Structural comparison of biological networks based on dominant vertices†

Cite this: *Mol. BioSyst.*, 2013, **9**, 1765

Beatriz Luna,^{*a} Edgardo Galán-Vásquez,^b Edgardo Ugalde^c and Agustino Martínez-Antonio^{*b}

It is a current practice to organize biological data in a network structure where vertices represent biological components and arrows represent their interactions. A great diversity of graph theoretical notions, such as clustering coefficient, network motifs, centrality, degree distribution, etc., have been developed in order to characterize the structure of these networks. However, none of the existent characterizations allow us to determine global similarity among networks of different sizes. It is the aim of the present paper to introduce a mathematical tool to compare networks not only with regard to their topological structure, but also in their dynamical capabilities. For this reason we aim to propose a pseudo-distance between networks, built around the notions of determination and dominance, concepts recently introduced in the context of regulatory dynamics on networks. We use our proposed pseudo-distance to compare networks from the following bacteria: *E. coli*, *B. subtilis*, *P. aeruginosa*, *M. tuberculosis*, *S. aureus* and *C. glutamicum*. We also use this pseudo-distance to compare these real bacterial networks with equivalent homogeneous, scale-free and geometric three dimensional random networks. We found that even when bacterial networks are characterized with different levels of detail, have different sizes and represent different aspects of the organisms, the proposed pseudo-distance captures all these characteristics, and indicates how similar they are or not from random networks.

Received 21st February 2013,
Accepted 9th May 2013

DOI: 10.1039/c3mb70077a

www.rsc.org/molecularbiosystems

1 Introduction

We consider gene expression as the process by which genes coded in DNA are transcribed into RNA, then, this messenger RNA is used by the translation machinery to synthesize proteins. Proteins are necessary in several functions performed by organisms. Frequently, the protein produced by a certain gene is responsible for the activation or inhibition of another gene; that is to say, it promotes (enhances) or suppresses (restricts) its expression. The overall gene interactions architecture is encoded in a network where vertices symbolize genes and edges represent their regulatory interactions. Therefore, regulatory networks represent a complex set of highly interconnected processes that govern the rate at which the genes in a cell are expressed.¹ Network representation allows working with a large quantity

of data, which can be quantified and be analyzed by means of the tools supplied by graph theory. Indicators like degree distribution, clustering coefficient, cycles distribution, and network motifs^{1–3} have been used to characterize and understand the topological complexity of a given network. In order to understand the relation between topology and function it is essential to compare the structure of networks performing similar functions. However, the lack of information and the size differences between networks with similar dynamical behavior makes this type of global comparison difficult. Recent studies have dealt with this problem proposing different approaches, for instance:

- Comparison of the interconnections of different sub-graphs into each network.⁴
- Identification of interacting patterns of regulation and genetic sequence similarities between nodes.⁵
- Recognition of modular network components, followed by projection of these modules onto networks of other species.⁶
- Quantification of the structural differences between two networks using the normalized graph spectra.⁷

On the other hand, comparative analyses have been also proposed for metabolic pathways considering:

- The pathway topology and the number of enzymes they contain.⁸

^a Scientific Computing and Applied Mathematics Laboratory, Universidad del Papaloapan, Av. Ferrocarril s/n 68400 Loma Bonita, Oaxaca, México. E-mail: bcluna@unpa.edu.mx

^b Departamento de Ingeniería Genética, Cinvestav, Km. 9.6 Libramiento Norte Carr. Irapuato León, México. E-mail: amartinez@ira.cinvestav.mx

^c Instituto de Física, Av. Manuel Nava 6, zona Universitaria, 78290 San Luis Potosí, SLP, México

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70077a

- The phylogenetic tree built from the structural information inherent in the metabolic pathways.⁹
- The compression of vertices and edges.¹⁰

More recently a technique based on the determination of rewiring rates has been proposed to compare genetic interacting networks, metabolic pathway networks, transcription, phosphorylation and social coauthorship networks.¹¹

The motivation behind this study is twofold: first, to provide an alternative way to compare available regulatory networks; second, to understand the information that comparisons can yield. Regarding the first point, until recently only the *E. coli* network had been reconstructed from the literature, however, new high-throughput methodologies allow the massive reconstruction of regulatory networks from bacteria and other organisms. Focusing on bacteria, under supposition that they have a common origin, one way to compare them might be based on phylogenetic analysis, which depends on orthologous genes, however, this approach is limited, since the phylogenetic divergency increases as the bacterial genus. Therefore it is necessary to appeal to methods developed from systems biology theory. From this perspective, we focus on the analysis of the elements of the regulatory genome, the transcription factors, particularly on those crucial elements for regulation. Regarding the second point, this analysis allows us to discern between real biological networks and random networks, but most important, it allows us to decipher the salient elements/arrangements on biological networks associated with interesting traits: *i.e.*, pathogenic *vs.* non-pathogenic, adaptation to different life conditions and intracellular *vs.* free/living styles, among others.

We propose a global measure, based on the concepts of determination and dominance, which have been recently introduced and explored. As shown by Luna and Ugalde,¹² the global state of the network can be reconstructed from the history of a sub-collection of vertices they call dominant vertices through a series of steps following a hierarchical organization of the network. In the same study, dominant vertices and vertex determination were used as indicators of the structural complexity of the network. Here, we define a pseudo-distance between networks by using the hierarchy imposed by the dominance and vertex determination. We use this pseudo-distance to compare bacterial regulatory networks and additionally to compare these real biological networks with random network models, as performed in ref. 10 and 13.

In the next section we present some preliminary concepts and the bacteria whose transcriptional networks will be examined. In Section 3 we give the definition of our main analytical tool and we describe the computational method that was applied to the six bacterial regulatory networks and equivalent random networks. Section 4 contains results and discussion. Finally, Section 5 is devoted to the final remarks and conclusions.

2 Preliminaries

2.1 Dominant vertices

The dynamics of regulatory networks is usually modeled by means of piecewise-contractive dynamical systems.^{14,15} For these kind of

dynamical models, it is possible to distinguish a set of vertices in a network that allows recovery of the global configuration of the network based just on the history of those vertices. Conversely, one can determine the future configurations the network will have by imposing the successive states of the systems on the dominant vertices. Every nonempty directed network admits a dominant set. The rigorous definition of the dominant vertices and the statement of their properties, were widely introduced by Luna and Ugalde.¹²

Here we use the word network as a synonym of finite directed graph. It consists of a finite set of vertices, denoted by V , and a finite collection of ordered pairs of vertices, which we identify with arrows and denote by A . We only consider simple networks, which means that given two vertices $u, v \in V$ there exists at most an arrow going from u to v , which corresponds to the ordered pair (u, v) . The input set of a vertex v is the collection $I(v)$ of all vertices included in an arrow ending at v , *i.e.*, $I(v) := \{u \in V : (u, v) \in A\}$. The input degree of $v \in V$ is simply the cardinality of $I(v)$. Similarly, the output set of v is the set $O(v)$ of all vertices included in an arrow starting at v , *i.e.*, $O(v) := \{u \in V : (v, u) \in A\}$. The output degree of $v \in V$ is nothing but the cardinality of $O(v)$.

A regulatory network is a dynamical system involving several interconnected units, whose dynamic rule is determined by the interconnected topology. It is therefore a dynamical system over a network. The interacting units (which in our case correspond to gene products) have a numerical value (which in our case corresponds to its relative concentration) at each time step. The time can be considered continuous, in which case the system is modeled by a system of piecewise affine differential equations, or can be considered discrete, in which case the pertinent model is a piecewise affine coupled map network. To fix these ideas, let us take the discrete time case. Here the global configuration of the system is recorded at times $t_0 < t_1 < \dots < t_n < \dots$, which by simplicity, we identify with the integer numbers progression. After the enumeration of the interacting units, $(v_1, v_2, \dots, v_N)$, the configuration of the system at any time t can be encoded in an N -dimensional array $(x_1^t, x_2^t, \dots, x_N^t)$, where N is just the number of interacting units, and $x_i^t \in \mathbb{R}$ denotes the level of activity of the i -th unit at time t . The dynamic rule governing the state of the system $(x_1^{t+1}, x_2^{t+1}, \dots, x_N^{t+1})$ at time $t + 1$, from its configuration at time t , reads, for vertex v_j , as

$$x_j^{t+1} = \alpha_j x_j^t + D_j(x_k^t : v_k \in I(v_j)), \quad (1)$$

where $\alpha_j \in (0, 1)$ accounts for the degradation rate of v_j , while D_j is a function which depends on the activity levels of the vertices in the set of inputs of the referred vertex. The function D_j takes a finite number of values, which can be related to the levels of activation affecting the vertex v_j . This function is analogous to the so-called regulatory phrase of Boolean and logical networks,¹⁶ this analogy allows to obtain similar results from these models. The dynamics so defined is such that the asymptotic activity level of the unit v_j is completely determined by the sequence of activating levels affecting this unit in the course of time. This characteristic allows the reconstruction of

the asymptotic dynamics of the whole system from solely the history of a well-chosen collection of vertices. The reconstruction goes as follows: let $V_0 \subset V$ be the collection of dominant vertices, and suppose that the activity level x_k^t is known for all the course of time t at every one of the dominant vertices $v_k \in V_0$. Now, consider the set $\partial V_0 := \{v \in V: I(v) \subset V_0\}$, i.e., the collection of vertices whose input set is a sub-collection of the dominant vertices. The dynamics of the system is such that, for each $v_k \in \partial V_0$ and all time τ sufficiently large, x_k^τ depends only on the sequence $(x_i^\tau: v_i \in V_0)_{i=0}^{\tau-1}$, i.e., x_k^τ can be determined, up to a negligible error, by the history of the dominant vertices up to time $\tau - 1$. Now, the history of the interacting units in the set $V_1 = V_0 \cup \partial V_0$ containing the dominant vertices and vertices determined by them, determines the state of vertices in the set $\partial V_1 := \{v \in V: I(v) \subset V_1\}$ and defines an enlarged set $V_2 = V_1 \cup \partial V_1$. We can continue this enlargement procedure, based on vertex determination, until no new vertices can be added. In order for V_0 to be dominant it is necessary that at the end of this iterative process all the vertices from the network are included, i.e., if there exists a nested sequence of vertex sets, $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, such that $V_{i+1} = V_i \cup \partial V_i$, $i = 1, \dots, d$. Here,

$$\partial V_i := \{v \in V: I(v) \subset V_i\} \quad (2)$$

for each $V_i \subset V$. The number d of steps needed to determine all the vertices from the dominant seed is the depth of the network with respect to the chosen dominant set V_0 , here referred to as hierarchical levels or dominance levels. It is worth noting that a given network can have several sets of dominant vertices. A trivial choice is $V_0 = V$. Obviously, a valid choice would be a dominant set of minimal cardinality. Luna and Ugalde¹² proposed an algorithm aimed to identify dominant sets of minimal cardinality, as well as the hierarchy of sets it determines. In what follows we will make extensive use of that algorithm. It is important to notice that this paper is dedicated to a structural comparison, for that reason we do not make use of the dynamical properties of the dominant set, however given its dynamical characteristics, already proved, it is possible to perform a dynamical comparative analysis.

2.2 Connected components

A network (V, A) is said to be simply connected if any couple of vertices $u, v \in V$ can be joined through an undirected path, i.e., if there are vertices $u \equiv u_1, u_2, \dots, u_\ell \equiv v \in V$ such that for each $1 \leq i \leq \ell - 1$, either (u_i, u_{i+1}) or (u_{i+1}, u_i) belong to the arrow set A . If (V, A) satisfies the stronger condition requiring that any two vertices can be joined through a directed path, then the network is said to be strongly connected. It is not difficult to prove that any network can be decomposed in a unique way into a collection of connected sub-networks called connected components (CC). This concept will be mentioned later.

2.3 Random networks

Formally speaking, a random network is a probability distribution over a specific set of networks, for instance, the set of all networks on N vertices, or the set of all finite networks, or the set of all the infinite networks with bounded input and

output degrees, etc. These objects were introduced by P. Erdős and A. Rényi.¹⁷ These authors were interested in the connectivity properties of random homogeneous graphs. Roughly speaking, they considered distributions over the set of networks on N vertices, giving the same or almost the same probability to all the simple symmetric networks having a fixed number of input–output degrees.

A class of random networks, intended as models of complex evolving structures was introduced by A. Barabási and R. Albert.² A random process that incorporates continuous growth and preferential attachment usually defines these structures, known as scale free networks. There are several ways to implement this kind of construction, leading to the similar statistical properties.¹⁸ One can for instance start with the complete undirected graph on a small number of vertices, then, at each time step, add a new vertex forming new edges (bidirectional arrows) with randomly chosen vertices in the preexisting network. In order to ensure the preferential attachment, a vertex is chosen to form a new edge with a probability proportional to its degree. The process would continue until a desired number of vertices are obtained.

More recently, random geometric networks were introduced as models of biological networks.^{4,19} We use the three dimensional model, where every vertex of a set is randomly placed at a unit cube, and vertices are connected if the Euclidean distance between them is less than a parameter r .

2.4 Description of the transcriptional networks here studied

In this study we consider six transcriptional regulatory networks pertaining to six bacteria: *E. coli*, whose information was obtained from RegulonDB;²⁰ *B. subtilis*, obtained from DBTBS;²¹ *P. aeruginosa*;²² *M. tuberculosis*;²³ *C. glutamicum* obtained from the CoryneRegNet database;²⁴ and *S. aureus*.²⁵ Below we briefly describe the principal attributes for each of these networks, also shown in Table 1.

- *Escherichia coli* (EC) is a Gram-negative bacterium and the bacterial model of study par excellence. The genome sequence of strain K-12 MG1655 was published in 1997; it has 4.6 Million base pairs (Mbp) and contains 4400 predicted genes.²⁶ The predicted transcription factors (TFs) are around 300, and 7 sigma factors.²⁷ The strain of reference is a benign commensal in the gut of mammals but some other strains could be pathogenic.

- *Bacillus subtilis* (BS) is the model of a Gram-positive bacterium. It is well known for its capability to produce, from vegetative growing cells, metabolically inactive spores.²⁸ The distinctive

Table 1 Components of the transcriptional regulatory networks of bacteria here analyzed

Bacteria	Genes on the network	Interactions	TF genes	Target genes	σ factors	CC
EC	1565	3530	176	1382	7	18
BS	1659	3121	106	1531	16	9
PA	690	1020	76	593	14	12
MT	1624	3212	83	1531	10	6
CG	574	806	71	499	4	27
SA	558	671	46	512	0	21

strain is the so named 168, its genome has 4.2 Mbp and contains 4422 predicted genes.²⁹ The estimated number of TFs is around 275, and the number of sigma factors is about 14.³⁰ This bacterium normally inhabits soil.

- *Pseudomonas aeruginosa* (PA) is a metabolically versatile Gram-negative bacterium. It expresses a wide range of virulence factors that allow it to be an opportunist pathogen of plants and animals.³¹ The genome sequence of strain PAO1 has 6.2 Mbp that contains 5570 predicted genes.³² The estimated number of TFs is around 475, and the number of sigma factors is about 26.²⁷

- *Mycobacterium tuberculosis* (MT) strain H37Rv is a pathogenic bacteria of the genus *Mycobacterium* and the causative agent of most cases of tuberculosis in humans, one of the most infectious diseases.³³ Its genome sequence was released in 1998, it consists of 4.41 Mbp and contains approximately 4047 predicted genes. The number of predicted regulatory proteins is approximately 190 TFs and 13 sigma factors.³⁴

- *Corynebacterium glutamicum* (CG) strain ATCC 13032 is a Gram-positive bacterium, and it is well known as the organism for the commercial production of monosodium glutamate. Its genome was published in 2003, it consists of 3.3 Mbp and contains 3002 predicted genes.³⁵ The number of predicted regulatory proteins is 182 TFs and 7 sigma factors.³⁶

- *Staphylococcus aureus* (SA) strain N315 is a Gram-positive bacterium and it is an important human pathogen. It has been intensively studied in virulence, biofilm formation and central metabolism. Its genome sequence was published in 2001, it consists of 2.8 Mbp and contains 2593 predicted genes.³⁷ The number of predicted regulatory proteins is 120 for TFs and 4 for sigma factors.³⁸

3 Methods

We began our analyses of bacterial networks by using the algorithm mentioned in ref. 12 (which we briefly describe below), then we computed the sets of dominant vertices of minimal cardinality, and we determined the associated hierarchical levels of these sets for each one of the six bacterial networks mentioned above. After that, we used the pseudo-metric M , defined below, to compare these networks.

In the second part of the study, we generated 100 homogeneous random networks, 100 scale free networks and 100 geometric random graphs with the same number of nodes and on average the same number of arrows of every one of the six bacterial networks considered. By using this random sample we estimated the mean M -distance for each bacterial network, as well as the variance of this distance. The results of these numerical experiments are presented in Section 4.

3.1 Computation of dominant sets

We determine dominant sets by using the following algorithm: first, we select all the vertices with input degree equal to zero, together with the vertices with the highest output-degree/input-degree ratio. Then we compute the hierarchy of sets of vertices so determined by this collection. If the hierarchy does not

include all the vertices, then we restart the iteration by adding vertices to the starting dominant set until all the vertices are included in the hierarchy (still considering the highest output/input-degree ratio criterion). At the end of this process some dominant vertices are eliminated in order to obtain a dominant set of minimal cardinality; elimination could be performed by different methods, for instance randomly or by ordering the vertices in the dominant set. It is recommended to apply the same algorithm to every set of networks to compare. In our case we chose the ordering criterion. In this way, we obtain a nested sequence of sets $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, such that $V_{\ell+1} = V_\ell \cup \partial V_\ell$, $\ell = 1, \dots, d$.¹²

3.2 Hierarchy obtained by dominant vertices

The nested sequence of sets $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, defines a hierarchy of V_0 . What we call the dominant set of vertices corresponds to the first level of the hierarchy. Then for each $1 \leq \ell \leq d$, the determined set ∂V_ℓ constitutes the $\ell + 1$ -th level of the hierarchy. The depth of this hierarchy corresponds to the number d of levels needed to reach all the vertices in the network, starting from the dominant set.

In order to illustrate these notions, consider the network depicted in Fig. 1. Here, a dominant set of minimal cardinality is $V_0 = V_1 = \{1, 4\}$, and the hierarchy determined for that set is the following: $\partial V_1 = \{2\}$, $V_2 = \{1, 4, 2\}$, $\partial V_2 = \{3, 5\}$, $V_3 = \{1, 2, 3, 4, 5\} = V$. The depth in this network is $d = 2$.

3.3 A pseudo-distance based on dominancy

The hierarchical decomposition of networks obtained from dominant sets of minimal cardinality provides us with a tool to compare networks of different sizes and topologies, but sharing global dynamical features, as follows. Given two networks $G := (V, A)$ and $G' := (V', A')$ with hierarchical decomposition $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, and $V'_0 \equiv V'_1 \subset V'_2 \subset \dots \subset V'_d \subset V'_{d+1} \equiv V'$, define

$$m(G, G') = \left| \frac{\#V_0}{\#V} - \frac{\#V'_0}{\#V'} \right| + \sum_{\ell=1}^D \left| \frac{\#\partial V_\ell}{\#V} - \frac{\#\partial V'_\ell}{\#V'} \right| \quad (3)$$

where $D := \max(d, d')$ and $\#S$ denotes the cardinality of any set S . It is worth noting that m depends on the chosen dominant set of vertices and that in general this choice is not unique. Nevertheless, for dominant sets of minimal cardinality the characteristics of the associated decomposition are very similar. In our case, since our algorithm chooses a particular dominant set of minimal cardinality, the pair function m is well defined.

Fig. 1 Illustrative network. Its hierarchy is shown according to dominant vertices.

It is symmetric, non-negative, and satisfies the triangle inequality. It can vanish for non-identical networks, and for this reason is not a distance but a pseudo-distance. It is not difficult to exhibit networks $G \neq G'$ for which $m(G, G') = 0$. In this case the networks G and G' have decompositions with the same number of levels and the same proportion of vertices at each level. This pseudo-distance allows for comparing networks regardless the respective number of vertices, arrows and connected components. It gives a global idea of the network structure.

A refinement of the pseudo-distance m is the following

$$M(G, G') = \frac{1}{2} \left(m(G, G') + \left| \frac{d+1}{\#V} - \frac{d'+1}{\#V'} \right| \right) \quad (4)$$

which weights in the pseudo-distance m as the relative number of levels in each decomposition. Clearly $M(G, G') = 0$ implies $m(G, G') = 0$, and whenever $m(G, G') > 0$, the pseudo-distance M magnifies the effect of networks having relatively many levels.

3.4 Random networks

As mentioned above, besides the comparison between the six bacterial networks, we also compare each one of them to random networks with the same number of vertices and approximately the same number of arrows. We proceed as follows:

- To generate homogeneous random networks we estimate the connection probability p by using the mean output degree, *i.e.*, for a given bacterial network (V, A) , we estimate $p = \#A/(\#V^2 - \#V)$. We obtain values $p = 0.001$ in the case of *E. coli*, *B. subtilis*, and *M. tuberculosis*, and $p = 0.002$ for *P. aeruginosa*, *C. glutamicum*, and *S. aureus*.

- For scale-free networks we compute the output degree distributions for each bacterial network, then we randomly connect vertices in order to obtain a network with the same output degree distribution as the real bacterial networks. The input degree distribution turns out to be Poissonian in this kind of network.

- To generate random geometric networks we consider every given bacterial network (V, A) . For every element of V we place randomly one point in the unit cube. For every pair of vertices u, v , an arrow (u, v) , or (v, u) , or both are placed in the cube with probability 0.4, 0.4 and 0.2 respectively, if $d < r$, where d is the Euclidean distance. Values of r were considered such that the number of arrows in the real network is approximately the same as that in real bacterial networks. For this study we used $r = 0.085$ in the case of *E. coli* and *M. tuberculosis*, $r = 0.098$ for *P. aeruginosa*, $r = 0.1$ for *C. glutamicum*, and $r = 0.094$ for *S. aureus*.

4 Results and discussion

4.1 Comparison of bacteria networks

By using the pseudo-distances m and M , we compare the six bacterial regulatory networks. The main results are shown in Table 2. As explained before, numbers close to zero indicate a high similarity between the network structures, according to the hierarchy provided by dominant vertices of the compared networks. Both pseudo-distances give similar results in most of cases.

Table 2 Comparison of the six bacterial networks by using the pseudo-distances m and M

	EC	BS	PA	MT	CG	SA
EC	0	$m = 0.35$ $M = 0.17$	$m = 0.53$ $M = 0.27$	$m = 0.59$ $M = 0.29$	$m = 0.86$ $M = 0.43$	$m = 1.12$ $M = 0.56$
BS		0	$m = 0.56$ $M = 0.28$	$m = 0.44$ $M = 0.22$	$m = 0.64$ $M = 0.32$	$m = 0.87$ $M = 0.44$
PA			0	$m = 0.59$ $M = 0.30$	$m = 0.69$ $M = 0.35$	$m = 0.86$ $M = 0.43$
MT				0	$m = 0.30$ $M = 0.15$	$m = 0.57$ $M = 0.29$
CG					0	$m = 0.44$ $M = 0.22$
SA						0

Table 3 Percentage of vertices by level for each bacterial network

EC	BS	PA	MT	CG	SA
4.5%	2.3%	7.3%	2.8%	8.3%	5.5%
20%	34.3%	32.0%	48.8%	52.7%	75.0%
34.9%	33.6%	11.3%	24.1%	19.6%	10.9%
9.4%	8.3%	11.1%	10.2%	16.0%	8.4%
13.2%	4.4%	13.9%	9.7%	3.1%	
5.3%	7.7%	5.0%	1.5%		
10.1%	6.6%	7.3%	2.7%		
1.4%	2.5%	1.7%			
0.74%		7.2%			
0.06%		2.7%			

As we can observe, *M. tuberculosis* and *C. glutamicum* attain the smallest M -distance. Although these networks are significantly different in size, they have similar global structures. The second smallest M -distance is obtained from the comparison between *E. coli* and *B. subtilis* networks. They have 4.53% and 2.35% of dominant vertices, respectively (Table 3). The largest M -distance, which indicates the greatest dissimilarity, was obtained by comparing *E. coli* and *S. aureus*.

Fig. 2 shows the network hierarchies determined by the dominant vertices. This characterization of the networks gives a global picture allowing a qualitative comparison between networks. Note the higher similarity between the regulatory

Fig. 2 Hierarchical distribution of vertices on bacterial networks. We show the hierarchies arranged according to the dominant vertices for the six bacterial transcriptional regulatory networks. Each circle size is proportional to the number of genes per level.

networks of *M. tuberculosis* and *C. glutamicum*, and *E. coli* and *B. subtilis*.

4.2 Comparison between real-bacteria and random networks

In Tables 4 and 5 we show the results of the comparison between bacterial networks and their corresponding random homogeneous, scale-free and geometric random networks. In all cases, homogeneous random networks have a larger set of dominant vertices, between 20% and 30% of all the vertices, and a higher number of levels. In contrast, scale-free random networks have a depth comparable to the corresponding real bacterial networks, but as the homogeneous random networks maintain a comparatively much larger set of dominant vertices with respect to the real ones. On the other hand geometric random networks have the largest set of dominant vertices with around 50%, but depth is not as big as in homogeneous random networks.

In Fig. 3 we show an example of the homogeneous (HR), scale free (SF), and geometric (GEO) random networks corresponding

Fig. 3 Hierarchy of random networks. Homogeneous (HR), scale free (SF), and geometric (GEO) networks corresponding to each real bacterial network.

Table 4 Comparison among bacterial regulatory networks and their corresponding random networks

	EC	BS	PA
Bacterial			
Percentage of dominant vertices	4.60%	3.90%	7.39%
Depth	9	7	9
Homogeneous			
Mean percentage of dominants	19.37%	20.62%	25.48%
Mean depth	159.36	141.44	541.92
Mean <i>m</i> -distance to bacterial	1.4	1.43	1.12
Mean <i>M</i> -distance to bacterial	0.7	0.75	0.59
Scale free			
Mean percentage of dominants	10.41%	13.89%	23.00%
Mean depth	14.65	10.8	10.04
Mean <i>m</i> -distance to bacterial	0.77	0.75	0.66
Mean <i>M</i> -distance to bacterial	0.39	0.38	0.33
Geometric			
Mean percentage of dominants	39.8%	39.66%	44.33%
Mean depth	10.9	11.69	7.43
Mean <i>m</i> -distance to bacterial	0.78	0.81	0.81
Mean <i>M</i> -distance to bacterial	0.39	0.40	0.41

Table 5 Comparison among bacterial regulatory networks and their corresponding random networks

	MT	CG	SA
Bacterial			
Percentage of dominant vertices	2.83%	8.36%	5.55%
Depth	6	4	3
Homogeneous			
Mean percentage of dominants	20.13%	27.37%	33.63%
Mean depth	152.89	43.97	27.89
Mean <i>m</i> -distance to bacterial	1.46	1.26	1.26
Mean <i>M</i> -distance to bacterial	0.77	0.66	0.65
Scale free			
Mean percentage of dominants	12.86%	24.00%	30.00%
Mean depth	10.08	8.09	5.89
Mean <i>m</i> -distance to bacterial	0.87	0.82	1.00
Mean <i>M</i> -distance to bacterial	0.43	0.41	0.50
Geometric			
Mean percentage of dominants	39.47%	46.63%	50.89%
Mean depth	11.24	6.73	5.99
Mean <i>m</i> -distance to bacterial	0.80	0.80	0.99
Mean <i>M</i> -distance to bacterial	0.40	0.40	0.50

to each bacterial network: *E. coli* (HR $d = 136$, SF $d = 13$, GEO $d = 11$), *B. subtilis* (HR $d = 98$, SF $d = 8$, GEO $d = 14$), *P. aeruginosa* (HR $d = 34$, SF $d = 9$, GEO $d = 7$), *M. tuberculosis* (HR $d = 114$, SF $d = 8$, GEO $d = 14$), *C. glutamicum* (HR $d = 22$, SF $d = 7$, GEO $d = 7$) and *S. aureus* (HR $d = 27$, SF $d = 5$, GEO $d = 5$).

In Fig. 4 we show the proportion of vertices by level in the hierarchical decomposition of each bacterial network, and the mean proportion of vertices in their corresponding homogeneous, scale-free and geometric random networks. This hierarchical decomposition allows us to clearly identify some salient topological characteristics on each kind of network and distinguish it between real bacterial networks from homogeneous, scale-free and geometric random networks.

4.3 Biological analysis

The bacterial networks here studied have different numbers of vertices, and in general, each network represents the description of different biological aspects, and at a different depth, on each bacterium. This is due to the motivations behind the research studies of these organisms. *E. coli* is possibly the bacteria where the study of different aspects of the biological organism are better balanced, in contrast with *P. aeruginosa* and *B. subtilis* where virulence and sporulation are the biological processes better described. *S. aureus* is the bacterium whose network more poorly describes its biology. Having said this, we warn our readers that these are not the final results and interpretations of the comparison of the regulatory networks in these organisms but we think this is a beginning to achieve that objective. Next we will describe the most representative biological aspects of each network, considering for this discussion, the biological role of the products of the genes regulated for the transcriptional and sigma factors on each level of dominance.

The first level in the TRN of *E. coli* is enriched with regulators for the transport and catabolism of carbon sources as well as for metal homeostasis and sensory systems of two-components. The second level is dominated by regulators of amino acids metabolism and for the response to exogenous stresses. The third level is enriched with regulators controlling the central metabolism. Regulators for exogenous response are

Fig. 4 Distribution of vertices by level on real and random networks. In the X axis we show the level of dominance according to the hierarchies, in the Y axis we show the proportion of vertices by level of dominance. The real bacterial networks (a), random homogeneous networks (b), scale-free random networks (c), and geometric random networks (d).

most representative in the fourth and fifth levels whereas regulators for response to endogenous stresses, motility and biofilm are located in the lowest hierarchical levels (ESI[†]). From the seven global regulators in *E. coli*,³⁹ 3 are positioned at the first level (*FNR*, *FIS* and *H-NS*), 3 in the second level (*CRP*, *ArcA* and *Lrp*) and *IHF*, the global regulator mostly expressed in the stationary phase, is located until the fourth level. The distribution of sigma factors is uniform; *RpoD* and *FecI* at the first level, *RpoE* and *RpoS* at the third, *RpoN* at the fourth, and *RpoH* and *FlaA* at the fifth and seventh levels. The most regulating and regulated TFs are shown in the ESI[†]. In a gross description the most regulating TFs are positioned on the upper part of the network hierarchy and correspond to the so called global regulators whose regulated genes make a great diversity of biological functions. On the contrary, the most regulated TFs are located on the lower part of the network hierarchy and correspond to regulators controlling the best defined biological processes; namely, response to drugs and endogenous stresses and the master regulators for biofilm and flagella formation.

For *B. subtilis*, the first level of the network hierarchy is enriched with regulators for response to exogenous stresses, sensory systems of two-components, regulators for amino acids metabolism and for the use of carbon sources. The second level is the most populated with regulators and is dominated by those regulators controlling responses to endogenous and exogenous stresses, the metabolism of amino acids and carbon sources, those regulating the secretion of proteolytic enzymes, competence and response to metals are also present on this level. In the third level the presence of anti-sigma factors and regulators of aerobic/anaerobic respiration is notable. The only biological process which has regulators in all the seven hierarchical levels is that for sporulation; being those which control the late spore coat and late mother cell on the lowest levels. The most regulating and regulated TFs are shown in the ESI[†]. Similarly to *E. coli*, the most regulating TFs are in the upper part of the network whereas the most regulated ones are in the lowest level of hierarchy. Remarkably, these last regulators are controlling genes for competence and sporulation processes.

In *P. aeruginosa* the first level is dominated by regulators controlling the alginate biosynthesis, virulence and response to external stresses, regulators of respiration, motility and biofilm formation are also concentrated on this level. At the second level there are regulators for response to external stresses and for quorum sensing. The regulators of siderophores are concentrated on the seventh level, being the regulators of the synthesis of pyoverdine and exotoxin A on the lowest levels. The regulators for alginate biosynthesis and quorum are located on the first three hierarchical levels whereas those for virulence, siderophores and anti-sigma factors are more distributed at all levels. The most regulating and regulated TFs are shown in the ESI[†]; unusually in this organism the most regulating gene is also one of the most regulated (*lasR*).

In *M. tuberculosis* the first level is dominated by regulators for response to exogenous stresses, diverse sigma factors, two-component sensory systems and for response to metals. On the second level there are regulators for response to exogenous stresses and metals. On the third level there are regulators for response to endogenous stress and for the production of virulence factors. In the ESI[†] are enumerated the most regulating and regulated TFs in *M. tuberculosis*. The most regulating genes are located in the upper part of the network.

In *C. glutamicum* the first level is enriched with regulators for response to external stresses, for the use of carbon sources and response to metals, the metabolism of amino acids, cofactors and nucleosides are also well represented at this level. The second level is mostly populated with regulators of amino acid metabolism. The ESI[†] shows the most regulating and regulated TFs in this bacterium.

In *S. aureus* the use of carbon sources, metals response and the metabolism of amino acids are controlling most of the regulators, as well as for the biosynthesis of cofactors, nucleosides and the production of virulence factors. In the second level there are regulators of carbon sources and amino acids metabolism. The ESI[†] shows the more regulating and regulated TFs in this bacterium.

A gross interpretation of the function of the dominant set of vertices in all six bacteria here analyzed is that they control fundamental aspects for survivability, that is to say, those functions related with the interaction with the surroundings by means of sensory systems of two-components, transport and efflux of small molecules and metals. In the middle part of the networks there are regulators for the control of the central metabolism and for the biosynthesis of cofactors (e.g. *biotin*) and nucleosides, it is also common to find regulators controlling the response to endogenous stresses as well as those produced by the bacterial metabolism (e.g. *organic acids*). The lowest part of the networks are populated with genes of the so called development processes in bacteria; namely, biofilm and flagella formation⁴⁰ and for the synthesis of virulence and competitive factors (*siderophores*, *exotoxins*).

The regulatory networks of *M. tuberculosis* and *C. glutamicum* which have the most similar global topologies are also the most related phylogenetically since both of them are Actinobacteria. It is adventurous to conclude more considering they are some of the least well-described regulatory networks and the livelihood of these organisms seems to be different (*pathogen vs. glutamate producer*); we should wait until the description of their regulatory networks are more complete. The regulatory networks of *E. coli* and *B. subtilis*, which are also the largest regulatory networks of the six here analyzed, have the second smallest *M*-distance. The known information contained in these networks is similar, but the network of *B. subtilis* is dominated by sporulation processes on all the levels. The regulatory network of *P. aeruginosa* is the least similar to the other bacterium networks, and this could be explained because in this bacterium the most studied biological processes are clearly biased to the virulence and pathogenic processes setting aside central biological processes.²²

5 Conclusions

The correct comparison of biological networks could have important implications in the understanding, at structural and functional levels, of diverse phenomena like evolution, diseases, modularity and robustness in organisms. This goal is important not only for biological networks but also for other types of complex networks, for instance social or web networks. Even for graph theory, where it has been tackled.

In this study we propose the pseudo-distance *M* to compare networks with different numbers of vertices, and arrows. This metric allows for the comparisons of networks retaining important structural and dynamical characteristics. Given several networks, we can determine the two of them having more similar global structures by using *M*.

The pseudo-distance *M* is based on the hierarchy provided by the concept of dominant vertices. Other hierarchies have been proposed in the literature, or implemented in computer programs related to networks analysis. However, the hierarchy obtained by the algorithm here used fulfils certain dynamical properties and provides insights into the biological network organization.

Other methods have been proposed to compare networks, however the comparison based on the hierarchy provided by

dominant vertices has the following advantages: it is performed by an algorithm of low complexity (polynomial time), in contrast with other methods that require the search of subgraphs, which is an NP complete problem;⁴ it allows the comparison of networks with different sizes and numbers of connected components;⁴¹ it is a global measure, converse to several heuristics measures which are focused on local properties; it is applied to directed networks, which is not allowed for most of the proposed methods;⁴¹ it returns only a number, such that networks are similar as the number closes to zero.

Consistent with other studies,^{4,13} we found that SF and GEO networks are closer to real biological networks than ER networks in a similar order to that established in those studies. Moreover, pseudo-distance makes clear that existing network models are far from real biological networks.

About biological networks here analyzed: *E. coli*, *B. subtilis*, *P. aeruginosa*, *M. tuberculosis*, *C. glutamicum* and *S. aureus*, we can manifest that these bacteria have distinctive life styles, however the information we have on their regulatory networks (close to a third of the total of their predicted number of genes) reveal a topology with approximately the same number of dominant vertices and depth, containing most of their vertices in the second and third levels of their hierarchies. The proposed hierarchies also capture the relationships in the known information, since regulatory networks are still in construction.

Lastly, the pseudo-distance *M* has proven to be effective for the comparison of real biological networks. It allows us to distinguish relevant biological features of the bacteria here analyzed and show them to be different from random networks.

Acknowledgements

We thank Irma Tristan, Zaira Luna and Marcelino Ramírez for their critical comments on the manuscript. This project was supported by NPTC project PROMEP 103.5/11/4818 given to B.L., and Conacyt grant 103686 given to A.M.-A. and Julio Collado-Vides.

References

- 1 B. H. Junker and F. Schreiber, *Analysis of biological networks*, Wiley Interscience, 1st edn, 2008.
- 2 B. Albert-László and N. O. Zoltán, *Nat. Rev.*, 2004, 101–114.
- 3 U. Alon, *Nat. Rev.*, 2007, 450–461.
- 4 N. Przulj, *Bioinformatics*, 2006, **23**, e177–e183.
- 5 J. Berg and M. Lässig, *BioSystems*, 2006, 186–196.
- 6 S. Erten, X. Li, G. Bebek, J. Li and M. Koyutürk, *BMC Bioinf.*, 2009, 1–14.
- 7 A. Banerjee, *BioSystems*, 2012, **107**, 186–196.
- 8 A. Mano, T. Tuller, O. Béja and R. Y. Pinter, *BMC Bioinf.*, 2010, **11**, 1–10.
- 9 M. Heyman and A. Singh, *Bioinformatics*, 2003, **19**, i138–i149.
- 10 M. Hayashida and T. Akutsu, *BMC Syst. Biol.*, 2010, **4**, 1–11.
- 11 C. Shou, N. Bhardwaj, H. Y. K. Lam, Y. Koon-Kiu, P. M. Kim, M. Snyder and M. B. Gerstein, *PLoS Comput. Biol.*, 2011, **7**, 1–14.

- 12 B. Luna and E. Ugalde, *Phys. D*, 2008, **237**, 2685–2695.
- 13 W. Hayes, K. Sun and N. Przulj, *Bioinformatics*, 2013, **29**, 483–491.
- 14 H. D. Jong, *J. Comput. Biol.*, 2002, **9**, 67–103.
- 15 R. Coutinho, B. Fernandez, R. Lima and A. Meyroneinc, *J. Math. Biol.*, 2006, **52**, 524–570.
- 16 D. Thieffry and R. Thomas, *Bull. Math. Biol.*, 1995, **57**, 277–297.
- 17 P. Erdős and A. Rényi, *Publ. Math.*, 1959, **6**, 290–297.
- 18 R. Durrett, *Random Graph Dynamics*, Cambridge Series in Statistical and Probabilistic Mathematics, 1st edn, 2007.
- 19 M. Penrose, *Geometric Random Graphs*, 2003, vol. 5.
- 20 S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñoz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett and J. Collado-Vides, *Nucleic Acids Res.*, 2011, D98–D105.
- 21 N. Sierro, Y. Makita, M. Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–D96.
- 22 E. Galán-Vásquez, B. Luna and A. Martínez-Antonio, *Microb. Inf. Exp.*, 2011, 1–11.
- 23 J. Sanz, J. Navarro, A. Arbués, C. Martín, P. C. Marijuán and Y. Moreno, *PLoS One*, 2011, **6**, 1–9.
- 24 J. Pauling, R. Röttger, A. Tauch, V. Azevedo and J. Baumbach, *Nucleic Acids Res.*, 2011, **40**, D610–D614.
- 25 D. A. Ravcheev, *J. Bacteriol.*, 2011, **193**, 3228–3240.
- 26 F. Blattner, G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, *Science*, 1997, **277**, 1453–1462.
- 27 E. Pérez-Rueda and J. Collado-Vides, *Nucleic Acids Res.*, 2000, **28**, 1838–1847.
- 28 D. Lopez, H. Vlamakis and R. Kolter, *FEMS Microbiol. Rev.*, 2009, **33**, 152–163.
- 29 F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessièrès, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, V. Bruschi, B. Caldwell, V. Capuano, N. Carter, S. Choi, J. Codani, I. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, D. Fuma, A. Galizzi, N. Galleron, S. Y. Chim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppe, B. J. Guy, K. Haga, J. Haieah, C. R. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Jumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Manuel, C. Medigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Poji, D. Portelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Pernelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Roche, M. Rose, Y. Sadaie, T. Sato, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Tognoni, V. Tosato, S. Uchiyama, M. Vandenbol, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, E. Wedler, H. Wedler, T. Weitzenecker, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H.-F. Yoshikawa, E. Zumstein, H. Yoshikawa and A. Danchin, *Nature*, 1997, **390**, 249–256.
- 30 I. Takahiro, Y. Ken-ichi, T. Goro, F. Yasutaro and N. Kenta, *Nucleic Acids Res.*, 2001, **29**, 278–280.
- 31 J. B. Lyczak, C. L. Cannon and G. B. Pier, *Microbes Infect.*, 2000, **2**, 1051–1060.
- 32 C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrenner, M. J. Hickey, F. S. L. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbing, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, L. T. Paulsen, J. Reizer, M. H. Saier, R. E. W. Hancock, S. Lory and M. V. Olson, *Nature*, 2000, **406**, 959–964.
- 33 I. Smith, *Clin. Microbiol. Rev.*, 2003, **16**, 463–496.
- 34 S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekai, K. Badcock and D. Basham, *Nature*, 1998, 537–544.
- 35 J. Kalinowski, B. Bathe, D. Bartels, N. Bischoff, M. Bott, A. Burkowski, N. Dusch, L. Eggeling, B. Eikmanns, L. Gaigalat, A. Goesmann, M. Hartmann, K. Huthmacher, R. Krämer, B. Linke, A. C. McHardy, F. Meyer, B. Möckel, W. Pfefferle, A. Pühler, D. A. Rey, C. Rückert, O. Rupp, H. Sahm, V. F. Wendisch, I. Wiegräbe and A. Tauch, *J. Biotechnol.*, 2003, **4**, 5–25.
- 36 T. M. Venancio and L. Aravind, *J. Biol.*, 2009, **29**, 1–5.
- 37 M. Kuroda, *Lancet*, 2001, 1225–1240.
- 38 J. S. Nielsen, M. H. G. Christiansen, M. Bonde, S. Gottschalk, D. Frees, L. R. Thomsen and B. H. Kallipolitis, *Arch. Microbiol.*, 2011, **193**, 23–34.
- 39 A. Martínez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.*, 2003, **6**, 482–489.
- 40 A. Martínez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
- 41 G. Jurman, S. Riccadonna, R. Visintainer and C. Furlanello, *arXiv preprint arXiv:1109.0220*, 2011.