK autophosphorylation (k_1). However, in systems with a monofunctional SK, a TC_{SK} causes an increase and a TC_{RR} causes a decrease in the size of the parametric region of bistability if the environment modulates the SK dephosphorylation (k_2). A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

4.4.5. Effect of changes in SK-independent RRP dephosphorylation and SKRR affinity on bistability

SK-independent RRP dephosphorylation and SKRR complex formation are needed for bistable responses to exist in Models A, B, and C. In order to investigate how quantitatively changing these features affects bistability we performed the following computational experiments (Table 5). We independently and simultaneously changed the values for k_8 (the reaction that regulates dephosphorylation by the SK) and k_9 (changing the rate of dissociation between SK and RR) between 10^{-6} and 10. Then, we calculated the steady state(s) for each system at different values of the signal represented by the parameters k_1 or k_2. k_1 and k_2 were independently and systematically scanned between 10^{-6} and 10 in logarithmic space at intervals of 0.01 units. The results are shown in Table 6 and Figure S3. Table 4 shows that, overall, bistability is possible in Model C in a smaller interval of parameter values than that for Models A and B. However, the picture changes when we analyze only the parameters that directly influence the necessary conditions for bistability (k_8, k_9, k_{10}). For these parameters, Model C is the system where overall bistability is possible in a wider range of parameter values, followed by Model B. Model A is the one where bistability is limited to a smaller region of parameter values. Nevertheless, when Model A is controlled to have signal-response curves that are as similar as possible to those of either Model B or Model C, Model A becomes the system where bistable responses can occur in a larger fraction of the space for k_8, k_9, and k_{10}. For values of k_8 below a threshold that depends on the system and is lower in Model B than in Model A, bistability is present in both models. Within the range of k_8 values that permit bistability, an increase in k_8 causes an increase in the k_2 range of bistability (up to approximately six orders of magnitude for k_2 at the threshold value for k_8). This is so, despite the enlargement of the fraction of RRP dephosphorylated by SK, because the increase in k_8 causes an increase in the concentration of the SKRR dead-end complex (see Figure S4). As k_8 decreases, the range of signal k_2 in which the models show bistability decreases steadily for a few orders of magnitude. Then, a lower boundary is reached and bistability is observed for one or less than one order of magnitude of k_2 signal, independently of the value for k_8.

Given that the formation of a dead-end complex between SK and RR is a necessary condition for bistability, we also want to understand the isolated effect of different fractions of RR and SK being sequestered into this complex on bistability. To understand the effect of changing the amount of SKRR dead-end complex on the signaling range in which the systems can be bistable we performed the following numerical experiment. First, we took each model from Figure 1. Then, we systematically scanned the values of the parameters k_9 and k_{10}, independently and simultaneously, between 10^{-6} and 10 in logarithmic space at intervals of 0.01 units. These parameters regulate the amount of SKRR that is formed. Finally, for each pair of values for k_9 and k_{10}, we independently calculated the steady state(s) for each system at different values of the signal represented by k_1 or k_2. Each of these parameters was independently and systematically scanned between 10^{-6} and 10 in logarithmic space at intervals of 0.01 units. The results are shown in Table 6 and Figure S3.

Bistability can be found only for intermediate steady state concentrations of SKRR. If too little or too much SKRR is formed, then no bistable response is possible. Overall, for bifunctional TCS, Model C has the largest range of SKRR steady state concentrations for which bistability is possible, followed by Model B. In its uncontrolled form Model A has the smallest interval of SKRR steady state concentrations where bistability is permitted. This interval of concentrations decreases further when Model A is controlled to be comparable to Model B. However, when Model A is controlled to be comparable to Model C, the range of SKRR steady state concentrations that enable bistability becomes the largest of the three systems. In monofunctional TCS, Model C has a smaller range of SKRR steady state concentrations for which bistability is possible than Model B.

The notion that Model C is the one in which bistable responses are less sensitive to changes in the steady state concentrations of SKRR (in consequence of changing the affinity between SK and RR) is misleading. Bistability is only found in this model if the affinity between the dephosphorylated forms of SK and RR is much larger than that between SKP and RR or SK and RRP. Given that the affinity between all forms of SK and RR was measured as similar, it is not likely that bistability can be found *in vivo* in systems that are represented by this model.

A similar experiment was made by changing independently and simultaneously the total amount of SK and RR, followed by independent calculation of the steady state(s) for each system at different values of the signal represented by k_1 or k_2. Again, each of the parameters was independently and systematically scanned between 10^{-6} and 10 in logarithmic space at intervals of 0.01 units. The results are shown in Table 6 and Figure S3. They are consistent with the situation described for changes in k_9 and k_{10}.

Table 5. Experiments to analyze the effect of changes in different parameter values and protein concentrations on the range of bistability for the alternative TCS modules [a].

Sensitivity to changes in	Parameter	Range of scanning	Parameter	Range of scanning
Formation of the SKRR dead-end complex	k_9	10^{-6}-10 s^{-1}	k_{10}	10^{-6}-10 $\mu M^{-1}s^{-1}$
Ratio between SK_{total} and RR_{total}.	SK_{total}	10^{-3} -$10^{3}\mu M$	RR_{total}	10^{-3} -$10^{3}\mu M$
Ratio between SK_{total} and $TC_{SK\ total}$.	$TC_{SK\ total}$	10^{-3} -$10^{3}\mu M$	SK_{total}	10^{-3} -$10^{3}\mu M$
Ratio between RR_{total} and $TC_{RR\ total}$.	RR_{total}	10^{-3} -$10^{3}\mu M$	$TC_{RR\ total}$	10^{-3} -$10^{3}\mu M$
Formation of the SKRR dead-end complex and rate of RRP dephoshoprylation by SK	k_8	10^{-6}-10	k_9	10^{-6}-10 s^{-1}

[a] The steady state(s) for the three models by scanning a)k_1 (SK autophosphorylation reaction rate constant) and b)k_2 (SKP autodephosphorylation reaction rate constant) between 10^{-6} and 10 at different values of the parameters named in the table (see text for details).

Table 6. Percentage of parameter space where a bistable response is possible for Models A, B, and C[a].

	Experiment	Model A [b]	Model A|B [c]	Model B [b]	Model A|C [c]	Model C [b]
Bifunctional						
	k_8,k_9,k_2	1.8	5.3	2.5	17.8	8.1
	k_9,k_{10},k_2	1.2	0.5	2.7	5.7	4.3
	SKt,RRt,k_2	0.6	NA	1.4	NA	1
	SKt,TCt,k_2	NA	NA	10.9	NA	3
	k_8,k_9,k_1	35.5	33.4	36.7	47.9	39
	k_9,k_{10},k_1	11.3	10.5	11.9	14.3	13.9
	SKt,RRt,k_1	14.1	NA	16	NA	14
	SKt,TCt,k_1	NA	NA	31.3	NA	26.4
Monofunctional						
	k_9,k_{10},k_2	11.9	8.2	15.6	20.9	13.1
	SKt,RRt,k_2	7.7	NA	9.2	NA	6.2
	SKt,TCt,k_2	NA	NA	4.4	NA	10
	k_9,k_{10},k_1	41.4	40.1	42.7	49.3	40.9
	SKt,RRt,k_1	31.2	NA	34	NA	27.9
	SKt,TCt,k_1	NA	NA	75.3	NA	30.7

[a] A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

k_i: kinetic constants for the reactions in the systems shown in Figure 1. SKt: total concentration of SK. RRt: total concentration of RR. TCt: total concentration of third component protein. The parameter space for k_i,k_j, and k_k was scanned between absolute values of 10^{-6} and 10 for each of the parameters. Sampling was uniform in logarithmic space.

[b] Percentage of the parameter space of k_i, k_j and k_k where bistability is found for Models A, B, and C respectively.

[c] Percentage of the parameter space where bistability is found in Model A controlled for B and for C, respectively.

NA Non Applicable. Mono functional systems have k_8=0. The concentration of TC=0 in Model A. Model A can not be scanned with respect to the concentration of SK in the controlled comparisons, because SK is independently fixed to make the dynamical response of Model A more similar to those of Models B and C.

4.4.6. Effect of the SK/TC$_{SK}$ and RR/TC$_{RR}$ concentration ratios on bistability

In order to understand how the relationship between the total amounts of SK (RR) and TC$_{SK}$ (TC$_{RR}$) influences the signaling range in which bistable responses are possible, we have performed a number of computational experiments. First, we took Models B and C from Figure 1. Then, we systematically, simultaneously and independently scanned the total amounts of SK (RR) and TC$_{SK}$ (TC$_{RR}$) in Model B (Model C), as described in Table 5. Finally, for each total amount of SK (RR) and TC$_{SK}$ (TC$_{RR}$), we calculated the steady state(s) for each system at different values of the signal represented by k_2. This parameter was also systematically scanned between 10^{-6} and 10 in logarithmic space at intervals of 0.01 units. The results are shown in Figure S3. We also performed similar test replacing k_2 by k_1.

The range of signal k_2 for which Model B can show a bistable response is observed to be dependent on the TC. Bistability is observed only within a narrow band of the SK-TC$_{SK}$ concentration space. Outside of this band, a bistable response cannot be observed. The range of total amount of SK in the system that may lead to a bistable response remains approximately constant for low total amounts of TC$_{SK}$. However, within the band of total SK and TC$_{SK}$ in which bistability is observed, as total TC$_{SK}$ increases, the range of total SK amount that can generate bistable responses also increases. At concentrations of TC$_{SK}$ between approximately 2 and 7 µM, we find bistability for total SK concentrations between 0.2 and 0.001 µM or lower. At higher total TC$_{SK}$ concentrations, only small amounts of SK are available in free form. This prevents formation of the SKRR dead-end complex that is required for bistability.

As is the case in Model B, bistability in Model C can be achieved in a narrow band of the concentration space. However, within the range of values of this simulation, whatever the concentration of TC_{RR}, the system can always show bistability.

4.5. Discussion

4.5.1. Summary of the comparisons

Tables 7 and 8 summarize our findings regarding the different physiological criteria that are relevant for TCS signal transduction and can be asserted from the analysis of our models. In general, if the signaling target is SK autophosphorylation Model C responds at lower signaling intensities, followed by Model A, and finally by Model B. If the signal enhances SK dephosphorylation, Model B is the one that responds at lower signal intensities, followed by Model A, and Model C. This causes Model C to be in an ON state for a wider signaling range, and Model B to be in an ON state for a narrower signaling range, in comparison with Model A.

The system with the largest range of signaling in which it can show a bistable response depends on both, the type of SK in the module and the SK activity (autophosphorylation or autodephosphorylation) that is targeted by the signal. For TCS with monofunctional SK, Model A has the largest signaling range for bistability, as well as the largest fraction of parameter space where such bistability can be observed, if the environment modulates SK phosphorylation. In contrast, Model B has the largest signaling range for bistability, as well as the largest fraction of parameter space where such bistability can be observed, if the environment modulates SK dephosphorylation. For TCS with bifunctional SK, Model B has the largest signaling range for bistability if the environment modulates SK phosphorylation. However, it is Model C that has the largest fraction of parameter space where bistability can be observed. In contrast, Model A has the largest signaling range for bistability, as well as the largest fraction of parameter space where such bistability can be observed, if the environment modulates SK dephosphorylation.

Modulation of SK dephosphorylation leads to responses that have an equally small amount of noise in all Models. However, modulation of SK phosphorylation leads to noisier responses in Model B, followed by Model A and finally Model C.

As is the case with bistability, the model with fastest response times depends on the type of SK in the module and on the SK activity (autophosphorylation or autodephosphorylation) that is targeted by the signal. Both in systems with monofunctional and bifunctional SK, Model A is the fastest to respond (Model C is the slowest) whether the signaling target is the autophosphorylation or the autodephosphorylation of the SK, with only one exception: Model B turns ON faster if SK autophosphorylation is modulated directly. The response times of Models A and B are similar, but Model C tends to be much slower than Model A.

Table 7. Summary of the comparison of physiologically relevant criteria between the alternative designs for monofunctional TCS [a].

		MONOFUNCTIONAL				
Signaling target	**Physiological criterion**	**Model A**	**Model B**	**Model C**	**Model A|B**	**Model A|C**
Phosphorylation of SK (k_1)	Sensitivity to signal	+++	++	+++++	++	++++
	Signaling range of bistability	+++	++	+	++	++++
	Fraction of parameter space with bistability	++++	+	++	+++	+++++
	Noisy response	+++	+++++	+	++++	++
	Fast OFF→ON response time	++++	++	+++	+	+++++
	Fast ON→OFF response time	+++	+	++++	++	+++++
		Model A	**Model B**	**Model C**	**Model A|B**	**Model A|C**
Dephosphorylation of SKP (k_2)	Sensitivity to signal	+++	+++++	++	++++	++
	Signaling range of bistability	++	++++	-	++++	-
	Fraction of parameter space with bistability	+++	++++	+	+++++	++
	Noisy response	+	+	+	+	+
	Fast OFF→ON response time	+++	++++	+	+++++	++
	Fast ON→OFF response time	+++	++++	+	+++++	++

[a] The model with the largest number of "+" signs for a given criterion is the one with the best performance with respect to that criterion.

A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

Table 8. Summary of the comparison of physiologically relevant criteria between the alternative designs for TCS with bifunctional SK [a].

				BIFUNCTIONAL		
Signaling target	**Physiological criterion**	**Model A**	**Model B**	**Model C**	**Model A|B**	**Model A|C**
Phosphorylation of SK (k_1)	Sensitivity to signal	++	+	++++	+	+++
	Signaling range of bistability	+++	++	+	++	-
	Fraction of parameter space with bistability	+++	++	+++++	++++	+
	Noisy response	+++	+++++	++	++++	+
	Fast OFF→ON response time	++++	++	+++	+	+++++
	Fast On→OFF response time	+++	+	++++	++	+++++
		Model A	**Model B**	**Model C**	**Model A|B**	**Model A|C**
Dephosphorylation of SKP (k_2)	Sensitivity to signal	++++	+	++	+	+++
	Signaling range of bistability	+++	-	-	-	-
	Fraction of parameter space with bistability	+++++	++	+++	++++	+
	Noisy response	+	+	+	+	+
	Fast OFF→ON response time	+++	++++	+	+++++	++
	Fast ON→OFF response time	++	+++	+	++++	+++++

[a] The model with the largest number of "+" signs for a given criterion is the one with the best performance with respect to that criterion.

A|B stands for Model A controlled for Model B. A|C stands for Model A controlled for Model C.

4.5.2. Biological Relevance

Bacteria often sense and adapt to changes in the environment through TCS and phosphorelays. A question that this work addresses is how variations to the prototypical TCS by means of an accessory third protein that either binds the SK or the RR affect the dynamical behavior of the TCS module.

TCS can, in principle, mediate both gradual and switch like (bistable) responses to environmental stimuli [32,33]. The switch-like response has typically been associated with the positive feedback introduced by genetic regulatory loops in the regulation of autogenous TCS. Nevertheless, such feedback does not necessarily imply the existence of bistability [34]. In fact, genetic positive feedback loops are not strictly necessary for the existence of bistable responses in prototypical TCS. Such responses can also come about through post-translational regulation of bacterial signal transduction networks [25,35]. Namely, bistability is possible in prototypical TCS if a reversible dead-end complex is formed between the dephosphorylated SK and RR and if a sufficient amount of RRP is dephosphorylated independently of the SK phosphatase activity [25].

TC proteins that regulate signal transmission to prototypical TCS have been known for years [36,37]. However, only recently have such interactions been proposed as a way to integrate non-cognate signals in the TCS regulated responses. In fact, these interactions have been reported in TCS that are responsible for regulating both, resistance to antibiotics and virulence [6,7,8,9,12,13,14,15].

Biological examples of the first situation can be found in the PmrB/PmrA/PmrD system. The third component PmrD binds and stabilizes the active form of the RR, PmrA. This system regulates antibiotic resistance in *Salmonella* and other bacteria.

Various studies of the PmrA/PmrB/PmrD system suggest that this TC_{RR} could be an intermediate evolutionary step to evolve indirect regulation of the TCS [12,16,17,22,38]. The feedforward connector loop formed by PmrD is presented as a design that speeds up activation and slows deactivation of the gene expression of the proteins in the TCS [17]. Our results suggest that this may not be so in non-autogenous TCS. If the TCS has a TC_{RR}, loss of this protein will make the corresponding prototypical TCS faster to turn ON and OFF (Table S1 and Figure S1). In fact, if the steady state response curve of the prototypical TCS is mathematically controlled to be as similar to that of the TCS with a TC_{RR} as possible, then that prototypical system is always faster. A TC_{RR} appears also to be a feature that decreases the fraction of parameter space in which bistable responses are possible (Tables 4 and 6), except in TCS with a bifunctional SK and when the environment modulates SK autophosphorylation. Thus, a TC_{RR} creates a TCS module that is less likely to show bistable responses and slower in responding to environmental signals, which it can sense at lower intensities than the prototypical TCS without any TC, if SK phosphorylation is modulated.

Antibiotic resistance is arguably a trait whose response should be gradual and proportional to the amount of antibiotic found by the bacteria to increase its survival chances. If this is not so, and a bistable response is possible, bacteria can be made more sensitive to antibiotics [39] and therefore their survival will be hindered. Given that bistability has been observed in the antibiotic resistance of some bacteria [39], a TC that binds the RR would reduce the possibility of such bistable response, potentiating adaptation and tolerance to threatening stress challenges. In addition, having such a TC could enable a response at low antibiotic concentrations, thus increasing the chances of survival for the organism.

The other well studied example of a TC interacting with the TCS is the RetS/GacS/GacA system, where RetS reversibly binds and inactivates the SK GacS. This system regulates virulence in *Pseudomonas aeruginosa*. Recently, it has been shown that the GacS/GacA TCS acts exclusively through the regulation of the transcription of two genes, *rsmY* and *rsmZ* [40]. The product of these genes are two untranslated small regulatory RNAs (sRNAs), RsmY and RsmZ, that counter translational repression exerted by the RNA-binding protein RsmA on target mRNAs encoding virulence factors. There is an additional SK, LadS, that appears to counter the action of RetS on GacS. However, this effect is indirect, as not direct physical interaction between GacS and LadS was observed [10]. It may be that LadS sequesters RetS, as RetS does with GacS. Our analysis of a TCS with a TC_{SK} reveals that this module will respond at signal intensities that are slightly higher (lower) than those of the prototypical TCS, if SK authophosphorylation (autodephosphorylation) is directly modulated. Furthermore, if one is to synthetically change a TCS module and create an artificial circuit with a TC_{SK}, the engineered circuit will typically respond faster to signals if the environment modulates SK dephosphorylation. However, evolution can eventually equalize response times by changing the SK concentration of the module and making both TCS modules have steady state response curves that are similar. A TC_{SK} can increase the signaling range in which a bistable response is possible (Table 6). Bistability could be advantageous when the system has to choose between two different operational states [35,41], as is often the case for virulent organisms. For example, *Mycobacterium tuberculosis* is a persistent organism in the lungs of 2/7 of the world population [42]. However, only under certain conditions that are not yet completely clear does this organism causes tuberculosis [42]. Bistability could provide populations with the capacity to sample which type of phenotype is more advantageous at different times and enhance survival of the organisms through bet-hedging strategies [43,44].

Experiments to test the existence of bistability in a TCS with a TC_{SK} could be as follows, taking the RetS/GaS/GacA system as an example. First, determine if the system can show bistable response: incubate two *Pseudomonas aeruginosa* strains (a wild type strain, with the TC protein RetS, and a RetS mutant strain, without the TC protein) at different environmental conditions of inducing signal intensity, allow the cells to approach a steady state and measure the levels of expression of the sRNAs RsmY and RsmZ. In a TCS module with a monostable gradual response, the level of expression of the output molecules should be proportional to the environmental inducing signal intensity: at intermediate signal intensities there is an intermediate amount of output molecule. However, if the RetS/ GacS/GacA response is bistable, we will find that, for intermediate intensities of inducing signal, the measured levels of RsmY and RsmZ in single cells are distributed in a bimodal manner, with low and high levels (but no intermediate levels) of this sRNAs. If bistability is present and the *in vivo* effect of RetS is to amplify the signaling range for which a bistable response is possible, we will find that this bimodal distribution of the measured levels of RsmY and RsmZ in single cells of the RetS mutant strain will be observed in a smaller range of signal values. To investigate if the results of our simulations are valid for *in vivo* conditions and if the RetS/GacS/GacA system could have an irreversible response (as observed in Figure 2 C), we can incubate both strains in a high-stimulus environment, allow the cells to approach a steady state and measure the levels of the sRNAs RsmY and RsmZ . If the in vivo system behaves as its *in silico* proxy, when the stimulus is removed (transfer the cells to a non-inducing environment), we will find that in RetS mutant cells the levels of RsmY and RsmZ shift from a low value to a high value, but in wild type cells the levels of RsmY and RsmZ remain at a low value.

The arguments discussed thus far explain part of the biological relevance of our work. Another interesting aspect of it regards the modulation of SK autophosphorylation and dephosphorylation. Currently the community is inclined to assume that dephosphorylation is the target of modulation by environmental signals in many cases. However, to our knowledge, conclusive experiments that decide the issue are still lacking in most systems and it is still unclear whether the physiological signal modulates SK autophosphorylation (k_1) or SKP dephosphorylation (k_2). That is why we have performed our simulations taking as a signal both changes in k_1 and k_2. An unexpected result of our simulations may shed some light on this issue, and allow us to hypothesize which one of the reaction rates is modulated by the signal in the case of TCS with a TC. We have found that, for TCS with a bifunctional SK, a TC decreases the possibility of a bistable response. For TCS with a monofunctional SK, the same effect is observed if the signal modulates k_1. However, if the signal modulates k_2, a TC_{SK} increases the range of signal intensities in which a TCS can show bistability, and a TC_{RR} decreases it. Thus, for TCS with a monofunctional SK, the results suggest that the physiological signal should modulate SK dephosphorylation (k_2) both when bistability is an advantageous feature in the function of a TCS with a TC_{SK} component, and when bistability is a disadvantageous feature in the function of a TCS with a TC_{RR}. Conversely, the physiological signal should modulate SK autophosphorylation (k_1) when bistability is a disadvantageous feature in the function of a TCS with a TC_{SK}.

The work presented in this chapter provides motivation for further analyses of the TCS responsible for regulating virulence and antibiotic resistance, providing clues as to possible mechanisms to both decrease virulence and antibiotic resistance. In the case of virulence, whenever it is regulated by a TCS of the type analyzed here, simultaneously targeting the TC and the SK appropriately could prevent the organism from becoming virulent. In the case of antibiotic resistance, targeting the TC and its interaction with the RR could be used to facilitate locking the bacteria in an antibiotic-sensitive state and facilitate treatment of infections.

4.6. References

1. Garcia Vescovi E, Sciara MI, Castelli ME (2010) Two component systems in the spatial program of bacteria. Curr Opin Microbiol 13: 210-218.
2. Wuichet K, Cantwell BJ, Zhulin IB (2010) Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol 13: 219-225.
3. Silversmith RE (2010) Auxiliary phosphatases in two-component signal transduction. Curr Opin Microbiol 13: 177-183.
4. Hazelbauer GL, Lai WC (2010) Bacterial chemoreceptors: providing enhanced features to two-component signaling. Curr Opin Microbiol 13: 124-132.
5. Atkinson MR, Ninfa AJ (1998) Role of the GlnK signal transduction protein in the regulation of nitrogen assimilation in Escherichia coli. Mol Microbiol 29: 431-447.
6. Buelow DR, Raivio TL (2010) Three (and more) component regulatory systems - auxiliary regulators of bacterial histidine kinases. Mol Microbiol 75: 547-566.
7. Goodman AL, Merighi M, Hyodo M, Ventre I, Filloux A, et al. (2009) Direct interaction between sensor kinase proteins mediates acute and chronic disease phenotypes in a bacterial pathogen. Genes Dev 23: 249-259.
8. Lapouge K, Schubert M, Allain FH, Haas D (2008) Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour. Mol Microbiol 67: 241-253.
9. Raghavan V, Groisman EA (2010) Orphan and hybrid two-component system proteins in health and disease. Curr Opin Microbiol 13: 226-231.
10. Workentine ML, Chang L, Ceri H, Turner RJ (2009) The GacS-GacA two-component regulatory system of Pseudomonas fluorescens: a bacterial two-hybrid analysis. FEMS Microbiol Lett 292: 50-56.
11. Yan Q, Wu XG, Wei HL, Wang HM, Zhang LQ (2009) Differential control of the PcoI/PcoR quorum-sensing system in Pseudomonas fluorescens 2P24 by sigma factor RpoS and the GacS/GacA two-component regulatory system. Microbiol Res 164: 18-26.
12. Kato A, Groisman EA (2004) Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor. Genes Dev 18: 2302-2313.
13. Gooderham WJ, Hancock RE (2009) Regulation of virulence and antibiotic resistance by two-component regulatory systems in Pseudomonas aeruginosa. FEMS Microbiol Rev 33: 279-294.
14. Goodman AL, Kulasekara B, Rietsch A, Boyd D, Smith RS, et al. (2004) A signaling network reciprocally regulates genes associated with acute infection and chronic persistence in Pseudomonas aeruginosa. Dev Cell 7: 745-754.
15. Eguchi Y, Utsumi R (2005) A novel mechanism for connecting bacterial two-component signal-transduction systems. Trends Biochem Sci 30: 70-72.
16. Chen HD, Jewett MW, Groisman EA (2011) Ancestral genes can control the ability of horizontally acquired loci to confer new traits. PLoS Genet 7: e1002184.
17. Mitrophanov AY, Jewett MW, Hadley TJ, Groisman EA (2008) Evolution and dynamics of regulatory architectures controlling polymyxin B resistance in enteric bacteria. PLoS Genet 4: e1000233.
18. Al-Khodor S, Kalachikov S, Morozova I, Price CT, Abu Kwaik Y (2009) The PmrA/PmrB two-component system of Legionella pneumophila is a global regulator required for intracellular replication within macrophages and protozoa. Infect Immun 77: 374-386.

19. Perez JC, Groisman EA (2007) Acid pH activation of the PmrA/PmrB two-component regulatory system of Salmonella enterica. Molecular Microbiology 63: 283-293.

20. McPhee JB, Bains M, Winsor G, Lewenza S, Kwasnicka A, et al. (2006) Contribution of the PhoP-PhoQ and PmrA-PmrB two-component regulatory systems to Mg2+-induced gene regulation in Pseudomonas aeruginosa. J Bacteriol 188: 3995-4006.

21. Cheng HY, Chen YF, Peng HL (2010) Molecular characterization of the PhoPQ-PmrD-PmrAB mediated pathway regulating polymyxin B resistance in Klebsiella pneumoniae CG43. J Biomed Sci 17: 60.

22. Kato A, Mitrophanov AY, Groisman EA (2007) A connector of two-component regulatory systems promotes signal amplification and persistence of expression. Proc Natl Acad Sci U S A 104: 12063-12068.

23. Salvado B, Karathia H, Chimenos AU, Vilaprinyo E, Omholt S, et al. (2011) Methods for and results from the study of design principles in molecular systems. Mathematical Biosciences 231: 3-18.

24. Alves R, Savageau MA (2003) Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function. Mol Microbiol 48: 25-51.

25. Igoshin OA, Alves R, Savageau MA (2008) Hysteretic and graded responses in bacterial two-component signal transduction. Mol Microbiol 68: 1196-1215.

26. Alves R, Vilaprinyo E, Hernandez-Bermejo B, Sorribas A (2008) Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. Biotechnology and Genetic Engineering Reviews, Vol 25 25: 1-40.

27. Batchelor E, Goulian M (2003) Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system. Proc Natl Acad Sci U S A 100: 691-696.

28. Bhattacharya M, Biswas A, Das AK (2010) Interaction analysis of TcrX/Y two component system from Mycobacterium tuberculosis. Biochimie 92: 263-272.

29. Stewart RC, Van Bruggen R (2004) Association and dissociation kinetics for CheY interacting with the P2 domain of CheA. J Mol Biol 336: 287-301.

30. Yoshida T, Cai S, Inouye M (2002) Interaction of EnvZ, a sensory histidine kinase, with phosphorylated OmpR, the cognate response regulator. Mol Microbiol 46: 1283-1294.

31. Alves R, Savageau MA (2000) Extending the method of mathematically controlled comparison to include numerical comparisons. Bioinformatics 16: 786-798.

32. Novick A, Weiner M (1957) Enzyme Induction as an All-or-None Phenomenon. Proc Natl Acad Sci U S A 43: 553-566.

33. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol 26: 389-401.

34. Tiwari A, Ray JC, Narula J, Igoshin OA (2011) Bistable responses in bacterial genetic networks: designs and dynamical consequences. Mathematical Biosciences 231: 76-89.

35. Igoshin OA, Price CW, Savageau MA (2006) Signalling network with a bistable hysteretic switch controls developmental activation of the sigma transcription factor in Bacillus subtilis. Mol Microbiol 61: 165-184.

36. Wolfe AJ, Conley MP, Kramer TJ, Berg HC (1987) Reconstitution of signaling in bacterial chemotaxis. J Bacteriol 169: 1878-1885.

37. Kamberov ES, Atkinson MR, Feng J, Chandran P, Ninfa AJ (1994) Sensory components controlling bacterial nitrogen assimilation. Cell Mol Biol Res 40: 175-191.

38. Perez JC, Groisman EA (2009) Evolution of transcriptional regulatory circuits in bacteria. Cell 138: 233-244.

39. Fange D, Nilsson K, Tenson T, Ehrenberg M (2009) Drug efflux pump deficiency and drug target resistance masking in growing bacteria. Proc Natl Acad Sci U S A 106: 8215-8220.

40. Brencic A, McFarland KA, McManus HR, Castang S, Mogno I, et al. (2009) The GacS/GacA signal transduction system of Pseudomonas aeruginosa acts exclusively through its control over the transcription of the RsmY and RsmZ regulatory small RNAs. Mol Microbiol 73: 434-445.

41. Boots M, Hudson PJ, Sasaki A (2004) Large shifts in pathogen virulence relate to host population structure. Science 303: 842-844.

42. Lin PL, Flynn JL (2010) Understanding latent tuberculosis: a moving target. J Immunol 185: 15-22.

43. Veening JW, Smits WK, Kuipers OP (2008) Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol 62: 193-210.

44. Minoia M, Gaillard M, Reinhard F, Stojanov M, Sentchilo V, et al. (2008) Stochasticity and bistability in horizontal transfer control of a genomic island in Pseudomonas. Proc Natl Acad Sci U S A 105: 20792-20797.

45. Wolfram Research I (2010) Mathematica. Champaign, Illinois: Wolfram Research, Inc.

46. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, et al. (2006) COPASI--a COmplex PAthway SImulator. Bioinformatics 22: 3067-3074.

47. Nalca Y, Jansch L, Bredenbruch F, Geffers R, Buer J, et al. (2006) Quorum-sensing antagonistic activities of azithromycin in Pseudomonas aeruginosa PAO1: a global approach. Antimicrob Agents Chemother 50: 1680-1688.

48. Mattison K, Kenney LJ (2002) Phosphorylation alters the interaction of the response regulator OmpR with its sensor kinase EnvZ. J Biol Chem 277: 11143-11148.

4.7. Supplementary materials

4.7.1. Supplementary Figures

Supporting Information Legends

Figure S1. Temporal responsiveness curves of Models A, B, and C. The systems are at an initial steady state and, at time zero, the signal, represented in the x axis, changes instantaneously and the time it takes for the system to get to within 90% of the new steady state is measured and plotted in the y axis. A-D: Response times of TCS with monofunctional SK. E-H: Response times of TCS with bifunctional SK. The OFF to ON plots start with the systems at an OFF steady state (low levels of RRP) corresponding to a low value of k_1 (A, C, E, G) or a high value of k_2 (B, D, F, H). The signal is then changed to increase the steady state level of RRP. The ON to OFF plots start with the systems at an ON steady state (high levels of RRP) corresponding to a high value of k_1 or a low value of k_2. The signal is then changed to decrease the steady state level of RRP. Peaks that indicate slower response times are located immediately outside the range of bistability. The lack of a peak in a curve can be due to monostability or irreversibility Absence of a dashed line indicates irreversible turning ON or OFF of the system (Model B in panel C) or absence of bistability (see the signal-response curves of Figure 2). The difference between this Figure and Figure 3 is that the time curves for Model A are calculated with the total concentration of SK being the same in the three Models. The overall response times (equivalent to the sum of all the transient response times for each curve) is shown in Table S1.

Figure S2. Effect of changing the parameter values on the range of bistability in the three TCS modules. In the panels, the x-axis represents values for k_1 (SK autophosphorylation rate constant) or k_2 (SK dephosphorylation rate constant), and the y-axis represents values for each of the other reaction rate constants that are common to the three models (from k_2 to k_{13}). The region where bistability is possible is shaded in blue. The number above each set of plots represents the summation of all areas of bistability in a given model, that is, is a measure of the size of the parametric space of bistability. A, B: Comparison between Models A and B, with a monofunctional SK. C, D: Comparison between Models A and B, with a bifunctional SK. E, F: Comparison between Models A and C, with a monofunctional SK. G, H: Comparison between Models A and C, with a bifunctional SK.

Figure S3. Percentage of parameter space where a bistable response is possible for Models A, B, and C. Experiments as described in Table 6. The x and y axis represent the values of the scanned parameters, while the z-axis represents the orders of magnitude of signal for which there is a bistable response. The red projection represents the area of parameter space where bistable responses are possible. A – Bifunctional system, signal modulating dephosphorylation of the SK.; B – Bifunctional system, signal modulating phosphorylation of the SK; C – Monofunctional system, signal modulating dephosphorylation of the SK.; D – Monofunctional system, signal modulating phosphorylation of the SK. See text for details and discussion. For higher resolution in this figure, please see the appendix SF3 in the digital version of this thesis.

Figure S4. Influence of the k_8 value (SK bifunctionality rate constant) on the k_2 range of bistability. Within a k_8 range of values, an increase in k_8 causes an increase in the k_2 range of bistability (panel a and b). This is so, despite an enlargement of the fraction of RRP dephosphorylated by SK (panel c), because of an increase in the SKRR concentration due to a higher value of k_8 (panel d). The simulations were performed using the system represented by Model A.

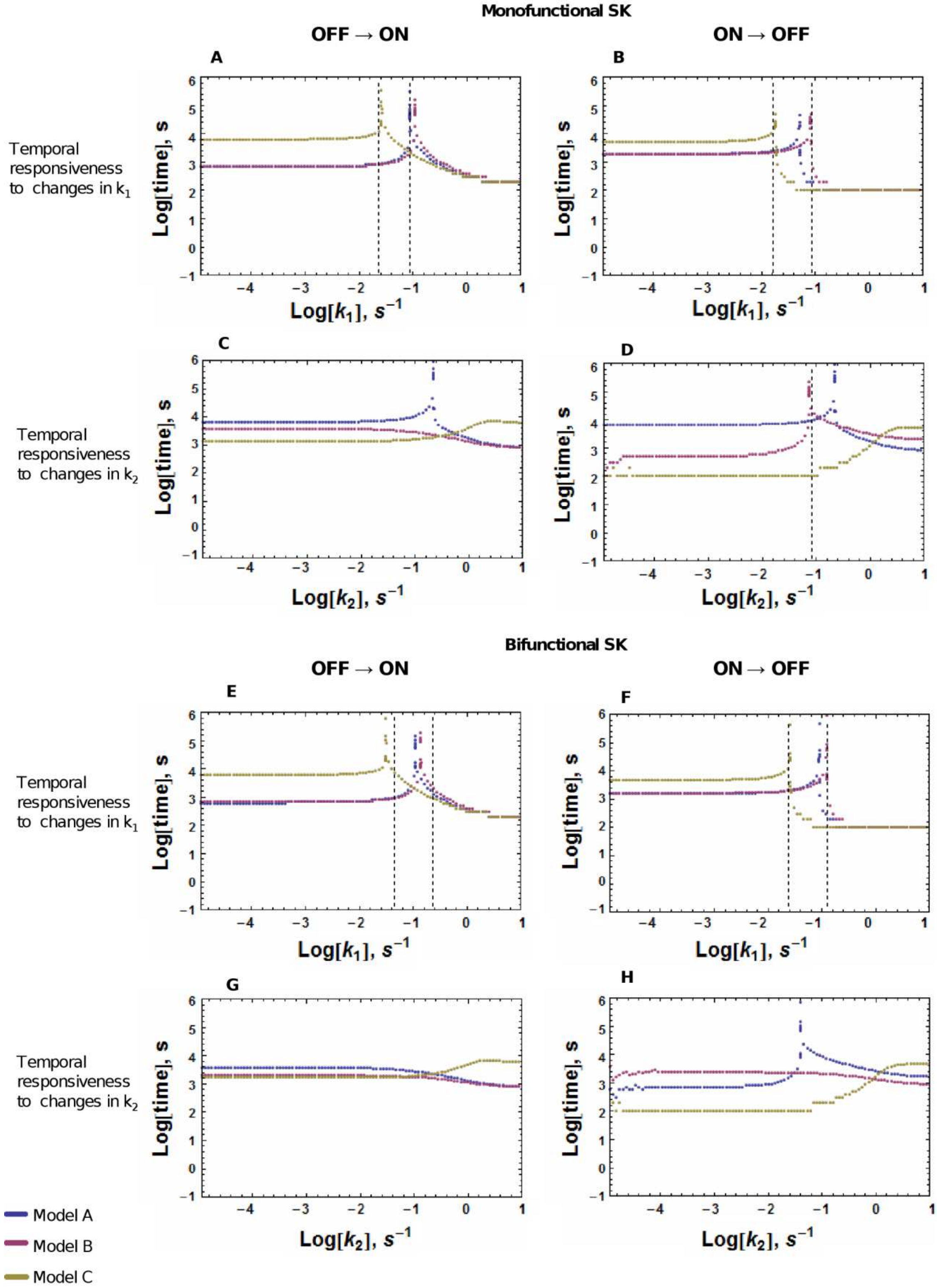

Monofunctional SK
OFF → ON
ON → OFF
A
B
C
D
Temporal responsiveness to changes in k_1
Temporal responsiveness to changes in k_2
Log[time], s
Log[k_1], s^{-1}
Log[k_2], s^{-1}
Bifunctional SK
OFF → ON
ON → OFF
E
F
G
H
Temporal responsiveness to changes in k_1
Temporal responsiveness to changes in k_2
Model A
Model B
Model C

Figure S1

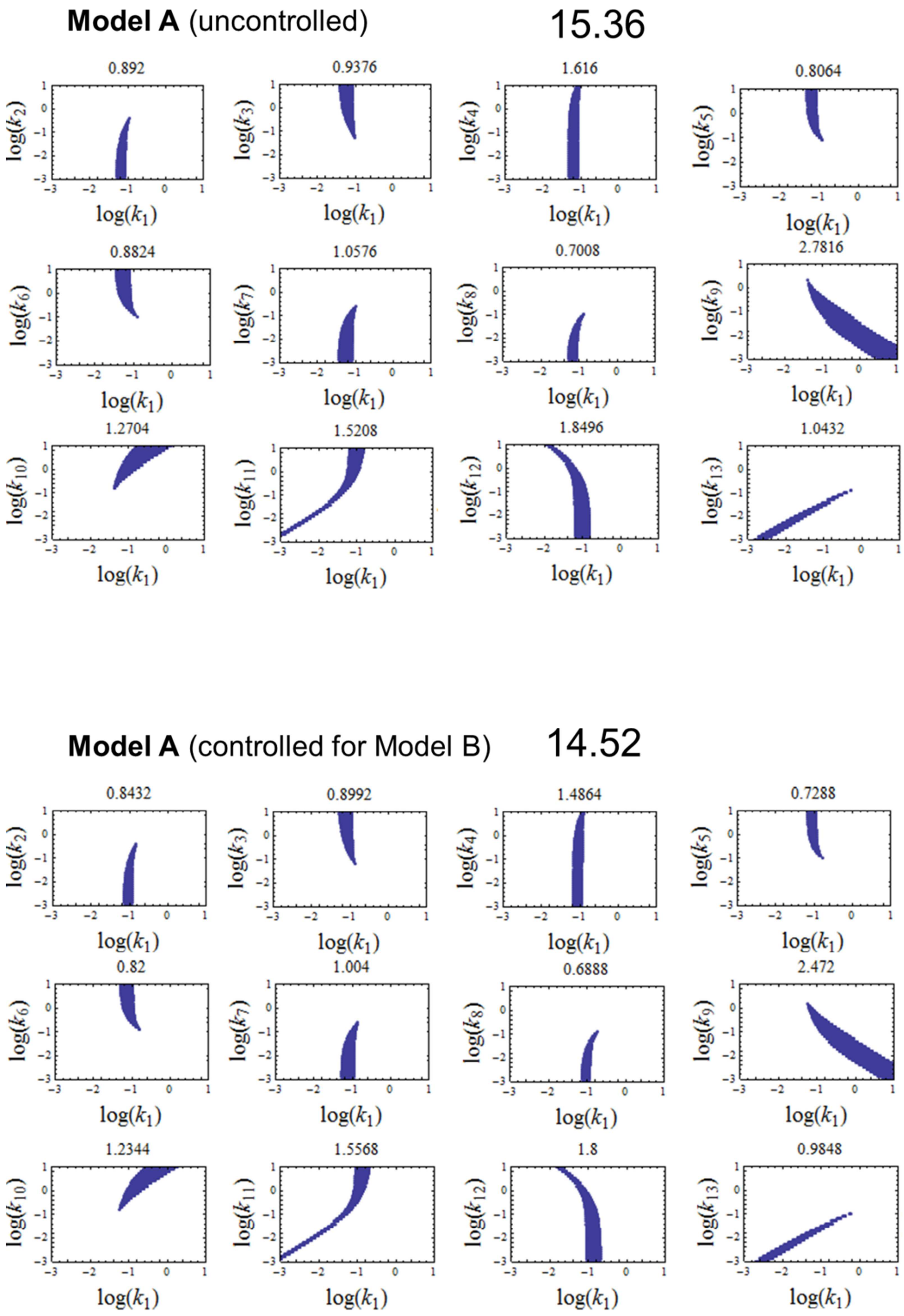


Figure S2 A. Systems with a monofunctional SK. Input signal = change in k_1

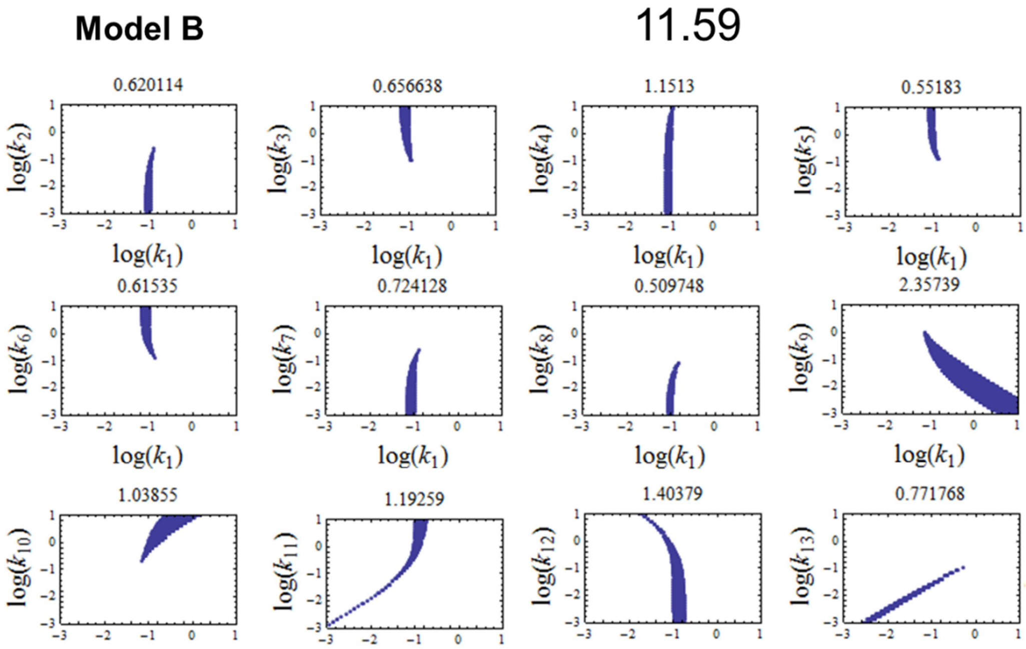


Figure S2 A (continued)

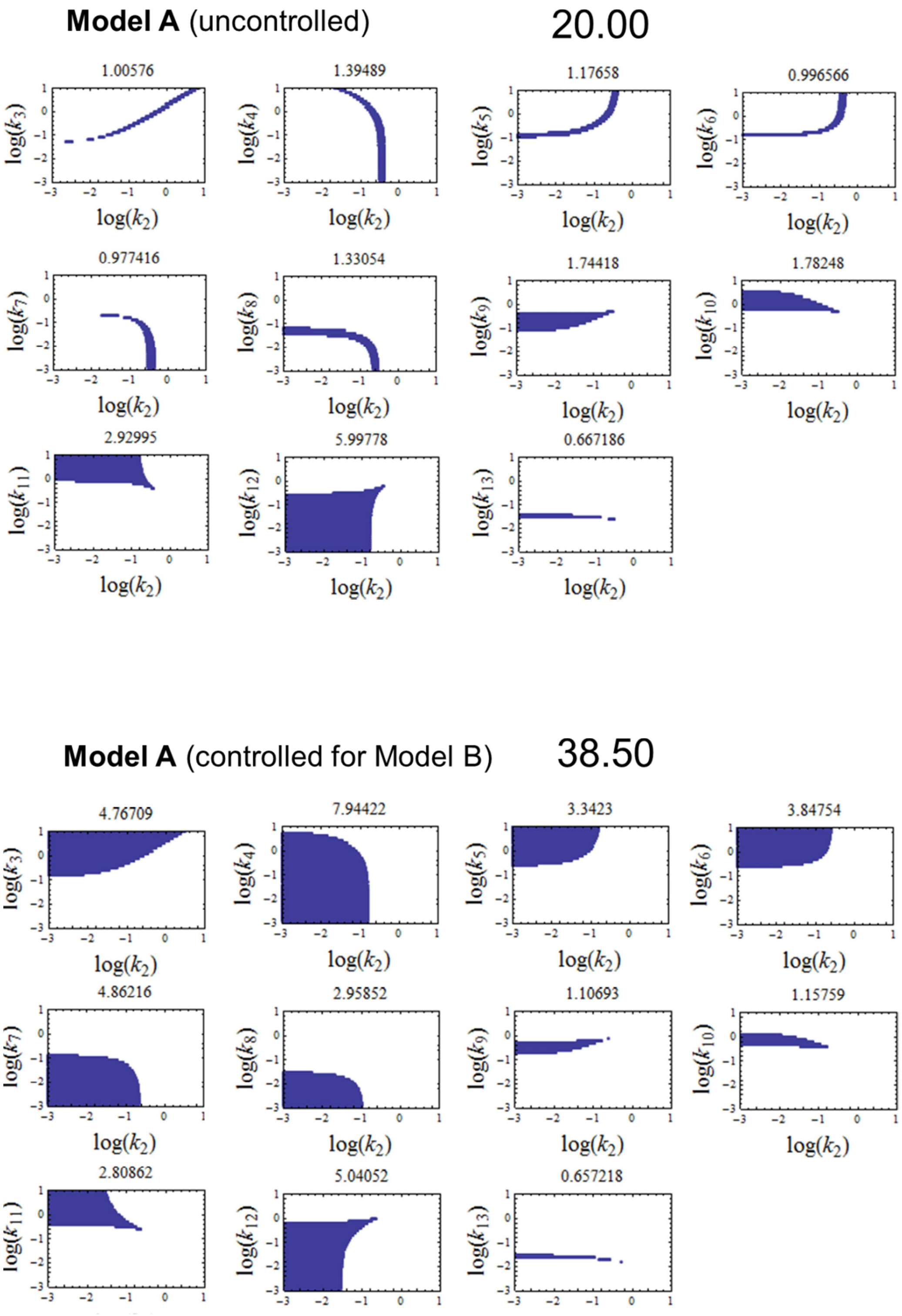


Figure S2 B. Systems with a monofunctional SK. Input signal = change in k_2

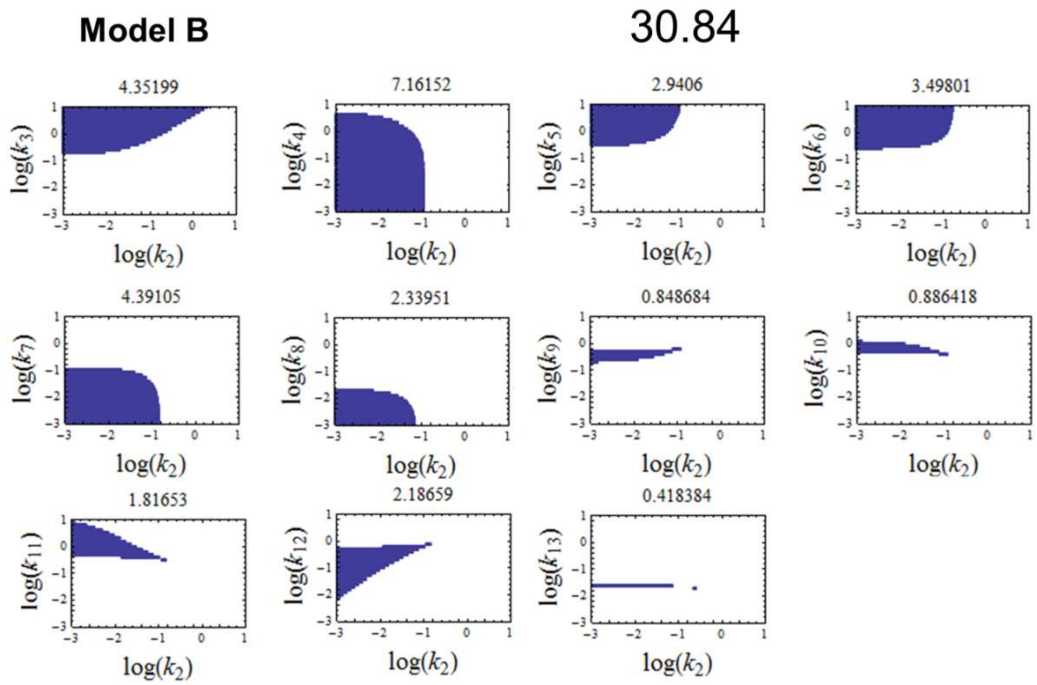


Figure S2 B (continued)

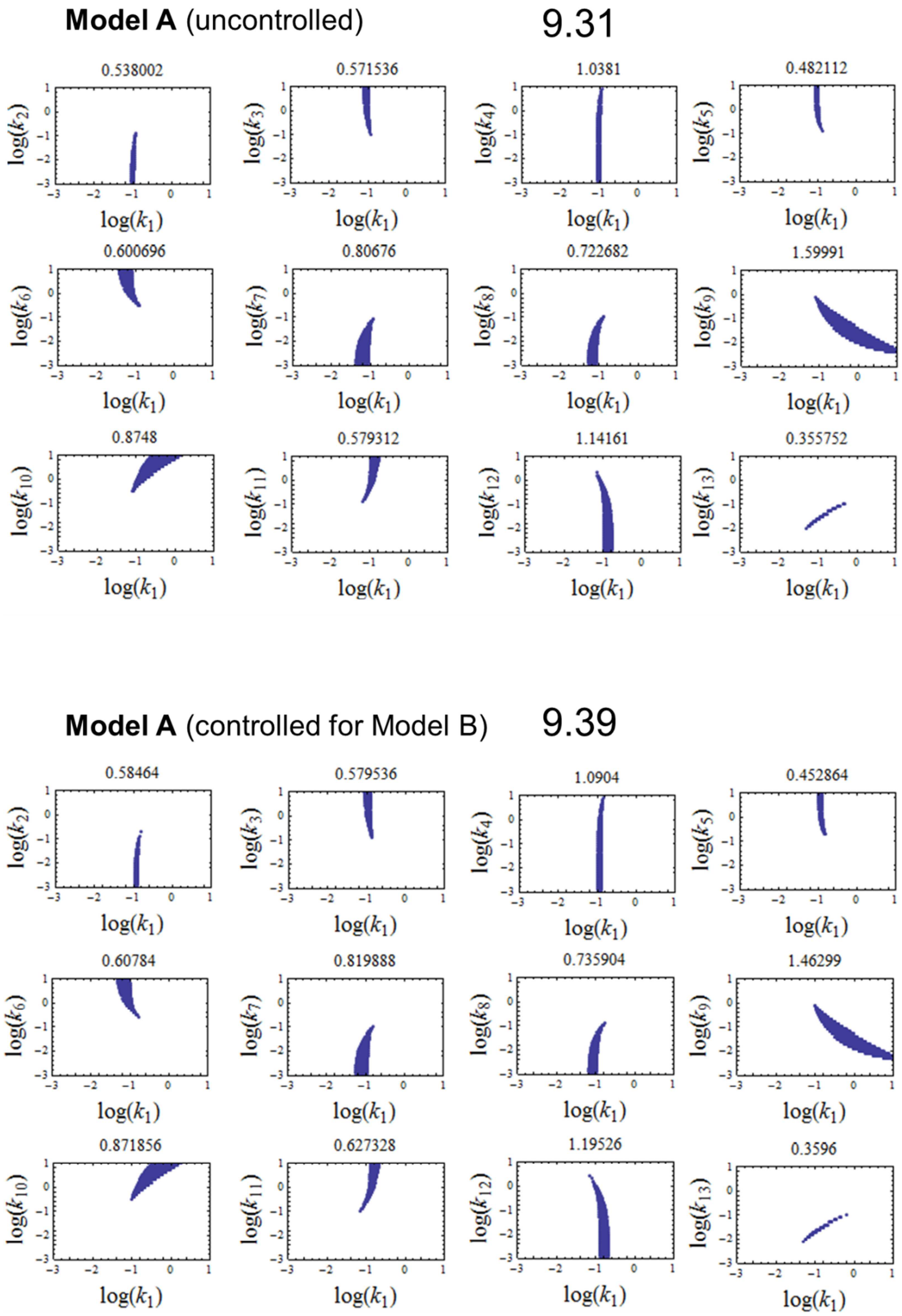


Figure S2 C. Systems with a bifunctional SK. Input signal = change in k_1

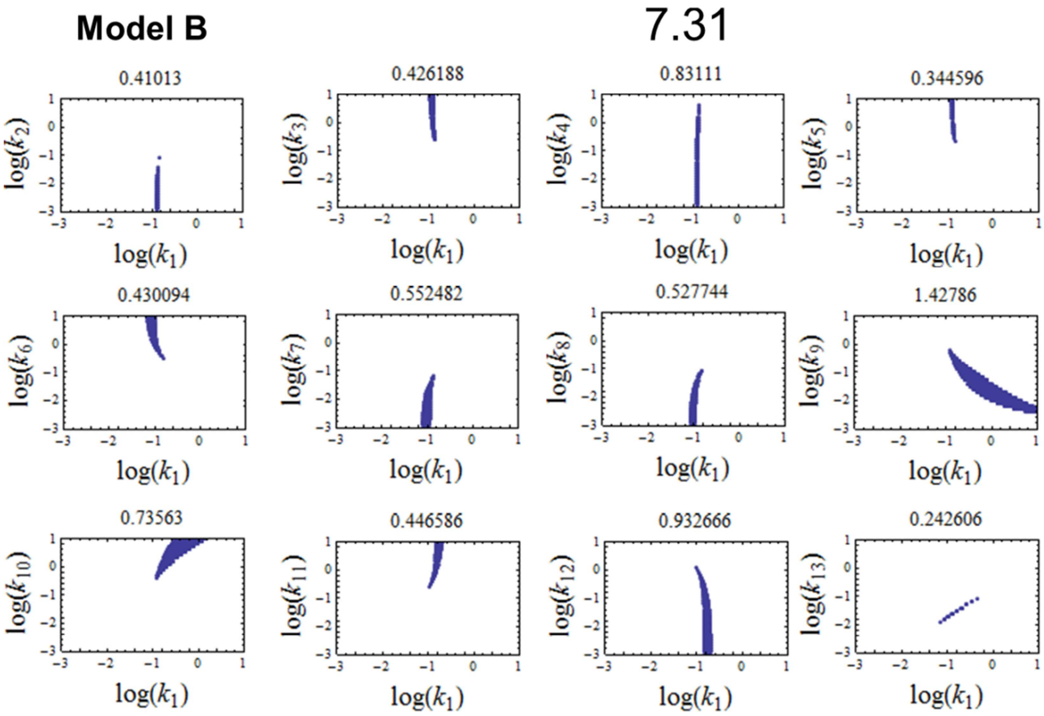


Figure S2 C (continued)

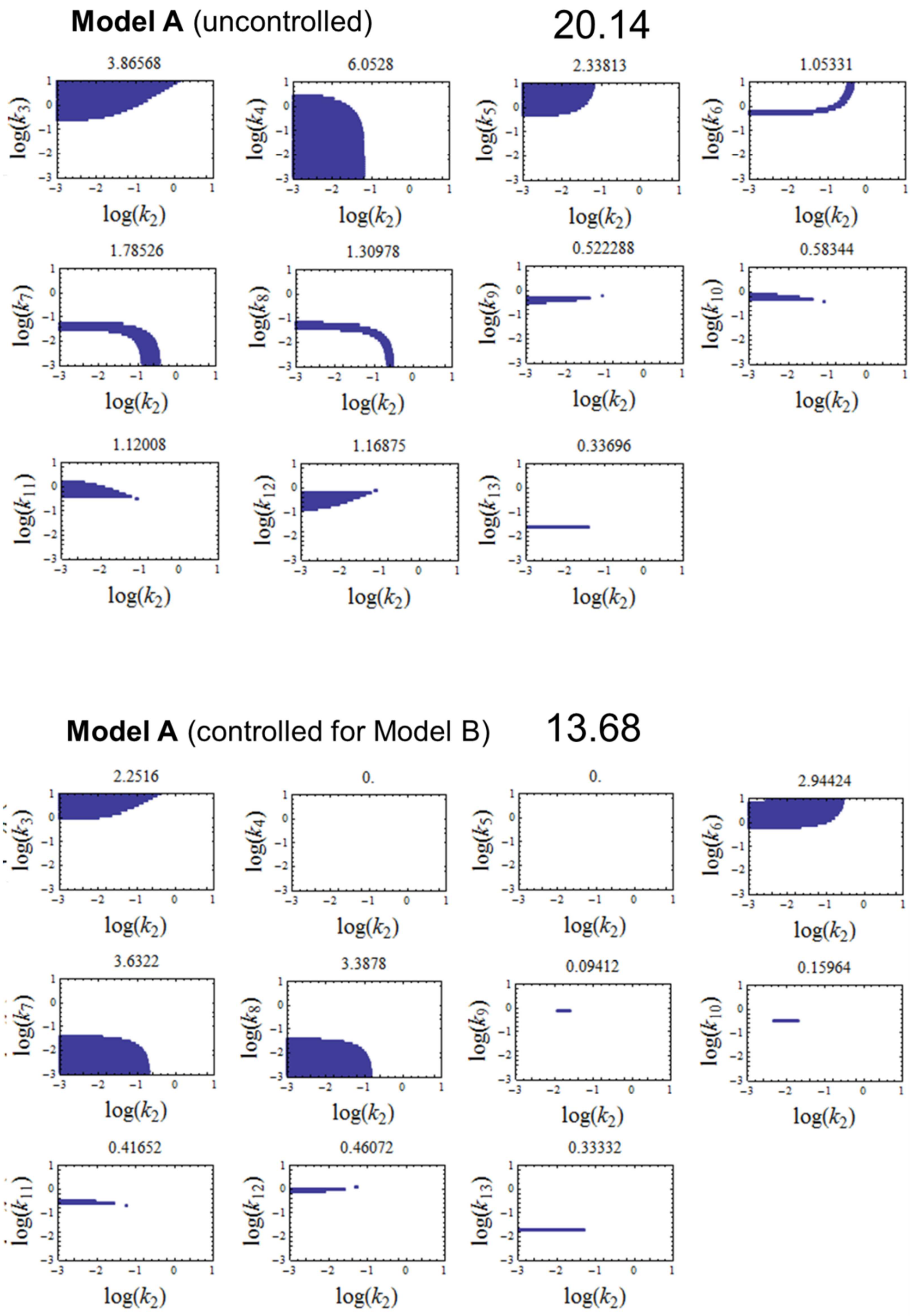


Figure S2 D. Systems with a bifunctional SK. Input signal = change in k_2

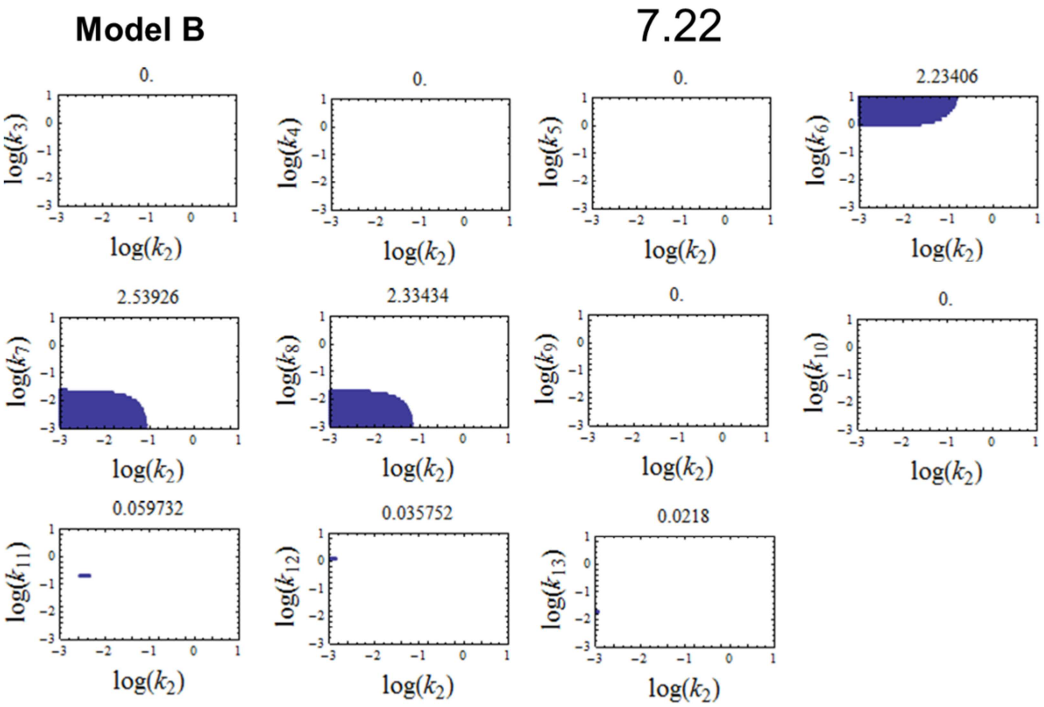


Figure S2 D (continued)

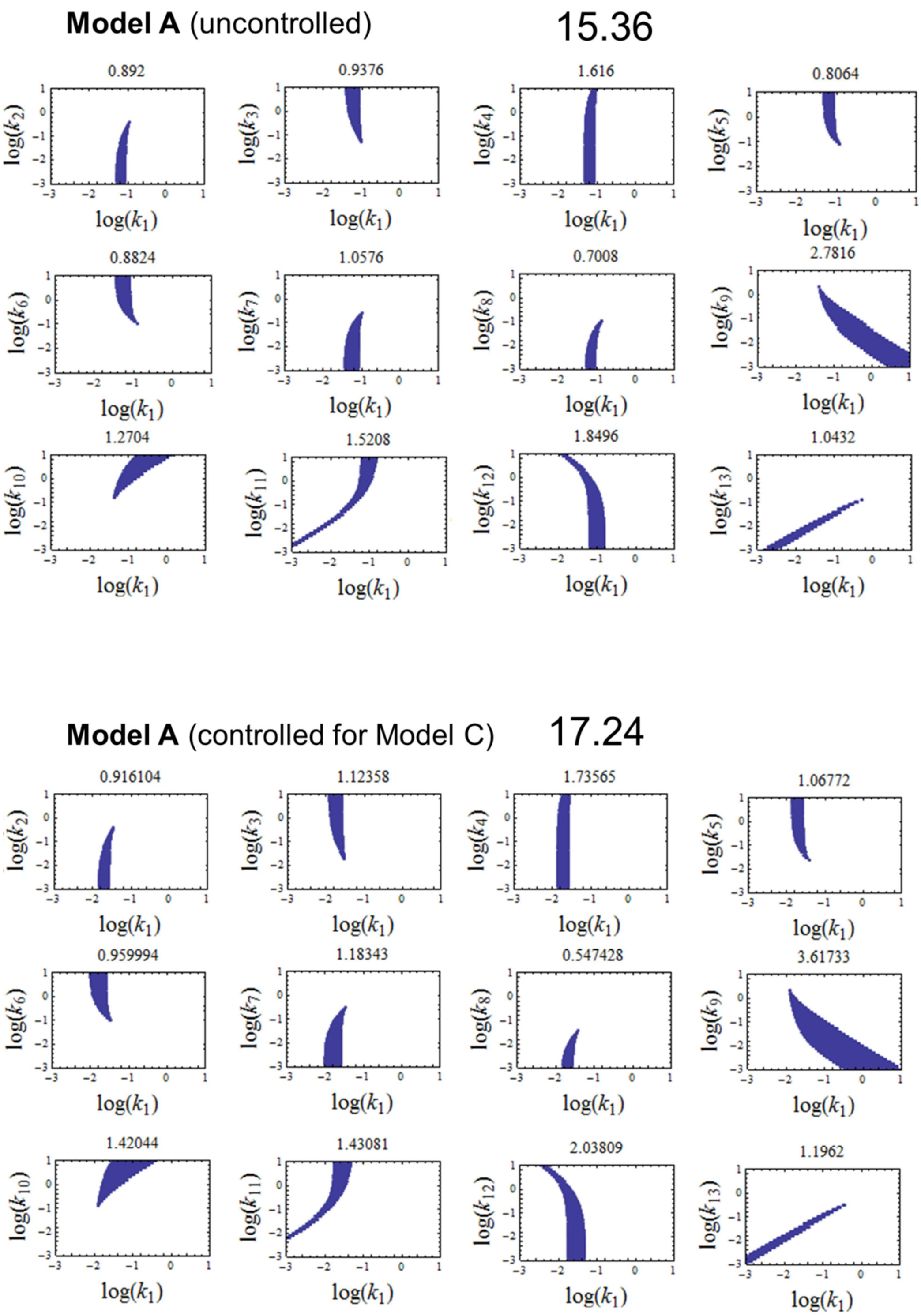


Figure S2 E. Systems with a monofunctional SK. Input signal = change in k_1

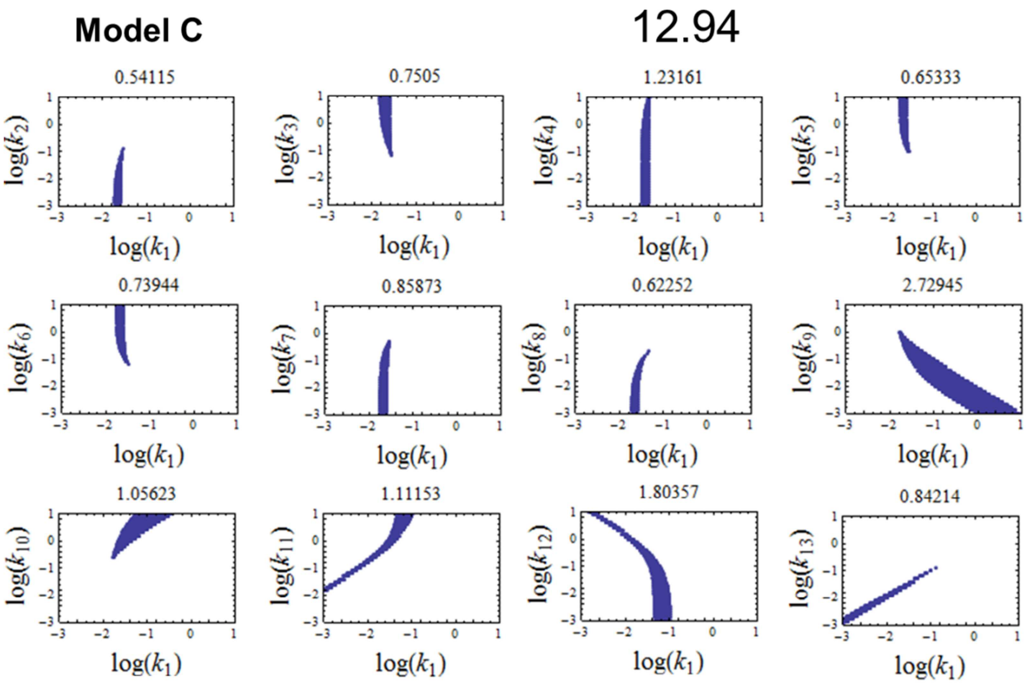


Figure S2 E (continued)

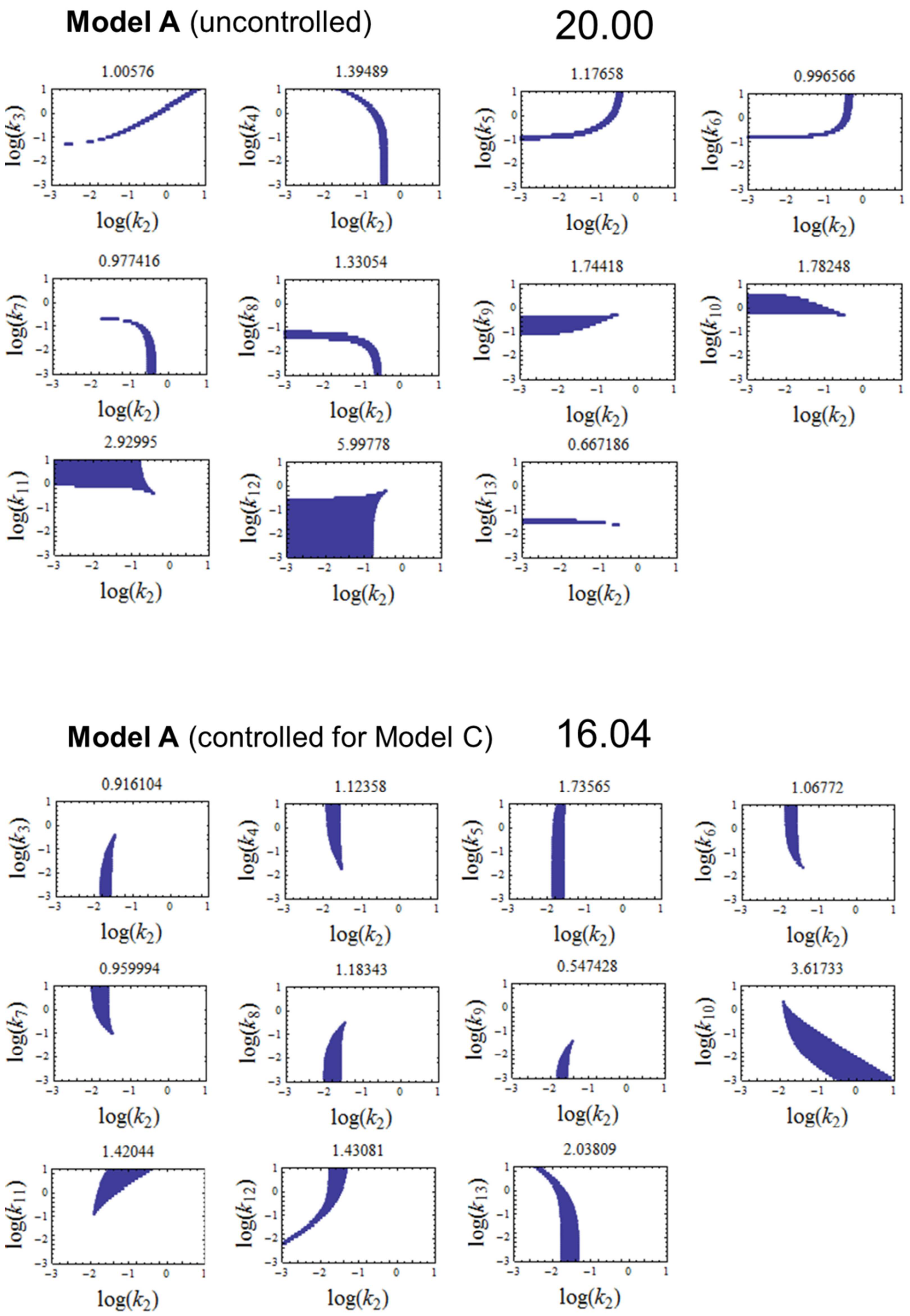


Figure S2 F. Systems with a monofunctional SK. Input signal = change in k_2

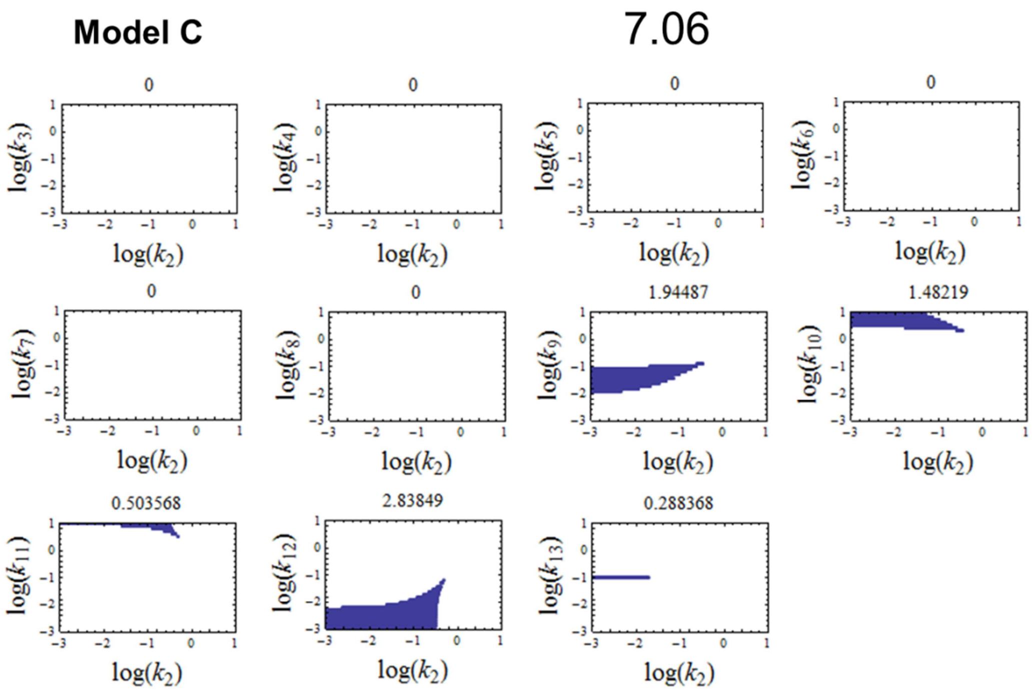


Figure S2 F (continued)

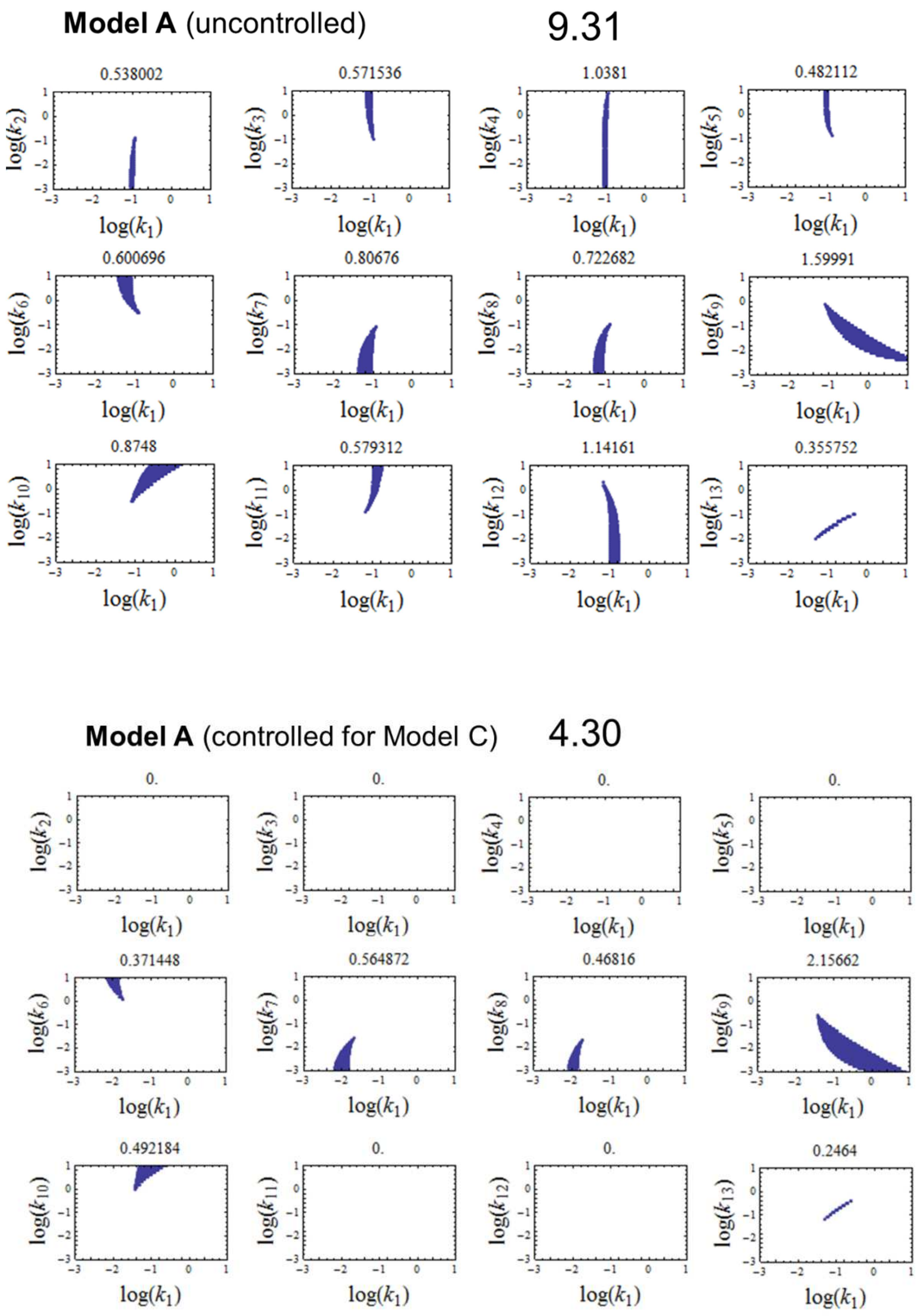


Figure S2 G. Systems with a bifunctional SK. Input signal = change in k_1

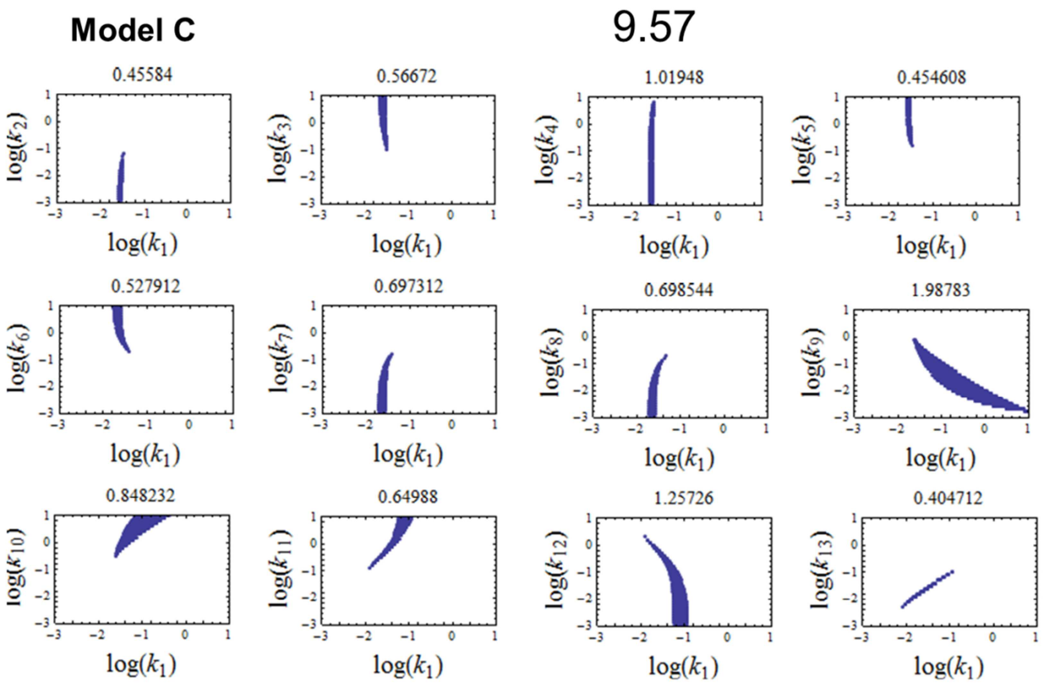


Figure S2 G (continued)

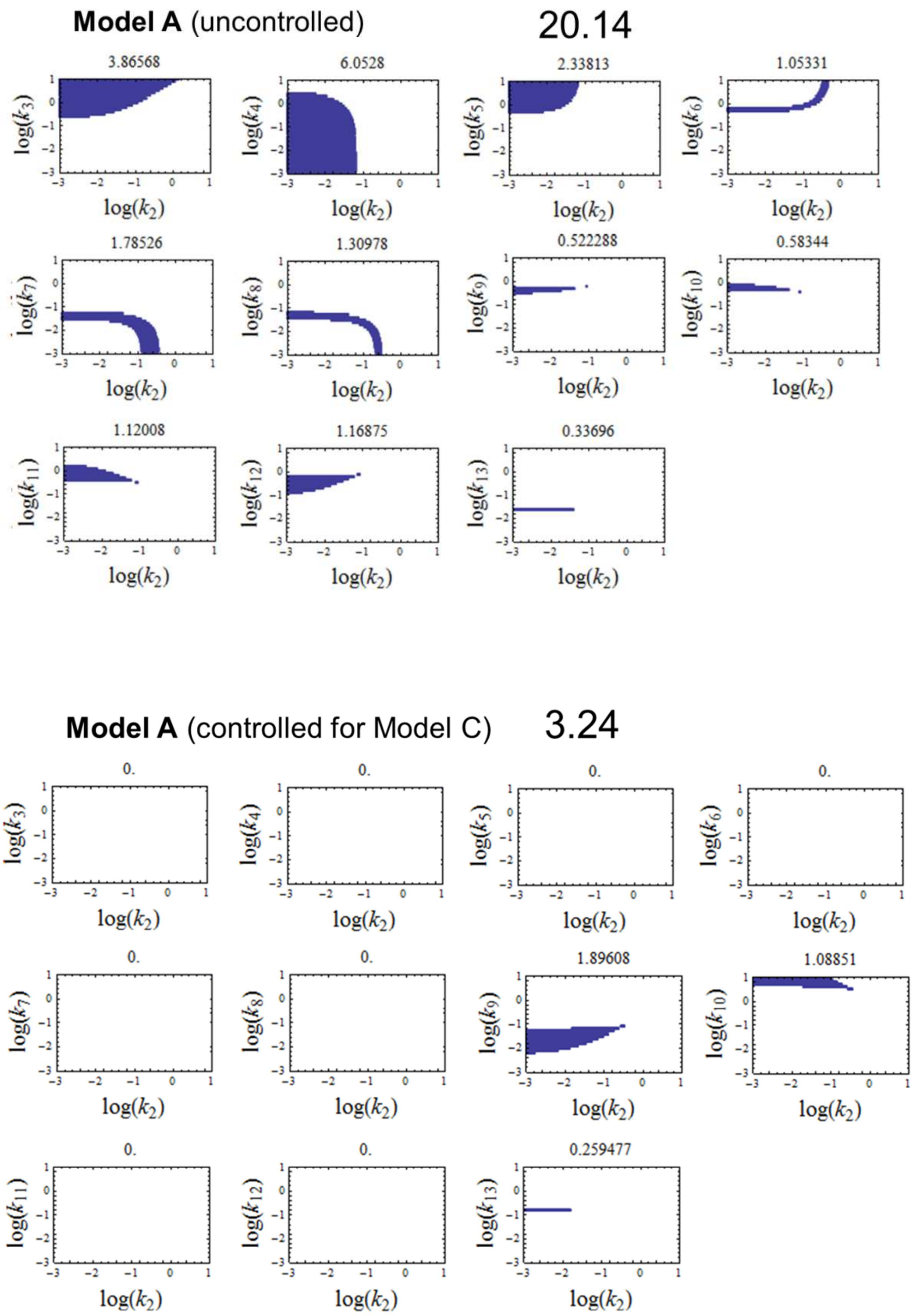


Figure S2 H. Systems with a bifunctional SK. Input signal = change in k_2

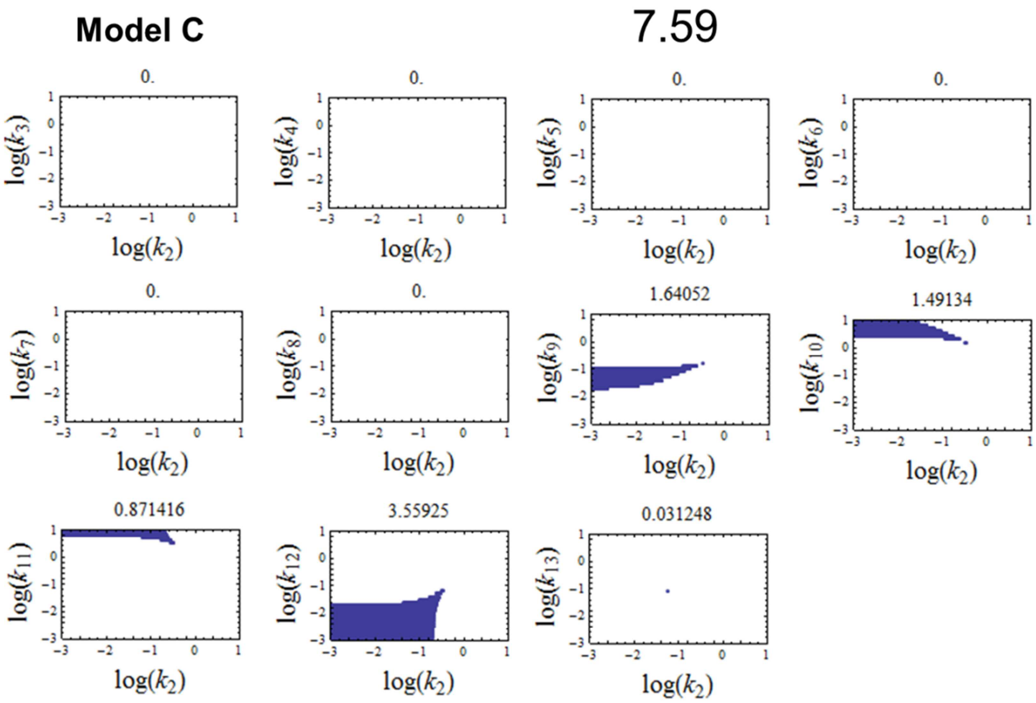


Figure S2 H (continued)

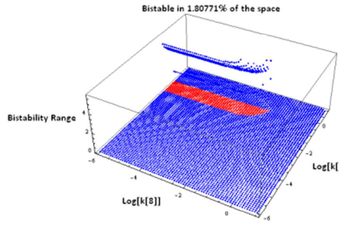
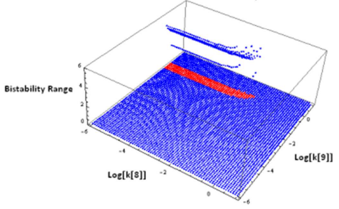
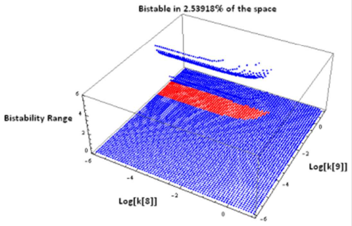
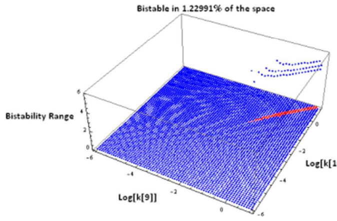
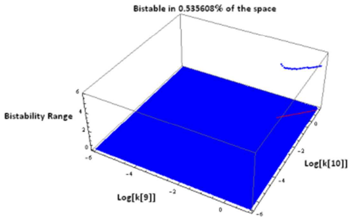
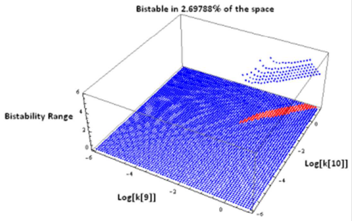
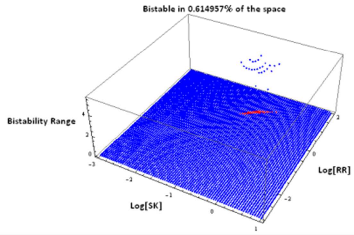
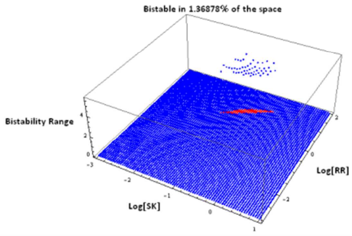
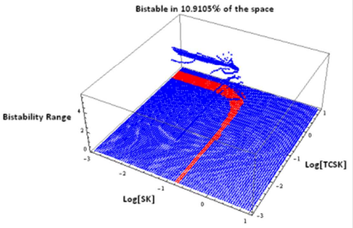
	Model A	Model A controlled for B	Model B
k8, k9	Bistable in 1.80771% of the space 	Bistable in 5.29657% of the space 	Bistable in 2.53918% of the space 
k9, k10	Bistable in 1.22991% of the space 	Bistable in 0.535608% of the space 	Bistable in 2.69788% of the space 
SK , RR	Bistable in 0.614957% of the space 	NON APPLICABLE	Bistable in 1.36878% of the space 
TCS, TC	NON APPLICABLE	NON APPLICABLE	Bistable in 10.9105% of the space 

Figure S3 A. Scanning for k_2, bifunctional TCS.

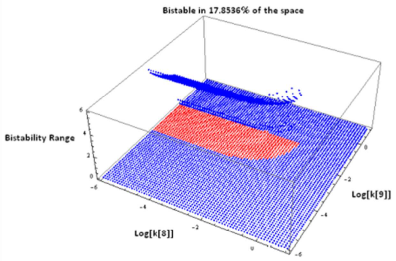
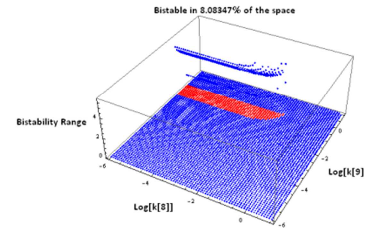
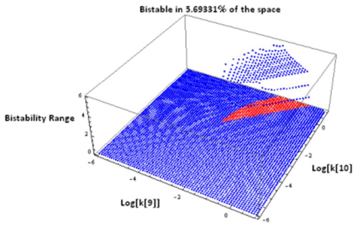
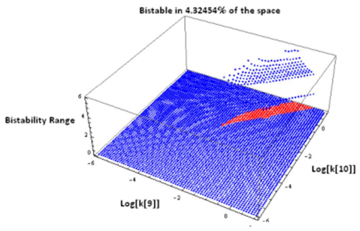
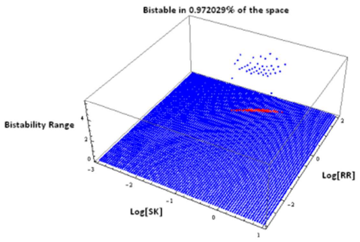
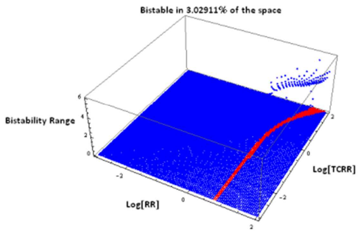
Model A controlled for C	Model C
 Bistable in 17.8536% of the space 	 Bistable in 8.08347% of the space
 Bistable in 5.69331% of the space 	 Bistable in 4.32454% of the space
NON APPLICABLE	 Bistable in 0.972029% of the space
NON APPLICABLE	 Bistable in 3.02911% of the space

Figure S3 A (continued)

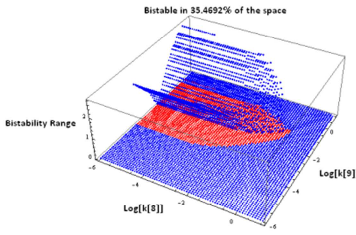
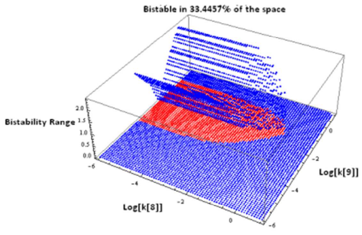
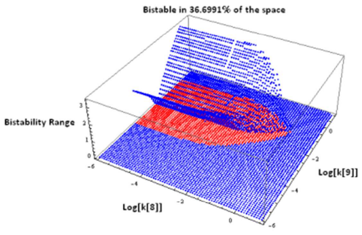
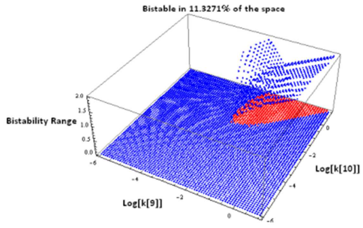
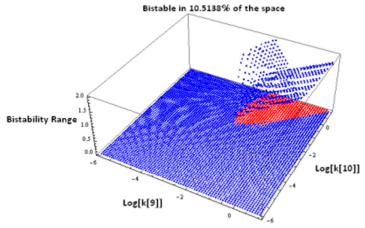
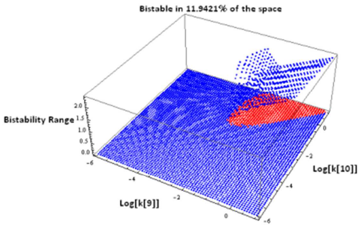
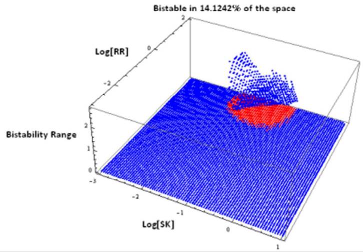
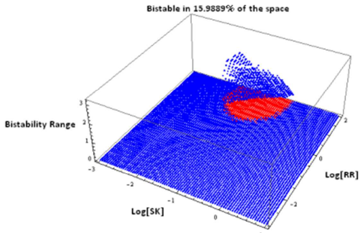
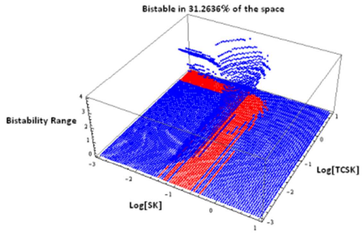
	Model A	Model A controlled for B	Model B
k8, k9			
k9, k10			
SK , RR		NON APPLICABLE	
TCS, TC	NON APPLICABLE	NON APPLICABLE	

Figure S3 B. Scanning for k_1, bifunctional TCS.

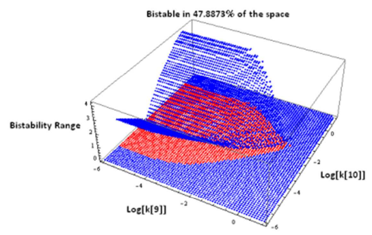
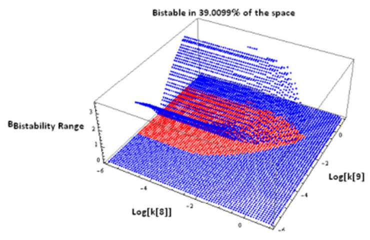
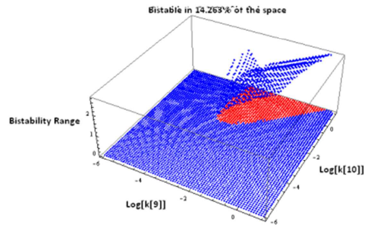
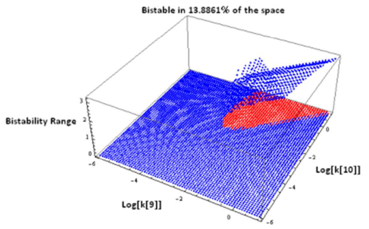
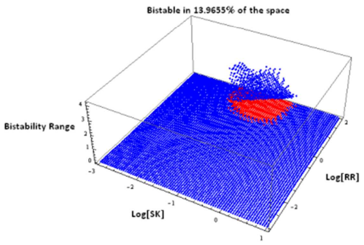
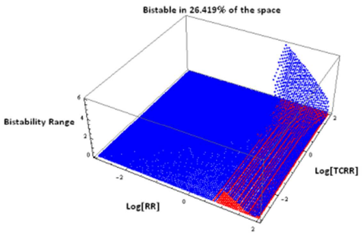
Model A controlled for C	Model C
Bistable in 47.8873% of the space 	Bistable in 39.0099% of the space 
Bistable in 14.263% of the space 	Bistable in 13.8861% of the space 
NON APPLICABLE	Bistable in 13.9655% of the space 
NON APPLICABLE	Bistable in 26.419% of the space 

Figure S3 B (continued)

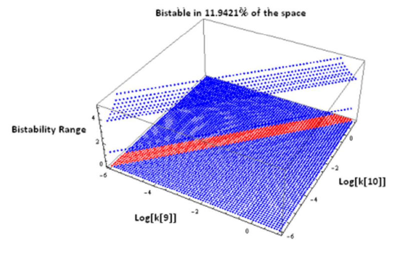
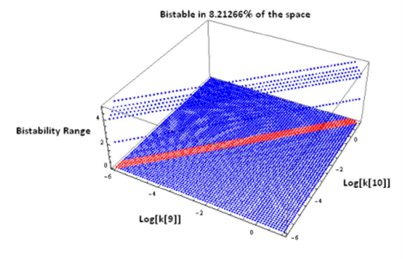
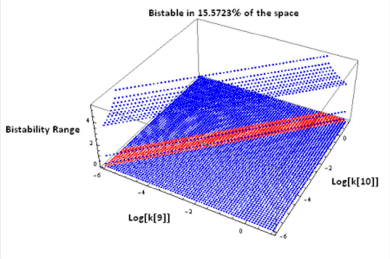
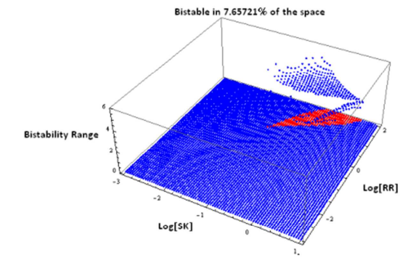
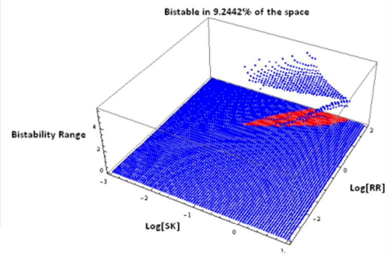
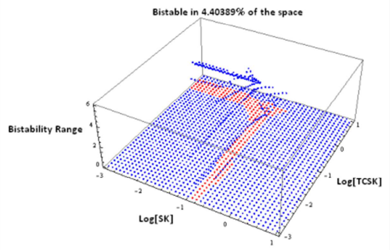
	Model A	Model A controlled for B	Model B
k9, k10	 Bistable in 11.9421% of the space	 Bistable in 8.21266% of the space	 Bistable in 15.5723% of the space
SK , RR	 Bistable in 7.65721% of the space	NON APPLICABLE	 Bistable in 9.2442% of the space
TCS, TC	NON APPLICABLE	NON APPLICABLE	 Bistable in 4.40389% of the space

Figure S3 C. Scanning for k_2, monofunctional TCS.

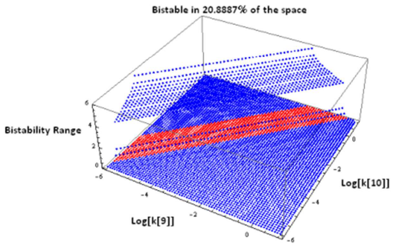
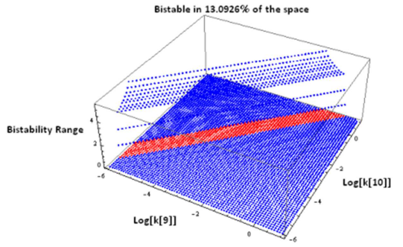
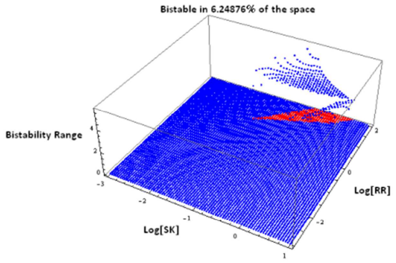
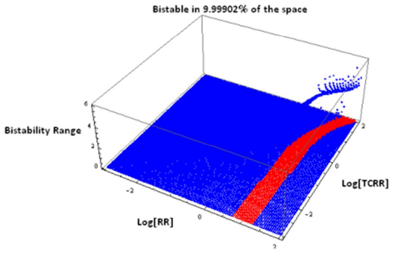
Model A controlled for C	Model C
 Bistable in 20.8887% of the space 	 Bistable in 13.0926% of the space
NON APPLICABLE	 Bistable in 6.24876% of the space
NON APPLICABLE	 Bistable in 9.99902% of the space

Figure S3 C (continued)

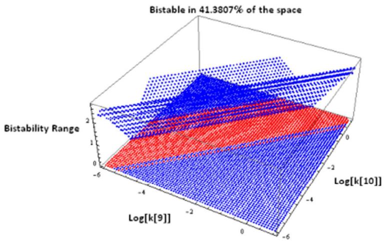
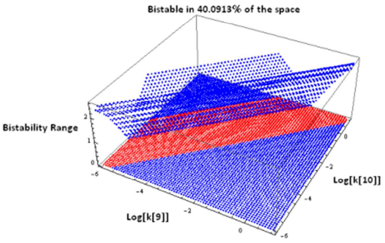
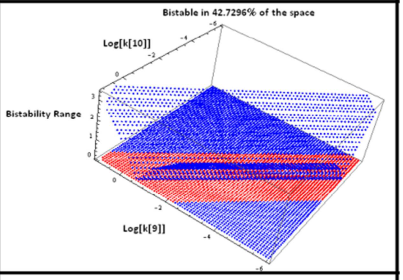
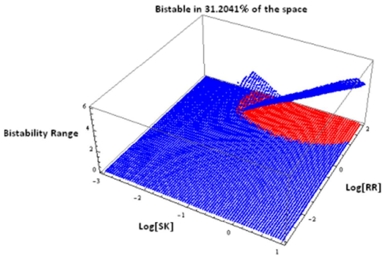
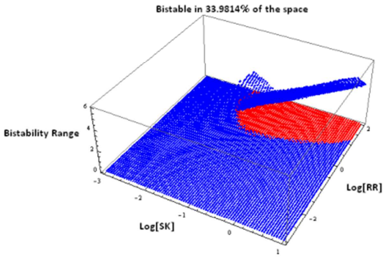
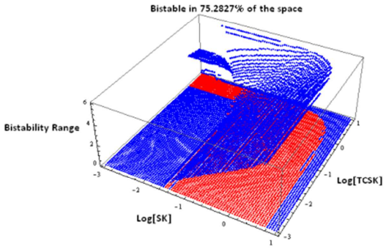
	Model A	Model A controlled for B	Model B
k9, k10	 Bistable in 41.3807% of the space	 Bistable in 40.0913% of the space	 Bistable in 42.7296% of the space
SK , RR	 Bistable in 31.2041% of the space	NON APPLICABLE	 Bistable in 33.9814% of the space
TCS, TC	NON APPLICABLE	NON APPLICABLE	 Bistable in 75.2827% of the space

Figure S3 D. Scanning for k_1, monofunctional TCS.

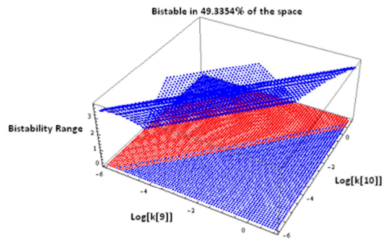
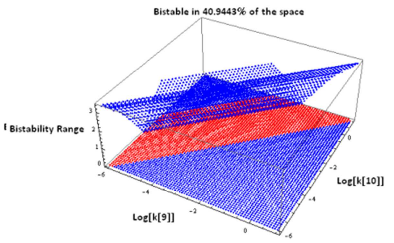
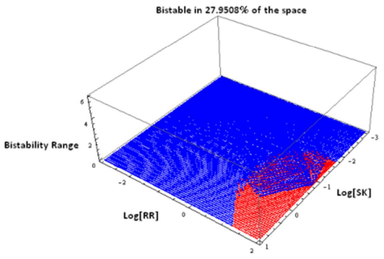
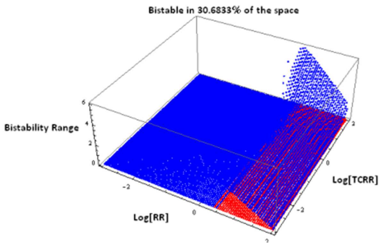
Model A controlled for C	Model C
	
NON APPLICABLE	
NON APPLICABLE	

Figure S3 D (continued)

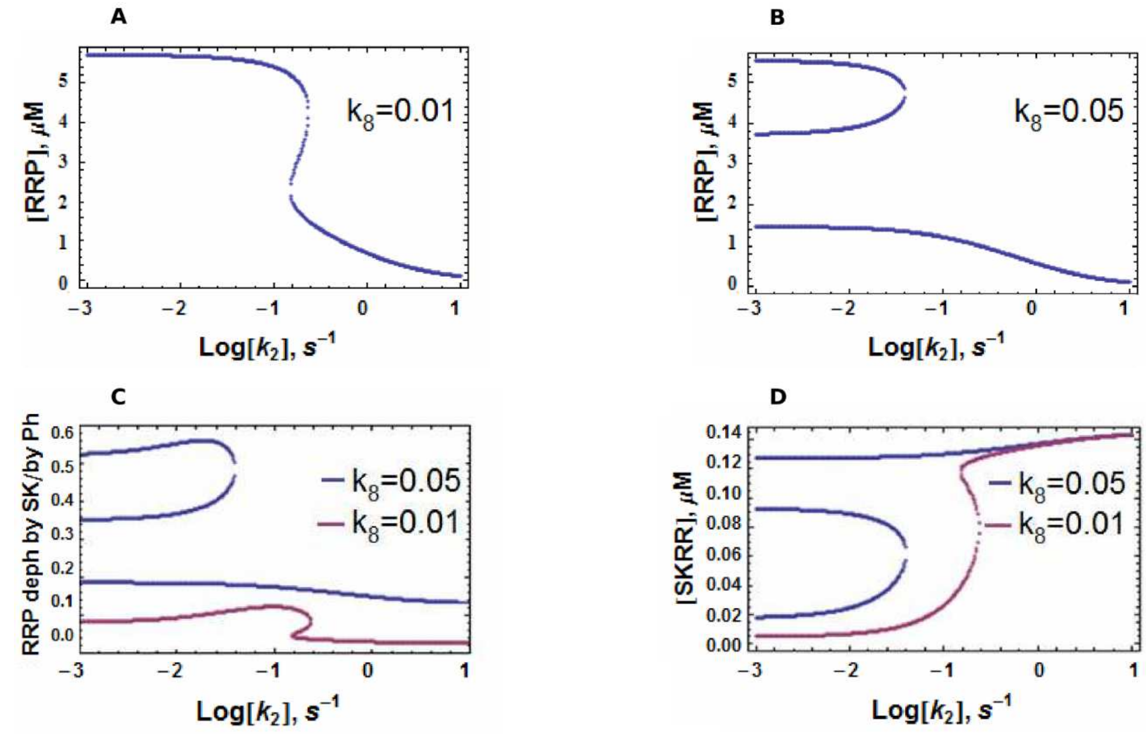

A
[RRP], μM
5
4
3
2
1
0
k_8=0.01
−3
−2
−1
0
1
Log[k_2], s^{-1}
B
[RRP], μM
5
4
3
2
1
0
k_8=0.05
−3
−2
−1
0
1
Log[k_2], s^{-1}
C
RRP deph by SK/by Ph
0.6
0.5
0.4
0.3
0.2
0.1
0.0
k_8=0.05
k_8=0.01
−3
−2
−1
0
1
Log[k_2], s^{-1}
D
[SKRR], μM
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
k_8=0.05
k_8=0.01
−3
−2
−1
0
1
Log[k_2], s^{-1}

Figure S4

4.7.2. Supplementary tables

Supplementary Table 1. Overall response times for the three systems modeled (uncontrolled comparison) [a].

	Modulation of SK autophosphorylation (k_1)		Modulation of SKP dephosphorylation (k_2)	
	OFF → ON	ON → OFF	OFF → ON	ON → OFF
Monofunctional				
Model A	3 011.49	1 143.31	15 515.40	27 816.30
Model B	3 406.48	1 337.95	9 467.02	24 801.00
Model C	3 125.05	1 091.73	57 574.80	43 048.20
Bifunctional				
Model A	3 346.30	1 378.56	9 336.50	20 907.90
Model B	3 672.27	1 739.08	8 695.38	10 672.20
Model C	3 358.06	1 195.35	57 212.80	40 114.40

[a] Results of the integral for the signal-response time function of Models A (uncontrolled), B and C. These values represent the area below each curve in Supplementary Figure 2, that is, the sum of the transient times for each response.

5 Discussion

5.1. Importance of studying design principles

Naturally evolved and artificially engineered networks[4] share some global properties, such as the “scale-free” and the “small world” properties, although they may be very different in size and the nature of their nodes [1]. In addition, certain patterns of interconnections, known as network motifs, are found in networks at much higher frequency than expected by chance [2, 3]. For example, one finds the same pattern of interconnections between the nodes (genes or proteins) of networks/circuits regulating transcription in different organisms, from bacteria and yeast, to plants and animals [4-6]. Some of those network motifs can also be overrepresented in other classes of networks. For instance, feedforward loops (X regulates Y, and Y regulates Z, which is also regulated by X, as shown in Figure 1a) are recurring regulation patterns present in both transcription and signal transduction networks.

Every molecular network that contains these and other motifs is responsible for a given type of function in the cell. Hence, it evolved under selective pressures to improve its performance. These networks seem to have converged into a restricted set of molecular solutions to the challenges imposed by their functional demands. In recent years, controlled experiments that test the effect of variations in motifs and other design elements of a circuit on its physiological response have identified circuit variants that maximize organismal fitness [7]. These experiments reinforce the idea that overrepresented patterns found in molecular networks may have become recurrent because they can provide functional advantages that are important for certain types of networks, even if we do not know what those advantages are to begin with. For example, signaling networks often contain a connectivity pattern known as

[4] We note that we will use the words **network** and **circuit** interchangeably throughout this discussion.

diamond (protein X regulates proteins Y_1 and Y_2, and both Y_1 and Y_2 regulate protein Z, as shown in Figure 1b), that is not found in transcription networks.

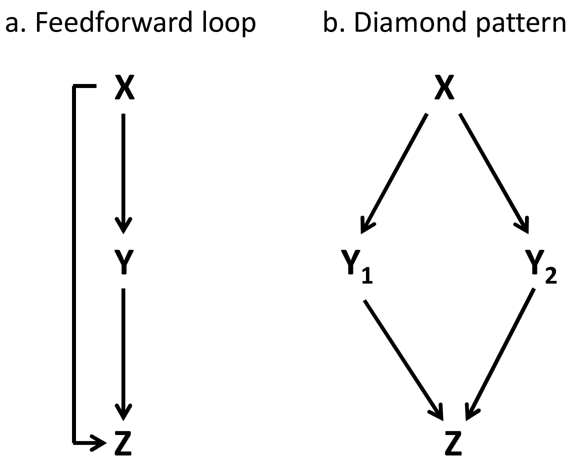


Figure 1. Two examples of patterns of node interconnections that are overrepresented in certain types of networks. X, Y and Z represent either proteins or genes, and arrows represent regulatory interactions between the nodes. The arrow points to the regulated molecule. Regulatory interactions can have either positive (activation) or negative (inhibition) sign. a) The feedforward loop is a recurrent motif found in both transcription and signaling networks. b) The diamond pattern is a network motif typical of signaling networks, but not of transcription networks.

Such variations between the topology of networks responsible for different functions (transcription regulation or signal transduction, for example) are probably due to differences in the timescale, spatial organization, and precision of the various types of responses that the different biological processes must orchestrate [2, 3]. For example, while a signal transduction cascade catalyzes a set of chemical reactions with response times on the range of seconds to minutes, gene expression adaptive responses require between minutes and hours to reach their peaks.

A recurrent feature of a network can also consist of a repeated quantitative relation between the system's components. For example, it has been determined that during adaptation of yeast to heat shock there are changes in gene expression that are quantitatively constrained to lead to an efficient response. When this well-defined program is activated, among other changes in gene expression, the hexose transporter genes are over-expressed by a factor of between 6 and 9, and the glucokinase gene is over-expressed at least by a factor of 4 [8].

All those qualitative or quantitative common patterns of molecular networks involved in certain type of cellular function appear as a result of physical and dynamical constraints imposed by the functional needs on the network's operation, and reflect the linkage between the network's architecture and its physiological behavior [9]. Globally, these repeated features seem to suggest the existence of biological design principles. Here, we define a biological design principle as a rule that explains the existence of a given biological feature based on the effect of that feature on the functional effectiveness of the network. Needless to say that the word design here is not related with any kind of intelligence or consciousness, but only with the pattern of interconnectivity between the network components. Among the daunting diversity observed in life, such a rule provides not only understanding about why a given feature is how it is, but it may also predict how that feature changes under changing conditions of the organism's milieu [10].

But how do patterns that fulfill biological design principles arise in natural ways? The variability of heritable biological traits is a random, unplanned event that arises through genetic processes such as mutation, recombination, and/or horizontal gene transfer. Subsequently, these features can either vanish or be passed to the offspring and become fixed in the population. Both situations can occur either by chance (especially in small effective populations) or as a result of natural selection acting on individuals and favoring those who carry traits that entail any advantage to the organism in that particular environment. If we can prove that natural selection enriches the networks observed in a population with respect to a given feature or pattern because that pattern leads to fitter organisms, then a design principle has been identified in nature. The functional rules that explain the improved fitness of an organism and justify the biological design principles can be described mathematically, independently of whether the design principle is general and applicable to various types of networks or specific and applicable only to a limited range of circuits.

Making sure that the repeated occurrence of a network feature is due to a design principle and to natural selection requires proving that they provide functional advantages to the network. In other words, we must prove that the feature is the best design among known alternatives to carry out the task required of the network. Such features are expected to also improve the fitness of the whole organism.

A circumstantial way to obtain such a proof starts with knowing the specific role of the network in the biological function it is involved in. This enables us to define the functional criteria that are required for the proper functionality of the network in that biological context. Then, mathematical models of alternative network designs can be built in order to compare their dynamical behavior. This strategy allows us to establish which alternative design has a better performance with respect to each functional criterion [11, 12].

If we find that the prevalent network's designs are the most efficient ones according to the defined functional criteria, this is consistent with a positive selection for those designs, given that they can be viewed as adaptive traits. However, if we observe that the prevalent designs are not the most efficient ones from the point of view of the functional criteria we chose, we can think that:

i) our functional criteria are not properly selected and there might exist other unidentified functional criteria that are more important in that biological context, or

ii) our hypothesis is wrong and the existence of that network's recurrent motif cannot be explained by a better functional performance. In this case, the trait in question might be the result of random processes such as genetic drift. It might also be present as a byproduct of the selection of some other correlated characteristic that increases global fitness of the organism in a Pareto optimal way [13].

We remark that in a living organism, some features can be associated with one another, and the selection of one of those features can lead to the occurrence of other (not necessarily adaptive) traits just because of the linkage between features [14]. This is known as a Saint Marcus spandrel. Saint Marcus spandrels can sometimes lead to the false positive identification of functional effectiveness criteria, as two or more of these criteria might be interdependent. As a consequence, highly correlated functional effectiveness criteria should be considered as one common functionality criterion, rather than as independent criteria to be tallied when the effectiveness of alternative designs is compared.

Some design principles are general principles independent of the network's specific function. One example of a global design principle, not restricted to a specific type of network, comes from the application of Reaction Network Theory. This theory proves that a mass-action network can only have more than one steady state if its species-reaction graph satisfy strict connectivity conditions that are described by what is called the Defficiency One Theorem [15]. More examples of global patterns that are observed in all types of molecular networks are: a positive feedback loop is a necessary condition for multistability and is functionally associated with switch-like behavior and a slower system's response [16, 17]; negative feedback loops reduce noise and speed up the network's response [18]; the more a feedback loop maximizes the correlation between input and output, the more the noise amplification [19].

Other design principles are system specific, closely related to the precise function of the molecular network. Section 2 of this thesis identifies several examples of system specific design principles characteristic of gene circuits, metabolic networks, cell cycle, and signal transduction networks. For instance, it has been reported that two-component signal transduction systems (TCS) mediating responses that require hysteresis must meet two conditions: a major flux channel for the response regulator (RR) dephosphorylation that is independent of the phosphatase activity of the unphosphorylated sensor kinase (SK), and the formation of a dead-end complex between the unphosphorylated forms of RR and SK [20]. Moreover, histidine kinase bifunctionality (SK catalizes both the phosphorylation of RR and the dephosphorylation of phosphorylated RR) minimizes crosstalk [21] and is necessary for input-output insensitivity to changes in the concentration of the system's components [22].

As stated above, the identification of design principles in biochemical systems could help making sense of the complexity observed in molecular systems, in the same way that the periodic table of the elements or knowing the properties of series of organic alcohols or acids allows making sense in the diversity of existing chemical substances.

Such a fundamental understanding of the molecular network's structure in terms of its function and evolution is enough to justify the importance of studying biological design principles. In addition, this knowledge can be applied to other fields of biological research and engineering: biomedical research can use those principles to identify new therapeutic strategies; and synthetic biology takes advantage of the knowledge derived from design principles to engineer new circuits within organisms which provide them with new physiological properties. This ability to tune organisms is a promising opportunity for multiple biotechnological applications such as bioremediation, production of substances or agriculture.

5.2. What have we accomplished?

In this work, after reviewing some of the methods for and results from the study of design principles in molecular systems, we have focused on the search of new design principles in signal transduction circuits, and more specifically in Two Component Systems and other histidine-aspartate Phosphorelays (TCS/PR). We analyzed all fully annotated organisms in the NIH genome database, studying the proteomic and genomic distribution of protein domains characteristic of TCS/PR cascades (SK, RR and HPt). As a result of this phylogenetic analysis, we confirm that, mostly, genes coding for proteins involved in TCS/PR cascades have a coordinated expression, but there are fundamental differences between prokaryotes and eukaryotes in the way in which they implement that gene expression coordination: prokaryotes tend to cluster functionally related genes in the genome forming operons, while eukaryotes tend to fuse the genes that should be coordinated in one single gene coding for a multidomain protein. Why this is so is a question to be answered in future investigations, although we suggest in the final discussion in section 3 that the reason could be related with signal amplification maximization in prokaryotes and noise minimization in eukaryotes.

The extensive census of genes coding for proteins involved in TCS/PR cascades we performed, in addition to identifying different strategies for their coordinated expression, allows deducing several network designs. Our ultimate aim is to use the data from this census to build a library of all alternative TCS/PR network topologies existing in nature. The alternative TCS/PR network designs collected in that library will be mathematically modeled and subjected to controlled comparisons in order to systematically identify the differences in their dynamic behavior. Then, we will try to correlate those dynamic differences with the specific functional demands each alternative network topology can more efficiently satisfy, with the purpose of finding new design principles in TCS/PR signaling pathways.

As a first step in this set of systematic comparisons between alternative TCS/PR network designs, we analyzed the effect of an auxiliary third protein in a canonical TCS. This third protein can either bind to the SK and inhibit its phosphorylation or bind to and stabilize the phosphorylated form of RR. The prototypical TCS (without any third component), as stated above, can show bistability if some conditions are met. Our mathematical simulations point out that a TCS with an RR binding third component has a smaller parameter space where a bistable response to signals is possible, when compared to a prototypical TCS. An SK binding third component also decreases the TCS range of bistability if SK is bifunctional. However, if SK is monofunctional, an SK binding third component increases the parametric range of bistability of the system, in comparison with the prototypical one. Bistability in the system's response to changes in the environmental inputs could be advantageous when a switch-like response to signals is required, as is often the case for virulent organisms which have to (irreversibly) choose between to different operational states. In contrast, bistability is a disadvantageous feature when the organism has to respond in a gradual, proportional way to changing intensities of an input signal, as is the case in the cell's adaptation to different environmental stresses (thermic, acidic or antibiotic stress).

This is only one example of how dynamical system properties can be modulated through changes in the network of interactions between the system components. Additional comparisons between the dynamical features of alternative TCS/PR network designs will provide an overall perspective of the physiological properties associated with each variation in the basic pattern of these biochemical signaling cascades. Then, perhaps this will give us some clues to understand why TCS/PRs are the main signaling pathway in prokaryotes, while signal transduction in eukaryotes is chiefly done through other molecular cascades such as the MAP kinase cascade.

5.3. How can our work be continued?

As a continuation of the present work, the following challenges we must face are:

- Complete an extensive library of the alternative variations in the design of TCS/PR circuits.
- Analyze the dynamical features of each alternative TCS/PR circuit, explaining how changes in circuit structure result in different system properties that correlate with their functional demands.
- Build a collection of design principles governing TCS/PR circuits, useful as a guide for engineering synthetic regulatory circuits and finding new therapeutic strategies.

- Compare the functional properties of TCS/PR circuits with those of other phosphorylation signaling cascades prevalent in eukaryotes, such as the MAP kinase cascades, so that we can explain why different signaling pathways are preferred in cells from different domains.

5.4. Possible future directions in design principles research

The biological design principles described in this thesis and, more generally, in the primary literature make a collection of isolated examples obtained through ad hoc strategies. If we want to promote the advance in the search of biological design principles, we should find a more systematic and large-scale enabling way to identify, organize, and classify them. Such a systematic identification and classification requires the definition of generic functional criteria, valid for all kind of molecular circuits, independently of their specific function.

Information theory could provide a suitable conceptual framework for the formulation of such general functional criteria, given that biological regulatory networks can be viewed as analog-to-electronic devices, in which a biochemical circuit design will be selected if it optimizes the correlation between the environmental signal and the system's output, while minimizing the effect of noise on the response. Such optimization allows cells to improve the reliability of their inferences about the state of their changing environment and improve the appropriateness of the adaptive responses that ultimately increase their fitness and probability of survival. This is a challenging task, considering that the environment fluctuates in a noisy way, and cells transduce those fluctuating signals through biochemical networks that are themselves stochastic and history dependent [23, 24]. Cells must discern the unknown stimulus from the result of their stochastic signal transduction mechanism, and choose the proper decision based on that uncertainty. In spite of this, cells are obviously able to thrive, either buffering noise or taking advantage of it for several biological activities such as generate phenotypic heterogeneity in the population [24-26]. The theoretical framework provided by information theory, sequential data processing and optimality arguments [27], along with the use of mathematical analytical techniques such as graph theory, sensitivity analysis, statistics and thermodynamic analysis could provide a way to systematize the way we study and understand how molecular circuits are shaped by evolution to allow cells transfer information from their environment to take the correct decisions. Such conceptual approach will contribute to develop a standard methodology for the systematic identification of biological design principles in all type of molecular networks.

5.5. References

1. Milo, R., et al., *Superfamilies of evolved and designed networks.* Science, 2004. **303**(5663): p. 1538-42.
2. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-7.
3. Alon, U., *Network motifs: theory and experimental approaches.* Nat Rev Genet, 2007. **8**(6): p. 450-61.
4. Eichenberger, P., et al., *The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis.* PLoS Biol, 2004. **2**(10): p. e328.
5. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
6. Odom, D.T., et al., *Control of pancreas and liver gene expression by HNF transcription factors.* Science, 2004. **303**(5662): p. 1378-81.
7. Bayer, T.S., et al., *Synthetic control of a fitness tradeoff in yeast nitrogen metabolism.* J Biol Eng, 2009. **3**: p. 1.
8. Vilaprinyo, E., R. Alves, and A. Sorribas, *Use of physiological constraints to identify quantitative design principles for gene expression in yeast adaptation to heat shock.* BMC Bioinformatics, 2006. **7**: p. 184.
9. Alon, U., ed. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. ed. C.a. Hall/CRC. 2006.
10. Savageau, M.A., *Design principles for elementary gene circuits: Elements, methods, and examples.* Chaos, 2001. **11**(1): p. 142-159.
11. Alves, R. and M.A. Savageau, *Extending the method of mathematically controlled comparison to include numerical comparisons.* Bioinformatics, 2000. **16**(9): p. 786-98.
12. Schwacke, J.H. and E.O. Voit, *Improved methods for the mathematically controlled comparison of biochemical systems.* Theor Biol Med Model, 2004. **1**: p. 1.
13. Shoval, O., et al., *Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space.* Science, 2012. **336**(6085): p. 1157-60.
14. Gould, S.J. and R.C. Lewontin, *The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme.* Proc R Soc Lond B Biol Sci, 1979. **205**(1161): p. 581-98.
15. Feinberg, M., *Chemical reaction network structure and the stability of complex isothermal reactors- II. Multiple steady states for networks of deficiency one.* Chemical Engineering Science, 1988. **43**(1): p. 1-25.
16. Veening, J.W., W.K. Smits, and O.P. Kuipers, *Bistability, epigenetics, and bet-hedging in bacteria.* Annu Rev Microbiol, 2008. **62**: p. 193-210.
17. Veening, J.W., et al., *Transient heterogeneity in extracellular protease production by Bacillus subtilis.* Mol Syst Biol, 2008. **4**: p. 184.
18. Nevozhay, D., et al., *Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(13): p. 5123-8.
19. Lestas, I., G. Vinnicombe, and J. Paulsson, *Fundamental limits on the suppression of molecular fluctuations.* Nature, 2010. **467**(7312): p. 174-8.
20. Igoshin, O.A., R. Alves, and M.A. Savageau, *Hysteretic and graded responses in bacterial two-component signal transduction.* Mol Microbiol, 2008. **68**(5): p. 1196-215.
21. Alves, R. and M.A. Savageau, *Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function.* Mol Microbiol, 2003. **48**(1): p. 25-51.

22. Shinar, G., et al., *Input output robustness in simple bacterial signaling systems.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 19931-5.
23. Elowitz, M.B., et al., *Stochastic gene expression in a single cell.* Science, 2002. **297**(5584): p. 1183-6.
24. Snijder, B. and L. Pelkmans, *Origins of regulated cell-to-cell variability.* Nat Rev Mol Cell Biol, 2011. **12**(2): p. 119-25.
25. Samoilov, M.S., G. Price, and A.P. Arkin, *From fluctuations to phenotypes: the physiology of noise.* Sci STKE, 2006. **2006**(366): p. re17.
26. Eldar, A. and M.B. Elowitz, *Functional roles for noise in genetic circuits.* Nature, 2010. **467**(7312): p. 167-73.
27. Bowsher, C.G. and P.S. Swain, *Environmental sensing, information transfer, and cellular decision-making.* Curr Opin Biotechnol, 2014. **28**: p. 149-55.

6 Conclusions

1. Numerous design principles have been identified so far for molecular networks. Some of them are system specific and closely related to the function of that molecular network, while others are independent of the network's specific function and represent constraints caused by the circuit structure upon its own dynamical behavior.
2. A way to systematically study and identify design principles in molecular circuits is still forthcoming.
3. TCS/PR proteins represent, on average, between 1 and 2% of a prokaryotic proteome (mean = 1.37%). Among the surveyed proteomes, Deltaproteobacteria is the group with the highest average percentage of TCS/PR proteins, while Tenericutes and Chlamydiae have the lowest percentage. In contrast, when a eukaryotic proteome contains TCS/PR proteins, they account for between 0.05 and 0.2% of the entire proteome (mean = 0.11%). These proteins are absent in animals.
4. Genes coding for TCS/PR proteins involved in the same pathway have a coordinated expression tens to hundreds of times more frequently than expected by chance. We derive this conclusion from the genomic and proteomic organization of HK, RR and HPt protein domains, which are found clustered either forming operons or multidomain proteins with a frequency much higher than the expected frequency if gene order and gene fusion events were random.
5. Prokaryotes and eukaryotes differ in the way in which they organize the protein domains responsible for internal signal transduction in TCS/PR cascades: prokaryotes tend to cluster functionally related genes in the genome forming operons tens to hundreds of times more often than expected by chance, while eukaryotes tend to fuse the genes that should be coordinated in one single gene coding for a multidomain protein.

6. We find 530 unique circuit designs for TCS and PR cascades, based on our analysis of TCS/PR operon composition.
7. We find 50 different combinations of HK, RR and HPt domains that can occur in a single polypeptide chain in the 7609 surveyed proteomes. RR, HK, HKRR, HKRRHPt, HKHPt, HPt, RR_1RR_2, $HKRR_1RR_2$, are the most abundant protein types, sorted by abundance. The number of HK and RR gene fusion events increases with the number of HK and RR domains in the genome, which is consistent with a positive selection for fused HKRR proteins.
8. The number of TCS/PR proteins in a proteome increases with the total number of proteins in that proteome. This relationship between number of TCS/PR proteins and proteome size is significantly different between prokaryotes and eukaryotes. R^2 of our linear model is 0.21 for prokaryotes and 0.49 for eukaryotes.
9. The presence in a TCS of an HK binding third component which prevents HK phosphorylation increases the parameter space where a bistable response of the TCS module is possible, when the HK is monofunctional, but decreases it if the HK is bifunctional.
10. The presence in a TCS of an RR binding third component which protects phosphorylated RR from dephosphorylation decreases the parameter space where a bistable response of the TCS module to signals is possible.