JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 7, 2014

Pp. 508–519
DOI: 10.1089/cmb.2013.0150

# New Approach for Phylogenetic Tree Recovery Based on Genome-Scale Metabolic Networks

DANIEL GAMERMANN,[1,2] ARNAUD MONTAGUD,[2] J. ALBERTO CONEJERO,[2] JAVIER F. URCHUEGUÍA,[2] and PEDRO FERNÁNDEZ DE CÓRDOBA[2]

## ABSTRACT

**A wide range of applications and research has been done with genome-scale metabolic models. In this work, we describe an innovative methodology for comparing metabolic networks constructed from genome-scale metabolic models and how to apply this comparison in order to infer evolutionary distances between different organisms. Our methodology allows a quantification of the metabolic differences between different species from a broad range of families and even kingdoms. This quantification is then applied in order to reconstruct phylogenetic trees for sets of various organisms.**

**Key words:** connectivity, genome-scale metabolic models, networks, phylogeny.

## 1. INTRODUCTION

Metabolic models at the genome scale are one of the prerequisites for obtaining insight into the operation and regulation of metabolism as a whole (Barrett et al., 2006; Morange, 2009; Patil et al., 2004; Stephanopoulos et al., 1998). Uses of metabolic models embrace all aspects of biotechnology, from food (Nielsen, 2001) to pharmaceutical (Boghigian et al., 2010) and biofuels (Montagud et al., 2010, 2011a). Genome-scale metabolic network reconstruction is, in essence, a systematic assembly and organization of all reactions that build up the metabolism of a given organism. It usually starts with genome sequences to identify reactions and network topology. This methodology also offers an opportunity to systematically analyze *omics* datasets in the context of cellular metabolic phenotype.

Reconstructions have now been built for a wide variety of organisms and have been used toward five major ends (Oberhardt et al., 2009): contextualization of high-throughput data (Stephanopoulos et al., 1998; Montagud et al., 2010; Edwards et al., 1999), guidance of metabolic engineering (Angermayr et al., 2009), directing hypothesis-driven discovery (Nevoigt, 2008), interrogation of multi-species relationships (Stolyar et al., 2007), and network property discovery (Guimera and Nunes Amaral, 2005).

Nowadays, phylogeny has become so popular that it's being used in almost every branch of biology (Yang and Rannala, 2012). Beyond representing the relationships among species in the tree of life, phylogeny is used to describe relationships between paralogues in a gene family (Maser et al., 2001), histories of populations (Edwards, 2009), the evolutionary and epidemiological dynamics of pathogens (Marra et al.,

[1]Cátedra Energesis de Tecnología Interdisciplinar, Universidad Católica de Valencia San Vicente Mártir, Valencia, Spain.
[2]Instituto Universitario de Matemática Pura y Aplicada, Universidad Politécnica de Valencia, Valencia, Spain.

2003; Grenfell et al., 2004), the genealogical relationship of somatic cells during differentiation and cancer development (Salipante and Horwitz, 2006), and even the evolution of language (Gray et al., 2009). More recently, molecular phylogenetics has become an indispensable tool for genome comparisons (Brady and Salzberg, 2011; Kellis et al., 2003; Green et al., 2010).

A phylogeny is a tree containing vertices that are connected by branches. Each branch represents the persistence of a genetic lineage through time, and each vertex represents the birth of a new lineage. If the tree represents the relationships among a group of species, then the vertices represent speciation events. Phylogenetic trees are not directly observed and are instead inferred from sequence or other data. Phylogeny reconstruction methods are either distance-based or character-based. In distance matrix methods, the distance between every pair of sequences is calculated, and the resulting distance matrix is used for tree reconstruction. For a very instructive review, please refer to Yang and Rannala (2012).

This work is organized as follows. In the next section, we explain the genome-scale models with which we work, how we define a parameter for comparing two models, and how we recover the phylogenetic tree from the comparison matrix obtained for many metabolic models. Additionally, we will account for the minimum spanning tree of a nondirected, connected, weighted network associated with these metabolic models. In the subsequent section, we present the results, a brief study of the sensibility of the comparison parameter, and a summary and overview.

## 2. COMPARISON BETWEEN METABOLIC MODELS

In a recent article (Reyes et al., 2012), a method has been presented for automatically generating genome-scale metabolic models from data contained in the KEGG database (Kanehisa and Goto, 2000). The method consists of searching the database for genes and pathways present in an organism and downloading the corresponding set of chemical reactions. The algorithm filters isosenzymes, or other repeated reactions, and may add missing reactions to a given pathway using a probabilistic criterion based on the comparison of the organism's pathway with the same pathway in other organisms. In this work, we use data obtained from this platform, but the method described can, in principle, be used with any set of metabolic models given that the compound names in the models follow the same standard (the same compound has the same name in all models).

The methodology we are about to describe will make use of two fundamentally different networks. One is the metabolic network build-up from the chemical reactions contained in an organism's metabolism. In this network, each metabolite represents a node (or vertex), and each link (or edge) is associated with a pair of nodes if their respective metabolites are connected as a substrate and product by some reaction. The second kind of network is the complete weighted network where each vertex represents an organism and each edge connecting two nodes is weighted by the parameter measuring the metabolic distance between the organisms' metabolism (note that this will be a complete network, where all vertices are connected to all others). In order to distinguish clearly the two networks in the text, we will talk about nodes and links for the metabolic network while for the organisms' network we will use the terms vertices and edges. As for the notation, we use capital letters (N, V, E) for the network, nodes, and links in the metabolic networks and curly letters ($\mathcal{N}$, $\mathcal{V}$, $\mathcal{E}$) for the network, vertices, and edges in the organisms' network. In the metabolic network we will use roman lowercase letters for indices representing single metabolites in sums, while for the organisms network we use Greek letters for the indices representing single organisms.

The first step in our work is to construct for every metabolic model $A$ a nondirected connected network $N_A = (V_A, E_A)$ from the information contained in it. Here, $V_A$ stands for the set of nodes of $A$, and $E_A$ stands for its set of links. A metabolic model comprises a set of chemical reactions. Each chemical reaction associates a set of substrates with a set of products. For constructing the network, first we define the set of nodes $V_A$ as the set of compounds in $A$ (metabolites present in the model), assigning a node to each metabolite. The chemical reactions in the model will define the links of the network. If two metabolites appear as a substrate and as a product, respectively, in a chemical reaction, a link connecting the correspondent nodes is added to the network. A typical metabolic model of a prokaryote, with around 1000 metabolites and the same number of chemical reactions, becomes through this process a nondirected connected network with 1000 nodes and approximately 3000 links.

The problem at hand is to elaborate a method to systematically compare and quantify the differences between two metabolic networks. For this purpose, we define a parameter that scales between zero and

infinity, zero meaning identical networks and infinity for networks that either share no node or no link in common. The definition of this parameter is based on the identity of the nodes (the compounds) but not directly on the chemical reactions of the metabolic models, only indirectly through the links of the network.

Here we start with the metabolic networks of two organisms $A = (V_A, E_A)$ and $B = (V_B, E_B)$. The set of all metabolites in between the two organisms $A \cup B = (V_A \cup V_B, E_A \cup E_B)$ can be divided into a partition of three disjoint sets: the set of metabolites only present in $A$, the set of metabolites only present in $B$, and the set of metabolites common to both organisms:

$$V_{A\cup B} = \underbrace{(V_A \setminus V_B)}_{\text{Only in A}} \cup \underbrace{(V_A \cap V_B)}_{\text{Common}} \cup \underbrace{(V_B \setminus V_A)}_{\text{Only in B}} \tag{1}$$

where $\setminus$ stands for the difference of sets. A representation of this situation is shown in Figure 1. As it is represented there, each metabolite may have connections to metabolites within its set and connections to metabolites in the other sets.

Suppose that $V_A \cup V_B = \{v_1, \ldots, v_n\}$. Fix an arbitrary node $v_i$, $1 \le i \le n$. We can consider its degree in $A \cup B$, that is, the total number of connections of $v_i$ to the rest of the metabolites of $V_A \cup V_B$, that we denote by $\deg(v_i)$. We can also consider the degree of $v_i$ when we restrict ourselves to the subnetwork generated by the node in $(V_A \setminus V_B)$, which we will call $\deg_{A\setminus B}(v_i)$. Similarly, we can also define $\deg_{A\cap B}(v_i)$ and $\deg_{B\setminus A}(v_i)$. With these degrees we can define, for each metabolite $v_i \in V_A \cup V_B$, the rate $p_{A\setminus B,i}$ of connections of $v_i$ to metabolites inside $A$ and not in $B$ with respect to the total number of connections of $v_i$, that is:

$$p_{A\setminus B,\,i} = \frac{\deg_{A\setminus B}(v_i)}{\deg(v_i)}.$$

Analogously, we can define

$$p_{B\setminus A,\,i} = \frac{\deg_{B\setminus A}(v_i)}{\deg(v_i)} \quad \text{and} \quad p_{A\cap B,\,i} = \frac{\deg_{A\cap B}(v_i)}{\deg(v_i)}.$$

The following weighted sum of the rates $p_{A\setminus B,i}$ provides a parameter of the differentiation of $A \cup B$ with respect to $A$:

$$\alpha = \left(\frac{1}{|V_A \setminus V_B|}\sum_{v_j \in V_A \setminus V_B} \deg(v_j)\right) \sum_{v_i \in V_{A\setminus B}} \frac{p_{A\setminus B,\,i}}{\deg(v_i)}$$

On the one hand, the rates $p_{A\setminus B,i}$ are multiplied by the inverse of the total number of connections of $v_i$ to give more importance to the metabolites with fewer connections. The reason to do this is that metabolic networks of all organisms usually share their hubs (metabolites with many connections), so in order to establish differences and similitude for different networks, one should focus on specific metabolites particular to only some organisms sharing common features. This weighting of $p_{A\setminus B,i}$ with the inverse of $\deg(v_i)$ will reduce the importance of very connected metabolites (hubs) that are common to most organisms and adds weight to specific metabolites that might be particular for a branch in the tree of life, helping in this way to differentiate the branches. Removing this inverse weighting results in a very mild difference between the organisms, which makes the second step in the reconstruction very hard, because the differences will appear as a small noise in the parameters.

On the other hand, the factor $\frac{1}{|V_A \setminus V_B|}\sum_{v_j \in V_A \setminus V_B} \deg(v_j)$ gives an average of the number of connections of the metabolites only present in $A$ with respect to the whole network. This is done in order to rescale the size

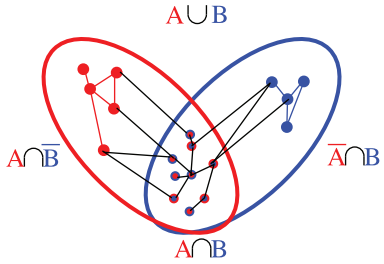


**FIG. 1.** Representation of the sets of metabolites between two organisms.

of the network and normalize (balance) the parameter after the inverse weighting done by the factor $\deg(v_i)$ for each metabolite in the set.

Analogously, we can define $\beta$ and $\gamma$ from the metabolites in the other two sets.

$$\beta = \left( \frac{1}{|V_B \setminus V_A|} \sum_{v_j \in V_B \setminus V_A} \deg(v_j) \right) \sum_{v_i \in V_{B \setminus A}} \frac{p_{B \setminus A, i}}{\deg(v_i)}$$

$$\gamma = \left( \frac{1}{|V_A \cap V_B|} \sum_{v_j \in V_A \cap V_B} \deg(v_j) \right) \sum_{v_i \in V_{A \cap B}} \frac{p_{A \cap B, i}}{\deg(v_i)}$$

For illustrating the process, let's consider three organisms, the *Synechocystis* sp. PCC 6803 (which we refer to as syn), *Synechococcus elongatus* PCC7942 (referred to as syf), and the *Escherichia coli* K-12 MG1655 (referred to as eco). In Table 1, you can see the number of metabolites and links in the networks of these organisms, and in Table 2, we show the number of elements in each one of the three sets of the partition in which we split the set of nodes of the network obtained from each pair of these three organisms.

Now let's focus on a few metabolites to see their contribution to the differentiation parameters (i.e., to the parameters $\alpha$, $\beta$, and $\gamma$). For this, we chose pyruvate (PYR), glyoxylate (GXL), and 2-dehydro-3-deoxy-6-phospho-D-gluconate (6PDG), which are respectively very, medium, and poorly connected metabolites present in these three organisms. In Table 3, we show the contribution of these metabolites to the parameters $\alpha$, $\beta$, and $\gamma$. Column $\delta_i$ of Table 3 shows, for each one of these metabolites, the value of

$$\delta_i = \left( \frac{\frac{1}{\deg(v_i)}}{\sum_{v_j \in V_A \cap V_B} \frac{1}{\deg(v_j)}} \right), \tag{2}$$

which is the weight proportion associated with the metabolite (with respect to all others) discussed above in the text. Note that this weight for PYR is very small, since pyruvate has many connections and is a very common metabolite in the metabolism of virtually any organism, and therefore is not a good candidate to help differentiate branches in the tree of life. On the other hand, 6PDG has few connections and they are different in cyanobacteria than in the *E. Coli*, potentially helping, in this way, to differentiate these two branches.

Finally, the comparison between the networks $A$ and $B$, namely $\zeta_{A,B}$, is defined as:

$$\zeta_{A,B} = \frac{\frac{|V_B|}{|V_A|}\alpha + \frac{|V_A|}{|V_B|}\beta}{2\gamma}$$

The parameters $\alpha$ and $\beta$ are balanced since some organisms have much smaller metabolic networks than others. If this is not corrected, it results in a disproportionate size between subnetworks generated by $V_{A \setminus B}$ and $V_{B \setminus A}$. In order to weaken this difference, the parameter factors $\frac{|V_B|}{|V_A|}$ and $\frac{|V_A|}{|V_B|}$ are introduced. For two identical networks, $\alpha$ and $\beta$ are zero, and so that $\zeta = 0$. For two networks that do not have a single metabolite in common we have $\gamma = 0$ and so $\zeta = \infty$.

## 3. CONSTRUCTION OF THE PHYLOGENETIC TREE

Given a set of $n$ organisms $\{A_1, A_2, \ldots, A_n\}$, we will see how to construct their phylogenetic tree taking into account the degrees of similarity between every pair of metabolic models.

TABLE 1. SETS OF NODES AND LINKS

| *Organism* | *No. nodes* | *No. links* |
|---|---|---|
| syn | 1001 | 2891 |
| syf | 979 | 2810 |
| eco | 1227 | 3801 |

Nodes and links in the networks of syn, syf, and eco.

TABLE 2. METABOLITES IN THE PARTITIONS

| | *syf* | *eco* |
|---|---|---|
| syn | $\lvert V_A \cap V_B \rvert = 911$ | $\lvert V_A \cap V_B \rvert = 778$ |
| | $\lvert V_A \setminus V_B \rvert = 90$ | $\lvert V_A \setminus V_B \rvert = 223$ |
| | $\lvert V_B \setminus V_A \rvert = 68$ | $\lvert V_B \setminus V_A \rvert = 449$ |
| syf | - | $\lvert V_A \cap V_B \rvert = 775$ |
| | - | $\lvert V_A \setminus V_B \rvert = 204$ |
| | - | $\lvert V_B \setminus V_A \rvert = 452$ |

Metabolites in the three sets of the partition when comparing three organisms.

Firstly, let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, w)$ be a nondirected, connected, complete weighted network, where $\mathcal{V} = \{A_1, A_2, \ldots, A_n\}$ is the set of vertices that represent the metabolic models of the aforementioned organisms, $\mathcal{E}$ is the set of edges $(A_\mu, A_\nu)$, $1 \leq \mu, \nu \leq n$, $\mu \neq \nu$, and $w : \mathcal{E} \to \mathbb{R}$ is a function that assigns to every edge $(A_\mu, A_\nu)$, the amount $w_{\mu,\nu} = \zeta_{A_\mu, A_\nu}$. Looking at the definition of $\zeta$, we observe that this network $\mathcal{N}$ must be symmetric. In particular, all the weights in our study are strictly positive.

Secondly, we will compute a minimum spanning tree of $\mathcal{N}$, that is, a tree that has $\mathcal{V}$ as the set of vertices, and such that the sum of the weights associated with the edges of this tree is minimum. In these trees, every vertex $A_\mu \in \mathcal{V}$ is connected with at least one of the other vertex of $\mathcal{V} \setminus \{A_\mu\}$ by an edge that has minimum weight among all the edges incident to $A_\mu$. The well-known Kruskal algorithm gives us a procedure for finding these trees (see, for instance, Gross and Yellen, 2005). We just have to follow the trace of the Kruskal algorithm in order to recover the phylogenetic tree of the organisms represented by the models $A_1, \ldots, A_n$.

In order to compute the phylogenetic tree of the models $\{A_1, A_2, \ldots, A_n\}$, consider the minimum spanning tree of $\mathcal{N}$, namely $\mathcal{T} = (\mathcal{V}, \mathcal{E}', w|_{\mathcal{E}'})$, where $\mathcal{E}' \subset \mathcal{E}$ and $w|_{\mathcal{E}'}$ denotes the restriction of the function $w$ to the elements in $\mathcal{E}'$. Let us take all the elements of $\mathcal{E}'$ in decreasing order of weights, that is, $\mathcal{E}' = \{e'_1, e'_2, \ldots, e'_{n-1}\}$ with $w(e'_1) \geq w(e'_2) \ldots \geq w(e'_{n-1})$. We are going to remove edges from $\mathcal{T}$ following this order. Every time an edge is removed, the number of connected components of the resulting graph is increased in one respect to the previous one. We can represent this division of connected components by a binary tree. The phylogenetic tree is generated taking into account how we divide $\mathcal{T}$.

There are two different situations depending on the size of the (new) connected components (if any of them consists on a single vertex or not). Let us start with the edge with maximum weight in $\mathcal{T}$ which we have denoted as $e'_1$. Suppose that $e'_1$ is adjacent to two vertices $A_{\mu 0}$ and $A_{\nu 0}$, with $1 \leq \mu_0, \nu_0 \leq n$, $\mu \neq \nu$. Then two possibilities can occur:

(a) One of these vertices, for instance $A_{\mu 0}$, is a leaf (vertex of degree 1),

TABLE 3. METABOLITE WEIGHTING

| *Metabolite* | *Organisms in comparison* | $p_{A \cap B, i}$ | $\delta_i$ | *Contribution (%)* |
|---|---|---|---|---|
| PYR | syn and syf | 0.98 | 0.127 | 0.0064 |
| | syn and eco | 0.73 | 0.117 | 0.0044 |
| | syf and eco | 0.75 | 0.113 | 0.0044 |
| GXL | syn and syf | 0.86 | 0.454 | 0.020 |
| | syn and eco | 0.87 | 0.550 | 0.024 |
| | syf and eco | 0.80 | 0.439 | 0.018 |
| 6PDG | syn and syf | 1.00 | 3.176 | 0.16 |
| | syn and eco | 0.80 | 1.762 | 0.072 |
| | syf and eco | 0.80 | 1.757 | 0.072 |

Contributions of different metabolites to the differentiation parameter ($\zeta$) between two networks. The column $\delta_i$ shows the weight of the metabolite in the calculation of $p_{A \cap B, i}$, which is the inverse of the degree of the metabolite divided by the sum of the inverses of the degrees of all metabolites contributing to the parameter.

TABLE 4. COMPARISON MATRIX

| *org* | *syf* | *syn* | *syc* | *mge* | *lpl* | *cbe* | *bcj* | *eco* | *tma* | *ypk* |
|---|---|---|---|---|---|---|---|---|---|---|
| syf | 0.0 | 0.019 | 0.0061 | 0.1628 | 0.1493 | 0.1239 | 0.1083 | 0.106 | 0.1567 | 0.1155 |
| syn | 0.019 | 0.0 | 0.0177 | 0.1821 | 0.1524 | 0.1269 | 0.1079 | 0.1116 | 0.161 | 0.1213 |
| syc | 0.0061 | 0.0177 | 0.0 | 0.1779 | 0.1616 | 0.1318 | 0.1067 | 0.1032 | 0.1572 | 0.112 |
| mge | 0.1628 | 0.1821 | 0.1779 | 0.0 | 0.1179 | 0.1351 | 0.1257 | 0.1252 | 0.1159 | 0.1266 |
| lpl | 0.1493 | 0.1524 | 0.1616 | 0.1179 | 0.0 | 0.0711 | 0.1098 | 0.1194 | 0.0668 | 0.111 |
| cbe | 0.1239 | 0.1269 | 0.1318 | 0.1351 | 0.0711 | 0.0 | 0.0979 | 0.0926 | 0.0674 | 0.1049 |
| bcj | 0.1083 | 0.1079 | 0.1067 | 0.1257 | 0.1098 | 0.0979 | 0.0 | 0.0592 | 0.1167 | 0.0557 |
| eco | 0.106 | 0.1116 | 0.1032 | 0.1252 | 0.1194 | 0.0926 | 0.0592 | 0.0 | 0.102 | 0.0294 |
| tma | 0.1567 | 0.161 | 0.1572 | 0.1159 | 0.0668 | 0.0674 | 0.1167 | 0.102 | 0.0 | 0.1044 |
| ypk | 0.1155 | 0.1213 | 0.112 | 0.1266 | 0.111 | 0.1049 | 0.0557 | 0.0294 | 0.1044 | 0.0 |

Comparison matrix for 10 organisms.

(b) Neither of these two vertices is a leaf (each vertex is still connected with the other vertex). This happens only if the former connected component has three or more vertices.

We point out that our phylogenetic tree will have two types of vertices: the leaves, which represent metabolic models, and the inner vertices, which represent two branches that each have more than one vertex.

We start our phylogenetic tree with a vertex $v_0$ that will be its root. Then two vertices $v_1$, $v_2$ are hanged from $v_0$. Each one of these vertices represents one of the two connected components of the network $\mathcal{T}\setminus\{e_1'\}$. Let us see what to do with $v_1$ and $v_2$ according to the case.

- If we are in case (a), one of these two vertices, for instance $v_1$, represents the vertex $A_{\mu 0}$, and $v_2$ represents the other connected component of $\mathcal{T}$ which is a subgraph of $\mathcal{T}$ generated by the vertex of $\mathcal{V}\setminus\{A_{\mu_0}\}$.
- If we are in case (b), one of the vertices, for instance $v_1$, represents the connected component of $\mathcal{T}\setminus\{e_1'\}$ that contains $A_{\mu 0}$, and the other vertex, $v_2$, represents the connected component of $\mathcal{T}\setminus\{e_1'\}$ that contains $A_{\nu 0}$.

This procedure is repeated again with $v_1$ and $v_2$ and by removing $e_2'$ from $\mathcal{T}\setminus\{e_1'\}$. When we remove $e_2'$, then either the connected component that represents $v_1$ or $v_2$ is split into two smaller ones, and the vertex associated with this component plays again the role of $v_0$. This process is repeated until we remove all the edges.

Let us see with two examples how it works:

1. In Table 4, we have the weights associated with a set of 10 organisms. We can represent them by a complete weighted network in which every organism is connected with the others. This is a weighted network, so that we can apply the Kruskal algorithm in order to get a minimum spanning tree of this network, which is represented in Figure 2. Following the aforementioned

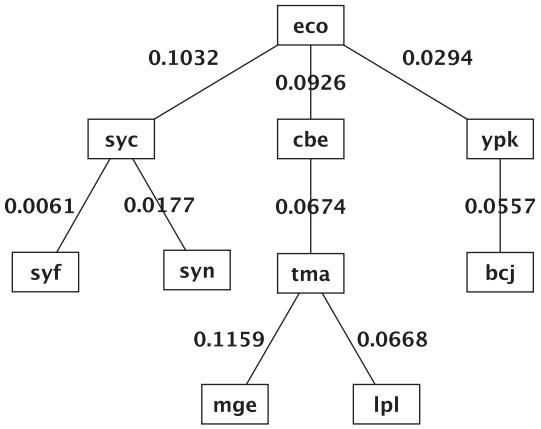


**FIG. 2.** A minimum spanning tree associated with 10 organisms.

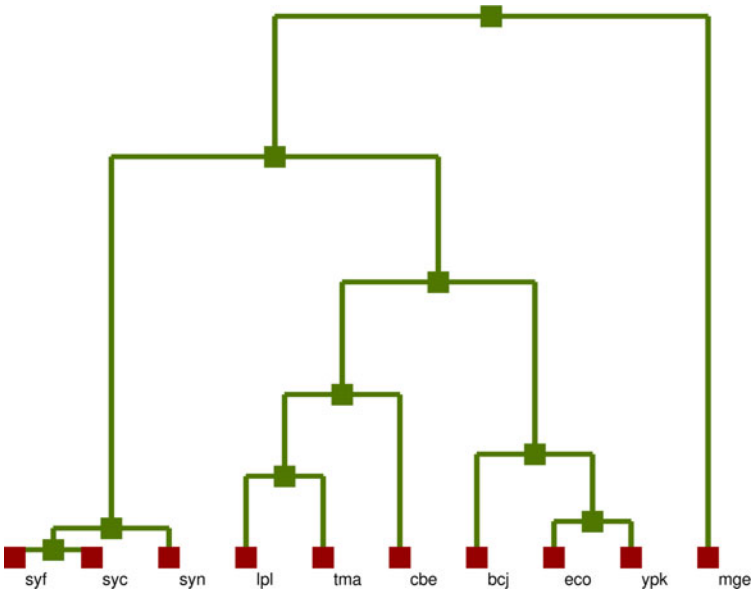


**FIG. 3.** A phylogenetic tree with 10 organisms.

notation, $e'_1$ corresponds to the edge that connects *mge* with *tma*, weighting 0.1159. We can see in Figure 3 that two vertices are hanging from the root of the tree. The one on the left represents the *mge;* the one on the right represents the subgraph associated with the rest of vertices, where tma can be found.

2. In the case of 38 organisms, when we remove from the minimum spanning tree the edge with maximum weight, we split this tree into two connected components: the one associated with the pair *mge* and *mpm*, and the one associated to the other vertices.

Finally, the vertices in the phylogenetic tree can keep more information concerning the aforementioned minimum spanning tree. Suppose that the height of our phylogenic tree is $w(e'_1)$, which represents the maximum weight in the minimum spanning tree (i.e., the weight associated with $e'_1$). We place the root of our phylogenetic tree at height $y = w(e'_1)$. Now, two vertex are hanged from the root. If one is associated with a single vertex, for instance, $v_1$ in case (a), then we place this vertex at height $y = 0$. We remember that this vertex represents the organism $A_{\mu 0}$. If not, for instance, $v_2$ in case (a) and either $v_1$ or $v_2$ in case (b), each one of these vertices represents a connected component with more than one vertex in which the minimum spanning tree is split. In order to know at which height we should put these vertices, we have to continue removing edges from the former tree. After removing $e'_2$, one of these connected components, for instance, the one represented by $v_2$, is split again into two smaller connected components. So we place the vertex $v_2$ at height $w(e'_2)$. We repeat this process recursively until the initial tree is just reduced to isolated vertices.

## 4. RESULTS AND DISCUSSION

We have reconstructed two phylogenetic trees, one with 10 bacteria and another one with both prokaryotes and eukaryotes. In Table 4 we show the parameter $\zeta$ for the pairwise comparison of the 10 prokaryotes in the first tree. The data for the comparison of the 33 organisms in the second tree is given in the Supplementary Material (available online at www.liebertonline.com/cmb).

The organisms in each comparison are:

- 10 organisms tree → *Mycoplasma genitalium* (mge), *Lactobacillus plantarum* WCFS1 (lpl), *Synechocystis* sp. PCC 6803 (syn), *Synechococcus elongatus* PCC7942 (syf), *Synechococcus elongatus* PCC6301 (syc), *Clostridium beijerinckii* (cbe), *Burkhoderia cenocepacia* J2315 (bcj), *Escherichia coli* K-12 MG1655 (eco), *Thermotoga maritima* (tma), and *Yersinia pestis* KIM10 (ypk).

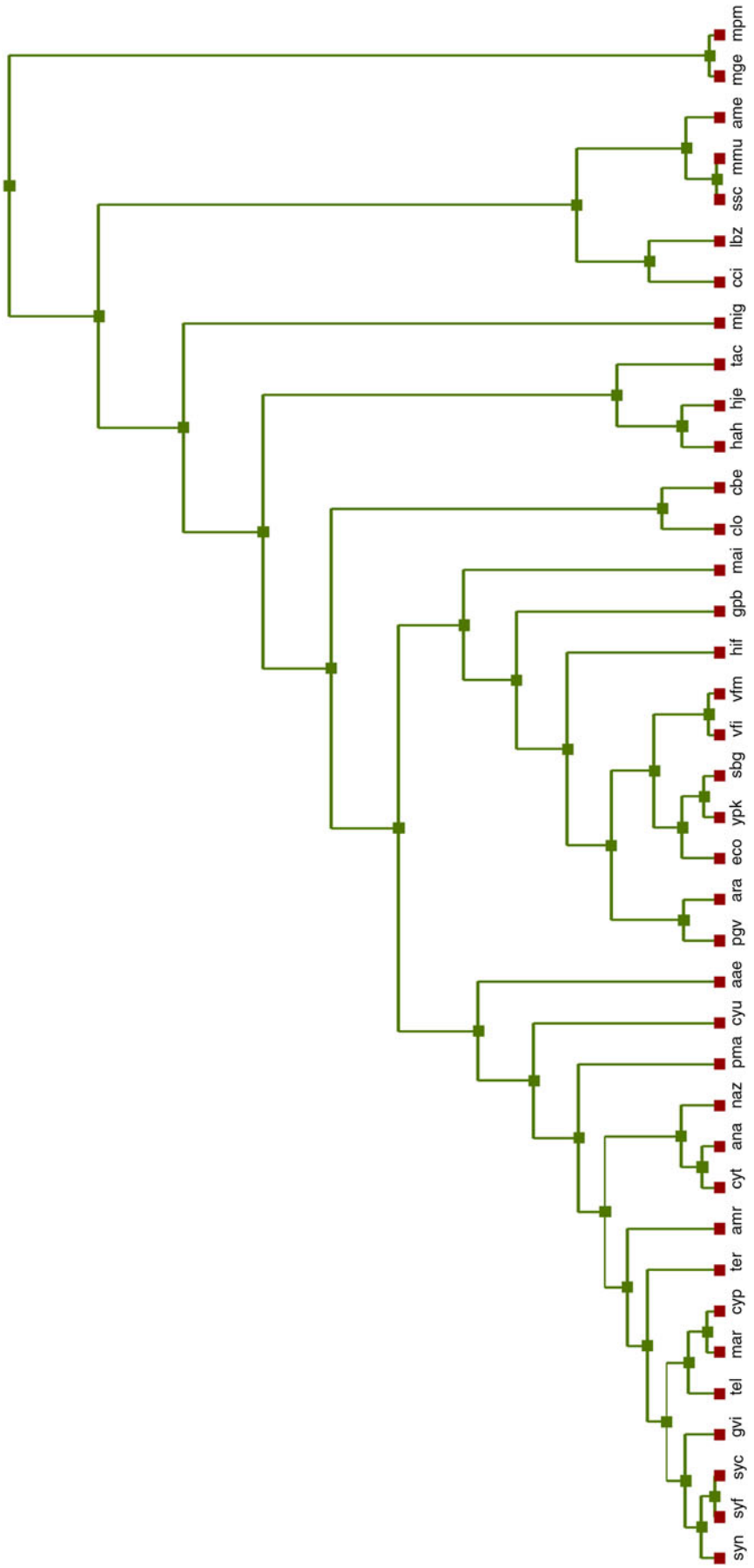


**FIG. 4.** A phylogenetic tree with 38 organisms.

TABLE 5. SENSIBILITY STUDY 1

| *org* \ *org* | *syn* | *syf* | *eco* | *mge* |
|---|---|---|---|---|
| syn | 0.0002 ± 0.0003 | 0.0184 ± 0.0005 | 0.0893 ± 0.0005 | 0.1600 ± 0.0014 |
| syf | 0.0184 ± 0.0004 | 0.0002 ± 0.0003 | 0.0857 ± 0.0006 | 0.1527 ± 0.0014 |
| eco | 0.0892 ± 0.0005 | 0.0856 ± 0.0005 | 0.0001 ± 0.0002 | 0.1278 ± 0.0009 |
| mge | 0.1597 ± 0.0025 | 0.1527 ± 0.0026 | 0.1283 ± 0.0015 | 0.0014 ± 0.0016 |

Sensibility calculation for $N_t = 500$ and $n_K = 5$. Each element in the table is the average of the parameter $\zeta$ in an ensemble plus (minus) its standard deviation ($\bar{\zeta} \pm \sigma_\zeta$).

- 38 organisms tree → *Mycoplasma genitalium* (mge), *Mycoplasma pneumoniae* 309 (mpm), *Synechocystis* sp. PCC 6803 (syn), *Synechococcus elongatus* PCC7942 (syf), *Synechococcus elongatus* PCC6301 (syc), *Clostridium beijerinckii* (cbe), *Salmonella bongori* (sbg), *Escherichia coli* K-12 MG1655 (eco), *Aquifex aeolicus* (aae), *Yersinia pestis* KIM 10 (ypk), *Cyanobacterium* UCYN-A (cyu), *Thermosynechococcus elongatus* (tel), *Microcystis aeruginosa* (mar), *Cyanothece* sp. ATCC 51142 (cyt), *Cyanothece* sp. PCC 8801 (cyp), *Gloeobacter violaceus* (gvi), *Anabaena* sp. PCC7120 (ana), *Anabaena azollae* 0708 (naz), *Prochlorococcus marinus* SS120 (pma), *Trichodesmium erythraeum* (ter), *Acaryochloris marina* (amr), *Halophilic archaeon* (hah), *Polymorphum gilvum* (pgv), *Micavibrio aeruginosavorus* (mai), *Agrobacterium radiobacter* K84 (ara), *Clostridiales genomosp.* BVAB3 (clo), *Gamma proteobacterium* HdN1 (gpb), *Vibrio fischeri* ES114 (vfi), *Vibrio fischeri* MJ11 (vfm), *Haemophilus influenzae* F3031 (hif), *Coprinopsis cinerea* (cci), *Sus scrofa* (ssc) and *Leishmania braziliensis* (lbz), *Mus musculus* (mmu), *Apis mellifera* (ame), *Methanotorris igneus* (mig), *Halalkalicoccus jeotgali* (hje), and *Thermoplasma acidophilum* (tac).

In Figures 3 and 4 we present the two phylogenetic trees that we have constructed. In the first tree, the only organism displaced in relation to what is expected from standard methods of phylogenetic tree reconstruction is the tma. In both trees mge (and mpm in the second one) diverges from other organisms at the beginning of the tree. This happens because of their minimalistic genomes, with only a couple hundred metabolites in their metabolomes. As a result, when compared with an organism without a reduced genome with almost a thousand metabolites, several hundred metabolites will not have a correspondent one, increasing hugely the value of $\alpha$ in the calculation of the parameter $\zeta$, and therefore distancing these organisms from the rest. The problem with these parasitic organisms has been noticed elsewhere (Fukami-Kobayashi et al., 2007), but unfortunately the solution found in this article did not yield better results in our present study. One should keep in mind that the present approach only considers genes (and proteins) associated with metabolic reactions and moreover, considers only the existence/absence of the enzymes (reactions). Our work yields results that are very close to the tree of life, in spite of using only a subset of all genome's information. It was not our intention to build trees that would address properly minimal organisms' phylogenies, but to prove the feasibility of building those trees using only reactome data. In any case, for the second study we used organisms from very different origins in the evolutionary history, and we found that the method is able to separate bacteria, archea, and eukaryotes. Different strains of the same species also appear closely related and share branches with organisms from the same family and order.

We have also studied the sensibility of the parameter $\zeta$. For this we performed a Monte Carlo analysis of $\zeta$. The procedure for this analysis is explained as follows. Given two organisms, one of them remains the wild type while, with the other, one builds an ensemble with $N_t$ elements, where each element is the result

TABLE 6. SENSIBILITY STUDY 2

| *org* \ *org* | *syn* | *syf* | *eco* | *mge* |
|---|---|---|---|---|
| syn | 0.0005 ± 0.0005 | 0.0186 ± 0.0006 | 0.0896 ± 0.0007 | 0.1604 ± 0.0018 |
| syf | 0.0187 ± 0.0006 | 0.0005 ± 0.0005 | 0.0860 ± 0.0007 | 0.1532 ± 0.0019 |
| eco | 0.0893 ± 0.0008 | 0.0857 ± 0.0007 | 0.0003 ± 0.0003 | 0.1281 ± 0.0011 |
| mge | 0.1602 ± 0.0035 | 0.1531 ± 0.0032 | 0.1288 ± 0.0023 | 0.0028 ± 0.0023 |

Sensibility calculation for $N_t = 500$ and $n_K = 10$. Each element in the table is the average of the parameter $\zeta$ in an ensemble plus (minus) its standard deviation ($\bar{\zeta} \pm \sigma_\zeta$).

TABLE 7. SENSIBILITY STUDY 3

| *org* \ *org* | *syn* | *syf* | *eco* | *mge* |
|---|---|---|---|---|
| syn | 0.0028 ± 0.0011 | 0.0209 ± 0.0014 | 0.0915 ± 0.0017 | 0.1652 ± 0.0045 |
| syf | 0.0207 ± 0.0013 | 0.0029 ± 0.0011 | 0.0879 ± 0.0016 | 0.1575 ± 0.0044 |
| eco | 0.0903 ± 0.0017 | 0.0868 ± 0.0016 | 0.0016 ± 0.0007 | 0.1301 ± 0.0029 |
| mge | 0.1638 ± 0.0080 | 0.1577 ± 0.0077 | 0.1343 ± 0.0055 | 0.0170 ± 0.0053 |

Sensibility calculation for $N_t = 500$ and $n_K = 50$. Each element in the table is the average of the parameter $\zeta$ in an ensemble plus (minus) its standard deviation ($\bar{\zeta} \pm \sigma_\zeta$).

of $n_K$ knock-outs (removal of $n_K$ randomly selected reactions from the metabolic model) in the organism. Then the calculation of $\zeta$ is performed between the wild-type organism and each organism in the knock-out ensemble. From this process one obtains an ensemble of $N_t$ values of $\zeta$ for the comparison (one from each version of the organism in the knock-out ensemble), from which one calculates its average and standard deviation. This standard deviation is treated as an indicator of the sensibility of the parameter (as a function of the number of knock-outs).

We performed this sensibility analysis for four organisms (syn, syf, eco, and mge) with ensembles of sizes $N_t = 500$ for $n_K = 5$, 10, 50, and 100. The results are shown in Tables 5 through 8. These four organisms have been chosen to observe the sensibility in the comparison between very similar organisms (syn and syf), more distant ones (syn and eco), and very different ones (syn and mge).

This sensibility analysis mainly reflects the uncertainties in the calculation of the metabolic distances. Since the distance parameter is based on metabolic models, one relies in the genome annotations for each organism and any annotation is usually faulty. One may miss enzymes or wrongly annotate existing ones. The models used in this study have been automatically generated from a database constructed from information downloaded from the KEGG database (Kanehisa and Goto, 2000), and since the beginning of this study the databases have been updated and most models have to be changed as well. The ''knocked-out'' models used for the sensibility parameter analysis simulate such imperfect annotations: one might consider the situation with $n_K = 5$ as the model constructed from a well-annotated genome, while the case with $n_K = 100$ is the model resulting from a very poor annotation. One can see that when only a few enzymes might be missing from the annotation, the error in the parameter can be expected to be less than 1%, except for the case of the minimalistic genomes like the parasitic mge, that has an error more than five times bigger than the other organisms. This error increases as the number of knock-outs increase, but it keeps below 5% even for 100 knockouts (or missing enzymes), except again in the case of the mge, but even for the mge it is below 10%. This shows that the methodology is robust and that one works here with an uncertainty of less than 5% in most of the cases.

## 5. CONCLUSIONS AND OVERVIEW

In this work, we have developed a methodology for comparing organisms based on their metabolic networks. This methodology has been successfully applied for the reconstruction of phylogenic trees for several organisms from a broad range of families and kingdoms. Resulting trees stand up well to their comparison with the so-called ''tree of life.'' The great majority of the branches in the tree fit their expected

TABLE 8. SENSIBILITY STUDY 4

| *org* \ *org* | *syn* | *syf* | *eco* | *mge* |
|---|---|---|---|---|
| syn | 0.0058 ± 0.0016 | 0.0239 ± 0.0020 | 0.0942 ± 0.0024 | 0.1715 ± 0.0062 |
| syf | 0.0238 ± 0.0018 | 0.0061 ± 0.0017 | 0.0907 ± 0.0023 | 0.1630 ± 0.0066 |
| eco | 0.0919 ± 0.0024 | 0.0883 ± 0.0022 | 0.0033 ± 0.0011 | 0.1329 ± 0.0040 |
| mge | 0.1694 ± 0.0120 | 0.1648 ± 0.0131 | 0.1433 ± 0.0092 | 0.0460 ± 0.0076 |

Sensibility calculation for $N_t = 500$ and $n_K = 100$. Each element in the table is the average of the parameter $\zeta$ in an ensemble plus (minus) its standard deviation ($\bar{\zeta} \pm \sigma_\zeta$).

positions well and their distance is in good correlation with evolutionary distances. The discrepancies found can be explained by particularities in these very few organisms not fitting the tree, such as tremendous genome reductions that caused reduced metabolisms.

Our methodology is innovative for it is not directly based on the structure and evolution of proteins or DNA but on the metabolism and the organisms' components and metabolic capabilities, allowing one to compare organisms very distant from the evolutionary point of view or organisms for which orthologs' comparison is difficult. In order to accomplish this, we make use of the correlation between evolutionary distances and metabolic network likelihood and propose our methodology as a starting point to study it.

Metabolism information is retrieved as a subset of the whole genome information. We hereby show that metabolic network connectivity can be used to build phylogenetic trees that are in accordance with gene-directed trees. It can be argued whether the selected construction parameter ($\zeta$) is the optimal one for this purpose (or even if there is an optimal one), but it stands clear that this is an innovative application for metabolic models, their curation, and cross-species evolutionary studies.

We have also performed a sensibility study in which we show that the methodology is robust even if the annotation information used to construct the metabolic models is faulty. This study also suggests an upper-bound for the uncertainty in the distance parameter of approximately 5%.

## ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement number 308518 (CyanoFactory).

## AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

## REFERENCES

Angermayr, S.A., Hellingwerf, K.J., Lindblad, P., and de Mattos, M.J. 2009. Energy biotechnology with cyanobacteria. *Curr. Opin. Biotechnol.* 20, 257–263.

Barrett, C.L., Kim, T.Y., Kim, H.U., et al. 2006. Systems biology as a foundation for genome-scale synthetic biology. *Curr. Opin. Biotechnol.*, 17, 488–492.

Boghigian, B.A., Seth, G., Kiss, R., and Pfeifer, B.A. 2010. Metabolic flux analysis and pharmaceutical production. *Metab. Eng.* 12, 81–95.

Brady, A., and Salzberg, S. 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods* 8, 367.

Edwards, J., Ramakrishna, R., Schilling, C., and Palsson, B. 1999. *Metabolic Flux Balance Analysis.* In *Metabolic Engineering*. Marcel Dekker Inc., New York.

Edwards, S.V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19.

Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y., and Nishikawa, K. 2007. A tree of life based on protein domain organizations. *Mol. Biol. Evol.* 24, 1181–1189.

Gray, R.D., Drummond, A.J., and Greenhill, S.J. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323, 479–483.

Green, R.E., Krause, J., Briggs, A.W., 2010 A draft sequence of the Neandertal genome. *Science* 328, 710–722.

Grenfell, B.T., Pybus, O.G., Gog, J.R., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332.

Gross, J.L., and Yellen, J. 2005. *Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications)*. Chapman and Hall/CRC.

Guimera, R., and Nunes Amaral, L.A., 2005. Functional cartography of complex metabolic networks. *Nature* 433, 895–900.

Kanehisa, M., and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kellis, M., Patterson, N., Endrizzi, M., et al. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254.

Marra, M.A., Jones, S.J., Astell, C.R., et al. 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300, 1399–1404.

Maser, P., Thomine, S., Schroeder, J.I., et al. 2001. Phylogenetic relationships within cation transporter families of Arabidopsis. *Plant Physiol.* 126, 1646–1667.

Montagud, A., Navarro, E., Fernandez de Cordoba, P., et al. 2010. Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst. Biol.* 4, 156.

Montagud, A., Zelezniak, A., Navarro, E., et al. 2011a. Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. *Biotechnol. J.*, 6, 330–342.

Morange, M. 2009. A new revolution? The place of systems biology and synthetic biology in the history of biology. *EMBO Rep.* 10, S50–S53.

Nevoigt, E. 2008. Progress in metabolic engineering of Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* 72, 379–412.

Nielsen, J. 2001. Metabolic engineering. *Appl. Microbiol. Biotechnol.* 55, 263–283.

Oberhardt, M.A., Palsson, B.O., and Papin, J.A. 2009. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320.

Patil, K.R., Akesson, M., and Nielsen, J. 2004. Use of genome-scale microbial models for metabolic engineering. *Curr. Opin. Biotechnol.* 15, 64–69.

Reyes, R., Gamermann, D., Montagud, A., 2012. Automation on the generation of genome-scale metabolic models. *J. Comput. Biol.* 19, 1295–1306.

Salipante, S.J., and Horwitz, M.S. 2006. Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5448–5453.

Stephanopoulos, G.N., Aristidou, A.A., and Nielsen, J. 1998. *Metabolic Engineering: Principles and Methodologies*. Academic Press.

Stolyar, S., Van Dien, S., Hillesland, K.L., et al. 2007. Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* 3, 92.

Yang, Z., and Rannala, B. 2012. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* 13, 303–314.

Address of correspondence:
*Dr. Daniel Gamermann*
*Instituto de Fisica*
*Universidade Federal do Rio Grande do Sul*
*Av. Bento Gonçalves 9500*
*Caixa Postal 15051, 91501-970*
*Porto Alegre RS, Brazil*

*E-mail:* daniel.gamermann@ucv.es